
Теоретические основы
ЗАЩИТЫ ИНФОРМАЦИИ

Мы должны быть благодарны Богу за то, что он сотворил мир таким, что все простое в нем истинно, а все сложное – ложно.

Григорий Сковорода
(украинский философ, XVIII ст.)

В любом случае есть только один способ правильно вести полемику: нужно сначала хорошо понять, о чем идет речь…
Безопасность – это предупреждение зла.

Платон
(древнегреческий философ, I ст. до н.э.)

National academy of sciences of Ukraine

Institute for safety problems of nuclear power plants

V.P. Babak, O.O.Klyuchnykov

Theoretical bases of INFORMATION PROTECTION

Textbook

Chornobyl 2012

Национальная академия наук Украины

Институт проблем безопасности атомных электростанций

В. П. Бабак, А. А. Ключников

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЗАЩИТЫ ИНФОРМАЦИИ

У ч е б н и к

Утверждено
Министерством образования и науки, молодежи и спорта Украины

Чернобыль 2012

УДК 004.056.5(075)
ББК 32.973-018.2я7
Б12

Рецензенты:
академикНАН Украины Б. С. Стогний,
д.т.н., проф. Ю. М. Туз

Утверждено Министерством образования и науки, молодежи и спорта Украины в качестве учебника для студентов высших учебных заведений (письмо № 1/11-4505 от 03.06.2011)

Бабак В.П.

Б12 Теоретические основы защиты информации: учебник / В. П. Бабак, А. А. Ключников; НАН Украины, Ин-т проблем безопасности АЭС. - Чернобыль (Киев.обл.) : Ин-т проблем безопасности АЭС, 2012. - 776 с.
ISBN978-966-02-6042-9

В учебнике изложены основные понятия и методы защиты информации, базирующиеся на математическом аппарате преобразования и исследования информационных сигналов, технологиях измерения, передачи и обработки информации, сигналов и данных, на помехоустойчивом кодировании, использовании современных информационно-коммуникационных каналов передачи информации, на алгоритмах шифрования и дешифрования, стегано- и криптографии, цифровой подписи и т.п. Наряду с алгоритмическими и техническими методами рассмотрены защита конфиденциальной и коммерческой информации, защита от несанкционированного доступа, защита интеллектуальной собственности, законодательное обеспечение защиты информации, а также международные стандарты в сфере защиты информации.

Для студентов технических специальностей высших учебных заведений, аспирантов, научных и инженерно-технических работников в области защиты информации.

УДК004.056.5(075)

ББК 32.973-018.2я7

ISBN978-966-02-6042-9

О Г Л А В Л Е Н И Е

Предисловие …… 10

Глава 1. Основные понятия защиты информации ……..…… 11
1.1. Термины и определения ……. 12
1.2. Классификация угроз безопасности информации и методы оценки её уязвимости..... 21
1.3. Каналы утечки информации …… 32
1.4. Принципы построения систем комплексной защиты информации.… …… 41
1.5. Критерии оценки помехоустойчивости информационных систем …….. 58
Основные выводы …… 63
Вопросы для самоконтроля …….. 66
The main conclusions…… 68

Глава 2. Количественные оценки информации ……. …… 71
2.1. Информация, её функции и свойства ……... 72
2.2. Количественные характеристики информации ……..… 75
2.3. Меры информации ……. 79
2.4. Энтропия и её свойства ……. 83
2.5. Производительность и избыточность источника информации…… 87
2.6. Связь информации с параметрами сигналов ……. 106
Основные выводы …… 111
Вопросы для самоконтроля …….. 112
The main conclusions…… 113

Глава 3. Информационные сигналы и их математические модели…. 115
3.1. Виды информационных сигналов и их математические модели ……..…….. 116
3.2. Случайные сигналы и помехи …… 142
3.3. Численные характеристики сигналов и помех ……. 149
3.4. Математические модели сигналов с ограниченным спектром …… 155
3.5. Дискретные сигналы …… 170
Основные выводы …… 195
Вопросы для самоконтроля …….. 198
The main conclusions ……... 199

Глава 4. Обработка информации ……. …… 203
4.1.Аналоговая обработка информации …… 204
4.2. Квантование и дискретизация …….. 272
4.3. Цифровая обработка информации …… . 285
Основные выводы …….. 329
Вопросы для самоконтроля …… 331
The main conclusions. ...…… 332

Глава 5. Передача информации ………….. 335
5.1. Информационные системы передачи данных…… 336
5.2. Виды информационных каналов, их математические модели и характеристики …… 342
5.3. Скорость передачи информации в каналах связи …… 359
5.4. Синтез элементов информационных систем. Оптимальный приёмник ……. 373
5.5. Многоканальные сети передачи данных. Разделение информационных каналов …… 387
5.6. Помехоустойчивость систем передачи информации …… 396
Основные выводы …… 403
Вопросы для самоконтроля …… 404
The main conclusions ……..... 405

Глава 6. Сети передачи информации ……. …….. 407
6.1. Информационно-коммуникационные сети …….. 408
6.2. Проектирование информационных сетей …….. 424
6.3. Системи беспроводной передачи информации и защита информационных ресурсов ..…… 434
6.4. Спутниковые каналы …….. 450
6.5. Множественный доступ к информационным ресурсам…… 460
Основные выводы …….. 479
Вопросы для самоконтроля …… 481
The main conclusions ……... 482

Глава 7. Кодирование информации …….…….. 485
7.1. Кодирование источника сообщения и сжатие данных …….. 486
7.2. Помехоустойчивое кодирование ……. 521
7.3. Блочное помехоустойчивое кодирование …….. 531
7.4. Свёрточное помехоустойчивое кодирование …… 555
Основные выводы …… 574
Вопросы для самоконтроля …….. 575
The main conclusions…… 576

Глава 8. Шифрование и дешифрование информации …… 577
8.1. Модели, задачи и системы шифрования …… 578
8.2. Шифрование в каналах связи …….. 584
8.3. Алгоритмы и системы симметричного и асимметричного шифрования (криптографической защиты информации)……. 588
8.4. Электронная цифровая подпись …….. 605
8.5. Стеганографические методы защиты информации …….. 624
8.6. Методы криптоанализа …… 638
Основные выводы ……..……. 642
Вопросы для самоконтроля ……. 643
The main conclusions…… 644

Глава 9. Защита информации от несанкционированного доступа…… 647
9.1. Методы несанкционированного доступа к ресурсам информационных систем …….. 648
9.2. Средства защиты от несанкционированного доступа ……... 671
9.3. Моделирование систем и процессов защиты информации ….. 677
9.4. Противодействие сетевому несанкционированному доступу... 693
Основные выводы ……... 711
Вопросы для самоконтроля ……. 714
The main conclusions…… 714

Глава 10. Законодательное обеспечение защиты информации…… 719
10.1. Правовые основы защиты информации ……..... 720
10.2. Организационные меры защиты информации …….. 741
10.3. Научно-методическое обеспечение защиты информации … 749
10.4. Международные стандарты информационной безопасности …….. 756
Основные выводы …… 762
Вопросы для самоконтроля …… 764
The main conclusions …… 765

Предметный указатель …… …….. 767
Список литературы **……**…… 774

CONTENTS

Introduction …… 10

Par**t 1. The main notions of protection and safety of information**…… 11

1.1. Terms and definitions …… 12
1.2 Threats of safety of the information and methods of an estimation of its vulnerability …… 21
1.3. Technical channels and the sources of the flowing out of the information…….. 32
1.4. Systems of complex information protection …… 41
1.5. Criteria of an estimation of noise immunity of information systems.. 58

Part 2. Quantitative estimations of the information …… 71

2.1. The information, its functions and qualities …..…….. 72
2.2. Quantitative characteristics of the informatio…… 75
2.3. Measures of the information ……. 79
2.4. Entropy and its qualities……..……. 83
2.5. Productivity, excessiveness and noise immunity of the source of information …… 87
2.6. Connection of the information with the parameters of signals……. 106

Part 3. Informational signals and their mathematical models…… 115

3.1. Kinds of informational signals and their mathematical models …… 116
3.2. Random signals and interferences……. 142
3.3. Numerical characteristics of signals and interferences……149
3.4. Mathematical models of signals with the limited spectrum……. 155
3.5. Discrete signals…… 170

Part 4. Information processing…… 203

4.1. Analogue information process. …… 204
4.2. Quantization and digitization…… 272
4.3. Digital information processing…….. 285

Part 5. Transmission of the information…… 335

5.1. Information systems of data communication…… . 336
5.2. Kinds of information channels, their mathematical models and characteristics …… 342
5.3. Speed of information transmission in communication channels….. 359
5.4. Synthesis of units of information systems. The optimal receiver… 373
5.5. Multi-channel networks of data communication. Division of information channels…….. …… 387
5.6. Noise immunity of systems of information transmission……. 396

Part 6. Networks of the information transmission……..…… 407

6.1. Information-communication networks…… 408
6.2. The designing of informational networks.…….. 424
6.3. The systems of wireless information transmission and protection of the informational resource……. 434
6.4. The satellite channels…….. 450
6.5. Plural access to the informational resources…….. 460

Part 7. The coding of the information….…… 485

7.1. The coding of the source of informing and compression of data… 486
7.2. Noise immunity coding ……..…….. 521
7.3. Block noise immunity coding ……. 531
7.4. Folding noise immunity coding ……. 555

Part 8. The enciphering and deciphering of the information …….. 577

8.1. Models, tasks and systems of the enciphering …… 578
8.2. The enciphering in the communication channels …….. 584
8.3. Algorithms and systems of a symmetric and asymmetric enciphering (Cryptographic information protection) …… 588
8.4. The electronic digital signature …… 605
8.5. Stenographic methods of information protection …… 624
8.6. Methods of cryptoanalysis …….. 638

Part 9. Information protection from unauthorized access …… 647

9.1. Methods of unauthorized access to resources of information systems …… 648
9.2. Safety devices from unauthorized access …….. 671
9.3. Modeling of systems and processes of information protection ….. 677
9.4. Ways of the warning of the flowing out of information through technical chanels……. 693

Part 10. Legislative securing of information protection…….. .. 719

10.1. Legal bases of the information …… 720
10.2. Organizational measures of protection …… 741
10.3. Scientifically-methodical securing of information protection …. 749
10.4. The international standards of informational safety……..…… 756

Subject index …….. 767
The list of the literature …… 774

ПРЕДИСЛОВИЕ

Как известно, на смену индустриальному обществу, сформировавшемуся во второй половине XX столетия, приходит общество нового типа – информационное. Его создание чрезвычайно ускорилось благодаря развитию сети Интернет и новых информационных технологий.

Информация заполнила современный мир, стала коммерческим продуктом, вошла во все сферы жизнедеятельности человека, и при этом, естественно, требования к методам и системам получения, передачи, приема и обработки информации резко возросли. Появилось мощное направление, связанное с защитой и безопасностью информации, во многом изменились требования к подготовке специалистов в направлениях электроники и информатики, радиотехники и информационной безопасности и т.д.

Поэтому возникла необходимость создания базового учебника по защите информации. При его написании авторы руководствовались афоризмом Бертрана Рассела: «Книга должна быть или ясной, или строгой; объединить эти два требования - невозможно», а материал старались преподнести в наиболее простой форме.

Учебник практически охватывает весь круг проблем защиты и безопасности информации, базирующихся на крипто- и стеганографии, цифровой подписи и математическом аппарате преобразования и исследования информационных сигналов, технологиях измерения, передачи и обработки информации, сигналов и данных, на помехоустойчивом кодировании, использовании современных телекоммуникационных и спутниковых каналов связи.

Наряду с алгоритмическими и техническими методами рассмотрены организационно – правовые методы законодательного обеспечения защиты информации от несанкционированного доступа, вопросы защиты интеллектуальной собственности, а также международные стандарты в сфере защиты информации. После каждого раздела помещены основные выводы, вопросы для самоконтроля, а также ключевые слова на русском и английском языках.

Обращаясь к читателям, отметим, что Вы можете всегда критиковать эту книгу за отсутствие в ней ваших любимых сюжетов. Но мы ни в коем случае не склонны просить за это прощение. По этому случаю мы выбрали в качестве девиза слова из песни Рикки Нельсона «Вечеринка в саду»: «Ты не можешь угодить всем, так что угоди самому себе».

Учебник основывается на курсах лекций, которые читались студентам Национального авиационного университета (г. Киев).

Авторы выражают искреннюю благодарность рецензентам - академику – секретарю НАН Украины Б. С. Стогнию и доктору технических наук, профессору Ю. М. Тузу за замечания, рекомендации и помощь в формировании рукописи. Особая благодарность за техническую поддержку при подготовке рукописи доценту С. В. Бабаку и инженеру Л.Мелащенко.

Киев, Чернобыль, осень 2011 ***В. П. Бабак, А. А. Ключников***

ОСНОВНЫЕ ПОНЯТИЯ ЗАЩИТЫ ИНФОРМАЦИИ

1

1.1 Термины и определения
1.2 Классификация угроз безопасности информации и методы оценки ее уязвимости
1.3 Каналы утечки информации
1.4 Принцип построения систем комплексной защиты информации
1.5 Критерии оценки помехоустойчивости информационных систем

1.1. Термины и определения

В широком смысле информация (от лат. information - разъяснение, высказывание) – это совокупность сведений о внешнем мире, получаемая нами в результате взаимодействия с ним. Информация – одна из важнейших категорий природоведения наряду с веществом и энергией.

Можно выделить три основных вида информации в обществе: личная, специальная и массовая. Личная информация касается тех или иных аспектов личной жизни человека. К специальной относится научно – техническая, интеллектуальная, деловая (коммерческая), производственная и другая информация. Массовая информация предназначена для большой группы людей и распространяется средствами массовой информации, такими, как газеты, журналы, радио, телевидение.

Информация в любой форме является объектом хранения, передачи и преобразования или обработки. Передают информацию в виде сообщений.

Сообщение *– это способ представления информации, или выраженная конкретным образом информация, предназначенная для передачи от источника к потребителю.*

Это условные знаки, при помощи которых мы получаем те или иные новости (информацию). Например, о ходе футбольной встречи можно узнать из рассказа очевидца (устное сообщение) или прочитать в газете (письменное сообщение). При передаче телеграммы сообщение – это текст в виде последовательности разных букв и знаков; при беседе – это последовательность звуков; во время телевизионных передач – изменение яркости и цветности элементов изображения, последовательность звуков. Но сообщения в таком виде не могут быть переданы в системах связи и потому преобразовываются в сигналы.

Сигнал *(от лат. signum - знак) – это процесс изменения во времени состояния некоторого объекта, используемый для отображения, регистрации или передачи сообщений.* ***Сигнал –*** *это материальный носитель сообщений (информации).*

В современной технике используются электрические, электромагнитные, световые, ультразвуковые и другие носители. В целом, для передачи сообщения необходимо применить такой носитель, который наилучшим образом может преодолеть расстояние от источника к пользователю.

То есть, сама по себе информация может быть отнесена к абстрактным категориям, подобным, например, математическим формулам. Но проявляется она всегда в материально – энергетической форме в виде сигнала. Методологическая схема формирования и материализации информации и возникновения сигналов приведена на рис.1.1.

В повседневной жизни миллионов людей и функционировании большинства предприятий и организаций широко используются средства компьютерной техники и соответствующие сети. В административном управлении

и предпринимательской деятельности ведущее место заняли информационно - коммуникационные системы и компьютерные технологии. Стремительное распространение современных методов обработки, передачи, накопления и сохранения информации способствовало возникновению проблемы, связанной с возможностью потери информации, раскрытия и модификации данных, принадлежащим конечным пользователям.

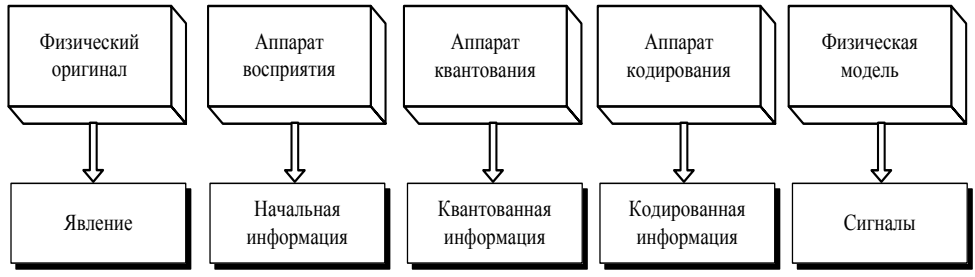


Рис. 1.1. Методологическая схема формирования и материализации информации

Данные *- это сообщения в виде цифровых сигналов.*

Упомянутая проблема получила название ***информационной безопасности****, под которой понимают состояние защищенности информации (сигналов), обрабатываемой, передаваемой или сберегаемой от незаконного (несанкционированного) доступа, преобразования и уничтожения, а также состояние защищенности информационных технических средств от действий, направленных на нарушение их работоспособности.*

Основной задачей защиты информации является обеспечение конфиденциальности, целостности, достоверности, оперативности доступа и юридической значимости.

Конфиденциальность *– свойство информации быть доступной только ограниченному количеству лиц – пользователей той или иной информационной системы, в которой циркулирует данная информация*

Целостность *- свойство информации или программного обеспечения сохранять структуру и содержание в процессе функционирования системы.*

Достоверность *- свойство принадлежности информации некоторому объекту (субъекту), являющемуся ее источником, или тому объекту, от которого эта информация принята.*

Оперативность доступа – *способность информации или информационных средств быть доступными для конечного пользователя в соответствии с его потребностями.*

Юридическая значимость *информации означает, что документ, содержащий эту информацию, имеет юридическую силу (например, электронная подпись).*

Кроме перечисленных задач защиты информации есть еще одна – задача специальной защиты, состоящая в предоставлении пользователям возможно-

сти исполнять действия в информационных системах незаметно (скрытно) для других пользователей. Особенно это актуально для электронных платежных систем. Обеспечение конфиденциальности платежных операций – защита от тотального контроля над пользователями информационных систем.

Задачи защиты информационных технических средств, решаемые в рамках информационной безопасности систем, направлены на обеспечение их работоспособности путем предотвращения:

воздействия на информационные каналы, средства сигнализации и управления, удаленную аппаратуру загрузки баз данных, на системное и прикладное программное обеспечение;

несанкционированного доступа к техническим средствам информационных систем, приводящего к нарушению их целостности, утечки информации, изменению параметров, обеспечивающих недоступность баз данных;

нарушения функционирования или вскрытия встроенных и внешних средств защиты;

неправомерных действий как пользователей, так и обслуживающего персонала.

Для решения заданий обеспечения безопасности информационно - коммуникационных систем необходимо выполнить ряд действий (рис. 1.2).

Рис. 1.2. Схема действий для решения задач безопасности информационно - коммуникационных систем

В связи с необходимостью разделения передаваемого сообщения и помех, используется понятие *обработки сигналов* (рис. 1.3).

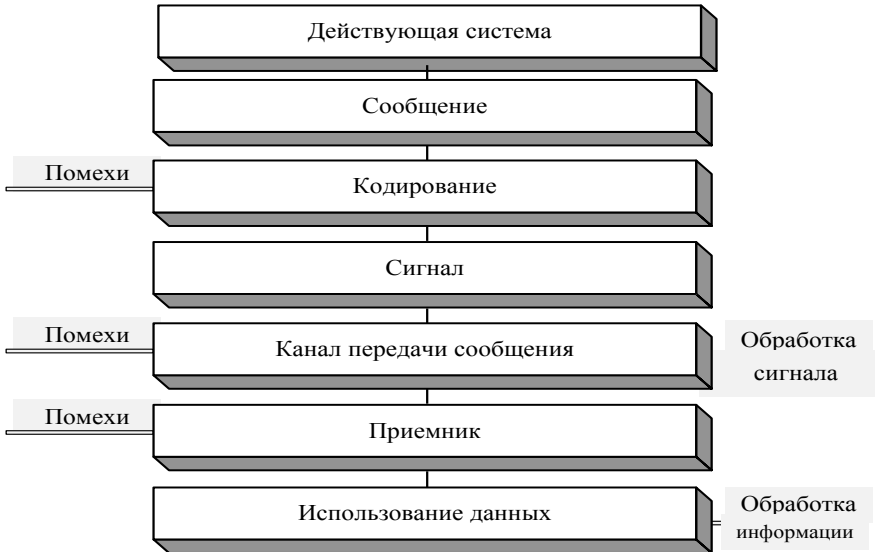


Рис.1.3. Схема пояснения принципа обработки сигналов

Одновременно, в процессе передачи от источника к пользователю, сообщение претерпевает ряд изменений, которые можно проиллюстрировать схемой передачи и принятия сообщений. Так называемая *симплексная* (однонаправленная) *система передачи сообщений* изображена на рис. 1.4.

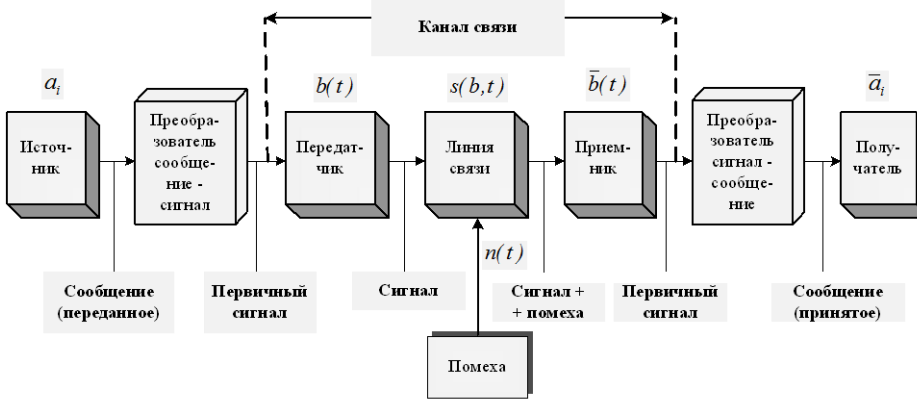


Рис. 1.4. Обобщенная структурная схема передачи и принятия сообщений

Более распространенной является *дуплексная* (двунаправленная) *система передачи сообщений*, в которой на каждом конце канала передачи и прие-

ма устанавливается как передатчик, так и приемник сообщений. При этом, обычно, преобразователь «сообщение – сигнал» (модулятор) и преобразователь «сигнал – сообщение» (демодулятор) объединяются в единое физическое устройство, называемое *модем*.

Сигнал передающей части (передатчика сигнала) от источника сообщений поступает на вход принимающей части (приемника сигналов) по линии связи. Линей связи может быть обычная пара электрических проводов, а материальным (физическим) носителем сообщения (информации) – низкочастотный электрический сигнал.

В системе радио - или мобильной связи линия связи – это пространство между передатчиком и приемником. Физическим носителем информации в этом случае есть высокочастотное электромагнитное колебание (электромагнитные волны).

Сигнал, проходя через линию связи, подвергается влиянию внешних помех. На электромагнитные волны накладываются помехи, созданные атмосферными электрическими разрядами, искрениями промышленных установок и т.д. Итак, сигнал на выходе линии связи отличается от сигнала на ее входе так же, как и сообщение, доставляемое получателю, отличается от сообщения, формируемого источником сообщений. Искажение сообщений обусловливается также влиянием на сигнал внутренних помех в передающей и принимающей частях системы. Источниками внутренних помех может быть тепловой шум (так называемый дробный эффект), создаваемый физическими элементами приемника-передатчика при прохождении по ним электрического тока, и шум, создаваемый за счет нестабильности источников питания и т.д.

С точки зрения обеспечения необходимой помехоустойчивости важным элементом является ***электромагнитная совместимость***, *под которой понимают способность радиоэлектронных приборов одновременно функционировать в реальных условиях эксплуатации с надлежащим качеством во время действия неслучайных помех и не создавать недопустимых помех другим радиоэлектронным средствам.*

На приемное устройство влияют полезное электромагнитное поле сигнала и электромагнитное поле помех. Последние создаются разными радиоэлектронными средствами, размещенными на различных расстояниях от приемника. Некоторые из этих электромагнитных полей помех вообще не влияют на качество принятия сигнала. Такие поля называются ***совместимыми*** с электромагнитным полем сигнала.

Но некоторые из полей помех при любых условиях влияют на принятие сигнала. Такие поля помех называют ***несовместимыми*** с полем сигнала. Большинство электромагнитных полей помех могут влиять на принятие сигнала в определенных условиях, тогда как при других условиях эти поля будут совместимыми с полем полезного сигнала. Такие поля помех называют условно совместимыми. Общую электромагнитную совместимость полей определяют по их отдельным совместимостям. Различают три вида совме-

стимостей: *амплитудную* (или энергетическую), *частотную* и *временную*.

Электромагнитные поля сигнала и помехи будут совместимыми, если обеспечивается хотя бы одна (любая) отдельная совместимость. Поля тогда несовместимы, когда они несовместимы сразу по трем отдельным совместимостям. Например, электромагнитные поля сигнала и помехи совместимы по частоте, если они занимают разные частотные интервалы, поскольку в этом случае обеспечивается прием сигналов с заданным качеством при наличии электромагнитных полей помех.

Конечная цель задач обеспечения электромагнитной совместимости – обеспечение совместного функционирования комплекса средств связи и другого электро- и радиооборудования. Эту цель можно реализовать, определив основные группы помех, формирующих электромагнитную обстановку, и найдя пути и способы обеспечения электромагнитной совместимости в конкретной электромагнитной обстановке.

Острейшей проблемой телекоммуникационных систем является необходимость учета взаимных помех, наиболее влияющих на электромагнитную совместимость. Это обусловлено рядом причин (рис. 1.5).

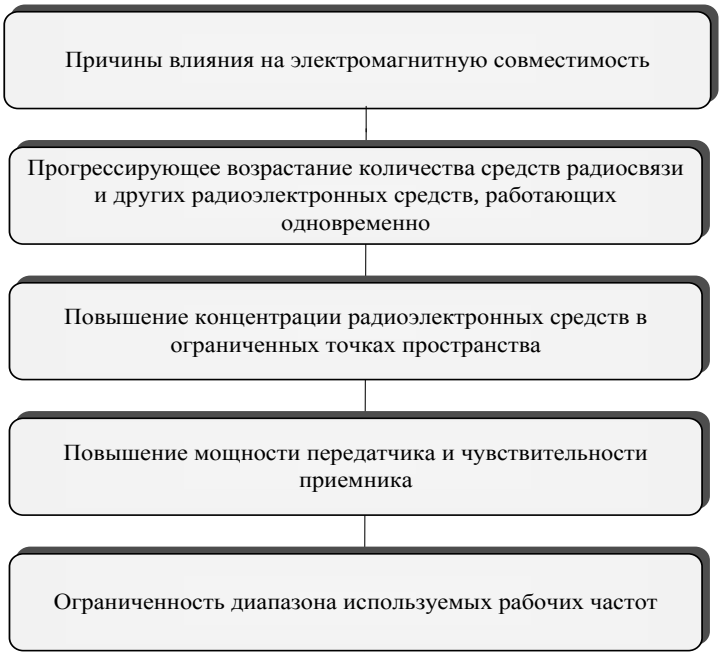


Рис. 1.5. Причины влияния внешних факторов на электромагнитную совместимость

Обеспечение электромагнитной совместимости требует:

поиска наиболее эффективных методов использования радиочастотного диапазона (регламентированное распределение рабочих частот между або-

нентами, порядок их изменений и т.д.);

рационального создания аппаратуры с учетом обеспечения электромагнитной совместимости (уменьшение уровня излучения, экранирование излучающих блоков, целесообразное размещение этих блоков, борьба с нелинейными искажениями и т.д.);

рациональной регламентации размещения и использования радиосредств (например, рациональное размещение их на одном объекте и на местности, порядок размещения излучающей аппаратуры и т.д.).

Под ***помехозащищенностью*** *системы передачи понимают ее способность предупреждать действие случайных помех.* Под ***случайными помехами*** *понимают специальные помехи, создаваемые системой радиоэлектронного глушения.*

В общем случае радиоэлектронное глушение содержит два последовательных этапа – радиотехническая разведка и радиопротиводействие. Цель радиотехнической разведки – установление факта работы (излучения) системы и определение ее параметров, необходимых для организации радиопротиводействия. Цель радиопротиводействия - создание помех, действие которых максимально усложнит работу системы или приведет к нарушению ее нормального функционирования. Очевидно, создание помех будет тем эффективнее, чем больше информации собрано на этапе разведки о системе, подлежащей глушению. В соответствии с функциями радиоэлектронного глушения помехозащищенность системы передачи определяется ее секретностью и помехоустойчивостью.

Секретность *системы передачи – ее способность противостоять действию радиоразведки.* Радиотехническая разведка предусматривает последовательное выполнение трех основных задач (рис. 1.6).

Рис. 1.6. Основные задачи радиотехнической разведки

В соответствии с этими задачами можно определить три вида секретности: энергетическую, структурную и информационную.

Помехоустойчивость *– способность системы передачи противостоять вредному действию помех.* Анализ помехоустойчивости проводят независимо от причин появления помех в системе передачи. Для повышения помехоустойчивости систем передачи нужно повышать их секретность и помехоустойчивость (рис. 1.7).

Рис. 1.7. Методы обеспечения повышения энергетической и структурной секретности

Для решения проблемы информационной секретности с целью предотвращения несанкционированного доступа к информационному содержанию сообщения используют засекречивание (шифрование) сообщений, рассматриваемое в следующих разделах.

Для того чтобы сигналы стали объектами теоретического изучения и расчетов, нужно указать способы их математического описания, то есть построить математическую модель исследуемого объекта. ***Математическая модель сигнала*** *– функциональная зависимость, адекватно описывающая изменение во времени физического состояния некоторого объекта.*

Таким образом, математическая модель сигнала – это функциональная зависимость, в которой аргументом есть время. В дальнейшем математическую модель сигналов будем обозначать символами латинского алфавита: $s(t)$, $u(t)$, $f(t)$ и др.

Описание сигнала при помощи математической модели дает возможность абстрагироваться от конкретной природы носителя сигнала. Например, в электро- или радиотехнике одна и та же математическая модель равнозначно описывает ток, напряжение, сопротивление и т.д. Кроме этого, математическая модель дает возможность описывать именно те свойства сигнала, которые объективно наиважнейшие. При этом пренебрегают численными вто-

ростепенными, несущественными признаками. Например, в большинстве случаев крайне сложно подобрать точные функциональные зависимости, соответствующие электрическим колебаниям, наблюдаемым в процессе эксперимента. Но исследователь, руководствуясь всей совокупностью доступных ему ведомостей о системе в целом, выбирает из имеющегося набора математических моделей сигналов именно те, которые в конкретной ситуации с достаточной точностью описывают физический процесс. Итак, *выбор модели в той или иной степени - процесс творческий*.

Сложность освещения проблемы безопасности информации связана с отсутствием до этого времени общепринятого толкования терминов, используемых для описания в этой области. Так, наряду с терминами «безопасность информации», «защита информации» в последнее время активно используется термин «информационная безопасность», определяемый толкованием того контекста, в котором он употребляется.

С учетом вышеизложенного приведем список основных понятий, используемых в дальнейшем, и раскроем их содержание.

Безопасность информации *– состояние защищенности данных, обрабатываемых, сохраняемых и передаваемых, от незаконного вмешательства с целью нарушения физической и логической целостности информации (ознакомления, преобразования или искажения и уничтожения) или несанкционированного использования.*

Угроза безопасности информации – *события или действия, могущие вызывать нарушение функционирование системы, связанное с уничтожением или несанкционированным использованием информации, которая в ней обрабатывается.*

Уязвимость информации *- возможность возникновения на одном из этапов жизненного цикла системы такого состояния, при котором создаются условия для реализации угроз безопасности информации.*

Защищенность информации – *поддержка в системе на заданном уровне параметров данных, характеризующих установленный статус их обработки, сохранения и использования.*

Защита информации – *процесс создания и использования в системе специальных механизмов, поддерживающих статус их защищенности.*

Комплексная защита информации – *целенаправленное регулярное использование в системах средств и методов, а также принятие мер с целью поддержки заданного уровня защищенности информации по всей совокупности показателей и условий, существенных с точки зрения обеспечения безопасности информации.*

Заведомо защищенная информационная технология – *унифицированная в широком спектре применений информационная технология, содержащая все механизмы для обеспечения необходимого уровня защиты как основного показателя качества информации.*

Качество информации *– совокупность свойств, обусловливающих спо-*

собность информации удовлетворять конкретные запросы в соответствии с ее назначением.

1.2. Угрозы безопасности информации и методы оценки ее уязвимости

*Под **угрозой безопасности** (конфиденциальности) **информации** понимают потенциальные или реально возможные действия по отношению к информационным ресурсам, приводящие к неправомерному овладению охраняемыми сведениями.*

Такими действиями являются:

ознакомление с конфиденциальной информацией разными путями и средствами без нарушения ее целостности;

модификация информации в криминальных целях – частичное или значительное изменение состава и содержания сведений;

разрушение (уничтожение) информации как акт вандализма с целью прямого нанесения материального ущерба.

Противоправные действия с информацией приводят, в конце концов, к нарушению ее конфиденциальности, полноты, достоверности и доступности (рис. 1.8), нарушая как режим управления, так и его качества при условии ошибочной или неполной информации.

Рис. 1.8. Проявление угроз безопасности информации

Каждая угроза влечет за собой конкретные убытки – моральные или материальные, а защита и противодействие угрозам должны снизить объемы ущерба, в идеале – полностью, реально – значительно или хотя бы частично.

Но и это удается не всегда. Сказанное дает возможность произвести классификацию угроз безопасности информации (рис. 19).

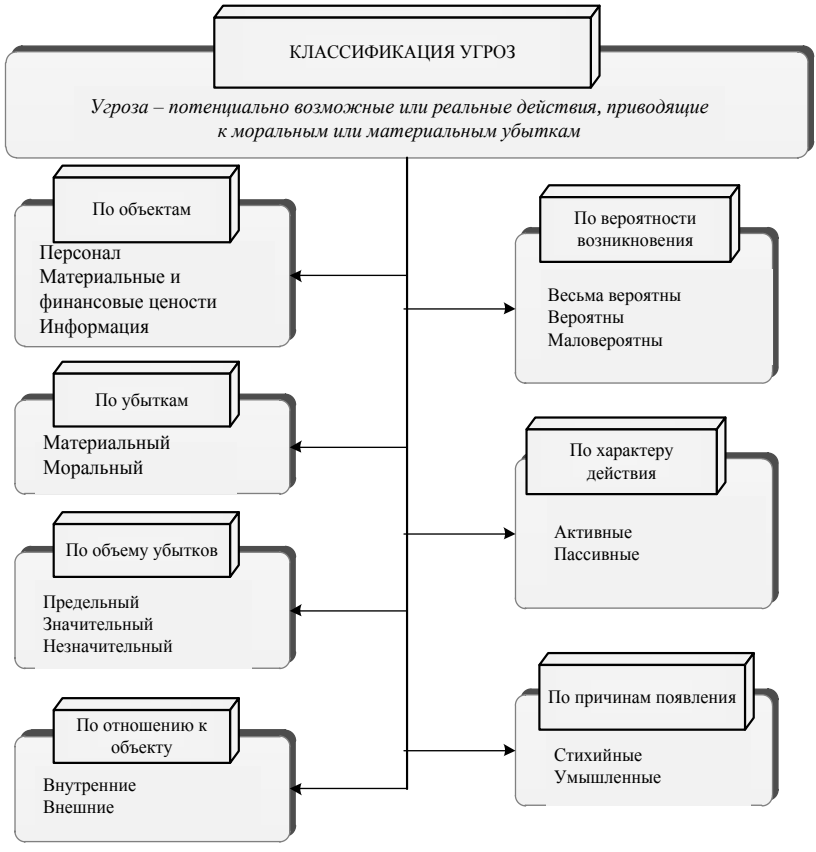


Рис.1.9. Классификация угроз безопасности информации

Систематизируем известные угрозы безопасности информации разнообразного происхождения и сведем их в табл. 1.1.

Приведем короткий комментарий использованных в табл. 1.1 определений классификации, их значений и содержания.

Виды угроз. Данный параметр является основополагающим, определяющим целевую направленность защиты информации.

Происхождение угроз. В табл. 1.1 выделены два значения этого параметра: случайное и преднамеренное. Под случайным понимается такое происхождение угроз, которое обусловливается спонтанными и не зависящими от воли людей обстоятельствами, возникающими в системе в процессе ее функционирования. Наиболее известными событиями данного плана являются отказы, ошибки, стихийные бедствия и побочные влияние.

Сущность перечисленных событий (кроме стихийных бедствий, сущность которых ясна) определяется следующим образом:

отказ – нарушение работоспособности какого-либо элемента системы, приводящее к невозможности выполнения им основных своих функций;

сбой – временное нарушение работоспособности какого-либо элемента системы, следствием чего может быть неправильное выполнение им в этот момент своей функции;

ошибка – неправильное (разовое или систематическое) выполнение элементом одной или нескольких функций, происходящее вследствие специфического (постоянного или временного) его состояния;

побочное влияние - негативное действие на систему в целом или отдельные ее элементы, оказываемое какими-либо явлениями, происходящими внутри системы или во внешней среде.

Таблица 1.1

Параметры классификации	Значения параметров	Содержание значения параметра
1. Виды	1.1. Физическая целостность 1.2. Логическая структура 1.3. Содержание 1.4. Конфиденциальность 1.5. Право собственности	Уничтожение (искажение) Искажение структуры Несанкционированная модификация Несанкционированное получение Присвоение чужого права
2. Природа происхождения	2.1. Случайная 2.2. Умышленная	Отказы, сбои, ошибки, стихийные бедствия, побочные явления Злоумышленные действия людей
3. Предпосылки появления	3.1. Объективные 3.2. Субъективные	Количественный недостаток элементов системы, качественный недостаток элементов системы Разведывательные органы иностранных государств, промышленный шпионаж, деятельность криминальных (преступных) элементов, действия недобросовестных сотрудников системы
4. Источники угроз	4.1. Люди 4.2. Технические устройства 4.3. Модели, алгоритмы, программы 4.4. Технологические системы обработки 4.5. Окружающая среда	Посторонние лица, пользователи, персонал Регистрация, передача, хранение, обработка, выдача Общего назначения, прикладные, вспомогательные Ручные, интерактивные, машинные, сетевые Состояние атмосферы, побочные шумы и помехи, побочные сигналы.

Преднамеренное возникновение угрозы обуславливается злоумышленными действиями людей.

Предпосылка появления угроз. В табл. 1.1 приведены две возможные разновидности предпосылок: объективные (количественная или качественная недостаточность элементов системы) и субъективные (деятельность разведывательных органов иностранных государств, промышленный шпионаж, деятельность уголовных элементов, действия недобросовестных сотрудников системы).

Перечисленные разновидности предпосылок интерпретируются следующим образом:

количественная недостаточность - физическая нехватка одного или нескольких элементов системы, вызывающая нарушение технологического процесса обработки данных и/или перегрузку имеющихся элементов;

качественная недостаточность - несовершенство конструкции (организации) элементов системы, в силу чего могут появляться возможности случайного или преднамеренного негативного воздействия на хранимую или обрабатываемую информацию;

деятельность разведывательных органов иностранных государств - специально организуемая деятельность государственных органов, профессионально ориентированных на получение необходимой информации всеми доступными способами. К основным видам разведки относятся: агентурная (несанкционированная деятельность профессиональных разведчиков, завербованных агентов и так называемых «доброжелателей») и техническая, охватывающая радиоразведку (перехват радиоэлектронными средствами информации, циркулирующей в коммуникационных каналах), радиотехническую разведку (регистрацию спецсредствами электромагнитных излучений технических систем) и космическую разведку (использование космических кораблей и искусственных спутников Земли) для наблюдения за территорией, ее фотографирования, регистрации радиосигналов и получения полезной информации любыми иными доступными методами;

промышленный шпионаж - негласная деятельность организаций (ее представителей) по добыванию информации, специально охраняемой от несанкционированной ее утечки или хищения, с целью создания для себя благоприятных условий и получения максимальных выгод (недобросовестная конкуренция);

злоумышленные действия уголовных элементов - хищение информации или компьютерных программ в целях наживы;

действия недобросовестных сотрудников - хищение (копирование) или уничтожение информационных массивов и/или программ по эгоистическим или корыстным мотивам, а также в результате несоблюдения установленного порядка работы с информацией.

Источники угроз. Под источником угроз понимают непосредственный их генератор или носитель. Таким источником могут быть люди, технические

средства, модели (алгоритмы), программы, внешняя среда.

Попытаемся теперь, опираясь на приведенную системную классификацию угроз безопасности информации, определить полное множество угроз, потенциально возможных в современных автоматизированных системах. При этом необходимо учесть не только все известные (ранее проявлявшиеся) угрозы, но и такие угрозы, которые ранее не проявлялись, но потенциально могут возникнуть при нынешних концепциях архитектурного построения автоматизированных систем и технологических схем обработки информации.

Классифицируем все возможные каналы несанкционированного получения информации (КНПИ) по двум критериям - необходимости доступа (физического или логического) к элементам системы для реализации какого-либо КНПИ и зависимости появления канала от состояния системы.

В соответствии с *первым критерием* КНПИ могут быть разделены на каналы, не требующие доступа, т.е. позволяющие получать необходимую информацию дистанционно (например, путем визуального наблюдения через окна помещений системы), и каналы, требующие доступа в помещения системы. В свою очередь КНПИ, воспользоваться которыми можно только получив доступ в помещение системы, делятся на каналы, не оставляющие следов в системе (например, визуальный просмотр изображений на экранах мониторов или документов на бумажных носителях), и на КНПИ, использование которых оставляет какие – либо следы (например, кража документов или машинных носителей информации).

По *второму критерию* КНПИ подразделяются на каналы, потенциально существующие независимо от состояния системы (например, похитить носитель информации можно независимо от нахождения системы в рабочем или нерабочем состоянии), и каналы, существующие только в рабочем состоянии системы (например, побочное электромагнитное излучение и наводка).

В соответствии с изложенным классификация КНПИ может быть представлена следующей таблицей (табл. 1.2.).

Таблица 1.2

Зависимость от доступа к элементам системы	Отношение к обработке информации	
	Проявляющиеся безотносительно к обработке	Проявляющиеся в процессе обработки
Не требующие доступа	1-й класс - общедоступные постоянные	2-й класс - общедоступные функциональные
Требующие доступа без замены элементов системы	3-й класс - узкодоступные постоянные без оставления следов	4-й класс - узкодоступные функциональные без оставления следов
Требующие доступа с изменением элементов системы	5-й класс - узкодоступные постоянные с оставлением следов	6-й класс - узкодоступные функциональные с оставлением следов

КНПИ 1-го класса – каналы, выявляемые безотносительно к обработке информации и без доступа злоумышленника к элементам системы. Сюда можно отнести подслушивание разговоров, провоцирование разговоров с лицами, имеющими отношение к информационным системам, и использование

злоумышленником визуальных, оптических и акустических средств. Такой канал может обнаружиться и путем похищения носителей в момент их нахождения за пределами помещения, в котором расположена система.

КНПИ 2-го класса - каналы, выявляемые в процессе обработки информации без доступа злоумышленника к элементам информационных систем. Сюда можно отнести электромагнитное излучение разнообразных устройств вычислительной техники, аппаратуры и линий связи; паразитные наводки в цепях питания, телефонных сетях, системах теплоснабжения, вентиляции и канализации, шинах заземления; подключение к информационно - вычислительной сети генераторов помех и регистрирующей аппаратуры. К этому же классу можно отнести осмотр отходов производства, поступающих за пределы контролируемой зоны.

КНПИ 3-го класса - каналы, выявляемые безотносительно обработки информации с доступом злоумышленника к элементам информационных систем, но без изменений последних. К ним относятся разнообразные виды копирования носителей информации и документов, а также похищение производственных отходов.

КНПИ 4-го класса - каналы, выявляемые в процессе обработки информации с доступом злоумышленника к элементам информационных систем, но без изменений последних. К ним относятся запоминание и копирование информации в процессе ее обработки, использование программных ловушек, недостатков языков программирования и операционных систем, а также поражение программного обеспечение вредоносными закладками, маскировка под зарегистрированного пользователя.

КНПИ 5-го класса - каналы, выявляемые безотносительно обработки информации с доступом злоумышленника к элементам информационных систем и с изменениями последних. Среди этих каналов - подмена и похищение носителей информации и аппаратуры, включения в программы блоков типа «троянский конь», «компьютерный червь» и т.д., чтение остаточной информации, содержащейся в памяти, после выполнения санкционированных запросов

КНПИ 6-го класса - каналы, выявляемые в процессе обработки информации с доступом злоумышленника к элементам информационных систем и с изменением последних. Сюда можно отнести незаконное подключение к аппаратуре и линиям связи, а также снятие информации на шинах питания различных элементов информационных систем.

При выполнении практических заданий защиты информации первостепенное значение имеет количественная оценка ее уязвимости. Рассмотрим возможные подходы к определению этой оценки.

Несанкционированное получение информации в системах возможно не только в результате непосредственного доступа к базам данных, но и многими иными путями, не требующими доступа. При этом основную опасность представляют действия злоумышленников. Воздействие случайных факторов само по себе не приводит к несанкционированному получению информации,

оно только способствует появлению КНПИ, которыми может воспользоваться злоумышленник.

Потенциально возможные несанкционированные действия могут иметь место в различных зонах (рис.1.10.):

внешней неконтролируемой зоне — территории вокруг информационной системы, на которой не используются никакие средства и не принимаются никакие меры по защите информации;

зоне контролируемой территории — территории вокруг помещений информационной системы, непрерывно контролируемой специальными средствами и персоналом;

зоне помещений информационной системы — внутреннего пространства помещений, в котором размещены средства системы;

зоне ресурсов информационной системы — части помещений, откуда возможен непосредственный доступ к ресурсам системы;

зоне баз данных — части ресурсов системы, в которых возможен непосредственный доступ к защищаемым данным.

Рис. 1.10. Территориальные зоны возможных несанкционированных действий

При этом для несанкционированного получения информации необходимо одновременное наступление следующих событий:

нарушитель должен получить доступ в соответствующую зону;

во время пребывания нарушителя в зоне в ней должен появиться (существовать) соответствующий КНПИ;

появившийся КНПИ должен быть доступен нарушителю соответствующей категории;

в КНПИ в момент доступа к нему нарушителя должна находиться защищаемая информация.

Попробуем теперь с учетом изложенного вывести формулу для ***оценки уязвимости информации***, обрабатываемой в информационных систе-

мах. Для этого введем такие обозначения:

$P_{ijl}^{Д}$ - вероятность доступа нарушителя *k*-й категории в *l*-ю зону *i*-го компонента информационной системы;

$P_{ijl}^{Н}$ - вероятность наличия (проявления) *j*-го КНПИ в *l*-й зоне *i*-го компонента информационной системы;

$P_{ijkl}^{П}$ - вероятность доступа несанкционированного получения информации *k*-й категории к *j*-му КНПИ в *l*-й зоне *i*-го компонента при условии доступа нарушителя в эту зону;

$P_{ijl}^{З}$ - вероятность наличия информации в *j*-м КНПИ в *l*-й зоне *i*-го компонента в момент доступа туда нарушителя для несанкционированного получения информации.

Тогда ***вероятность несанкционированного получения*** информации нарушителем *k*-й категории по *j*-му КНПИ в *l*-й зоне *i*-го структурного компонента информационной системы можно определить такой зависимостью:

$$P_{ijkl} = P_{ikl}^{Д} P_{ijl}^{Н} P_{ijkl}^{П} P_{ijl}^{З}. \qquad (1.1)$$

Вероятность несанкционированного получения информации в одной компоненте информационной системы одним злоумышленником одной категории по одному КНПИ назовем ***базовым показателем уязвимости информации*** (учитывая несанкционированное получение информации). С учетом (1.1) выражение для базового показателя примет такой вид:

$$P_{ikl}^{Б} = 1 - \prod_{l=1}^{5} \left[1 - P_{ijkl} \right] = 1 - \prod_{l=1}^{5} \left[1 - P_{ikl}^{Д} P_{ijl}^{Н} P_{ijkl}^{П} P_{ijl}^{З} \right]. \qquad (1.2)$$

Рассчитанные таким образом базовые показатели уязвимости сами по себе имеют ограниченное практическое значение. Для решения задач, связанных с разработкой и эксплуатацией систем защиты информации, необходимы значения показателей уязвимости, обобщенные по какому-либо индексу (*i*, *j*, *k*) или по их комбинации.

Рассмотрим возможные подходы к определению таких частично обобщенных показателей. Пусть (*K**) – интересующее нас подмножество из полного множества потенциально возможных нарушителей. Тогда вероятность нарушения защищенности информации множеством нарушителей по *j*-му фактору в *i*-й компоненте информационной системы $P_{ij\{K^*\}}^{r}$ определится как

$$P_{ij\{K^*\}}^{r} = 1 - \prod_{\forall K^*} \left[1 - P_{ijr}^{Б} \right], \qquad (1.3)$$

где $\prod_{\forall K^*}$ означает перемножение выражений в скобках для всех *k*, входящих в подмножество $\{K^*\}$. При этом верхний индекс *r* будет принимать значения *i*, *j* или *k* в зависимости от того, какие базовые показатели (нарушение целостно-

сти информации, ее несанкционированное получение или размножение) используются при расчетах.

Аналогично, если $\{J^*\}$ — подмножество интересующих нас КНПИ, то уязвимость информации в i-й компоненте по данному подмножеству факторов относительно k-го нарушителя определится выражением

$$P^r_{i\{J^*\}k} = 1 - \prod_{\forall J^*} \left[1 - P^{\text{Б}}_{ijkl} \right]. \tag{1.4}$$

Наконец, если $\{I^*\}$ — подмножество интересующих нас структурных компонент информационной системы, то уязвимость информации в них по j-му каналу несанкционированного получения информации k-м нарушителем

$$P^r_{\{I^*\}jk} = 1 - \prod_{\forall I^*} \left[1 - P^{\text{Б}}_{ijk} \right]. \tag{1.5}$$

Каждое из приведенных выражений позволяет проводить обобщения по одному параметру. Можно получить и общее выражение, если нас интересуют подмножества $\{I^*\}$, $\{J^*\}$, $\{K^*\}$ одновременно. В этом случае

$$P^r_{\{I^*\}\{J^*\}\{K^*\}} = 1 - \prod_{\forall I^*} \left[1 - P^{\text{Б}}_{ijr} \right] \prod_{\forall J^*} \left[1 - P^{\text{Б}}_{ijr} \right] \prod_{\forall K^*} \left[1 - P^{\text{Б}}_{ijr} \right]. \tag{1.6}$$

Очевидно, при таком подходе общий показатель уязвимости P^r определяется выражением

$$P^r = 1 - \prod_{\forall i} \left[1 - P^{\text{Б}}_{ijk} \right] \prod_{\forall j} \left[1 - P^{\text{Б}}_{ijk} \right] \prod_{\forall k} \left[1 - P^{\text{Б}}_{ijk} \right]. \tag{1.7}$$

На практике наибольший интерес представляют экстремальные показатели уязвимости, характеризующие наиболее неблагоприятные условия защищенности информации: самый уязвимый структурный компонент информационной системы (i'), наиболее опасный КНПИ (j'), наиболее опасная категория нарушителей (k').

Рассмотрим далее ***методы расчета показателей уязвимости информации*** с учетом интервала времени, на котором оценивается уязвимость. При этом следует учитывать, что с увеличением интервала времени увеличиваются возможности нарушителя для осуществления злоумышленных действий и тем выше вероятность изменения состояния информационной системы и условий автоматизированной обработки информации.

Можно определить такие интервалы времени (не сводимые к точке), на которых процессы, связанные с нарушением защищенности информации, были бы однородными. Назовем такие интервалы малыми. Каждый малый интервал, в свою очередь, может быть разделен на очень малые интервалы, уязвимость информации на которых определяется независимо от других. При этом, в силу однородности происходящих процессов, уязвимость информации на каждом из выделенных очень малых интервалов определяться по од-

ной и той же зависимости.

Тогда, обозначив через P_t^m интересующий нас показатель уязвимости в точке (на очень малом интервале), а через P^{μ} — тот же показатель на малом интервале, получим:

$$P^{\mu} = 1 - \prod_{t=1}^{n_t} \left[1 - P_t^m \right], \tag{1.8}$$

где t - переменный индекс очень малых интервалов, на которые поделен малый интервал; n_t - общее количество очень малых интервалов.

Нетрудно видеть, что рассмотренный подход можно распространить и на другие интервалы, а именно: большой интервал представить некоторой последовательностью малых, очень большой - последовательностью больших, а бесконечно большой - последовательностью очень больших.

Однако *приведенные выражения будут справедливы лишь в том случае, когда на всем рассматриваемом интервале времени условия для нарушения защищенности информации остаются неизменными.* В действительности эти условия могут изменяться, причем наиболее важным фактором здесь есть активное действие самой системы защиты информации.

Желающих подробнее ознакомиться с оценкой уязвимости информации на различных временных интервалах адресуем к книге В. А. Герасименко [8], в которой приведен ряд моделей определения показателей уязвимости для наиболее распространенных технологических маршрутов обработки информации.

Завершая рассмотрение вопросов уязвимости информации, мы не можем не остановиться на такой широко обсуждаемой на разных уровнях проблеме, как *информационное оружие*. В данный момент мировая цивилизация находится на переходном периоде от индустриального этапа своего развития к информационному, на котором главным стратегическим национальным ресурсом становятся информация и информационные технологии. Уже сегодня информационная зависимость всех сфер жизнедеятельности общества и государства необычайно велика.

Так (по оценкам американских экспертов), нарушение работы компьютерных систем, используемых в управлении государственными и банковскими структурами США, состоящее в выходе из строя этих систем и средств связи или уничтожении сберегаемой информации, может нанести убытки экономике страны, сравнимые по серьезности с убытками от применения против США ядерного оружия.

Отсюда следует вывод о том, что наблюдаемые в последние годы тенденции в развитии информационных технологий могут уже в недалеком будущем привести к появлению качественно новых (информационных) форм борьбы, получивших название ***информационной войны***. Следует отметить, что устоявшегося, признанного на международном уровне, определения информационной войны пока нет. Но эксперты определяют суть информаци-

онной войны как достижение какой-либо страной (или группой стран) подавляющего преимущества в информационной области, позволяющего с достаточно высокой степенью достоверности моделировать поведение «противника» и оказывать на него, в явной или скрытой форме, выгодное для себя влияние. Такое определение позволяет утверждать, что страны, проигравшие информационную войну, проигрывают ее навсегда, поскольку их возможные шаги по изменению ситуации, которые сами по себе требуют колоссальных материальных и интеллектуальных затрат, будут контролироваться и нейтрализоваться победившей стороной.

Одним из компонентов ведения информационной войны является так называемое ***информационное оружие***, которое эксперты определяют как совокупность методов и средств, позволяющих похищать, искажать или уничтожать информацию, ограничивать или прекращать доступ к ней законных пользователей; нарушать работу или выводить из строя телекоммуникационные сети или компьютерные системы, используемые в обеспечении жизнедеятельности общества и государства.

Основные объекты применения информационного оружия как в мирное, так и в военное время иллюстрирует рис. 1.11.

Рис. 1.11. Основные объекты применения информационного оружия

Помимо наиболее часто упоминаемой в публикациях военной области применения информационного оружия (для нарушения работы систем командования и управления войсками и боевыми средствами), специалисты также выделяют экономическую, банковскую, социальную и иные сферы его потенциального использования (рис. 1.12).

Сегодня мы уже можем привести примеры использования информационного оружия в различных международных конфликтах. Так, после окончания войны в Персидском заливе («Буря в пустыне, 1991 г.») и Ираке (2005–2006 гг.) США открыто заявили о широкомасштабном использовании информационного оружия в ходе боевых действий. Тогда массированное использование американцами радиоэлектронной разведки и борьбы, а также высокоточного оружия против иракских систем командования и управления

привело к значительному снижению эффективности работы этих систем, существенно повлияв на исход войны.

В качестве информационного оружия рассматриваются возможности создания принципиально новых вирусов и средств внедрения их в компьютерные системы противника. Разрабатываются технологии создания специальных электронных ловушек в микросхемах, используемых как элементная база систем вооружения противника. Такие микросхемы-ловушки, получив конкретную команду, смогут контролировать использование этих систем или нарушать их работу. Аналогичные микросхемы могут устанавливаться и в системах гражданского назначения.

Рис. 1.12. Основные цели применения информационного оружия

Изложенное выше свидетельствует о реальной возможности создания систем и средств ведения информационной войны, что представляло бы серьезную угрозу интересам национальной безопасности. Таким образом, рассматривая проблемы уязвимости защищаемой информации, следует не забывать о подобной возможности и проектировать системы обеспечения безопасности информации с учетом таких перспективных КНПИ.

1.3. Каналы утечки информации

При обработке информации в автоматизированных системах возможна ее утечка по так называемым побочным техническим каналам.

Под ***техническим каналом утечки информации*** понимается совокупность: физических полей, несущих конфиденциальную информацию; кон-

структивных элементов, взаимодействующих с ними; технических средств нарушителя для регистрации поля и снятия информации.

Конфиденциальная информация в таком техническом канале подается в виде сигналов (акустических, виброакустических, электрических, электромагнитных), получивших название *опасных.* В зависимости от физической природы возникающих полей и типа конструктивных элементов, взаимодействующих с ними, можно выделить основные виды технических каналов утечки информации (рис. 1.13).

Рис. 1.13. Виды технических каналов утечки информации

Созданию технических каналов утечки информации способствуют определенные обстоятельства и причины технического характера (рис. 1.14).

Канал побочных электромагнитных излучений и наведения, благодаря своей стабильности и неявной форме получения информации, является одним из основных каналов, по которым технические разведки стараются получить ту или иную закрытую информацию. Как известно, у всех ЭВМ существует проблема излучения высокочастотной электромагнитной энергии, которая может быть перехвачена. Особенно чувствительны к перехвату видеотерминалы.

При этом перехваченные сигналы достаточно просто интерпретировать, отобразив информацию на своем дисплее. Электромагнитные излучения присущи также принтерам, накопителям на магнитных дисках, графическим устройствам и каналам связи компьютерных сетей.

Сигналы от ПЭВМ наводятся в линиях электропитания и во внешних проводниковых линиях. По мере усовершенствования техники роль отдельных ка-

налов утечки изменяется и наблюдаются попытки злоумышленников создавать и использовать новые каналы просачивания информации.

Рис. 1.14. Основные причины создания технических каналов утечки информации

Именно поэтому важно определить источники просачивания информации и возможности ее снятия, имеющиеся у злоумышленника (рис. 1.15).

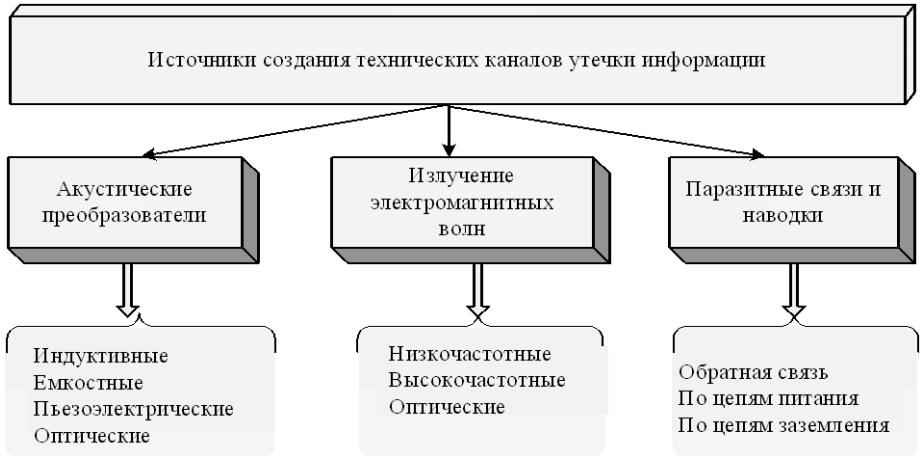


Рис. 1.15. Основные источники создание технических каналов утечки информации

Практически в каждой автоматизированной системе существуют два вида объектов, могущих создавать опасные сигналы и способствующих их распространению, т. е. быть источниками утечки информации. Это технические средства, в которых обрабатывается конфиденциальная информация, а также человек, в разговорной речи которого может содержаться (в виде акустических сигналов) конфиденциальная информация, доступная злоумышленнику по акустическому каналу, каналу проводниковой или радиосвязи при использовании конкретных технических средств.

Из сказанного следует, что все технические средства автоматизированных систем подразделяются на основные (рис. 1.16) и вспомогательные (рис. 1.17).

Рис. 1.16. Перечень основных технических средств автоматизированных информационных систем

Рис. 1.17. Вспомогательные технические средства автоматизированных информационных систем

Приведенный перечень основных технических средств составлен с учетом того факта, что конфиденциальная информация обрабатывается прежде всего этими средствами. При этом, как видим, не упоминаются специальные средства и системы защиты информации, эффективность применения которых настолько высока, что рассматривать их как возможные источники просачивания информации не имеет смысла.

Рассмотрим теперь некоторые технические возможности злоумышленника, учитывая приведенную выше классификацию источников информации.

Контроль акустической информации. Если источник информации - человеческая речь, то для несанкционированного снятия информации могут применяться как различные виды микрофонов, так и технические средства, использующие проводниковую или радиосвязь.

Узконаправленные микрофоны, электронные стетоскопы, лазерные детекторы и записывающие диктофоны могут быть использованы для непосредственного слухового контроля и записи конфиденциальных бесед. При этом при помощи специальных технических средств, таких как, например, аналоговый процессор обработки речевых сигналов, можно значительно улучшить качество снятия информации, повысить комфортность прослушивания зашумленных речевых сигналов. При использовании радиоканала и применении линий электропитания для передачи звуковой информации возможен так называемый электронный контроль речи.

В первом случае для акустического контроля помещений используют миниатюрные радиопередатчики, устанавливающиеся обычно в местах, на которые человек редко обращает внимание. Встречаются микропередатчики, закамуфлированные под обычные предметы, присутствие которых не вызывает подозрений, например под зажигалку, спичечный коробок, пепельницу, настольные письменные принадлежности, авторучку, калькулятор и т.д. Широко известны и миниатюрные карманные передатчики, носимые злоумышленниками. Приборы, в некоторых случаях осуществляющие передачу радиосигналов, устанавливают капитально во время строительства или ремонта стен помещений или их облицовки. Режим работы микропередатчиков может быть непрерывным и с контролируемым временем включения.

Во втором случае звуковая информация при помощи специальных технических средств передается за пределы помещений по линиям электропитания. Сигнал от передатчика к приемнику передается по цепям электропитания в ультразвуковом диапазоне частот. При этом дальность действия системы «передатчик - приемник» ограничивается, естественно, одной трансформаторной развязкой линии электропитания.

Несмотря на кажущуюся простоту обращения с радиомикрофонами (установка, включение) необходимо учитывать, что факту их применения предшествует большая, сложная и хорошо спланированная работа. Предусматривается определение соответствующих помещений, предварительный выбор места установки, подбор исполнителей и размещение принимающей

фиксирующей аппаратуры.

Контроль информации техническими средствами в каналах телефонной связи. Следует заметить, что каналы телефонной связи наиболее уязвимы. Прослушивание разговоров в помещениях, а также телефонных бесед осуществляется разными методами (рис. 1.18).

Рис. 1.18. Основные методы прослушивания разговоров в помещениях (или телефонных разговоров)

Контроль информации, обрабатываемой средствами вычислительной техники. Современный уровень развития электроники и радиотехники позволяет изготавливать устройства перехвата информации, выделяемой из побочных электромагнитных излучений работающего компьютера (в частности, через незащищенные цепи питания и заземление), а также устройства, обеспечивающие секретность их работы. Применяя эти устройства, можно фиксировать информацию, циркулирующую в компьютере, на расстоянии до нескольких сотен метров.

Существующие системы снятия информации учитывают особенности структуры сигналов в канале связи, адресные сигналы, особенности информации, что существенным образом способствует решению проблем промышленного шпионажа. При этом следует иметь в виду, что проще всего перехватить информацию, отображаемую на экране дисплея компьютера. Эта информация может быть восстановлена даже с помощью обычного черно-белого телевизионного приемника.

В случае перехвата информации с использованием побочных электромаг-

нитных излучений и наведений злоумышленнику, кроме хорошего технического оснащения, нужны еще определенные алгоритмические (аналитические) навыки. Речь идет о разработке алгоритмов предварительной обработки принятых сигналов, состоящих из совокупности полезных сигналов и помех, с целью выделения сигналов, несущих информацию.

К основным способам предотвращения просачивания информации по техническим каналам можно отнести организационные факторы и использование различных технических средств защиты. При этом эффективная защита достигается благодаря комплексному применению указанных подходов.

Рассмотрим сначала технические средства защиты, поскольку вопросам организационно-правового характера посвящен отдельный раздел этого учебника.

Все технические средства используются, чтобы или выявить факт снятия информации, или его предотвратить. В данный момент существуют три направления реализации указанного:

выявление активных средств секретного снятия акустической информации (радиомикрофонов, микрофонов с передачей информации по цепям электросети переменного тока, по радиотрансляционным и другим проводниковым сетях, телефонных передатчиков с передачей информации по радиоканалу, радиостетоскопов и т.д.);

постоянный или периодический контроль загрузки радиодиапазона (радиомониторинг), выявление и анализ новых излучений, потенциально и специально организованных радиоканалов просачивания информации (например, цифровых устройств с радиозакладкой или устройств с накоплением и дальнейшей передачей);

проведение специальных исследований систем обработки конфиденциальной информации с целью определения каналов утечки, уровня защищенности информации и дальнейшей реализации мер по обеспечению выполнения требований защиты информации.

Рассмотрим подробнее соответствующие технические решения.

Защита от просачивания информации по акустическому каналу. Следует сразу указать, что выявить наличие акустического контроля с помощью узконаправленных микрофонов, электронных стетоскопов и лазерных детекторов довольно сложно. Поэтому с целью защиты чаще всего используются средства предотвращения снятия информации. К ним относятся генераторы аудиопомех, вырабатывающие шумовой сигнал-помеху с изменяемыми амплитудой и частотой; они могут быть портативными (карманными) и стационарными.

Для защиты от контроля акустической информации в исключительных случаях используются специальные, прозрачные защитные кабины, гарантирующие защиту от любых видов прослушивания. В качестве материала для указанных кабин и внутренней мебели используют прозрачный пластик. Такие кабины применяются как наиболее эффективное средство защиты от про-

слушивания в посольствах ведущих стран мира. Существенно шире спектр технических средств, предназначенных для выявления специальных устройств электронного контроля речи (рис. 1.19).

Рис. 1.19. Основные технические средства выявления специальных устройств электронного контроля речи

Защита информации в каналах связи. К техническим средствам защиты информации в каналах связи можно отнести приборы, устанавливающие факт подключения к телефонным каналам подслушивающих устройств, а также спектральные анализаторы каналов связи и устройства защиты конфиденциальных разговоров по телефонным каналам. Современные спектральные анализаторы каналов связи, как правило, являются комбинированными приборами, которые также решают задачи радиомониторинга.

В речевых системах связи известны два основных метода закрытия каких-либо сигналов; они различаются способом передачи по каналам связи: аналоговое скремблирование и дискретизация речи с дальнейшим шифрованием. Под *скремблированием* понимается изменение характеристик речевого сигнала таким образом, что полученный модулированный сигнал, имея свойства неразборчивости и нераспознанности, занимает такую же полосу частот, как и начальный открытый речевой сигнал.

В системах дискретизации речевые компоненты с помощью аналого-цифрового преобразователя превращаются в цифровой поток данных, который

смешивается по определенным алгоритмам с псевдослучайной последовательностью, вырабатываемой ключевым генератором по одному из криптографических алгоритмов; полученное таким образом закрытое речевое сообщение передается с помощью модема в канал связи. На приемной стороне проводится обратное преобразование с целью получения открытого речевого сигнала.

Защита информации от утечки по каналу побочных электромагнитных излучений. Для защиты информации от утечки за счет побочных электромагнитных излучений применяют пассивный, активный и комбинированный методы. Пассивная защита состоит в снижении уровней излучения до значений, соизмеримых с естественными шумами, с помощью специальной элементной базы и конструктивной доработки техники, обрабатывающей конфиденциальную информацию. Существуют различные способы реализации этого метода. Одно из простейших технических решений заключается в размещении всего оборудования в безопасной среде, экранирующей радиоизлучение. К таким мерам прибегают в случае малогабаритной аппаратуры, когда это не приводит к чрезмерным расходам. Для больших систем экранирование целых залов и даже зданий может быть чрезвычайно дорогим, поэтому проблемы обеспечения электронной защиты для них рассматриваются на стадии проектирования. Например, для системы связи определяются требования относительно безопасности отдельных компонентов каждой ее секции. Разработчик может спланировать экранирование отдельных устройств системы с помощью металлического защитного покрытия или воспользоваться стандартными экранированными корпусами для блоков аппаратуры. Там, где экранирование компонентов нецелесообразно, предполагается достаточная изоляция линий данных и питания за счет различных объединений фильтров, устройств глушения сигнала, низкоимпедансного заземления, трансформаторов развязки. Должны экранироваться также кабели. При этом наилучший вариант защиты линий связи - применение отповолоконной технологии. Надежное экранирование абонентской аппаратуры связи чрезвычайно усложняет задачи электронного подслушивания.

Допустимые уровни излучений аппаратуры и меры защиты информации регламентируются специальными стандартами.

Активная защита предусматривает утаивание информационных сигналов за счет шумовой или заградительной помехи с помощью специальных генераторов шума. Активная радиотехническая маскировка состоит в формировании и излучении маскировочного сигнала в непосредственной близости от маскированной системы. При этом различают энергетический и неэнергетический методы активной радиотехнической маскировки.

В случае энергетической маскировки образовывается широкополосный шумовой сигнал с уровнем, во всем частотном диапазоне существенным образом превышающим уровень излучения системы. Одновременно происходит наводка шумовых колебаний в создаваемых цепях. Энергетическую маскировку можно реализовать только тогда, когда уровень излучений существенным образом меньше уровня, установленного действующими стандартами на элек-

тромагнитную совместимость, а также медицинскими требованиями. Иначе устройство маскировки либо будет создавать помехи различным радиоприборам, расположенным вблизи подлежащей защите системы, либо его нельзя будет использовать из-за опасности для здоровья человека.

Неэнергетический (статистический) метод активной радиотехнической маскировки заключается в изменении вероятностной структуры сигнала, который может быть принят приемником злоумышленника. Для такого изменения сигнала необходимо специальное устройство, встраиваемое непосредственно в систему или размещенное рядом с ней. Уровень излучаемого этим устройством маскировочного сигнала не превышает уровня информативного излучения системы, а потому такие устройства не создают ощутимых помех для других электронных приборов, расположенных рядом, и безопасны для здоровья оператора системы.

Комбинированная защита - это снижение уровней излучения до заданных значений с одновременным использованием как пассивной, так и активной защиты.

1.4. Принципы построения систем комплексной защиты информации

Эффективное обеспечение защиты информации в автоматизированных системах возможно только на основе комплексного использования всех известных методов и подходов к решению этой проблемы. *Концепция* такой *комплексной защиты* должна удовлетворять рассмотренную ниже совокупность требований.

Во-первых, должны быть разработаны и приведены к уровню регулярного использования все необходимые механизмы гарантированного обеспечения требуемого уровня защиты информации.

Во-вторых, должны существовать механизмы практической реализации требуемого уровня защиты информации.

В-третьих, нужно иметь в своем распоряжении средства рациональной реализации всех необходимых мероприятий по защите информации на базе достигнутого уровня развития науки и техники.

В-четвертых, должны быть разработаны способы оптимальной организации и обеспечения проведения всех мер по защите информации в процессе ее обработки.

С целью построения концепции, удовлетворяющей всю совокупность перечисленных требований, ученые активно разрабатывают теорию защиты информации, которая предлагает систему концептуальных решений (рис. 1.20), содержание которых в общем виде состоит в следующем.

Под ***функцией защиты*** понимают совокупность функционально однородных мер, регулярно осуществляемых в автоматизированных системах различными средствами и методами с целью создания, поддержки и обеспечения условий, объективно необходимых для надежной защиты информации. Для того чтобы множество функций отвечало своему назначению, они должны удовлетворять требованию полноты. В этом случае, при обеспечении

надлежащим образом соответствующего уровня выполнения каждой из функций указанного множества, можно гарантировано достичь необходимого уровня защиты информации.

Рис. 1.20. Концептуальные решения теории защиты информации

Проанализировав конкретные, возможные в процессе защиты информации ситуации, получим перечень *полного множества функций защиты:*

предотвращение возникновения условий, оказывающих содействие порождению (возникновению) дестабилизирующих факторов;

предотвращение непосредственного проявления дестабилизирующих факторов;

выявление проявившихся дестабилизирующих факторов;

предотвращение действия на подлежащую защите информацию:

проявленных и выявленных дестабилизирующих факторов;

проявленных, но не выявленных дестабилизирующих факторов;

выявление действия дестабилизирующих факторов на подлежащую защите информацию;

локализация (ограничение):

выявленного действия на информацию дестабилизирующих факторов;

не выявленного действия на информацию дестабилизирующих факторов;

ликвидация последствий:

локализованного выявленного действия на информацию дестабилизирующих факторов;

локализованного не выявленного действия на информацию дестабилизирующих факторов.

Чтобы доказать полноту множества перечисленных функций, рассмотрим приведенные на рис. 1.21 объединение событий, которые потенциально возможны при осуществлении всех функций. Каждый из результатов является событием случайным, а все вместе они образовывают полную группу несовместимых событий.

Как известно из теории вероятностей, сумма вероятностей таких событий равна 1. Благоприятными с точки зрения защиты информации есть результаты 1 - 6, поэтому сумма их вероятностей будет вероятностью того, что защищенность информации обеспечена. Обозначив эту вероятность через $P_З$, получим

$$P_З = \sum_{m=1}^{6} P_m^i. \tag{1.9}$$

Обозначим через $P_k^{(У)}$ вероятность успешного осуществления r-й функции, тогда для P_k^i будут выполняться следующие выражения:

$$P_1^{(i)} = P_1^{(У)}; \tag{1.10}$$

$$P_2^{(i)} = (1 - P_1^{(У)}) P_2^{(У)}; \tag{1.11}$$

$$P_3^{(i)} = (1 - P_1^{(У)})(1 - P_2^{(У)}) P_3^{(У)} P_{4а}^{(У)}; \tag{1.12}$$

$$\begin{aligned} P_4^{(i)} &= (1 - P_1^{(У)})(1 - P_2^{(У)}) P_3^{(У)} (1 - P_{4а}^{(У)}) + \\ &+ (1 - P_3^{(У)})(1 - P_{4б}^{(У)}) P_5^{(У)} P_{6а}^{(У)} P_{7а}^{(У)}; \end{aligned} \tag{1.13}$$

$$\begin{aligned} P_5^{(i)} &= (1 - P_1^{(У)})(1 - P_2^{(У)}) P_3^{(У)} (1 - P_{4а}^{(У)}) + \\ &+ (1 - P_3^{(У)})(1 - P_{4б}^{(У)})(1 - P_5^{(У)}) P_{6б}^{(У)} P_{7б}^{(У)}; \end{aligned} \tag{1.14}$$

$$P_6^{(i)} = (1 - P_1^{(У)})(1 - P_2^{(У)})(1 - P_3^{(У)}) P_{4б}^{(У)}. \tag{1.15}$$

Подставив выражения (1.11) - (1.15) в (1.10), получим

$$P_З = F(\{P_r^{(У)}\}). \tag{1.16}$$

Итак, *защищенность информации целиком и полностью определяется вероятностью успешного осуществления функций защиты*.

Поэтому, чтобы обеспечить уровень защищенности информации, который равен $\overline{P}_З$, необходимо выбрать такие совокупности мер по осуществлению каждой из функций защиты, при которых

$$F(\{P_r^{(У)}\}) \geq \overline{P}_З. \tag{1.17}$$

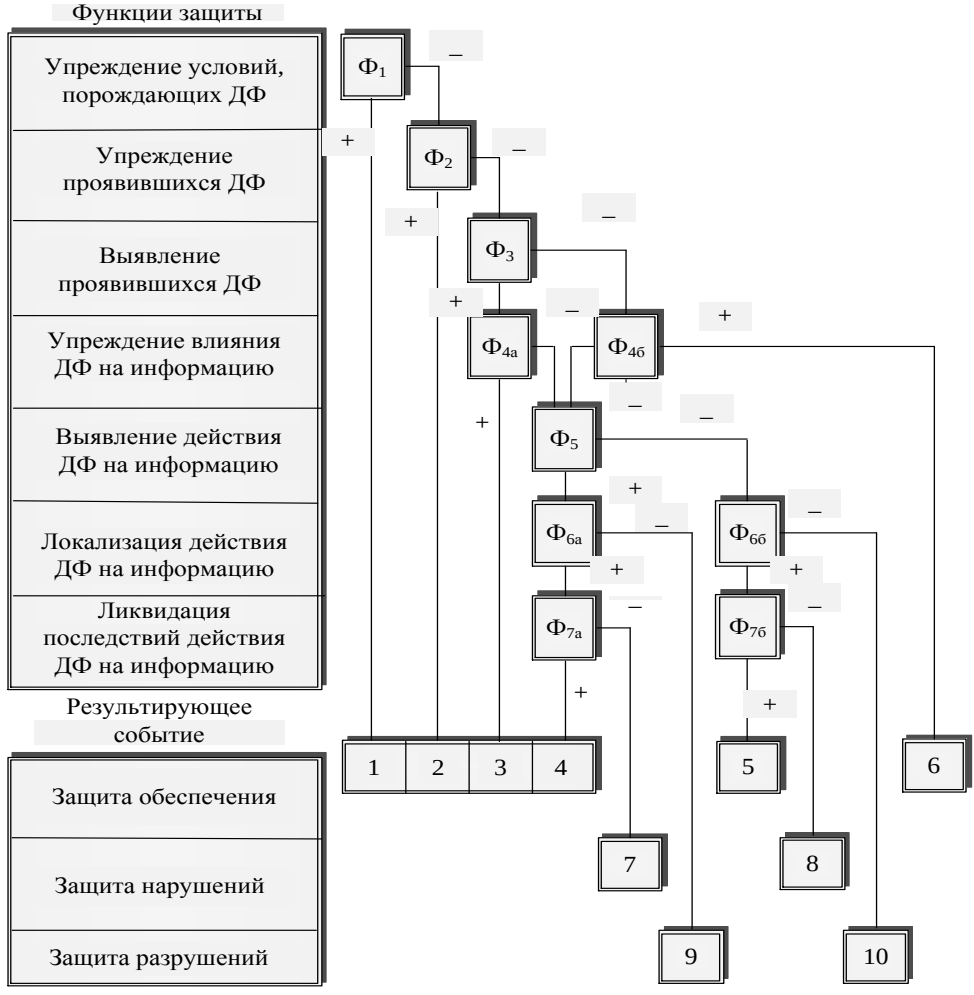


Рис. 1.21. Схема функций и результатов защиты информации
(ДФ – дестабилизирующий фактор)

Таким образом, можно считать доказанной *полноту множества функций защиты информации*, но с точностью до условия, что для каждой функции можно выбрать необходимые меры.

Полнота множества функций защиты имеет принципиальное значение еще и с точки зрения создания предпосылок для *оптимизации систем защиты информации*. Осуществление функций защиты информации связано с расходованием тех или иных ресурсов. Поэтому уровень осуществления каждой из функций защиты, при других равных условиях, будет зависеть от количества затрачиваемых ресурсов. Если количество ресурсов (например, в стоимостном выражении), затрачиваемых на осуществление *r*-й функции, обозначить через

$C_r^{(\text{у})}$, то

$$P_r^{(\text{у})} = \varphi_r C_r^{(\text{у})}. \tag{1.18}$$

Тогда зависимость (1.16) можно представить в таком виде:

$$P_{\text{з}} = F\left[\varphi_r(\{C_r^{(\text{у})}\})\right]. \tag{1.19}$$

Учитывая это, задачу защиты информации можно сформулировать как оптимизационную задачу: найти такие $C_r^{(\text{у})}$, при которых выполняются условия

$$\left.\begin{aligned} F\left[\varphi_r(\{C_r^{(\text{у})}\})\right] &\geq P_{\text{з}} \\ C = \sum_{\forall r} C_r^{(\text{у})} &\Rightarrow \min \end{aligned}\right\}, \tag{1.20}$$

или

$$\left.\begin{aligned} C = \sum_{\forall r} C_r^{(\text{у})} &\leq \bar{C} \\ P_{\text{з}} = F\left[\varphi_r(\{C_r^{(\text{у})}\})\right] &\Rightarrow \max \end{aligned}\right\}. \tag{1.21}$$

Здесь C — допустимый уровень затрат на защиту информации.

Несложно увидеть, что первая постановка адекватна тому варианту, когда заданный уровень защиты информации непременно должен быть достигнут, причем желательно при минимально возможных затратах; вторая постановка - вариант, когда затраты на защиту информации ограничены некоторым уровнем, а естественное желание при этом - достижение максимально возможного уровня защищенности информации.

Осуществление функций защиты в автоматизированных системах достигается решением задач защиты, под которыми понимаются возможности средств, методов и мер, осуществляемых в автоматизированных системах с целью полной или частичной реализации одной или нескольких функций защиты. Основными требованиями, выдвигаемыми к множеству таких задач, являются *репрезентативность* и *реализуемость*. Под *репрезентативностью* понимается достаточность указанных задач для обеспечения необходимого уровня и эффективности осуществления всех функций; а под *реализуемостью* - возможность выполнения имеющимися средствами и методами.

Все задачи, необходимые для осуществления функций обеспечения защиты, можно объединить в 10 классов (рис. 1.22).

Весьма сложной и практически нерешенной является проблема оценки эффективности функций защиты информации при выполнении той или иной задачи защиты или некоторой совокупности этих задач. Учитывая зависимость самого процесса защиты информации от влияния случайных факторов, существуют различные неформальные методы оценивания эффективности такой защиты и поиска решений.

Рис. 1.22. Основные задачи обеспечения защиты в информационных системах

Рассмотрим *содержание третьего концептуального решения теории защиты информации*.

Для решения любой задачи в автоматизированной системе должны и быть предусмотрены адекватные по смыслу и достаточные по количеству средства. Сейчас уже разработан весьма представительный по номенклатуре арсенал различных средств защиты информации. Множество разнообразных возможных средств защиты определяется, прежде всего, способами действия на дестабилизирующие факторы или порождающие их причины. Выделяют следующие классы средств защиты (рис. 1.23).

Рис. 1.23. Основные классы средств защиты

Физические средства — это механические, электрические, электромеханические, электронные, электронно-механические и другие устройства и системы, функционирующие автономно, создавая разного рода помехи дестабилизирующим факторам.

Аппаратные средства — различные электронные, электронно-механические и другие устройства, схематически встраиваемые в аппаратуру автоматизированных систем или соединяемые с ними специально для выполнения задач защиты информации.

Программные средства — специальные пакеты программ или отдельные программы, используемые для выполнения задач защиты.

Организационные средства — организационно-технические меры, специально предполагаемые в автоматизированных системах с целью выполнения задач защиты.

Нормативно-правовые средства — законы и другие нормативно-правовые акты, регламентирующие права и обязанности всех лиц и подразделений; касающиеся функционирования автоматизированной системы, в которой присутствует информация ограниченного доступа; устанавливающие ответственность за действия, следствием которых может быть нарушение защищенности информации.

Инженерно-технические средства — это совокупность и взаимодействие специальных подразделений, технических средств и мер.

К специфическим средствам защиты информации принадлежат ***криптографические методы.*** В автоматизированных системах криптографические методы защиты информации могут использоваться как для защиты обрабатываемой информации в компонентах системы, так и для защиты информации, передаваемой по каналам связи. Собственно преобразование информации может осуществляться аппаратными или программными средствами.

Итак, существуют все объективные предпосылки для разработки необходимого арсенала средств защиты. А чтобы разрешить вопрос о достаточном их количестве, необходимо иметь данные об эффективности использования различных средств при выполнении разнообразных задач. Указанные данные можно получить, организовав широкомасштабный сбор и статистическую обработку информации по этим вопросам на реально существующих и функционирующих системах защиты, а также обратившись к экспериментам или экспертным оценкам.

Четвертым концептуальным решением теории защиты информации является введение понятия ***системы защиты информации***, которая *определяется как организованная совокупность всех средств, методов и мер, назначаемых в автоматизированной системе для выполнения в ней тех или иных задач защиты*. Введением понятия *системы* защиты информации делается акцент на том, что все ресурсы, предназначенные для защиты информации, должны объединяться в единую, целостную систему, являющуюся функционально самостоятельной подсистемой автоматизированной системы.

Важнейшим концептуальным требованием к системе защиты информации (СЗИ) является требование адаптированности, т.е. способности приспосабливаться к изменениям структуры, технологических схем или условий функционирования автоматизированной системы.

Кроме общего концептуального требования к СЗИ выдвигается еще ряд конкретных целевых требований (рис. 1.24):

функциональные — обеспечение выполнения необходимой совокупности задач защиты, удовлетворение всех требований защиты;

эргономичные — минимизация препятствий пользователям, удобство для персонала СЗИ;

экономические — минимизация затрат на систему, максимальное использование серийных средств;

технические — комплексное использование средств защиты, оптимизация архитектуры;

организационные — структурированность всех компонентов, простота эксплуатации.

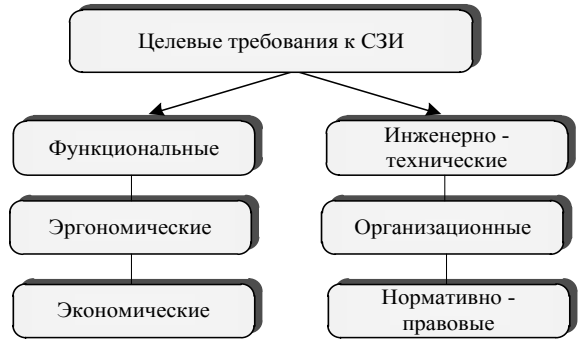


Рис. 1.24. Основные требования к СЗИ

Большое значение имеют типизация и стандартизация СЗИ. При этом с целью создания наилучших предпосылок для их практической реализации целесообразно выделить три уровня стандартизации: высший — уровень СЗИ в целом, средний — уровень компонентов СЗИ и низший — уровень проектных решений относительно средств и механизмов защиты.

Последним концептуальным решением является введение понятия *управления деятельностью системы защиты информации*, определяемый как частный случай управления в системах организационно-технологического типа. Для систем такого типа полное множество образовывают следующие четыре функции управления (рис. 1.25):

планирование, т.е. разработка рациональной программы будущих действий;

оперативно-диспетчерское управление, т.е. регулирование быстротекущих процессов в реальном масштабе времени;

календарно-плановое управление, т.е. периодический контроль выполнения плана и принятия (при необходимости) управленческих решений;

обеспечение повседневной деятельности системы управления, т.е. предоставление органам этой системы ресурсов, необходимых для эффективного управления.

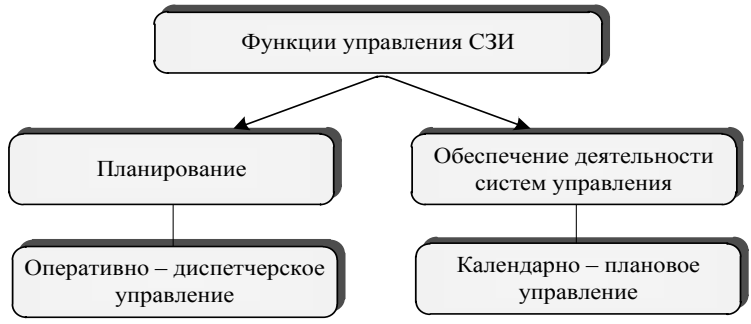


Рис. 1.25. Функции управления систем защиты информации

Проектирование систем защиты информации заключается в том, чтобы для заданной автоматизированной системы (или ее проекта) создать оптимальные механизмы обеспечения защиты информации и механизмы управления ими. При этом оптимальность СЗИ понимается в общепринятом смысле: достижение заданного уровня защищенности информации при минимальных затратах, или достижение максимально возможного уровня защищенности при заданном уровне затрат на защиту.

Собственно ***методология проектирования СЗИ*** полностью вписывается в общую методологию проектирования больших систем организационно-технологического типа. Общее правило, которым при этом следует руководствоваться, заключается в необходимости по возможности шире использовать типичные проектные решения.

Начальным пунктом проектирования является формирование требований по защите информации в автоматизированных системах. Прежде всего для каждого элемента автоматизированной системы, имеющей самостоятельное территориальное размещение, должна быть определена категория по необходимой защищенности: *слабая*, *сильная*, *очень сильная* и *особая защита*. Критериями для такого определения категории является степень секретности подлежащей защите информации, ее объем и условия обработки с точки зрения потенциальных возможностей несанкционированного доступа.

Затем определяются конкретные требования по совокупности следующих факторов:

характер обрабатываемой информации;

объем обрабатываемой информации;

продолжительность пребывания информации в автоматизированной систе-

ме;

структура автоматизированной системы;

вид защищаемой информации;

технология обработки информации;

организация информационно-вычислительного процесса в автоматизированной системе;

этап жизненного цикла автоматизированной системы.

Кроме этого, необходимо определить также значения важнейших параметров самой информации. К таким параметрам можно отнести (рис. 1.26):

адекватность, т.е. соответствие текущему состоянию описываемых информацией объектов или процессов;

релевантность информации, т.е. соответствие ее целевому назначению;

толерантность, т.е. удобство использования с точки зрения выполняемых задач;

важность, значимость с точки зрения тех задач, для выполнения которых используется оцениваемая информация;

значимость, или полнота, информации для информационного обеспечения выполняемых задач;

способ кодирования информации;

объем информации.

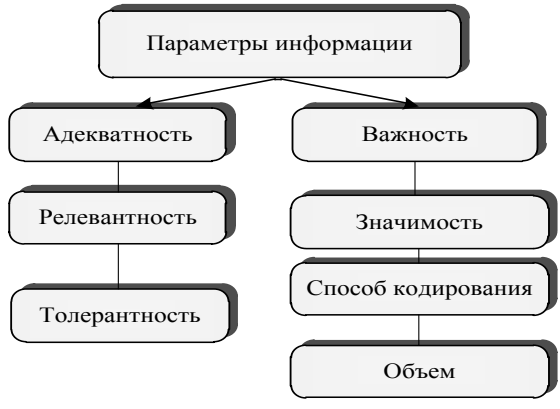


Рис.1.26. Важнейшие параметры информации

Для определения указанных показателей разработаны достаточно строгие алгоритмические процедуры, дающие возможность получать количественные оценки, хотя и в относительном выражении [8]. Необходимый уровень защиты информации должен определяться с учетом значений всех описанных прежде показателей. При этом можно воспользоваться такой процедурой:

все показатели делятся на три категории: определяющие, весомые и второстепенные;

необходимый уровень защиты рассчитывается по определяющим показателям;

выбранный уровень при необходимости можно скорректировать с учетом значений весомых показателей.

Значение второстепенных показателей при этом могут игнорироваться. Вариант классификации показателей подлежащей защите информации приведен в табл. 1.3.

Функционирование СЗИ организовывается соответственно принципам управления защитой информации. Как известно из теории управления, *процессы управления могут быть классифицированы как коротко-, средне- и долгосрочные.*

Отличительные особенности указанных видов управления относительно управления защитой информации такие:

для *краткосрочного управления* — использовать можно лишь те средства защиты, которые включены в состав СЗИ и находятся в работоспособном состоянии; в общей совокупности процессов управления большой удельный вес должны иметь процедуры оперативного реагирования в случае возникновения непредусмотренных ситуаций; архитектура СЗИ и самой автоматизированной системы, а также технологические схемы их функционирования изменению не подлежат, возможное только включение или исключение уже имеющихся компонентов;

для *среднесрочного управления* — использовать можно весь арсенал имеющихся средств защиты, которые вводятся в планированный период; основными процедурами управления будут планирование и обеспечение защиты; значительное место в процессах управления будет занимать анализ эффективности управления и разработка на этой основе предложений относительно развития средств и методов защиты; архитектура СЗИ и автоматизированной системы важным изменениям не подлежит, тем не менее возможны некоторые изменения в пределах имеющихся структурных элементов; технология функционирования СЗИ и автоматизированной системы может изменяться; могут формироваться предложения относительно усовершенствования и развития архитектуры СЗИ и автоматизированной системы;

для *долгосрочного управления* — основными процессами является перспективное планирование развития и использование средств защиты; большое внимание должно отводиться развитию и усовершенствованию концепции защиты; при необходимости может существенным образом изменяться как архитектура, так и технология функционирования СЗИ и автоматизированной системы.

Как уже подчеркивалось, управление защитой информации является отдельным случаем управления в системах организационно-технологического типа, причем основными функциями управления являются планирование, оперативно-диспетчерское управление, календарно-плановое руководство и обеспечение повседневной деятельности СЗИ.

Основными показателями оперативно-диспетчерского управления при этом должны быть:

беспрерывное слежение за функционированием механизмов защиты;

систематический анализ текущего состояния защищенности информации;

своевременное принятие мер в случае нарушения или угрозы нарушения защищенности информации;

сбор и накопления информации о функционировании СЗИ.

С целью обеспечения высокой эффективности оперативно-диспетчерского управления все его процедуры должны быть заранее по возможности полнее структурированы, а правила сбора и обработки соответствующих решений - строго определены.

Таблица 1.3

Показатель информации	Категория показателя			
	Вид тайны содержащейся в информации			Защита информации как товара
	Военная, государственная, научная	Промышленная, коммерческая	Конфиденциальная	
Важность	Определяющий	Определяющий	Определяющий	Определяющий
Полнота	Важный	Важный	Определяющий	Определяющий
Адекватность	Важный	Важный	Важный	Определяющий
Релевантность	Второстепенный	Важный	Важный	Важный
Толерантность	Второстепенный	Второстепенный	Второстепенный	Важный
Способ кодирования	Второстепенный	Второстепенный	Второстепенный	Важный
Объём	Второстепенный	Важный	Важный	Определяющий

Календарно-плановое управление защитой информации имеет целью организацию и обеспечение выполнения плановых мероприятий по защите информации, а также при необходимости - корректирование плана. Важнейшей задачей календарно-планового управления является контроль защищенности информации.Под обеспечением повседневной деятельности СЗИ понимается совокупность мер, осуществляемых с целью следующих выполнения задач (рис. 1.27). Дополнительно укажем, что исключительное значение имеет задача сбора, накопления и аналитико-синтетической обработки всех данных, касающихся защиты информации.

Итак, защита информации в современных автоматизированных системах является крупномасштабной и весьма сложной проблемой. Естественно, что такой масштабной постановке задачи должна отвечать не менее масштабная программа реализации концепции защиты, которая опирается на развитую инфраструктуру как отдельных предприятий и организаций, так и государства в целом. Возникает необходимость создания различных в функцио-

нальном плане органов защиты, основу системы которых могут представлять *отраслевые и территориальные центры защиты информации* (ЦЗИ).

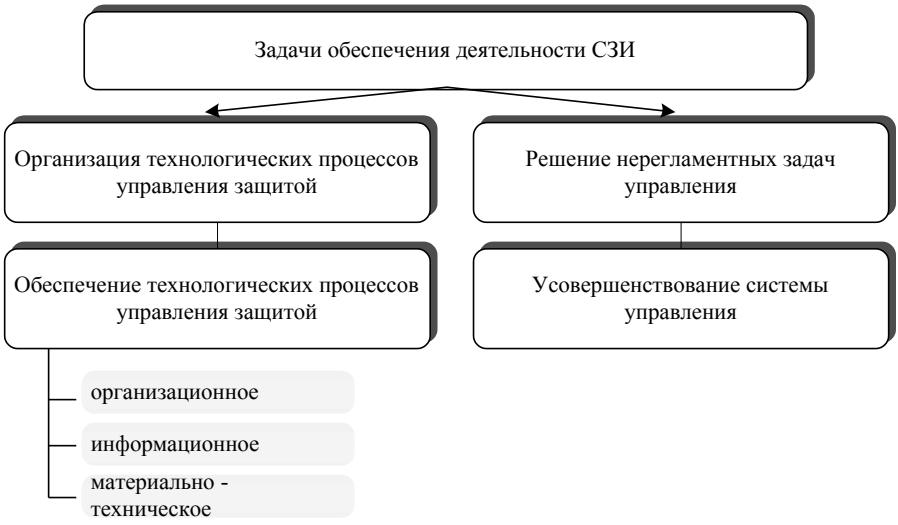


Рис. 1.27. Основные задачи обеспечения деятельности СЗИ

Вся работа сети ЦЗИ с целью повышения ее эффективности должна координироваться центром, роль которого может выполнять одна из государственных структур, ответственных за обеспечение информационной безопасности государства, или объединение таких структур (рис. 1.28).

Рис. 1.28. Основные функции главного ЦЗИ

Территориальный, или ведомственный, ЦЗИ должен быть специализированным научно-производственным предприятием, профессионально ориентированным на разработку, практическую реализацию и внедрение концептуальных решений в сфере защиты информации, а также конкретных средств, методов и мер защиты (рис. 1.29).

Рис. 1.29. Основные функции территориальных ЦЗИ

Центральной функцией всей системы должна стать работа с абонентами, направленная на предоставление им услуг по защите информации (рис. 1.30).

Одним из важнейших условий эффективного решения проблем защиты информации является создание регулярной системы сбора, накопления и аналитико-синтетической обработки данных, касающихся указанных проблем.

Такую работу должны выполнять все ЦЗИ при координирующей роли главного центра. Совокупность указанных данных должна быть достаточной для выполнения всех рассмотренных в предыдущих подразделах задач, связанных с защитой информации любого уровня конфиденциальности.

Важную роль в практической реализации концепции комплексной защиты информации сыграет также *организационно-правовое обеспечение*, *которое должно быть высокоупорядоченной совокупностью организационных решений*, *законов*, *нормативов и правил*, *регламентирующих не только общую организацию работ по защите информации*, *а и создание и функционирование СЗИ в конкретных автоматизированных системах*.

При этом первостепенное значение приобретает разработка типичных документов по защите, основное назначение которых:

обеспечивать научно-методологическое и концептуальное единство при решении всех вопросов защиты информации;

создавать условия для однозначного понимания и практической реализации основных положений концепции защиты информации;

обеспечивать необходимыми методиками и данными всех органов и лиц, принимающих участие в решении вопросов защиты информации;

обеспечение нормативно-правового регулирования процессов защиты информации.

Рис. 1.30. Основные направления услуг ЦЗИ

Все *документы должны образовывать единую систему, основными группами которой являются справочно-информационные документы, стандарты, руководящие методические материалы, инструкции.*

К *группе справочно-информационных* принадлежат такие документы, которые в систематизированном виде содержат полную совокупность сведений, необходимых и достаточных для создания четкого и однозначного представления обо всех аспектах проблемы защиты и, одновременно, о признанных специалистах-профессионалах в сфере защиты информации.

К *группе стандартов* принадлежат, естественно, документы, являющиеся образцом, эталоном в смысле совершенства по всем основным параметрам, имеющие соответствующий сертификат и утвержденные полномочным органом.

К *руководящим методическим материалам* принадлежит совокупность документов, содержащих полное и систематизированное описание соответствующих, касающихся защиты информации, вопросов, и утвержденных полномочными органами (тем не менее, в отличие от стандартов, утвержде-

ние имеет лишь рекомендательный характер).

К *инструкции* отнесены систематизированные наборы типичных, касающихся защиты информации, инструкций для различных категорий подразделов и лиц.

Одним из принципиальных положений концепции комплексной защиты информации является *наличие обратной связи* от конструктивных компонентов концепции к ее начальной основе, т.е. к концепциям построения и организации функционирования автоматизированной системы (АС). Основное содержание указанной обратной связи представляют условия, выполнение которых объективно оказывает содействие наиболее эффективному выполнению задач защиты. Всю совокупность этих условий можно подразделить на три класса: общеметодологические, организационные и конструктивные.

Общеметодологические условия создают общие предпосылки для повышения эффективности управления качеством информации. В этом классе выделяют две группы условий: первую — осознание проблемы, вторую — наличие предпосылок относительно принятия решения.

Организационные условия предусматривают разработку и реализацию четкой организации как архитектурного построения АС, так и технологии автоматизированной обработки информации. Основными группами условий этого класса являются структурно-функциональная однозначность компонентов АС и организационное единство управления обработкой информации.

Конструктивные условия предусматривают учет требований защиты в архитектурном построении АС и технологических схемах ее функционирования. Условия этого класса наиболее важны, они делятся на три группы: концептуальная стандартизация, структуризация компонентов АС и структуризация технологии обработки информации.

Под *концептуальной стандартизацией* понимается стандартизация на уровне концепций, общих принципов и правил организации и обеспечения соответствующего вида деятельности. Если, например, речь идет о такой сфере деятельности, как организационно распорядительное управление, то концептуальная стандартизация должна предусмотреть:

во-первых, структуризацию концепции управления;

во-вторых, структуризацию концепций комплексной автоматизации обработки данных в системах управления;

в-третьих, структуризацию концепций организации ресурсов вычислительной техники, необходимой для комплексной автоматизации.

При этом структуризация перечисленных концепций может осуществляться не изолировано, а взаимосвязано.

Структурированность компонентов АС является одним из важнейших конструктивных условий. При этом основное внимание отводится структуризации математического, программного и информационного обеспечения.

Принципиальным является также условие структурированности технологических схем автоматизированной обработки данных в АС. Объективные предпосылки такой структуризации состоят в том, что любую технологическую схему обработки информации можно представить как совокупность участков трех типов: линейных, разветвленных и циклических. Используя это положение, можно структурировать практически все возможные в современных АС схемы обработки информации. Больше того, проведенные в последнее время исследования показали, что на этой основе можно построить полностью структурированную и универсальную технологию, пригодную для широчайшего практического использования. Такую технологию можно назвать ***унифицированной технологией автоматизированной обработки информации*** (УТАОИ).

Чтобы быть достаточно универсальной, УТАОИ должна удовлетворять определенную совокупность требований (рис. 1.31). Чтобы быть полностью структурированной, УТАОИ, кроме общих условий структурированности, должна удовлетворять еще ряду условий:

сквозное модульное построение на всех уровнях (от самой обобщенной схемы к блок-схеме элементарных технологических операций);

автономность организации всех модулей на всех уровнях;

структуризация всех видов обеспечения.

Рис. 1.31. Основные требования к УТАОИ

При таком построении УТАОИ создает наиболее благоприятную среду для защиты информации. Строгая функциональная определенность, однозначность и специализация каждого этапа УТАОИ создают предпосылки для целенаправленной организации защиты информации на каждом этапе обработки и в интерфейсах. Другими словами, УТАОИ создает предпосылки для реализации предупредительной стратегии защиты информации. На базе УТАОИ можно создать эталонную защищенную информационную технологию, в которой будут синтезированы все достижения в сфере обработки и защиты информации.

1.5. Критерии оценки помехоустойчивости информационных систем

Помехоустойчивость *определяется как способность информационной системы противостоять воздействию разнообразных помех.* В результате действия помех принятое сообщение в определенной мере будет отличаться от переданного.

Поэтому *помехоустойчивость можно характеризовать как степень соответствия принятого сообщения переданному при заданной помехе*. При сравнении нескольких систем более помехоустойчивой будет та, которая при одинаковой помехе обеспечит меньшее отличие между принятым и переданным сообщениями.

Для характеристики степени соответствия принятого сообщения переданному введена *количественная мера, которая выбирается в зависимости от характера сообщений.*

При передаче беспрерывных сообщений, как правило, используется *критерий среднеквадратичного отклонения принятого сообщения* $Y(t)$ *от переданного* $X(t)$:

$$\sigma = \sqrt{\overline{\left|Y(t) - X(t)\right|^2}}. \tag{1.22}$$

Применяется также *критерий абсолютного отклонения*

$$\delta_{\text{абс}} = \overline{\left|Y(t) - X(t)\right|} \tag{1.23}$$

и *критерий наибольшего отклонения*

$$\delta_{\text{абс}} = \max |Y(t) - X(t)|. \tag{1.24}$$

Обычно, для удобства сравнительной оценки помехоустойчивости различных систем рассматриваются *относительные отклонения*:

$$\gamma_\sigma = \frac{\sigma}{X_э};\ \ \gamma_{\text{абс}} = \frac{\delta_{\text{абс}}}{X_э};\ \ \gamma_{\max} = \frac{\delta_{\max}}{X_э}, \tag{1.25}$$

а также сведенные *отклонения:*

$$\alpha_\sigma = \frac{\sigma}{L_x};\ \ \alpha_{\text{абс}} = \frac{\delta_{\text{абс}}}{L_x};\ \ \alpha_{\max} = \frac{\delta_{\max}}{L_x}, \tag{1.26}$$

где $X_е$ - эффективное значение сообщения; L_x - динамический диапазон сообщений, которые передаются.

Предположим, что канал передачи информации имеет идеальную П-образную АЧХ и линейную ФЧХ. Тогда, при наличии флуктуационной помехи типа белого (гауссового) шума, *среднеквадратичное значение отклонения принятого сообщения от переданного равняется корню квадратному из средней мощности помехи на выходе приемника; относительная величина этого отклонения определяется как корень квадратный из отношения средних мощностей помехи и сигнала на выходе приемника*:

$$\gamma_\sigma = \sqrt{(P_\xi / P_x)_{\text{вых}}}. \qquad (1.27)$$

Для сравнительной оценки систем и практических расчетов часто за *критерий помехоустойчивости* берут «*выигрыш системы*»

$$\text{В} = \frac{(P_x / P_\xi)_{\text{вых}}}{(P_x / P_\xi)_{\text{вх}}}, \qquad (1.28)$$

где $(\text{P}_\text{x} / \text{P}_\zeta)_{\text{вых}}$ и $(P_x / P_\zeta)_{\text{вх}}$ - отношение средних мощностей сигнала и помехи на выходе и входе устройства.

В случае передачи дискретных сообщений, а также беспрерывных сообщений с кодоимпульсной модуляцией сигналов как критерий правильности целесообразно использовать вероятность правильного приема

$$P_{\text{пр}} = 1 - P_{\text{ош}}, \qquad (1.29)$$

где $P_{\text{ош}}$ - вероятность ошибки в воспроизведении сообщения.

В реальных условиях вероятность ошибки $P_{\text{ош}}$ очень мала и значительно меньше единицы. Поэтому довольно часто для осуществления *оценки помехоустойчивости используют логарифмическую величину*

$$S = \lg \frac{1}{P_{\text{пом}}} = \lg \frac{1}{1 - P_{\text{пр}}}. \qquad (1.30)$$

По смыслу определения помехоустойчивость понимается как свойство передачи информации в целом. Тем не менее, оценить помехоустойчивость системы в целом — довольно сложная задача. Поэтому, по обыкновению, говорят о помехоустойчивости отдельных звеньев системы: о *помехоустойчивости передачи* (в частности, о помехоустойчивости кода или вида модуляции) и о *помехоустойчивости приема*. Помехоустойчивость кода можно оценить величиной (1.30), где $P_{\text{ош}}$ — вероятность искажения кодовой комбинации под влиянием помех.

Оценивая влияние модуляции на помехоустойчивость системы, сравнивают по обыкновению различные виды модуляции с амплитудной. С этой целью часто используют коэффициент

$$R_{\text{М}} = \frac{(X_\text{е} / \sigma_\xi)_{\text{вых}}}{(X_\text{е} / \sigma_\xi)_{\text{выхАМ}}} \qquad (1.31)$$

или

$$R_{\text{М}} = \frac{(P_x / P_\xi)_{\text{вых}}}{(P_\text{е} / P_\xi)_{\text{выхАМ}}}, \qquad (1.32)$$

где $(X_\text{е} / \sigma_\zeta)_{\text{вых}}$ и $(P_x / P_\zeta)_{\text{вых}}$ - отношение эффективного значения сигнала к среднеквадратичному значению помехи и отношение средней мощности сигнала к средней мощности помехи на выходе приемника при произвольном виде модуляции; $(X_\text{е} / \sigma_\zeta)_{\text{выхАМ}}$ и $(P_x / P_\zeta)_{\text{выхАМ}}$ - аналогичные отношения на выходе приемного устройства при амплитудной модуляции.

Во время приема в зависимости от назначения сигналов могут воз-

никать три вида задач: выявление сигналов, различение сигналов и восстановление сообщений.

Задача выявления заключается в том, чтобы по результатам обработки принятого сигнала, который может быть или только помехой, или суммой полезного сигнала и помехи, выяснить, содержится ли полезный сигнал в принятом. При этом возможны ошибки двух видов:

при отсутствии полезного сигнала принимается ошибочное решение о наличии сигнала;

при наличии полезного сигнала принимается ошибочное решение об отсутствии сигнала.

Первая ошибка называется *ошибкой первого рода* или *ошибочной тревогой*. Вторая ошибка называется *ошибкой второго рода* или *пропуском сигнала*. Количественно ошибки первого и второго рода оцениваются условными вероятностями α и β ошибочных решений о наличии полезного сигнала, когда на самом деле он отсутствует, и об отсутствии сигнала, когда на самом деле он есть.

Полная вероятность ошибочного решения определяется выражением

$$P_{\text{ош}} = q\alpha + p\beta, \qquad (1.33)$$

где α и β — априорная вероятность соответственно отсутствия и наличия полезного сигнала.

Помехоустойчивость приемника, который выполняет задачу выявления сигнала, можно оценить с помощью выражения (1.29), где $P_{\text{ош}}$ определяется из (1.33).

Когда априори известно, что передано один из двух сигналов $x_1(t)$ и $x_2(t)$, то ставится задача различения двух сигналов, т.е. определяется наличие на входе приемника сигнала $x_1(t)$ плюс помеха или сигнала $x_2(t)$ плюс помеха. Очевидно, помехоустойчивость приемника в этом случае также определяется выражениями (1.30) и (1.33), причем величины α и β в формуле (1.33) являются условными вероятностями ошибочных выводов о наличии сигналов $x_1(t)$ или $x_2(t)$, когда на самом деле на вход приемника поступает соответственно сигнал $x_2(t)$ или $x_1(t)$, а q и p являются априорными вероятностями поступления сигналов $x_1(t)$ и $x_2(t)$.

Случай различения многих сигналов принципиально мало отличается от случая различения двух сигналов. Помехоустойчивость приемника в этом случае также можно оценить формулой (1.30), причем вероятность ошибки

$$P_{\text{ош}} = \sum_{i=1}^{n} p_i \alpha_i, \qquad (1.34)$$

где p_i — априорная вероятность поступления i-го сигнала; α_i - условная вероятность ошибочного вывода о наличии i-го сигнала, когда на самом деле на вход приемника поступает любой другой из n сигналов.

Задача восстановления сообщения значительно отличается от задач выявления и различения сигналов. Оно сводится к получению исходного сооб-

щения $Y(t)$, наименее отличающегося от сообщения, передаваемого согласно выбранному критерию правильности. Помехоустойчивость приемника в этом случае можно оценить с помощью критериев отклонения, которые определяются выражениями (1.22) - (1.27).

Помехоустойчивость такого приемника можно также оценить с помощью критерия (1.28). Тем не менее, нужно иметь в виду, что выигрыш системы не всегда можно определить однозначно, причем эта величина не дает возможности объективно сравнивать между собой различные системы, если помеха имеет более широкий спектр, чем сигнал и, в частности, если она - типа белого шума. Мощность помехи на входе приемника в этом случае определяется полосой пропускания канала связи. Разные системы могут существенным образом отличаться полосой пропускания каналов связи. При одном и том же канале связи могут применяться приемники с разными полосами пропускания.

Для устранения имеющейся неоднозначности можно сравнивать на входе и на выходе приемника отношения мощности сигнала не к мощности помех, а к удельной мощности помех. Тогда выигрыш приемника оценивается соотношением

$$B' = \frac{P_x / P'_{\xi\text{вых}}}{P_x / P'_{\xi\text{вх}}}, \qquad (1.35)$$

где $P'_{\xi\text{вых}} = P_{\xi\text{вых}} / F_{\text{вых}}$ - удельная мощность помех на выходе приемника; $P'_{\xi\text{вх}} = P_{\xi\text{вх}} / F_{\text{вх}}$ - удельная мощность помех на входе приемника; $F_{\text{вх}}$ и $F_{\text{вых}}$ — полосы частот, в которых измеряется мощность помех на входе и выходе приемника.

Очевидно, что выигрыш приемника теперь можно определить как

$$B' = \beta/\rho, \qquad (1.36)$$

где $\rho = F_{\text{вх}}/F_{\text{вых}}$.

В основу всех способов повышения помехоустойчивости информационных систем положено использование определенных отличий между полезным сигналом и помехой. Поэтому для борьбы с помехами необходимые априорные сведения о свойствах помехи и сигнала.

Известно много способов повышения помехоустойчивости систем. Их можно поделить на две группы. Первая группа способов базируется на выборе метода передачи сообщений. Вторая группа способов связана с построением помехоустойчивых приемников.

Простым и часто применяемым способом повышения помехоустойчивости передачи является *увеличение отношения сигнал /помеха за счет увеличения мощности передатчика*. Тем не менее этот метод, несмотря на его простоту, может быть экономически невыгодным, поскольку связан с существенным возрастанием сложности и стоимости оборудования. Кроме того, увеличение мощности передачи сопровождается усилением действия соот-

ветствующего канала с помехами.

Важным способом повышения помехоустойчивости передачи непрерывных сигналов является *рациональный выбор вида модуляции сигналов*. Применяя виды модуляции, обеспечивающие значительное расширение полосы частот сигнала, можно достичь важного повышения помехоустойчивости передачи. Оценка помехоустойчивости отдельных видов модуляции рассматривается в гл. 3.

Радикальным способом повышения помехоустойчивости передачи дискретных сигналов является *использование специальных помехоустойчивых кодов*. При этом существуют два пути повышения помехоустойчивости кодов. Первый состоит в выборе таких способов передачи, которые обеспечивают по возможности меньшую вероятность искажения кода, второй — в улучшении свойств кодовых комбинаций, т.е. их корректировании.

Исследование первого пути показали высокую помехоустойчивость кодов с большим количеством частотных признаков. Второй путь связан с использованием кодов, которые дают возможность обнаруживать и устранять искажение в кодовых комбинациях. Такой способ кодирования связан с введением в код дополнительных, избыточных символов, что сопровождается увеличением времени или частоты передачи символов кода. Это приводит к расширению спектра сигнала. Основные положения теории помехоустойчивого кодирования изложены в гл. 4.

Повышение помехоустойчивости передачи можно также достичь *путем повторной передачи одного и того же сообщения*. На приемной стороне сравниваются получаемые сообщения и как истинные принимаются те, которые имеют наибольшее количество совпадений. Чтобы устранить неопределенность при обработке принимаемой информации и обеспечить отбор по критерию большинства, сообщение должно повторяться не менее чем трижды. Очевидно, этот способ повышения помехоустойчивости связан с увеличением времени передачи.

Системы с повторением передачи дискретной информации делятся на системы с групповым суммированием, в которых сравнения осуществляется по кодовым комбинациям, и на системы с посимвольным суммированием, в которых сравнение осуществляется по символам кодовых комбинаций. Исследования показали, что посимвольная проверка эффективнее, чем групповая. Разновидностью систем, в которых повышение помехоустойчивости достигается за счет увеличения времени передачи, являются *системы с обратной связью.*

При наличии искажений в переданных сообщениях информация, поступающая по обратному каналу, обеспечивает повторение передачи. Наличие обратного канала приводит к усложнению системы. Тем не менее, в отличие от систем с повторением передачи, в системах с обратной связью повторение передачи будет происходить лишь в случае выявления искажений в переданном сигнале, т.е. избыточность в целому будет меньшей.

Помехоустойчивый прием состоит в использовании избыточности, а также априорных сведений о сигналах и помехах для выполнения оптимальным образом задачи приема: *выявление сигнала, различение сигналов или восстановление сообщений. Ныне для синтеза оптимальных приемников широко используется аппарат теории статистических решений* [9].

Ошибки приема информации уменьшаются с увеличением отношения сигнал/помеха на входе приемника. Именно поэтому часто проводят предварительную обработку принятого сигнала с целью увеличения отношения полезной составляющей к помехе. К таким методам предварительной обработки сигналов принадлежат, например, объединение широкополосного усилителя, ограничителя и узкополосного усилителя, селекция сигналов по продолжительности, метод компенсации помех, метод фильтрации, корреляционный метод, метод накопления и т.д.

Основные выводы

Информация - это совокупность сведений об окружающем мире, которые мы получаем в результате взаимодействия с ним.

Сообщение - форма представления информации или выраженная в определенном измерении информация, предназначенная для передачи от источника к потребителю.

Сигнал - процесс изменения во времени физического состояния некоторого объекта, используемый для отображения, регистрации или передачи сообщений. Сигнал - это материальный носитель сообщений (информации).

Данные - это сообщение в виде цифровых сигналов.

Информационная безопасность - это состояние защищенности информации, которая обрабатывается, передается или сохраняется, от незаконного (несанкционированного) доступа, преобразования и уничтожения, а также состояние защищенности информационных технических средств от влияний, направленных на нарушение их работоспособности.

Под помехозащищенностью системы понимают ее способность предотвращать действие случайных помех. Под случайными помехами понимают специальные помехи, которые создаются системой радиоэлектронного глушения.

В телекоммуникационных системах острейшей проблемой является учет взаимных помех, которые более всего влияют на электромагнитную совместимость.

Секретность системы передачи - это ее способность противостоять действию радиоразведки.

Помехоустойчивость - это способность системы передачи противостоять вредному влиянию помех. Анализ помехоустойчивости проводят независимо от причин появления помех в системе передачи.

Для решения проблемы информационной секретности по предотвращению несанкционированного доступа к информационному содержанию сообщения используют засекречивание (шифрование) сообщений.

Для теоретического исследования сигналов необходимо построить их математические модели. Математическая модель сигнала - функциональная зависимость, которая адекватно описывает изменение во времени физического состояния некоторого объекта.

Безопасность информации – состояние защищенности информации (сигналов), которая обрабатывается, сохраняется и передается, от незаконного вмешательства, т.е. нарушения физической и логической целостности информации (ознакомление, преобразование или искажение и уничтожение) или несанкционированного ее использования.

Угрозы безопасности информации - события или действия, которые могут вызвать нарушение функционирования системы, связанное с уничтожением или несанкционированным использованием обрабатываемой в ней информации.

Уязвимость информации - возможность возникновения на любом этапе жизненного цикла системы такого ее состояния, при котором создаются условия для реализации угроз безопасности информации.

Защищенность информации - поддержание в системе на заданном уровне тех параметров сигналов, которые характеризуют установленный статус их хранения, обработки и использования.

Защита информации - процесс создания и использования в системах специальных механизмов, которые поддерживают статус ее защищенности.

Комплексная защита информации - целенаправленное регулярное использование в системах средств и методов, а также принятие мер с целью поддержания заданного уровня защищенности информации по всей совокупности показателей и условий, которые являются существенными с точки зрения обеспечения безопасности информации.

Качество информации - совокупность свойств, которые предопределяют пригодность информации удовлетворять определенные потребности согласно ее назначению.

Под техническим каналом утечки информации понимается совокупность физических полей, несущих конфиденциальную информацию, конструктивных элементов, взаимодействующих с ними, и технических средств злоумышленника для регистрации поля и снятие информации.

Конфиденциальная информация в таком техническом канале подается в виде сигналов (акустических, виброакустических, электрических, электромагнитных), которые получили название опасных сигналов.

Практически в каждой автоматизированной системе существуют объекты, могущие создавать опасные сигналы и оказывать содействие их распространению, т.е. быть источником просачивания информации. К ним относятся технические средства, в которых обрабатывается конфиденциальная ин-

формация, а также человек, в речи которой может содержаться конфиденциальная информация (речь в виде акустических сигналов), доступная злоумышленнику по акустическому каналу или по каналу проводниковой или радиосвязи при использовании определенных технических средств.

К техническим средствам защиты информации в каналах связи можно отнести приборы, устанавливающие факт подключения к телефонным каналам подслушивающих приборов, спектральные анализаторы каналов связи и устройства защиты конфиденциальных разговоров по телефонным каналам.

Для защиты информации от утечки за счет побочных электромагнитных излучений применяются пассивный, активный и комбинированный методы.

Под функцией защиты понимается совокупность однородных в функциональном понимании мер, регулярно осуществляемых в автоматизированных системах различными средствами и методами с целью создания, поддержки и обеспечения условий, объективно необходимых для надежной защиты информации.

К специфическим средствам защиты информации принадлежат криптографические методы. В автоматизированных системах криптографические методы защиты информации могут использоваться как для защиты обрабатываемой информации в компонентах системы, так и для защиты информации, передаваемой по каналам связи. Собственно преобразование информации может осуществляться аппаратными или программными средствами.

Важнейшим концептуальным требованием к СЗИ является адаптация, т.е. способность целеустремленно приспосабливаться к изменениям структуры, технологических схем или условий функционирования АС.

Проектирование систем защиты информации заключается в том, чтобы для заданной автоматизированной системы (или ее проекта) создать оптимальные механизмы обеспечения защиты информации и управление. При этом оптимальность СЗИ понимается как достижение заданного уровня защищенности информации при минимальных затратах или достижение максимально возможного уровня защищенности при заданном уровне затрат на защиту.

Защита информации в современных автоматизированных системах требует создания различных в функциональном плане органов защиты, основу системы которых могут представлять отраслевые и территориальные центры защиты информации.

Все документы, касающиеся защиты информации, должны образовывать единую систему, основными группами которой являются справочно-информационные документы, стандарты, руководящие методические материалы, инструкции.

Структурированность компонентов автоматизированных систем является одним из важнейших конструктивных условий. При этом основное внимание отводится структуризации математического, программного и информационного обеспечения.

Помехоустойчивость можно характеризовать как степень соответствия принятого сообщения переданному при заданной помехе.

Для характеристики степени соответствия принятого сообщения переданному введена количественная мера, которая выбирается в зависимости от характера сообщений.

При передаче дискретных сообщений или непрерывных сообщений с кодоимпульсной модуляцией сигналов целесообразно как критерий правильности использовать вероятность правильного приема.

Помехоустойчивый прием состоит в использовании избыточности, а также априорных сведений о сигналах и помехах для выполнения оптимальным образом задачи приема: выявление сигнала, различение сигналов или восстановление сообщений.

Вопросы для самоконтроля

1. *Приведите схему формирования и материализации информации.*
2. *Перечислите главные задачи защиты информационных технических средств.*
3. *Дайте объяснение понятия «обработка сигналов».*
4. *Разъясните понятие «помехоустойчивость системы».*
5. *Назовите методы повышения энергетической и структурной секретности систем передачи информации.*
6. *Дайте определение математической модели сигнала.*
7. *Дайте определение комплексной защиты информации.*
8. *Назовите известные подходы к классификации угроз безопасности информации.*
9. *Сравните их между собой с точки зрения наибольшего соответствия практическим нуждам создания систем защиты информации.*
10. *Охарактеризуйте основные принципы системной классификации угроз безопасности информации.*
11. *Приведите классификационную структуру каналов несанкционированного получения информации.*
12. *Выведите формулу для оценки уязвимости информации, обрабатываемой в информационных системах.*
13. *В чем заключается опасность разработки и применение информационного оружия? Какие мероприятия международного характера необходимо было бы провести с целью предотвращения информационных войн?*
14. *Дайте определение понятия «технический канал утечки информации». Назовите основные виды технических каналов.*
15. *Дайте классификацию источников утечки информации.*
16. *Приведите известные вам методы и средства контроля акустической информации.*
17. *Раскройте содержание методов контроля информации техниче-*

скими средствами в каналах телефонной связи.

18. Назовите методы контроля информации, обрабатываемой средствами вычислительной техники.

19. Охарактеризуйте основные способы предотвращения утечки информации по техническим каналам.

20. Охарактеризуйте способы защиты информации в каналах связи.

21. Сформулируйте основные концептуальные положения теории защиты информации.

22. Раскройте содержание функции защиты информации. Какие из функций образовывают полное множество функций защиты?

23. Сформулируйте задачу защиты и назовите 10 классов задач, которые образовывают репрезентативное множество задач защиты.

24. Приведите классификацию средств защиты информации. Какие преимущества и недостатки программных, аппаратных и организационных средств защиты информации?

25. Дайте определение системы защиты информации и сформулируйте основные концептуальные требования, которые выдвигаются к ней.

26. Как влияют показатели информации, подлежащей защите, на структуру системы защиты информации и подходы к ее проектированию?

27. Раскройте содержание коротко-, средне- и долгосрочного управления процессами защиты информации.

28. Какие задачи можно была бы возложить на региональные центры защиты информации как на структуры, призванные обеспечить практическую реализацию концепции комплексной защиты информации?

29. Какими должны быть функции центров защиты информации для максимальной реализации задач, возникающих перед ними?

30. Сформулируйте основные требования к системе нормативно-правовых документов, регламентирующих процессы комплексной защиты информации.

31. Раскройте понятие «помехоустойчивость информационных систем».

32. Какие критерии помехоустойчивости используются при передаче непрерывных сообщений?

33. Каким образом оценивают влияние модуляции на помехоустойчивость системы?

34. Что такое ошибки первого и второго рода?

35. Как устранить неоднозначность при оценивании помехоустойчивости информационных систем?

36. Что лежит в основе всех способов повышения помехоустойчивости информационных систем?

37. Назовите способы повышения помехоустойчивости передачи дискретных сигналов.

38. Раскройте суть помехоустойчивого приема сообщений.

The main conclusions (Part 1)

The information is a collection of data about the world surrounding us that we receive as a result of interaction with it.

The message is the form of representation of the information or it is the information intended for transmission from a source to the consumer that is expressed in the certain measurement.

The signal is the process of changing in time the physical state some object and it serves for displaying, registration or transmitting of messages. The signal is a material carrier of messages (information).

Data are the messages in the form of digital signals.

Informational safety is a state of protectability of the information that is treated, transmitted or saved from illegal (unauthorized) access, transformation and destruction and it is also a state of protectability of informational technical facilities from the influences directed on violation of their efficiency.

The noise immunity of the system is its ability to prevent the action of random interferences. Random interferences are special interferences that are created by the system of radio-electronic suppression.

The most nagging problem in telecommunication systems is the registration of mutual interferences that have the biggest influence on electromagnetic compatibility.

The secrecy of system of transmission is its ability to resist the action of radio intelligence.

The noise immunity is the ability of system of transmission to resist the harmful influence of interferences. The analysis of noise immunity is conducted regardless to the reasons of appearing of interferences in system of transmission.

The classification enciphering of messages are used for solution of a problem of informational secrecy in order to prevent the unauthorized access to the informational contents of the message.

It is necessary to construct the mathematical models of the signals for their theoretical research. The mathematical model of a signal is functional dependence that describes a change in time of a physical state of some object adequately.

Safety of the information is a state of protectability of the information (signals) that is treated, saved and transmitted, from illegal interference from the point of view of violation of physical and logical integrity of the information (acquaintance, transformation or distortion and destruction) or unauthorized use.

Threats of safety of the information such as events or operations that can call the violation of functioning of the system connected with destruction or unauthorized use of the information which is treated in it.

Vulnerability of the information is possibility of occurrence on any stage of life cycle of the system of such state when conditions for realization of threats to safety of the information are created.

Protectability of the information is the support on back level of the parameters of the signals that characterize the installed status of their storage, treatment and use in system.

Protection of the information is the process of creation and usage in systems of special mechanisms that support the status of its protectability.

Complex protection of the information is purposeful regular using of facilities and methods in systems and also realization of actions with the purpose of support of the set level of protectability of the information regarding all totality of characteristics and conditions that are important from the point of view of providing the safety of the information.

The information technology protected in advance is unified in a wide spectrum of the functional additions information technology which contains all mechanisms for providing of a necessary level of protection as a main characteristic of quality of the information.

Quality of the information is a collection of qualities that determine the fitness of the information to satisfy the certain needs according to its purpose

The technical channel of a flowing out of the information is the collection of physical fields that carry the confidential information, structural units that cooperate with them, and technical facilities of the malefactor for registration of the field and removal of the information.

The confidential information in such technical channel is presented as signals (acoustic, vibroacoustic, electrical, electromagnetic) that have received the name of dangerous signals.

Practically in each automated system there are the objects that can create dangerous signals and assist their spreading, that is to be a source of information leakage. They are the technical facilities where the confidential information is treated and also a man whose language contains the confidential information (language in a form of acoustic signals), accessible to the malefactor through the acoustic channel or on the wire or radio channel while using certain technical facilities.

Instruments that install the fact of connection to telephone channels of eavesdropper devices, spectral analyzers of connection and devices of protection of private talks through the telephone channels can be referred to the technical facilities of information protection.

The passive, active and combined methods are used for protection of the information from flowing out due to side electromagnetic radiations.

Function of protection is a collection of homogeneous actions in functional relationship regularly performed in computer based systems with the help of different facilities and methods with the purpose of creation, support and providing the conditions objectively necessary for reliable protection of the information.

Cryptography methods belong to specific facilities of protection of the information. In the computer based systems cryptography methods of protection of the information can be used for both protection of the treated information in the components of system and protection of the information which is transmitted through

communication channel. It is the transformation of the information that can be performed by hardware or software.

The major conceptual requirement to information protection facilities is adaptation that is ability to purpose fuladaptatio during change structures, technological schemes or conditions of functioning of the computer based system.

The designing of system of information protection is to create optimal mechanisms of providing of information protection and control for the set computer based system (orits project) . Thus optimum of information protection facilities is achievement of the set level of protectability of the information at the minimum expenses or achievement of the greatest possible level of protectability at the set level of expenditures on protection.

Information protection in the modern computer based systems demands creation of versatile bodies of protection in functional relation a basis of system of which can make the branch and territorial centers of information protection.

All documents on information protection should form a united system, the main groups of which are reference and information standards, supervising methodical materials, instructions.

Structuredness of components of the computer based systems is one of the major constructive conditions. Thus the main attention is given to structurization of mathematical providing, soft ware and data ware.

Noise immunity can be characterized as a degree of correspondence of the accepted message to the transmitted at the set obstacle.

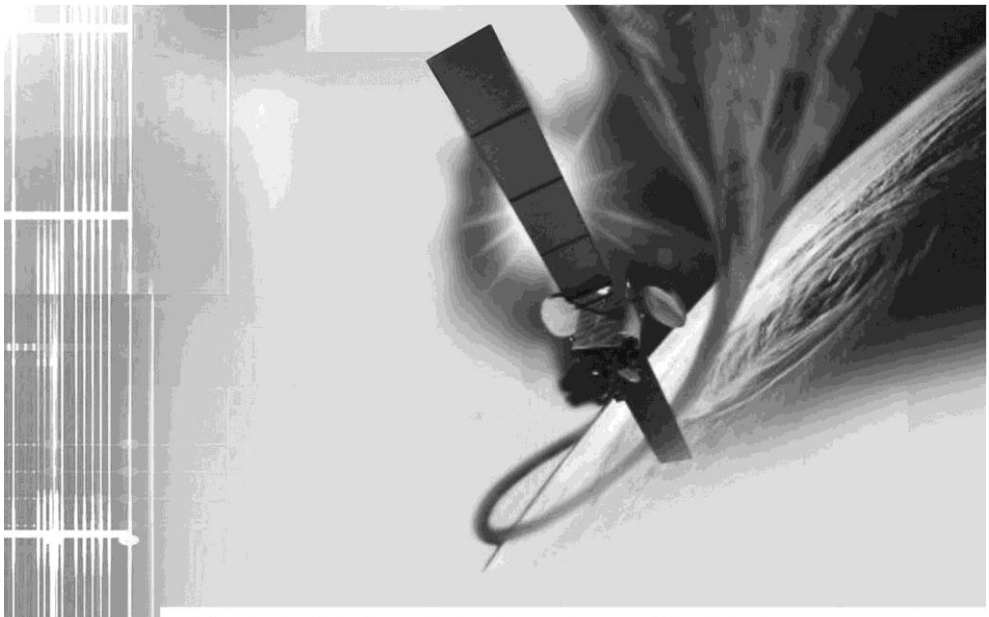
The quantitative measure which is chosen depending on character of messages, is brought in for the characteristic of a degree of correspondence of the accepted message to the transmitted.

It is expedient to use probability of correct reception during transmission of discrete messages or continuous messages with pulse - code modulation of signals as criterion of correctness.

The noiseless reception is the use of excessiveness and apriori information about signals and interferences for solution the task of reception in optimal way: revealing of signal, distinguishing of signals or renewal of messages.

Ключевые слова

Русский	**Английский**
информация	information
сообщение	report
сигнал	signal
безопасность	safety
уязвимость информации	vulnerability of information
канал несанкционированного доступа	channel of unauthorized division
защита информации	information protection
помехоустойчивость	noise immunity



КОЛИЧЕСТВЕННЫЕ ОЦЕНКИ ИНФОРМАЦИИ

2

2.1. Информация, ее функции и свойства
2.2. Количественные характеристики информации
2.3. Меры информации
2.4. Энтропия и ее свойства
2.5. Продуктивность, избыточность и помехозащищенность источника информации
2.6. Связь информации с параметрами сигналов

2.1. Информация, ее функции и свойства

Термин *информация*, как определено в гл. 1, происходит от латинского *informatio*, что означает разъяснение, информирование, изложение. В широком значении информация - это общенаучное понятие, которое охватывает обмен сведениями между людьми, обмен сигналами между живой и неживой природой, людьми и устройствами. В этом смысле информация является отражением реального мира с помощью сведений (данных, сообщений).

Понятие «информация» имеет несколько определений, зависимых от взглядов на нее. Мы будем придерживаться компромиссного толкования, согласно которому понятие «информация» объединяет характеристики объектов живой и неживой природы, являющихся потенциальным источником сведений разного рода для людей, а также и сами сведения, которые получают люди.

Как рабочее предлагаем менее короткое определение: *информация - это совокупность сведений об окружающем мире, объектах и явлениях окружающей среды, а также процессы, которые происходят в них, сведения об их параметрах, свойствах и состоянии, которые уменьшают степень неопределенности, неполноты знаний о них.*

Методы (подходы) анализа информации иллюстрирует рис. 2.1.

Рис. 2.1. Методы анализа информации

Согласно синтаксическому подходу осуществляется структурный анализ информации (количество сообщений, форма их представления, статистические характеристики появления и т.д.).

Семантический подход заключается в анализе содержательности информации и влияния ее на человека.

Прагматический подход связан с оцениванием полезности, ценности полученной информации.

Важно отметить *динамические свойства информации*. Объекты информации окружающего мира постоянно изменяются с разной скоростью. Изменения эти имеют случайный характер относительно моментов проявления, параметров, диапазона изменения параметров и законов изменения. Информация как отображение этих процессов также имеет динамический, случайный

характер. Всегда необходима обновленная информация, призванная устранять неопределенность ситуации в данный момент.

Обобщенной функцией информации есть обеспечение выживания, развития человека и общества в сложном и непостоянном окружающем мире. Она состоит из ряда частичных функций, которые имеют разную степень выявления в зависимости от внешних условий и характера деятельности человека. Чаще всего мы сталкиваемся с проявлениями таких *функций информации*:

управленческой, которая присутствует во всех сферах повседневной жизни и в трудовой деятельности и призвана помочь человеку в выборе варианта собственного поведения или целенаправленного влияния на объекты и процессы реального мира;

коммуникационной, которая присутствует во время обмена информацией между людьми и направленная на организацию взаимодействия между ними;

познавательной, которая предопределяется потребностью в информации ради общего развития, получении специальности и, в общем, для удовлетворения влечения к новому безотносительно к его прагматическому значению;

психологической, которая сказывается при формировании определенного эмоционального расположения духа с помощью некоторых видов информации, способов ее представления человеку.

Обобщенным свойством информационных функций является их направленность на устранение неопределенности нашего представления о состоянии объектов и процессов реального мира, которые нас интересуют, или на данный момент, или после некоторых влияний на них внутренних или внешних сил. Именно на устранение неопределенности сориентированы процессы сбора, передачи и обработки информации. Степень изменения неопределенности ситуации положена в основу *количественной меры информации.*

Адекватность информации - это определенный уровень соответствия образа, который создается с помощью полученной информации, реальному объекту, процесса, явлению и др. На практике вряд ли или можно достичь полной адекватности информации. Всегда существует некоторая степень неопределенности. От степени адекватности информации реальному состоянию объекта или процесса зависит правильность принимаемых человеком решений.

Адекватность информации может выражаться в трех формах: синтаксической, семантической, прагматической.

Синтаксическая адекватность - отражает формально-структурные характеристики информации и не затрагивает ее содержательное наполнение. На синтаксическом уровне учитываются тип носителя и способ представления информации, скорость передачи и обработки, размеры кодов представления информации, надежность и точность преобразования этих кодов и т.п. Формализованную информацию, которая рассматривается только из синтаксических позиций, обычно называют данными, поскольку при этом абстрагируются от содержательного аспекта. Эта форма адекватности содействует

восприятию внешних структурных, т.е. синтаксических, характеристик информации.

Семантическая (*содержательная*) *адекватность* - определяет другую степень соответствия образа исследуемого объекта самому объекту. Семантический аспект предусматривает учет содержательного наполнения информации. На этом уровне анализируются те сведения, которые отображает информация, и рассматриваются содержательные связи между ними.

Эта форма адекватности дает возможность формировать понятие и представление, обнаруживать значение, содержание информации и обобщать ее.

Прагматическая (*потребительская*) *адекватность* - отображает соотношение между информацией и ее потребителем, степень соответствия информации той цели управления, которая на ее основе реализуется. Обнаруживаются прагматические свойства информации только при наличии единства информации (объекта), пользователя и цели управления. Прагматический аспект рассмотрения связан с ценностью, полезностью информации при подготовке потребителем решения для достижения поставленной цели. В соответствии с этим анализируются потребительские свойства информации.

Эта форма адекватности непосредственно связана с практическим использованием информации и ее соответствием целевой функции работы системы.

Информация характеризует соотношение между *источником информации* (объектом исследования), сообщением и его потребителем. При отсутствии потребителя, хотя бы потенциального, говорить об информации нет смысла.

Если речь идет об автоматизированной работе с информацией с помощью некоторых технических устройств, прежде всего интересуются не содержанием, а *источником сообщения* и количеством символов, которое содержит это сообщение.

Информационные объекты - это предметы, процессы, явления материальной или нематериальной природы, которые рассматриваются с точки зрения их информационных свойств.

Источник информации или сообщение - это физический объект (система или явление), что формирует переданное сообщение. Само сообщение - это значение или изменение некоторой физической величины, которая отображает состояние информационного объекта (системы или явления).

При работе с информацией всегда существуют ее источник и потребитель (получатель). Пути и процессы, которые обеспечивают передачу сообщений от источника информации к ее потребителю, называются *информационно-коммуникационными системами* (*ИКС*). Этот термин был введен в более широком понимании - как комплекс организационно-технических мероприятий, информационных технологий и информационных ресурсов, предназначенных для обеспечения информационных процессов, в частности, создания, распространения, использования, хранения и уничтожения информации.

Относительно компьютерной обработки данных под информацией понимают некоторую последовательность символических обозначений (букв, цифр, закодированных графических образов и звуков и т.п.), которая несет содержательную нагрузку и подается в понятном для компьютера виде. Каждый новый символ в такой последовательности увеличивает информационный объем сообщения.

Данные содержат *информацию* о событиях, которые состоялись в материальном мире. Для того чтобы из данных можно было получить информацию, необходимо наличие *метода ее обработки.*

2.2. Количественные характеристики информации

Предположим, что вы получили какое-то сообщение (например, прочитали статью в любимом журнале). В этом сообщении содержится некоторая информация. Как оценить информацию, которую вы получили? Другими словами, как измерить информацию? Можно ли сказать, что чем больше статья, тем больше информации она содержит? Разные люди, которые получили одно и то же сообщение, по-разному оценивают информацию, которая содержится в нем. Это происходит из-за того, что знание людей об этих событиях, явлениях к получению сообщения были разными. Поэтому те, кто знал об этом мало, признают, что получили много информации, а те, кто знал больше, чем написано в статье, скажут, что не получили информации вообще. Итак, можно сказать, что оценка информации зависит от того, насколько новой или полезной есть эта информация для получателя.

При таком подходе непонятно, по каким критериям можно ввести единицу оценки информации. Итак, с точки зрения информации как носителя новизны мы не можем оценить информацию, которая содержится в научном открытии, новой теории общественного развития.

В технике информацией считается любая последовательность символов, которая сохраняется, обрабатывается или передается. Часто используют простой способ определения количества информации, который можно назвать объемным. Он основывается на подсчете количества символов в сообщении, т.е. связан лишь с его длиной и не учитывает содержания.

Длина сообщения зависит от количества разных символов, использованных для записи сообщения. В вычислительной технике применяются две стандартных единицы измерения: *бит* (двоичный знак двоичного алфавита {0,1}) — минимальная единица измерения информации и *байт* (который равняется восьми *битам* и является одним символом, т.е. при введении из клавиатуры этого символа машине передается 1 байт информации).

В теории информации под *количеством информации* понимают меру уменьшения неопределенности знания. Нахождение такой меры нуждается в оценивании и учете количества переданной информации.

В этом смысле количество информации - это числовая характеристика

Джозеф Леонард Уолш (Joseph LeonardWalsch, 1895 - 1973), американский математик. Работал в Гарвардском и Чикагском университетах. Основные работы касаются теории функций и топологии. Всего он опубликовал свыше 300 работ. На русский язык переведен его книгу "Интерполяция и аппроксимация рациональными функциями в комплексной области", а также "Теория сплайнов и ее применение" (вместе из Дж. Альбергом и Е. Нильсоном).

сигнала, которая не зависит от его формы и содержания и характеризует неопределенность, которая исчезает после получения сообщения в виде данного сигнала. В таком случае количество информации зависит от вероятности получения сообщения о том или другом событии.

Информационный объем (*информационная вместительность*) *сообщения* - количество информации в сообщении, измерена в битах, байтах или производных единицах.

Для абсолютно достоверного события (т.е. такого, что непременно состоится, а потому его вероятность равняется единице) количество информации в сообщении о ней равняется нулю. Чем невероятнее событие, тем большую информацию о ней несет сообщение. Только если ответы равновероятностные, то ответ «да» или «нет» несет один бит информации.

Г.Хартли предложил формулу для вычисления количества I информации об объекте, который может находиться в одном из равновероятностных N состояний:

$$I = \log_2 N. \qquad (2.1)$$

Формулу для вычисления количества информации для событий с разной вероятностью предложил К. Шеннон в 1948 г. В этом случае количество I информации определяется по формуле

$$I = -\sum_{i=1}^{N} p_i \log_2 p_i, \qquad (2.2)$$

где N - количество возможных событий; p_i - вероятность отдельных событий.

Количественная мера информации. При введении количественной меры информации (Г.Хартли и К.Шеннон) было принято содержательное наполнение сообщений (семантику) не учитывать, а ограничиться только формальными признаками, важными с точки зрения передачи сообщений по каналам связи. В результате учитываются только количество N сообщений, которые подлежат передаче, и вероятности $p(x_i)$ поступления их на вход канала.

Всю совокупность сообщений подают в виде некоторой системы X с состояниями x_i:

$$X = \frac{x_1\ x_2 \ldots x_n}{p(x_1)\,p(x_2)\ldots p(x_n)}, \quad \sum_{i=1}^{N} p(x_1) = 1, \tag{2.3}$$

где x_i - отдельные сообщения (или их типы, классы); $p(x_i)$ - априорная вероятность появления сообщений x_i.

В результате передачи сообщения x_i будет получено сообщение y_j. Оно с некоторой вероятностью может быть похожим на любое из сообщений (x_1, x_2,...x_N) в частности на переданное сообщение x_i. Апостериорная вероятность присутствия x_i в y_j равняется $p(x_i/y_j)$.

В основу меры количества информации, положены обусловленные искажениям информации в канале связи изменения вероятности появления сообщений - от априорной $p(x_i)$ на входе канала к апостериорной $p(x_i/y_j)$ на выходе канала.

Сравнивая вероятности $p(x_i)$ и $p(x_i/y_j)$, можно установить меру количества информации, переданной к потребителю. Удобной мерой оказался логарифм отношения апостериорной вероятности к априорной.

Количество информации, которое содержится в событии y_j относительно события x_i, определяется по формуле

$$I(x_i; y_j) = \log \frac{p(x_i / y_j)}{p(x_i)}. \tag{2.4}$$

В качестве основания логарифма чаще всего берут 2, $e \approx 2{,}72$, или 10 (может быть 8, 16 и т.п.). В зависимости от основания изменяются единицы измерения количества информации (бит - двоичные единицы; нит - натуральные единицы; хартли (дит) - десятичные единицы).

Рассмотрим свойства количества информации.

1. *Свойство симметрии*. Информация, что содержится в y_j относительно x_i, равняется информации, что содержится в x_i относительно y_j.

Учитывая это свойство, величину $I(x_i, y_j)$ называют *количеством взаимной информации между* x_i и y_j.

2. *Свойство аддитивности.* Информация, что содержится в паре символов y_j, z_k относительно x_i, равняется сумме информации, что содержится в y_j относительно x_i, и информации, что содержится в z_k относительно x_i, при условии, что значение y_j известное:

$$I(x_i; y_j z_k) = I(x_i; y_j) + (x_i; z_k / y_j). \tag{2.6}$$

Количество собственной информации в x_i определяется из формулы (2.2) при $p(x_i y_j) = 1$:

$$I(x_i; y_j) = -\log p(x_i). \tag{2.7}$$

Эта величина определяет количество информации, необходимое для однозначного определения x_i на выходе канала.

С учетом введенного понятия (2.6) можно преобразовать выражение (2.4) к виду

Клод Элвуд Шеннон (Claude Elwood Shannon, 1916 - 2001), американский математик и электротехник, один из творцов математической теории информации, в значительной мере определил своими результатами развитие общей теории дискретных автоматов, которые являются важными составляющими кибернетики. В 1948 г. опубликовал фундаментальную работу "Математическая теория связи", в которой сформулированы основы теории информации. Большую ценность представляет и другая работа "Теория связи в секретных системах" (1949), в которой изложены математические основы криптографии.

$$I\left(x_i; y_j\right) = I\left(x_i\right) - I\left(x_i / y_j\right), \qquad (2.8)$$

где $I(x_i / y_j) = -\log p(x_i / y_j)$ - условная собственная информация.

Среднее количество взаимной информации образовывается усреднением (2.4) за всеми i и j:

$$I\left(x_i; y_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} p\left(x_i, y_j\right) = \log\frac{p\left(x_i / y_j\right)}{p\left(x_i\right)}. \qquad (2.9)$$

Это становится очевидным, если числитель и знаменатель в (2.4) умножить на $p(y_j)$ и выполнить преобразование:

$$I\left(x_i; y_j\right) = \log\frac{p\left(x_i / y_j\right)p\left(y_j\right)}{p\left(x_i\right)p\left(y_j\right)} = \\ = \log\frac{p\left(x_i, y_j\right)}{p\left(x_i\right)p\left(y_j\right)} = I\left(y_j; x_i\right), \qquad (2.5)$$

поскольку $p(x_i, y_j) = p(y_j)$, $p(x_i / y_j) = p(x_i)$, $p(y_j / x_i)$ - вероятность одновременного появления y_j и x_i.

Учитывая это свойство, величину $I(x_i, y_j)$ называют *количеством взаимной информации между* x_i и y_j.

2. *Свойство аддитивности.* Информация, что содержится в паре символов y_j, z_k относительно x_i, равняется сумме информации, что содержится в y_j относительно x_i, и информации, что содержится в z_k относительно x_i, при условии, что значение y_j известное:

$$I\left(x_i; y_j z_k\right) = I\left(x_i; y_j\right) + \left(x_i; z_k / y_j\right). \qquad (2.6)$$

Количество собственной информации в x_i определяется из формулы (2.2) при $p(x_i y_j) = 1$:

$$I\left(x_i; y_j\right) = -\log p\left(x_i\right). \qquad (2.7)$$

Эта величина определяет количество информации, необходимое для однозначного определения x_i на выходе канала.

С учетом введенного понятия (2.6) можно преобразовать выражение (2.4) к виду

$$I\left(x_i; y_j\right) = I\left(x_i\right) - I\left(x_i / y_j\right), \qquad (2.8)$$

где $I(x_i / y_j) = -\log p(x_i / y_j)$ - условная собственная информация.

Среднее количество взаимной информации образовывается усреднением (2.4) за всеми *i* и *j*:

$$I\left(x_i; y_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} p\left(x_i, y_j\right) = \log\frac{p\left(x_i / y_j\right)}{p\left(x_i\right)}. \quad (2.9)$$

2.3. Меры информации

Для количественной оценки информации довольно часто применяют синтаксическую, семантическую и прагматическую меры информации (рис. 2.2).

Синтаксическая мера информации - оперирует с обезличенной информацией, которая не выражает содержательной связи с объектом.

На синтаксическом уровне учитываются тип носителя и способ представления информации, скорость ее передачи и обработки, размеры кодов представления информации.

Для определения такой количественной меры информации вводятся два параметра: *объем данных* $V_д$ и *количество информации I*.

Объем данных ($V_д$) - информационный объем сообщения или объем памяти, необходимый для хранения сообщения без любых перемен.

Объем данных $V_д$ в сообщении измеряется количеством символов (разрядов). Единица измерения зависит от системы исчисления.

Количество I информации в сообщении об объекте, который может находиться в одном из равновероятностных *N* состояний, определяют согласно приведенной раньше формуле Р. Хартли (2.1).

Из этой формулы следует, что чем неопределеннее была ситуация к получению сообщения, т.е. чем большего количества состояний мог приобрести объект, тем большее количество информации несет данное сообщение.

Единицы измерения информации зависят от применяемой системы исчисления: в двоичной системе единица измерения - бит (bit - binary digit - двоичный разряд), в десятичной - дит.

Бит - количество информации, нужной для различения двух равновероятностных сообщений. Бит - очень малая единица измерения. На практике чаще

Ральф Винтон Лион Хартли (Ralph Vinton Lyon Hartley, 1888 - 1970), исследователь в области электроники. Изобрел гетеродин и преобразования, названные в его честь. Сделал вклад в основы теории информации. Родился в штате Невада. Получил степень бакалавра наук в Оксфордском университете (1913). В 1915 г. был ответственным за разработку радиопередатчика для трансатлантического радиотелефонного теста системы Белл. Для этого разработал гетеродин Хартли, а также схему нейтрализации для устранения паразитного самовозбуждения триода, который возникает в результате внутреннего соединения.

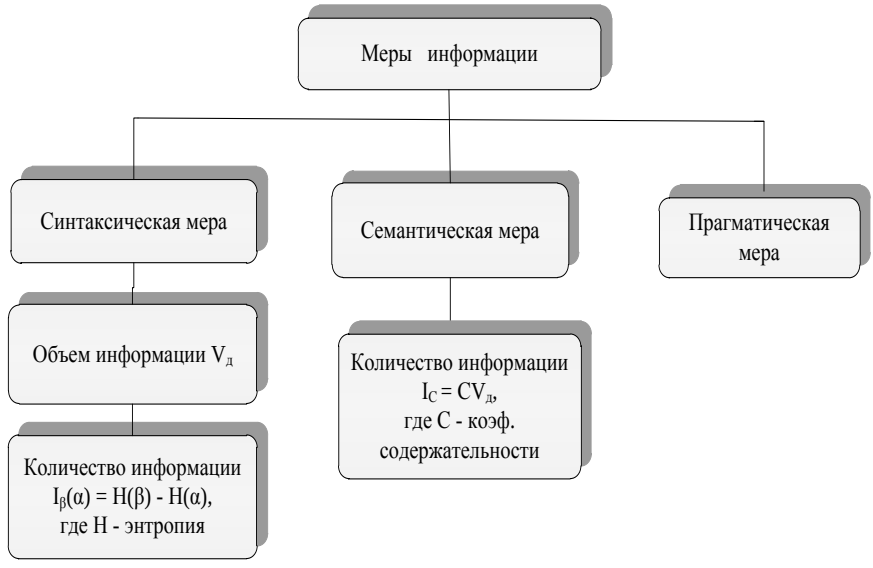


Рис. 2.2. Классификация мер информации

применяется единица *байт*, который равняется *восьми битам***.** Именно восемь бит нужно для того, чтобы закодировать любой из 256 символов алфавита клавиатуры компьютера ($256 = 2^8$).

Широко используются также еще большие производные единицы измерения информации:

1 килобайт (кбайт) = 1024 байт = 2^{10} байт;

1 мегабайт (Мбайт) = 1024 кбайт = 2^{20} байт;

1 гигабайт (Гбайт) = 1024 Мбайт = 2^{30} байт;

1 терабайт (Тбайт) = 1024 Гбайт = 2^{40} байт;

1 петабайт (Пбайт) = 1024 Тбайт = 2^{50} байт.

В качестве единицы информации можно было бы взять количество информации, необходимой для различения, например, десяти равновероятностных сообщений. Это будет не двоичная (бит), а десятичная (дит**)** единица информации. В компьютерной практике слово «бит» используется также как единица измерения объема памяти. Элемент памяти размером в 1 бит может находиться в одном из двух состояний («включено» и «выключено») и в него можно записать одну цифру (0 или 1).

Количество I информации на синтаксическом уровне определяется через понятие *энтропии системы.*

Пусть до получения информации потребитель имеет некоторые предыдущие (априорные) сведения о системе *y.* Мерой его неосведомленности о системе есть функция *H*(*y*), которая вместе с тем есть и мерой неопределенности состояния системы.

После получения некоторого сообщения x получатель приобрел некоторую дополнительную информацию $I(x)$, что уменьшила его априорную неосведомленность так, что неопределенность состояния системы после получения сообщения x стала равняться $H(x)$. Тогда *количество* $I(x)$ *информации* о системе, полученной в сообщении β, определится как

$$I(x) = H(x) - H(x). \tag{2.10}$$

т.е. *количество информации* измеряется величиной изменения (уменьшения) неопределенности состояния системы.

Если конечная неопределенность $H(x)$ стремится к нулю, то первичное неполное знание заменится полным знанием, и количество информации $I(x) = H(x)$. Другими словами, *энтропия системы* $H(y)$ может рассматриваться как мера информации, которой недостает.

Энтропия системы $H(x)$, которая имеет N возможных состояний, согласно формуле К. Шеннона равняется

$$H(y) = -\sum_{i=1}^{N} P_i \log P_i, \tag{2.11}$$

где P_i - вероятность того, что система находится в i-м состоянии. Для случая, когда все состояния системы равновероятны, т.е. их вероятность равняется

$$P_i = 1/N, \tag{2.12}$$

ее энтропия определяется соотношением

$$H(y) = -\sum_{i=1}^{N} (1/N) \log (1/N) = \log N \text{ - формула Р. Хартли} \tag{2.13}$$

Пример. Часто информация кодируется числовыми кодами в той или другой системе исчисления, особенно это актуально в случае представления информации в компьютере. Естественно, что одно и то же количество разрядов в разных системах исчисления может передавать разное количество состояний отображаемого объекта, который можно подать в виде соотношения

$$N = m^n, \tag{2.14}$$

где N - количество отображаемых состояний; m - основа системы исчисления (количество символов, употребляемых в алфавите); n - количество разрядов (символов) в сообщении.

Предположим, что по каналу связи передается *n-разрядное* сообщение, которое использует m разных символов. Принимая во внимание, что количество кодовых комбинаций будет $N = m^n$, приходим к выводу: при равновероятностном появлении любой из комбинаций количество информации, приобретенной абонентом в результате получения сообщения

$$I = \log N = n \log m. \tag{2.15}$$

Если за основу логарифма взять m, то $I = n$. В этом случае количество информации (при условии полного априорного незнания абонентом содержания сообщения) будет равно объему данных $I = V_д$, полученных по каналу связи.

Коэффициент (*степень*) *информативности* (*лаконичность*) *сообщения* определяется отношением количества информации к объему данных:

$$Y = I / V_д . \qquad (2.16)$$

С увеличением Y уменьшаются объемы работы по преобразованию информации (данных) в системе. Поэтому стараются повышать информативность, для этого разрабатываются специальные методы оптимального кодирования информации.

Семантическая мера информации - применяется для измерения содержания информации. Для этого используется *тезаурусная мера*, которая учитывает способность получателя сообщения его воспринять.

Тезаурусом называют совокупность сведений, которые имеет в своем распоряжении пользователь или система.

Максимальное количество семантической информации $I_с$ пользователь получает, если ее содержание S будет согласованно с его тезаурусом S_p, т.е. когда информация, которая поступает, понятна пользователю и несет ему не известные раньше (отсутствуют в его тезаурусе) сведения. Количество семантической информации в сообщении является относительной и зависит от подготовленности получателя. Одна и та же информация может иметь смысл для специалиста и быть лишенной смысла для дилетанта.

В зависимости от соотношений между содержательным наполнением информации S и тезаурусом пользователя S_p изменяется количество семантической информации $I_с$, которую воспринимает пользователь и включает в дальнейшем в свой тезаурус. Характер такой зависимости делает наглядным рис. 2.3.

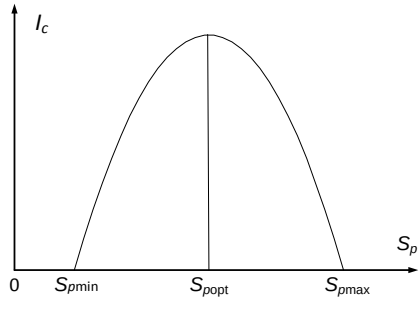


Рис. 2.3. Зависимость количества семантической информации, воспринимаемой потребителем, от его тезауруса

Рассмотрим два предельных случая, когда количество семантической информации $I_с$ равняется нулю:

при $0 \le S_p \le S_{p\min}$ пользователь не воспринимает, не понимает информацию, которая поступает;

при $S_p > S_{p\min}$ пользователь знает все, а поэтому информация, которая поступает, ему не нужна.

Максимальное количество семантической информации I_c потребитель получит, как уже отмечалось, согласовав ее содержательное наполнение S со своим тезаурусом $S_p = S_{p\text{opt}}$.

Относительной мерой количества семантической информации может быть коэффициент содержательности C, который определяется как отношение количества семантической информации к ее объему:

$$C = I_c / V_д. \tag{2.17}$$

Прагматическая мера информации - определяет полезность информации для достижения пользователем поставленных целей. В частности, полезность экономической информации можно определить за ростом экономических показателей организации, обусловленных использованием указанной информации. Это может быть сокращение товарных запасов, увеличение скорости оборота средств, повышение качества принятия управленческих решений и т.д. *Ценность информации* измеряется в тех самых (или близких к ним) единицах, в которых измеряется целевая функция.

2.4. Энтропия и ее свойства

Энтропия определяет меру неопределенности всего множества сообщений на входе системы и вычисляется как среднее количество собственной информации во всех сообщениях:

$$I(X) = -\sum p(x_i) \log p(x_i) = H(X). \tag{2.18}$$

Свойства энтропии:

1. Энтропия $H(X)$ положительная: $H(X) > 0$.

2. Энтропия $H(X) < \log N$.

3. Величина $\log N = D$ называется *информационной пропускной способностью* алфавита (информационной вместительностью алфавита).

4. Если $N = 2$, то $p(x_1) = p$, $p(x_2) = 1 - p$,

$$H(X) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$

5. Максимум $H(X) = -\log_2 0{,}5 = \log_2 2 = 1$ - вместительность двоичного алфавита равняется 1 бит.

Зависимость $H(X)$ от значения p иллюстрирует рис. 2.4.

Рассмотренные характеристики источника информации - количество информации и энтропия - касались одного источника, который вырабатывает поток независимых или простых сообщений, т.е. *источника без памяти.*

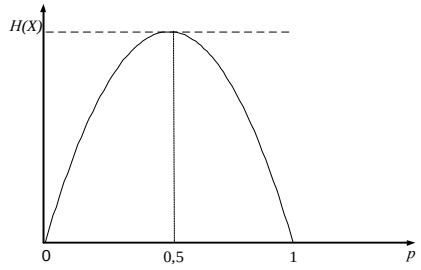


Рис. 2.4. Зависимость энтропии от величины *p*

Тем не менее в реальных условиях независимость элементарных сообщений, которые вырабатываются источником, - явление довольно редкостное. Чаще бывает именно наоборот: существует сильная детерминированная или статистическая связь между элементами сообщения с одного или нескольких источников.

При передаче и хранении данных часто имеют дело с несколькими источниками, которые формируют статистически связанные одно за другим сообщение. Сообщения, которые вырабатываются такими источниками, называются *сложными сообщениями*, а непосредственно эти источники - *источниками с памятью*.

Очевидно, что при определении энтропии и количества информации в сообщениях, элементы которых статистически взаимосвязанные, нельзя ограничиваться только безусловными вероятностями - необходимо учитывать также условные вероятности появления отдельных сообщений.

Определим *энтропию сложного сообщения*, которое вырабатывается двумя зависимыми источниками (так же определяется энтропия сложного сообщения, которое вырабатывается одним источником с памятью):

Пусть сообщение первого источника приобретают значения x_1, x_2, x_3, ..., x_k с вероятностями соответственно $P(x_1)$, $P(x_2)$, ..., $P(x_k)$, а сообщение второго - значений y_1, y_2, ..., y_m с вероятностями $P(y_1)$, $P(y_2)$, ..., $P(y_m)$. Общую энтропию двух источников X и Y можно определить как

$$H(X,Y) = -\sum_{i=1}^{\hat{a}} \sum_{j=1}^{m} P(x_i, y_j) \log P(x_i, y_j), \tag{2.19}$$

где $P(x_i, y_j)$ - вероятность общего появления сообщений x_i и y_j. Поскольку общая вероятность $P(x_i, y_j)$ за формулой Байеса определяется как

$$P(x_i, y_j) = P(x_i) P(y_j / x_i)) = P(y_j) P(x_i / y_j), \tag{2.20}$$

эту формулу для общей энтропии можно записать в виде

$$\begin{aligned} H(X,lY) &= l - \sum_{i=1}^{k} \sum_{j=1}^{m} P(x_i) P(y_j / x_i) \log \left\{ P(x_i) P(y_j / x_i) \right\} = \\ &= -\sum_{i=1}^{k} P(x_i) \log P(x_i) \sum_{j=1}^{m} P(y_j / x_i) - \sum_{i=1}^{k} P(x_i) \sum_{j=1}^{m} P(y_j / x_i) \log P(y_j / x_i). \end{aligned} \tag{2.21}$$

Поскольку передача сообщения x_i непременно отвечает передаче одного из сообщений (каждого) из ансамбля Y, то

$$\sum_{j=1}^{m} P\left(y_j / x_i\right) = 1, \tag{2.22}$$

причем *общая энтропия источников сообщений* $H(X, Y)$ определяется как

$$\begin{aligned} H\left(X,Y\right) &= -\sum_{i=1}^{k} P\left(x_i\right)\log P\left(x_i\right) - \sum_{i=1}^{k} P\left(x_i\right)\sum_{j=1}^{m} P\left(y_j / x_i\right)\log P\left(y_j / x_i\right) = \\ &= H\left(X\right) + \sum_{i=1}^{k} P\left(x_i\right) H\left(Y / x_i\right), \end{aligned} \tag{2.23}$$

где $H(Y/x_i)$ - так называемая *частичная условная энтропия*, которая отбивает энтропию сообщения Y при условии, что сообщение x_i поступило. Второе слагаемое представляет собой усреднение $H(Y/x_i)$ по всем сообщениям x_i и называется *средней условной энтропией источника Y при условии передачи сообщения X*. И окончательно:

$$H(X, Y) = H(X) + H(Y / X). \tag{2.24}$$

Общая энтропия двух сообщений равняется сумме безусловной энтропии одного из них и условной энтропии второго.

Можно отметить такие основные свойства энтропии сложных сообщений.

1. В случае *статистически независимых сообщениях X и Y общая энтропия* равняется сумме энтропии каждого из источников:

$$H(X, Y) = H(X) + H(Y), \tag{2.25}$$

при этом учтено, что $H(Y / X) = H(Y)$.

2. В случае *полной статистической зависимости сообщений X и Y общая энтропия* равняется *безусловной энтропией одного* из сообщений. Второе сообщение при этом информации не прибавляет.

В самом деле, при полной статистической зависимости сообщений условные вероятности $P(y_j / x_i)$ и $P(x_i / y_j)$ равняются или нулю, или единице, тогда

$$P(x_i / y_j) \log P(x_i / y_j) = P(y_j / x_i) \log P(y_j / x_i) = 0 \tag{2.26}$$

и

$$H(X, Y) = H(X) = H(Y).$$

3. Условная энтропия изменяется в пределах: $0 < H(Y/X) < H(Y)$.

4. Для общей энтропии двух источников всегда осуществляется соотношение $H(X, Y) \le H(X) + H(Y)$, при этом равенство выполняется только для независимых источников сообщений.

Итак, *при наличии связи между элементарными сообщениями энтропия источника снижается*, причем тем больше, чем сильнее связь между элементами сообщения.

Энтропия систем с непрерывным множеством состояний вычисляется по правилам анализа дискретных систем с предыдущим квантованием плотности вероятности $w(x)$ с шагом Δx.

Тогда количество состояний в системе будет $N = (x_{\max} - x_{\min})/\Delta x$, а вероятность состояний $p(x_i) = w(x_i)\,\Delta x$.

Воспользовавшись известными формулами данного раздела, можно найти энтропию суммы дискретных сообщений:

$$H_{\Delta x}(x) = \sum_{i=1}^{N} w(x_i)\Delta x \log\{w(x_i)\Delta x\}.$$

После преобразований при условии, которое Δx □> 0, имеем:

$$H_{\Delta x}(x) = H^*(X) - \log \Delta x.$$

Величина $H^*(X)$ называется *сведенной энтропией*:

$$H^*(X) = -\int_{-\infty}^{\infty} w(x)\log w(x)\,dx. \qquad (2.27)$$

Итак, приходим к таким выводам относительно степени информативности источников сообщений:

1. Энтропия источника и количество информации тем больше, чем больше размер алфавита источника.

2. Энтропия источника зависит от статистических свойств сообщения.

3. Энтропия максимальная, если сообщение источника равновероятностное и статистически независимое.

4. Энтропия источника, который вырабатывает не равновероятностные сообщения, всегда меньшая, чем максимально достижимая.

5. При наличии статистических связей между элементарными сообщениями (памяти источника) его энтропия уменьшается.

Пример. Рассмотрим источник с алфавитом, который состоит из букв *а*, *б*, *в*, ..., *ю*, *я*. Будем считать для упрощения, что размер алфавита источника $K = 2^5 = 32$.

Если бы все буквы алфавита имели одинаковую вероятность и были статистически независимыми, то средняя энтропия, которая приходится на один символ, представляла бы $H(\lambda)_{\max} = \log_2 32 = 5$ бит/букву.

Если теперь взять во внимание лишь разную вероятность букв в тексте (а нетрудно проверить, что так оно и есть), расчетная энтропия будет представлять $H(\lambda) = 4{,}39$ бит/букву.

С учетом корреляции (статистической связи) между двумя и тремя соседними буквами (после буквы «п» чаще случается «а» и почти никогда - «ю» и «ц») энтропия уменьшится соответственно $H(\lambda) = 3{,}52$ бит/букву и $H(\lambda) =$

= 3,05 бит/букву. В конце концов, если учесть корреляцию между восьмью и больше символами, энтропия уменьшится к $H(\lambda) = 2{,}0$ бит/букву и дальше будет оставаться без перемен.

Поскольку реальные источники с тем самым размером алфавита могут иметь совсем разную энтропию (а это не только тексты, но и язык, музыка, изображения и т.п.), вводят такую характеристику источника, как *чрезмерность*

$$\rho_u = 1 - H(\lambda) / H(\lambda)_{\max} = 1 - H(\lambda) / \log K,$$

где $H(\lambda)$ - энтропия реального источника; $\log K$ - максимально достижимая энтропия для источника с объемом алфавита из K символов.

Тогда чрезмерность, например, литературного текста представляет

$$\rho_u = 1 - (2 \text{ бит / букву}) / (5 \text{ бит / букву}) = 0{,}6.$$

Иначе говоря, при передаче текста по каналу связи каждые шесть букв из десяти переданных не несут никакой информации и могут без потерь просто не передаваться. Такую же, если не большую ($\rho_u = 0{,}9...0{,}95$), чрезмерность имеют и другие источники информации - язык и особенно музыка, телевизионные изображения и т.п.

Возникает правомерный вопрос: есть ли смысл занимать носитель информации или канал связи передачей символов, которые практически не несут информации, т.е. возможно ли такое преобразование исходного сообщения, в результате которого информация «втискивалась» бы в минимально необходимое для этого количество символов?

2.5. Производительность и избыточность источника информации

Рассмотрим источник двух событий s_1 и s_2. Если событие s_1 случается редко, а событие s_2 - часто, то количество информации о реализации события s_1 будет значительно большей, чем о реализации события s_2: $I(s_1) >> I(s_2)$, где $I(s)$ - количество информации. Очевидно, что количество информации источника двух событий значительно меньше, чем источника 10 или 20 событий, т.е. чем больше разных событий характеризует то или другое явление, тем больше необходимо информации для его описания. Таким образом, информация есть характеристикой такого общего свойства материального мира, как его разнообразие.

Информация всегда подается в виде сигналов, физическая природа которых зависит от типа источника сообщений. Сигналы как носители информации представляют собой механические колебания в твердых материалах, жидкостях, газах (инфразвук, звук, ультразвук), электрические и электромагнитные колебания или волны (радио, оптические). Сигналы воспринимаются приемочными устройствами, в частности органами чувства живых организмов и человека. Из сигналов добывается информация, которая дальше пре-

вращается и запоминается или передается по линиям связи, превращаясь и отображаясь в виде, удобном для восприятия, осмысления и использования человеком при принятии решений.

В *источниках информации* формируются сообщения. *Сообщение* - это последовательность знаков (символов) или непрерывные сигналы, которые содержат те или другие сведенья, данные, результаты измерений. Множество разных знаков, используемых для формирования сообщений, называют *алфавитом источника сообщений*, а количество знаков - *объемом алфавита*. В частности, знаками могут быть буквы естественного языка, цифры, иероглифы.

Непрерывные сообщения не делятся на элементы, являются функциями времени. Типичными примерами могут быть языковые сигналы из выхода микрофона, непрерывно измеренные данные о температуре, давлении, направлении и скорости ветра и т.п. В последнее время они, как правило, превращаются в цифровые с целью повышения качества передачи, хранения и защиты информации.

Производительность источников информации. По обыкновению источники передают сообщения с некоторой скоростью, затрачивая в среднем время T на передачу одного сообщения.

Производительностью источника $H'(X)$ называется суммарная энтропия сообщений, переданных в единицу времени: $H'(X) = H(X)/T$.

Производительность измеряется в битах в секунду.

Представим величину $1/T$ как скорость $v_c = 1T$ (элементов в секунду), получим $H'(X) = v_c \cdot H(X)/T$.

Аналогично, поделив значение энтропии и количества информации на T и представив $H'(X/Y) = H(X/Y)/T$, $I'(X/Y) = I(X/Y)/T$, получим соответствующее равенство для условных энтропий и количества информации, рассчитанных на одно сообщение в единицу времени.

Скоростью передачи информации называется количество информации $I'(X,Y)$ алфавитного ансамбля сигналов на входе системы, отнесенное к единице времени. Если, например, X - ансамбль сигналов на входе дискретного канала, а Y - ансамбль сигналов на его выходе, то скорость передачи информации по каналу

$$I'(X,Y) = H'(X) - H'(X/Y) = H'(Y) - H'(Y/X). \qquad (2.28)$$

где $H'(X/Y)$ - производительность источника переданного сигнала X, а $H'(Y)$ - «производительность» канала, т.е. полная собственная информация в принятом сигнале за единицу времени.

Это соотношение делает наглядным рис. 2.5.

Величина $H'(X/Y)$ является потерей информации, или ненадежностью канала за единицу времени, а $H'(X/Y)$ - скорость создания ошибочной, посторонней информации в канале, которая не касается X и обусловлена присутствующими в канале помехами. По определению К. Шеннона, ненадежность

канала является энтропией входа, когда выход известный, т.е. ее можно считать мерой средней неопределенности принятого сигнала. Что касается величины $H'(X/Y)$, то она представляет собой энтропию выхода, когда вход известный, т.е. является мерой средней неопределенности переданного сигнала.

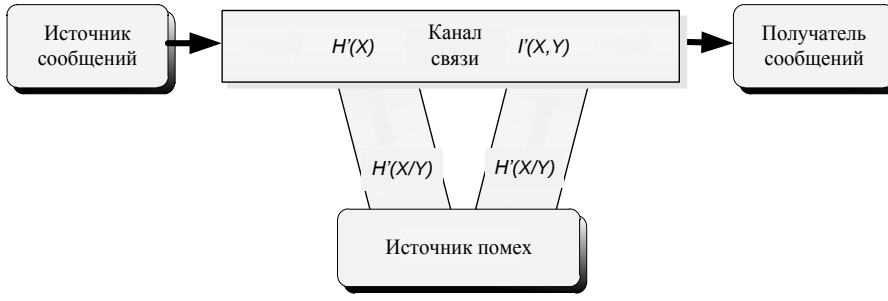


Рис. 2.6. Соотношение между характеристиками канала и скоростью передачи информации

Соотношение между $H'(X/Y)$ и $H'(X/Y)$ зависит от свойств канала. Например, при передаче звукового сигнала по каналу с узкой полосой пропускания, недостаточной для высококачественного воспроизведения сигнала, и с низким уровнем помех теряется часть полезной информации, но почти не получается лишняя информация, т.е. в этом случае $H'(X/Y) >> H'(X/Y)$. Если же сигнал воссоздается на высоком уровне, качественно, но при этом прослушиваются наводки от соседнего радиоканала, то это означает, что почти без потерь полезной информации мы получили много лишней, т.е. избыточной, информации, которая мешает обработке. В этом случае исполняется соотношение $H'(X/Y) >> H'(X/Y)$.

Эффективность и избыточность источников информации. При передаче непрерывных сообщений переданные сигналы являются непрерывными функциями времени $A(t)$, что принадлежат некоторому множеству, а принятые сигналы $X(t)$ будут их обезображенными вариантами. Все реальные сигналы имеют спектры с ограниченной полосой F. В соответствии с теоремой В. А. Котельникова такие сигналы определяются их значениями в точках отсчета, которые содержатся одна от другой на расстоянии $\Delta t = 1/2F$.

В реальных условиях на сигнал накладываются помехи, вследствие чего количество заметных уровней сигнала в точках отсчета будет конечным.

Итак, совокупность значений, которые определяют непрерывный сигнал, эквивалентна некоторой дискретной конечной совокупности. Это дает возможность определить количество информации и пропускную способность канала при передаче непрерывных сообщений на основании результатов, полученных для дискретных сообщений.

Владимир Александрович Котельников (1908 - 2005),
Выдающийся советский и русский ученый в области радиотехники, радиосвязи и радио-астрономии. Основные работы посвящены проблемам усовершенствования методов радиоприема, изучению радиопомех и разработке методов борьбы с ними. К его наибольшим научным достижениям следует отнести открытие теоремы отсчетов, которая носит его имя, создание теории потенциальной помехоустойчивости, а также разработку планетарных радиолокаторов и проведение с их помощью фундаментальных астрономических исследований.

Если переданный сигнал $A(t)$ содержит n точек отсчета a_1, a_2,..., a_n, а принятый $X(f)$ - такое же количество точек отсчета x_1 x_2,..., x_n, то соответственно количество информации, которая содержится в принятом сигнале X относительно переданного A, определится выражением

$$I(X,A)=\log\frac{p(a_1,a_2,...,a_n/x_1,x_2,...,x_n)}{p(a_1,a_2,...,a_n)},$$

где $p(a_1, a_2, ..., a_n) = P(A)$ - распределение вероятностей переданного сигнала (априорное распределение); $p(a_1, a_2, ..., a_n / x_1, x_2, ..., x_n) = P(A / X)$ - распределение условных вероятностей, что характеризует статистические свойства принятых сигналов (апостериорное распределение). После усреднения формулы (2.29) по всем значениям A и X получаем выражение для определения среднего количества информации

$$I(X,\ A)=\int\int P(A,\ X)\log\frac{P(A/X)}{P(A)}, \qquad (2.30)$$

или, после преобразования:

$$I(X,\ A)=H(A)-H(A/X)=H(X)-H(X/A). \qquad (2.31)$$

При этом скорость передачи информации

$$R=\frac{1}{T}\left|H(A)-H(A/X)\right|=\frac{1}{T}\left|H(X)-H(X/A)\right|,$$

где

$$H(A)=-\int\int...\int p(a_1,\ a_2,\ ...,\ a_n)\times \times\log p(a_1,\ a_2,\ ...,\ a_n)da_1da_2...da_n \qquad (2.33)$$

- энтропия переданного сигнала;

$$H(A/X)=-\int\int...\int p(a_1,\ a_2,\ ...,\ a_n;\ x_1,\ x_2,\ ...,\ x_n)\times$$

$$\times\log p(a_1,\ a_2,\ ...,\ a_n/x_1,\ x_2,\ ...,\ x_n)\times da_1...da_ndx_1...dx_n$$

- условная энтропия сигнала.

Поскольку отсчеты a_1, a_2, ..., a_n - независимые случайные величины, то $p(a_1, a_2, ..., a_n) = p(a_1)\ p(a_2)$, ..., $p(a_n)$, и согласно выражению (2.33) после интегрирования находим

$$H(A)=\sum_{i=1}^{n}H(a_i), \qquad (2.34)$$

где $H(a_i)= -\int p(a_i)\log p(a_i)\,da_i$ - энтропия i-го отсчета.

Аналогичными соотношениями определяются $H(X)$ и $H(X/A)$. Пропускная способность канала C определяется как максимум за всеми возможными ансамблями переданных сигналов:

$$C = \max_{P(A)} \frac{1}{T}\left|H(A) - H(A/X)\right| = \max \frac{1}{T}\left|H(X) - H(X/A)\right|. \qquad (2.35)$$

Если сигнал и помехи, которые влияют на него, независимые, то $p(X/A)$ есть функция только разности $X - A = N$ и $p(X/A) = p(N)$. В этом случае

$$H(X/A) = H(N)$$

и

$$R = \frac{1}{T}\left|H(X) - H(N)\right|, \qquad (2.36)$$

$$C = \max \frac{1}{T}\left|H(X) - H(N)\right|. \qquad (2.37)$$

Шум (помехи) в полосе F, равно как и сигнал, может быть представленный по теореме Котельникова n отсчетами $n_1, ..., n_2, ..., n_n$, где n, как и раньше, равняется $2TF$. Величины n_i распределены по нормальному закону, независимые и имеют одинаковую дисперсию σ_3^2. Общее распределение вероятностей этих отсчетов определяется *n-мерным* нормальным распределением. В этом случае энтропия шума

$$H(N) = -\int\int ...\int p(n_1,\ n_2,\ ...,\ n_n) \times$$
$$\times \log p(n_1,\ n_2,\ ...,\ n_n)\, dn_1,\ ...,\ dn_n = nH(n), \qquad (2.38)$$

где $H(n) = -\int p(n) \log p(n)\, dn$ — энтропия одного отсчета. С учетом только что изложенного имеем

$$H(N) = TF \log(2\pi eN), \qquad (2.39)$$

где $N = \sigma_{ш}^2$ — средняя мощность шума.

Можно показать, что это наиболее возможная энтропия любого колебания при заданных значениях F, T и N.

Легко убедиться, что с увеличением ширины полосы пропуска F пропускная способность C возрастает, и при $F \to \infty$ имеем

$$C_\infty = 1{,}443 \frac{P}{N_0}. \qquad (2.40)$$

В реальных каналах связи скорость передачи информации R намного меньше пропускной способности C. Чем большую скорость передачи предполагает данная система, тем эффективнее используется канал.

Наиболее общей оценкой эффективности системы передачи информации является *коэффициент использования канала*, который равняется отношению скорости передачи к пропускной способности:

$$\eta = R / C\,. \qquad (2.41)$$

Роберт Марио Фано (Robert Mario Fano, 1917), заслуженный профессор в отставке по электротехнике и компьютерным наукам в Массачусетском технологическом институте. Фано известен благодаря его работе в области теории информации, изобретению (вместе с Клодом Шенноном) кодирования Шеннона - Фано. В 1947г. получил степень доктора наук в Массачусетском технологическом институте за диссертацию "Теоретические ограничения широкополосного согласования импеданса". Получил награду имени Клода Елвуда Шеннона 1976 г. за свою работу в области теории информации.

Этот коэффициент называют также η-эффективностью канала.

Для идеальной системы $\eta = 1$, для реальных систем $\eta < 1$. При передаче непрерывных сообщений, соответственно, имеем

$$\eta = \frac{F_m \log\left(P_\infty / N_\infty + 1\right)}{F \log\left(P/N + 1\right)}. \qquad (2.42)$$

В общем случае выражение для η - эффективности можно записать в виде произведения двух величин:

$$\eta = \eta_1 \eta_2, \qquad (2.43)$$

где η_1 - эффективность системы кодирования, которое равняется коэффициенту сжатия сообщения,

$$\eta_1 = 1 - r, \qquad (2.44)$$

а η_2 - эффективность системы модуляции.

Введя величину избыточности, представим:

$$\eta_1 = 1 - r_1; \eta_2 = 1 - r_2, \qquad (2.45)$$

где r_1 - избыточность сообщения; r_2 - чрезмерность сигнала.

После подстановки выражения (2.45) в (2.43) имеем

$$\eta = \eta_1 \eta_2 = \left(1 - r_1\right)\left(1 - r_2\right) = 1 - r,$$

где

$$r = r_1 + r_2 - r_1 r_2 \qquad (2.46)$$

- полная чрезмерность системы.

Таким образом, η-*эффективность системы передачи информации полностью определяется значением ее избыточности.* Отсюда задачи повышения эффективности передачи сводятся к задаче уменьшения избыточности сообщения и сигнала, точнее, к рациональному ее использованию.

Избыточность сообщения обусловлена тем, что элементы сообщения не является равновероятностными, и между ними существует статистическая связь.

При кодировании можно перераспределить вероятности исходного сообщения так, чтобы распределе-

ние вероятностей символов приближалось к оптимальному (к равномерному - в дискретном случае или к нормальному - при передаче непрерывных сообщений). Такое перераспределение дает возможность отстранить чрезмерность, которая зависит от распределения вероятностей элементов сообщения. Примером такого кодирования есть код Шеннона - Фано.

Ранее были рассмотрены некоторые способы повышения эффективности передачи за счет устранения избыточности сообщения. Что касается сигнала, то его избыточность зависит от способа модуляции и от вида носителя. Процесс модуляции обычно сопровождается расширением полосы частот сигнала сравнительно с полосой частот переданного сообщения. Это расширение полосы и является избыточным. Частотная избыточность также увеличивается при переходе от синусоидального носителя к носителю импульсному или шумоподобному. С точки зрения повышения эффективности передачи следовало бы выбирать такие способы модуляции, которые имеют малую избыточность. К таким способам принадлежит, в частности, однополосная передача, когда переданные сигналы не содержат частотной избыточности, - они являются просто копиями переданных сообщений.

Однако, говоря об эффективности системы связи, нельзя забывать о ее помехоустойчивости. Устранение избыточности повышает эффективность передачи, но снижает при этом ее вероятность (помехоустойчивость), и наоборот, сохранение или введение избыточности обеспечивает высокую вероятность передачи. Итак, эффективность и помехоустойчивость систем передачи информации нельзя рассматривать отдельно друг от друга. Вопрос повышения эффективности и помехоустойчивости систем передачи информации представляют единую проблему. Задача состоит в том, чтобы отыскать умный компромисс при ее решении.

Измерение информации источников двусимвольных сообщений. Исследуем источник сообщений, состоящий из последовательности двух случайных независимых событий. Обозначим их символами s_1 и s_2. Рассмотрим длинное сообщение вида $s_1 s_2 s_2 s_1 s_1 s_2 \ldots s_1$, в котором n_1 событий s_1 и n_2 событий s_2, так что размер сообщения $n = n_1 + n_2$. Очевидно, что сообщение этого источника такого размера будут отличаться одно от другого, так как количество событий n_1 и n_2 являются случайными величинами. Количество вариантов равняется количеству комбинаций из n по n_1:

$$C_n^{n_1} = \binom{n}{n_1}. \qquad (2.47)$$

Рассмотрим такой же источник сообщений, которое состоит из последовательности событий D_1, D_2, D_3, …, D_m, в которой количество символов D_1 и D_2 равняется m_1 и m_2, получим $m = m_1 + m_2$. Тогда для второго источника:

$$C_m^{m_1} = \binom{m}{m_1}. \qquad (2.48)$$

Джеймс Стирлинг (James Stirling, 1692 - 1770), выдающийся шотландский математик, член Лондонского королевского общества (1729). Образование получил в Баллиол - колледже в Оксфорде. Важнейшая работа - "Разностный метод" (1730), в которой он впервые дал асимптотическое разложение логарифма гамма-функции (так называется ряд Стирлинга), рассмотрел бесконечные произведения. Некоторые из его открытий сделал Л.Эйлер в своих более общих исследованиях. Так называемая формула Стирлинга легко выводится из ряда Стирлинга, но у него самого в явном виде не встречается.

Количество информации каждого из источников зависит от значений $C_n^{n_1}$ и $C_m^{m_1}$. Обозначим количество информации первого источника $I(s)$ и второго $I(D)$. Объединим эти два независимых источника в одно (третье) с символами-событиями Y. Очевидно, что количество информации третьего источника $I(Y)$ должно равняться сумме $I(s) + I(D)$.

Поскольку логарифм произведения равняется сумме логарифмов, то

$$I(Y) = \ln\left(C_n^{n_1} C_m^{m_1}\right) = \ln\left(C_n^{n_1}\right) + \ln\left(C_m^{m_1}\right) = I(s) + I(D). \quad (2.49)$$

Возьмем за меру количества информации источников двусимвольных сообщений число, пропорциональное логарифму ожидаемого количества вариантов сообщений

$$I = k \ln\left(C_n^{n_1}\right) = k \ln\left(\frac{n!}{n_1! n_2!}\right), \quad (2.50)$$

где k - постоянная, зависящая от выбора единицы измерения информации.

Рассмотрим достаточно длинное сообщение, в котором $n \to \infty$, и представим его в виде

$$I = k \ln(n!) - k \ln(n_1!) - k \ln(n_2!). \quad (2.51)$$

Из теории факториалов известно, что когда m - большое число, то оправдывается формула Стирлинга, как

$$\ln(m!) \approx m \ln m - m.$$

Преобразуем (2.51), воспользовавшись формулой Стирлинга:

$$I = k\left(n \ln n - n - n_1 \ln n_1 + n_1 - n_2 \ln n_2 + n_2\right) =$$

$$= -kn\left(\frac{n_1}{n} \ln n_1 + \frac{n_2}{n} \ln n_2 - \ln n\right).$$

Выражение в скобках преобразуем так:

$$\frac{n_1}{n}\left(\ln \frac{n_1}{n} + \ln n\right) + \frac{n_2}{n}\left(\ln \frac{n_2}{n} + \ln n\right) - \ln n = \frac{n_1}{n} \ln \frac{n_1}{n} +$$

$$+ \frac{n_2}{n} \ln \frac{n_2}{n} + \frac{n_1 + n_2}{n} \ln n - \ln n.$$

В результате получим:

$$I = -kn\left(\frac{n_1}{n}\ln\frac{n_1}{n} + \frac{n_2}{n}\ln\frac{n_2}{n}\right). \qquad (2.52)$$

Формула (2.52) дает возможность оценить количество информации источника двусимвольных сообщений по экспериментальным данным: по количеству символов n_1 и n_2 в соответствующем сообщении. Если события s_1 и s_2 являются случайными, то при $n \to \infty$ отношение n_1/n и n_2/n равняются вероятностям появления событий $P(s_1)$ и $P(s_2)$. Тогда формула (3.52) превращается в вид:

$$I = -nk\left[P(s_1)\ln P(s_1) + P(s_2)\ln P(s_2)\right]. \qquad (2.53)$$

Количество информации источника, который приходится на один символ (событие), равняется его производительности:

$$i = \frac{I}{n} = -k\left[p_1 \ln p_1 + p_2 \ln p_2\right], \qquad (2.54)$$

где $p_1 = P(s_1)$; $p_2 = P(s_2)$.

Выражение (2.54) для вычисления информации двусимвольного источника случайных событий называется *формулой К. Шеннона*. Рассмотрим ее подробнее. Поскольку $p_1 + p_2 = 1$, то, обозначив $p = p_1$ и $p_2 = 1 - p$, получим:

$$i = -k\left[p\ln p + (1-p)\ln(1-p)\right] = i(p).$$

Информация, которая приходится на символ, зависит от вероятности одной из событий, и при $p = 0$, а также при $p = 1$ равняется нулю, т.е. $i(0) = i(1) = 0$. Определим условия максимума функции $i(p)$:

$$\frac{di}{dp} = -k\left[1 + \ln p - \ln(1-p) - 1\right] = \ln p - \ln(1-p) = 0.$$

Итак, $$\ln\frac{p}{1-p} = 0, \quad \frac{p}{1-p} = 1, \quad p_{\max} = 0{,}5.$$

Максимальное количество информации, которую выдает источник, достигается с вероятностью появления событий $p_1 = p_2 = 0{,}5$, т.е. $i_{\max} = k \ln 2$. Возьмем такой источник, как эталон, предположив, что его информация на один символ равняется единице ($i_{\max} = 1$). Тогда $k = (\ln 2)^{-1}$ и формула К. Шеннона запишется в виде:

$$i = -\frac{1}{\ln 2}\left(p_1 \ln p_1 + p_2 \ln p_2\right). \qquad (2.55)$$

Запись формулы К. Шеннона можно упростить, изменив основу логарифма. Обозначим $x = \ln p \ln 2$. Тогда $x \ln 2 = \ln p$, $\ln 2^x = \ln p$, $2^x = p$.

Итак,

$$x = \log_2 p = \ln p / \ln 2.$$

Для формулы К. Шеннона получаем новую запись

$$i = -p_1 \log_2 p_1 - p_2 \log_2 p_2,$$

где $\log_2 p$ - двоичный логарифм числа 2.

Пример. Если $p_1 = 0{,}125 = 2\text{-}3$ и $p_2 = 0{,}875$, то $i = 0{,}54356$, что значительно меньше единицы. Это означает, что когда два источника выдали одинаковое количество информации $I_1 = i_1 N_1$, $I_2 = i_2 N_2$ и $I_1 = I_2$, то отношение размеров сообщений обратно пропорционально их производительности, т.е. одну и ту же информацию можно передавать за меньшее время или занимать меньшую память при ее хранении, если превратить сообщение так, чтобы производительность источника была максимальной (или близкой к ней).

Исследуя дальше формулу К. Шеннона, запишем ее в виде

$$\begin{aligned} i &= -\left[P(s_1)\log_2 P(s_1) + P(s_2)\log_2 P(s_2) \right] \\ &= \left[P(s_1) i(s_1) + P(s_2) i(s_2) \right]. \end{aligned} \tag{2.56}$$

Здесь $i(s_1)$ и $i(s_2)$ - количество информации при реализации случайных событий s_1 и s_2, а $P(s_1)$ и $P(s_2)$ - закон распределения их вероятностей.

Вероятностно-статистическое описание источника двусимвольных сообщений i - это математическое ожидание количества информации источника на один символ (одно событие), а $i(s_1)$ и $i(s_2)$ - количество информации при реализации соответственно случайных событий s_1 и s_2:

$$i(s_1) = -\log_2 P(s_1), \quad i(s_2) = -\log_2 P(s_2).$$

Формула (2.56) представляет собой вероятностно-статистическое описание источника двусимвольных сообщений при реализации случайных событий s_1 и s_2.

Измерение информации источников многосимвольных сообщений. Рассмотрим источник информации, сообщения которого состоят из m разных символов $s_1, s_2, s_3, \ldots, s_m$... Допустим известно, что они могут появляться в сообщении независимо одно от другого из вероятностями $p_1, p_2, p_3, \ldots, p_m$... Общее количество возможных сообщений длиной n с количеством символов $n_1, n_2, n_3, \ldots, n_m$ можно вычислить по формуле

$$N = \frac{n!}{\prod_{i=1}^{m} (n_i)!}, \text{ где } n = \sum_{i=1}^{m} n_i. \tag{2.57}$$

Определим количество информации в этом сообщении:

$$I = K \ln N = K\left(\ln (n!) - \sum_{i=1}^{m} \ln (n_i!) \right). \tag{2.58}$$

В частном случае при $m = 2$ получаем формулу (2.56). Воспользовавшись формулой Стирлинга, преобразуем формулу (2.58). В результате получим:

$$I \approx -Kn\sum_{i=1}^{m}\frac{n_i}{n}\ln\left(\frac{n_i}{n}\right).$$

Формула К. Шеннона для вычисления информации источников многосимвольных сообщений при $n \to \infty$ приобретает вид

$$i = -\left(\ln 2\right)^{-1}\sum_{j=1}^{m} P_j \ln P_j. \qquad (2.59)$$

Пример. Рассмотрим сообщения на украинском языке при использовании 32 символов (букв). Вероятности их появления в текстах приведено в табл. 2.1.

Таблица 2.1

Буква	*P*	Буква	*P*	Буква	*P*	Буква	*P*
А	0,062	И	0,016	Р	0,040	Щ	0,003
Б	0,014	И, Ї	0,062	С	0,045	Ь	0,014
В	0,038	Й	0,010	Т	0,053	Ю	0,006
Г	0,013	К	0,028	У	0,021	Я	0,018
Д	0,025	Л	0,035	Ф	0,002		
Е	0,072	М	0,026	Х	0,009		
Є	0,003	Н	0,053	Ц	0,004		
Ж	0,007	О	0,090	Ч	0,012		
З	0,016	П	0,023	Ш	0,006		

Допустим, что вероятности появления букв в тексте отвечают независимым случайным событиям, и получаем по формуле (2.59) количество информации на одну букву текста: $u = 4{,}42$ бит.

Покажем, что производительность источника многосимвольных сообщений будет максимальной, если $p_j = 1/m,\ j = 1, 2, \ldots, m$, т.е. когда появление символов равновероятностное. Поскольку $p_m = 1 - \sum_{j=1}^{m-1} p_j$, то функция (2.59) зависит от $(m-1)$ переменной:

$$i\left(p_1,\ p_2,\ \ldots,\ p_{m-1}\right) = -\frac{1}{\ln 2}\sum_{j=1}^{m-1} p_j \ln p_j - \frac{1}{\ln 2} p_m \ln p_m.$$

Необходимое условие *существования внутреннего максимума многомерных функций* формируется таким образом: все частичные производные в точке максимума должны превращаться в ноль:

$$\frac{\partial i}{\partial p_1} = 0;\ \frac{\partial i}{\partial p_2} = 0, \ldots, \quad \frac{\partial i}{\partial p_{m-1}} = 0.$$

В рассмотренном случае частичные производные

$$\frac{\partial i}{\partial p_j} = -\frac{1}{\ln 2}\left(\ln p_j + 1 + \frac{\partial p_m}{\partial p_j}\ln p_m + \frac{\partial p_m}{\partial p_j}\right).$$

В свою очередь, $$\frac{\partial p_m}{\partial p_j} = -\frac{\partial}{\partial p_j}\left(\sum_{j=1}^{m-1} p_j - 1\right) = -1.$$

В результате получаем

$$\frac{\partial i}{\partial p_j} = -\frac{1}{\ln 2}\ln\left(\frac{p_j}{p_m}\right) = 0, \qquad \text{j = 1, 2,…, m–1...} \tag{2.60}$$

Итак, $p_j = p_m$ для всех $j = 1, 2,\ldots, m–1...$ Но это возможно только в том случае, когда $p_j = 1 / m$, $j = 1, 2,\ldots, m...$ Таким образом, *необходимым условием максимума производительности источника сообщений есть равновероятность появления символов*.

Достаточным условием максимума многомерных функций есть выполнения неравенства

$$\sum_{i=1}^{m-1}\sum_{j=1}^{m-1}\left(\frac{\partial^2 i}{\partial p_i \partial p_j}\right) < 0 \tag{2.61}$$

в точке $p_j = 1/m,\ j = 1, 2, \ldots, m–1...$

Вторые производные определим из формулы (2.60):

$$\frac{\partial^2 i}{\partial p_j^2} = -\frac{1}{\ln 2}\left(\frac{1}{p_j} + \frac{1}{p_m}\right), \qquad \frac{\partial^2 i}{\partial p_i \partial p_j} = -\frac{1}{\ln 2}\frac{1}{p_m}.$$

В точке экстремума имеем

$$\frac{\partial^2 i}{\partial p_j^2} = -\frac{2}{\ln 2}m, \quad \frac{\partial^2 i}{\partial p_i \partial p_j} = -\frac{1}{\ln 2}m.$$

Итак, условия достаточности выполняются, поскольку все производные меньше нуля.

Если бы буквы украинского языка были равновероятностными, то производительность источника сообщения представляла бы

$$i_{\max} = -\sum_{i=1}^{32}\frac{1}{32}\log_2\left(\frac{1}{32}\right) = \log_2 2^5 = 5.$$

Поскольку появление букв не равновероятностное, то информация уменьшается и исчисляется по формуле (2.58).

Заметим, что на самом деле между парами букв существует статистическая связь (корреляция), т.е. если некоторая буква в сообщении есть, то вероятности появления других букв после нее сильно отличаются.

Измерение информации источника двух зависимых случайных событий. Как математическую модель двухсимвольной коррелируемой последовательности используем цепь Маркова.

Если предыдущее событие было s_2, то условные вероятности записываются в виде

$$P(s_1 / s_2),\ P(s_2 / s_2).$$

Обозначим вероятности появления событий s_1 и s_2 на k-м шаге $P_1(k)$ и $P_2(k)$. Предположим, что заданы начальные вероятности $P_1(0)$ и $P_2(0)$, причем

$$P_1(0) + P_2(0) = 1,\ P_1(k) + P_2(k) = 1,$$

а также известные условные вероятности перехода

$$P(s_1 / s_1),\ P(s_2 / s_1),\ P(s_1 / s_2),\ P(s_2 / s_2).$$

Тогда исполняются рекуррентные соотношения как формулы полной вероятности:

$$\begin{aligned} P_1(k) &= P_1(k-1)P(s_1/s_1) + P_2(k-1)P(s_1/s_2), \\ P_2(k) &= P_1(k-1)P(s_2/s_1) + P_2(k-1)P(s_2/s_2). \end{aligned} \qquad (2.62)$$

Для Марковских дискретных случайных последовательностей характерным есть то, что после нескольких переходов вероятности появления событий практически становятся независимыми от начальных условий и, соответственно, от номера шага.

Покажем это на простых примерах**.** Предположим, что $P_1(0)$ =1; $P_2(0)$ = 0; $P(s_1/s_1)$ = 0,5; $P(s_2/s_1)$ = 0,5; $P(s_1/s_2)$ = 0,6; $P(s_2/s_2)$ = 0,4. Результаты расчетов по формуле (2.62) приведены в табл. 2.2.

Таблица 2.2

K	0	1	2	3	4	5
$P_1(k)$	1	0,5	0,55	0,545	0,5454	0,54545
$P_2(k)$	0	0,5	0,455	0,455	0,4545	0,45455

При измененных начальных условиях ($P_1(0)$ = 0, $P_2(0)$ = 1) результаты приведены в табл. 2.3.

Таблица 2.3

K	0	1	2	3	4	5
$P_1(k)$	0	0,6	0,54	0,546	0,5454	0,54546
$P_2(k)$	1	0,4	0,46	0,454	0,4546	0,45454

Из анализа данных табл. 2.2 и 2.3 следует, что независимо от начальных условий уже на s-м шаге устанавливаются значение P_1 = 0,5454 и P_2 = 0,4546.

Поскольку при больших k вероятности не зависят от k, то, считая $P(k)$ = = $P(k$ - 1), из выражения (2.62) получаем формулы для оценки предельных значений вероятностей P_1 и P_2

Марков Андрей Андреевич (1856 - 1922), русский математик, представитель петербургской математической школы. Родился в Рязани. В 1884г. Марков защитил докторскую диссертацию, посвященную непрерывным дробям, в которой доказал и обобщил некоторые неравенства Чебышева, опубликованные раньше без доказательств. Маркову принадлежат также многочисленные работы из разнообразных разделов математического анализа. С конца 1890- х лет главным предметом исследований ученого стала теория вероятностей, ввел новый объект исследования - последовательности зависимых случайных величин, которые со временем получили название марковских цепей.

$$P_1 = \frac{P(s_1/s_2)}{P(s_1/s_2)+P(s_2/s_1)} = P(s_1),$$
$$P_2 = \frac{P(s_2/s_1)}{P(s_1/s_2)+P(s_2/s_1)} = P(s_2). \qquad (2.63)$$

Используя те или иные переходные вероятности, можно формировать модели двусимвольных сообщений с разными статистическими свойствами и исследовать влияние корреляции на количество информации источников сообщений.

Коэффициент корреляции между случайными событиями A и B определяется как соотношение

$$r = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)\left[1-P(A)\right]P(b)\left[1-P(B)\right]}}. \qquad (2.64)$$

Для марковских последовательностей $P(s_1s_2) = P(s_1)\,P(s_2 / s_1) = P(s_2)P(s_1 / s_2)$, $1 - P(s_1) = P(s_2)$, $1 - P(s_2) = P(s_1)$, а соответственно, для коэффициента корреляции получаем формулу

$$r = \frac{P(s_1/s_2) - P(s_1)}{P(s_1)} = \frac{P(s_2/s_1) - P(s_2)}{P(s_2)}. \qquad (2.65)$$

Выведем формулы для определения количества информации источника марковских сообщений. Если состоялось событие s_1 или s_2, то условные математические ожидания $i(s_1)$ и $i(s_2)$ определим за формулами:

$$i(s_1) = P(s_1/s_1)\,i(s_1/s_1) + P(s_2/s_1)\,i(s_2/s_1),$$
$$i(s_1) = P(s_1/s_2)\,i(s_1/s_2) + P(s_2/s_2)\,i(s_2/s_2). \qquad (2.66)$$

Ожидаемое количество информации источника на один символ

$$i = P(s_1)\,i(s_1) + P(s_2)i(s_2). \qquad (2.67)$$

Подставив формулу (2.66) в (2.67), получим:

$$i = P(s_1)P(s_1/s_1)\,i(s_1/s_1) + P(s_1)\,P(s_2/s_1)\,i(s_2/s_1) +$$
$$+P(s_2)P(s_1/s_2)\,i(s_1/s_2) + P(s_2)\,P(s_2/s_1)\,i(s_2/s_2).$$

Учитывая то, что $i(s_1 / s_1) = -\log_2 P(s_1/s_1)$, $i(s_2/s_1) = -\log_2 P(s_2/s_1)$, $i(s_1/s_2) = -\log_2 P(s_1/s_2)$, $i(s_2/s_2) = -\log_2 P(s_2/s_2)$, имеем

$$i = -\sum_{i=1}^{2}\sum_{j=1}^{2} P(s_i s_j)\log_2 P(s_i/s_j). \qquad (2.68)$$

Для Марковских сигналов вероятность произведения двух событий определяется так:

$$P(s_1s_1) = P(s_1)P(s_1/s_1) = \frac{P(s_1/s_1)P(s_1/s_2)}{P(s_1/s_2)+P(s_2/s_1)},$$

$$P(s_1s_2) = P(s_2s_1) = \frac{P(s_1/s_2)P(s_2/s_1)}{P(s_1/s_2)+P(s_2/s_1)}, \qquad (2.69)$$

$$P(s_2s_2) = P(s_2)P(s_2/s_2) = \frac{P(s_2/s_2)P(s_2/s_1)}{P(s_1/s_2)+P(s_2/s_1)}.$$

Пример. Рассмотрим марковское сообщение. Предположим, что

$$P(s_1/s_1) = 3/4;\ P(s_2/s_1) = 1/4;\ P(s_1/s_2) = 1/8\ ;\ P(s_2/s_2) = 7/8.$$

В таком сообщении $P(s_1)=1/3$; $P(s_2)= 2/3$. Если бы символы были независимые, то

$$i = \tfrac{1}{3}\log_2 3 + \tfrac{2}{3}\log_2 \tfrac{3}{2} = 0{,}91183.$$

Условная информация на символы s_1 и s_2 равняется:

$$i(s_1) = \tfrac{3}{4}\log_2 \tfrac{3}{4} + \tfrac{1}{4}\log_2 4 = 0{,}8113,$$

$$i(s_2) = \tfrac{1}{8}\log_2 8 + \tfrac{7}{8}\log_2 \tfrac{8}{7} = 0{,}5428.$$

Ожидаемая информация источника на один символ равняется их среднему значению

$$i = \tfrac{1}{3}i(s_1) + \tfrac{2}{3}i(s_2) = 0{,}6323.$$

Как и ожидалось, она меньшая, чем у источника сообщений с независимыми символами.

Измерение информации при объединении нескольких источников сообщений. Рассмотрим два источника сообщений с алфавитом символов v_1, v_2, …, v_n и u_1, u_2, …, u_m и объединенный источник с парами символов $v_j u_j$, i = 1, 2, …, n; j = 1, 2, …, m... Известный многомерный закон распределения вероятностей $P(v_1, v_2, \ldots, v_n, u_1, u_2, \ldots, u_m) = P(|v|,|u|)$,где $|v| = |v_1, v_2, \ldots, v_n|^T$ - вектор символов первого источника; $|u| = |u_1, u_2, \ldots, u_m|^T$ - вектор символов второго источника; T - знак транспонирования.

Законы распределения $P(|v|) = P(v_1, v_2, \ldots, v_n)$ и $P(|u|) = P(u_1, u_2, \ldots, u_m)$ получаем с объединенного закона:

$$P(|v|) = \sum_{j=1}^{m} P(v_1,\ v_2,\ ...,\ v_n,\ u_1,\ u_2,\ ...,\ u_j,\ ...,\ u_m),$$

$$P(|u|) = \sum_{i=1}^{n} P(v_1,\ v_2,\ ...v_i,\ ...,\ v_n,\ u_1,\ u_2,\ ...,\ u_m). \qquad (2.70)$$

Количество информации каждого из источников и объединенного источника определяем как

$$i\left(|v|\right) = -\sum_{i=1}^{n} P\left(v_i\right)\log_2 P\left(v_i\right), \qquad (2.71а)$$

$$i\left(|u|\right) = -\sum_{j=1}^{m} P\left(u_j\right)\log_2 P\left(u_j\right), \qquad (2.71б)$$

$$i\left(|v|,|u|\right) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(v_i,u_j\right)\lî g_2 P\left(v_i,u_j\right). \qquad (2.71в)$$

Если источники статистически независимые, то $P\left(v_i,u_j\right) = P\left(v_i\right)P\left(u_j\right)$ и тогда

$$i\left(|v|,|u|\right) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(v_i\right)P\left(u_j\right)\log_2\left[P\left(v_i\right)P\left(u_j\right)\right] =$$

$$= -\sum_{i=1}^{n} P\left(v_i\right)\log_2 P\left(v_i\right)\sum_{j=1}^{m} P\left(u_j\right) - \sum_{j=1}^{m} P\left(u_j\right)\log_2 P\left(u_j\right)\sum_{j=1}^{m} P\left(u_j\right)$$

Учитывая то, что $\sum_{i=1}^{n} P\left(v_i\right) = 1$ и $\sum_{j=1}^{m} P\left(u_j\right) = 1,$ получаем:

$$i\left(|v|,|u|\right) = i\left(|v|\right) + i\left(|u|\right), \qquad (2.72)$$

т.е. информация объединенного источника при независимых символах равняется сумме информации источников, которые входят в объединение.

Если источники сообщений статистически зависимые, то двумерный закон распределения записывается в виде:

$$P\left(v_i\ u_j\right) = P\left(v_i\right)P\left(u_j/v_i\right) = = P\left(u_j\right)P\left(v_i/u_j\right).$$

Тогда формулу для вычисления количества информации объединенного источника (2.71*в*) можно преобразовать в такой способ:

$$i\left(|v|,|u|\right) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P\left(v_i\right)P\left(u_j/v_i\right)\left[\log_2 P\left(v_i\right) + \log_2 P\left(u_j/v_i\right)\right].$$

Учитывая то что $\sum_{j=1}^{m} P\left(u_j/v_i\right) = 1$, имеем:

$$i\left(|v|,|u|\right) = -\sum_{i=1}^{n} P\left(v_i\right)\log_2 P\left(v_i\right) + \sum_{i=1}^{n} P\left(v_i\right) i\left(|u|/v_i\right),$$

где $i\left(|u|,v_i\right) = -\sum_{j=1}^{m} P\left(u_j/v_i\right)\log_2 P\left(u_j/v_i\right)$ - условное количество информации второго источника, если имел место символ v_i первого источника. Очевидно,

что $\sum_{i=1}^{n} P(v_i) i(|u|/v_i) = i(|u|/|v|)$ - условное количество информации второго источника относительно первого. Итак,

$$i(|v||u|) = i(v) + i(|u|/|v|). \tag{2.73}$$

Можно показать также, что исполняется соотношение

$$i(|v||u|) = i(|\grave{e}|) + i(|v|/|u|). \tag{2.74}$$

Поскольку $P(v_i) < 1,$ это выполняется неравенство

$$\sum_{i=1}^{n} P(v_i) i(|u|/v_i) < i(|u|).$$

Итак, всегда имеем:

$$i(|v|,|u|) < i(|v|) + i(u),$$

т.е. количество информации объединенного источника меньше суммы информации исходных статистически зависимых источников сообщений.

Измерение информации источников непрерывных сообщений. Интуитивно понятно, что многозначность непрерывных сигналов, которые используются для передачи информации, очень большая. Достаточно изменить значение сигнала в пределах одной точки, чтобы это был уже другой сигнал. Поделим диапазон изменения случайного сигнала *X* как непрерывной, случайной величины на конечное количество *n* малых интервалов ΔX так, чтобы $X_{i+1} = X_i + \Delta X$.

Будем считать, что реализовано значения X_i, если *X* удовлетворяет неравенство

$$X_i - \Delta X/2 \le X \le X_i + \Delta X/2.$$

Вероятность того, что это будет X_i, равняется

$$P(X_i) = \int_{U_i - 0.5\Delta U}^{U_i + 0.5\Delta U} W(X)\, dX \approx W(X_i)\Delta X, \tag{2.75}$$

где $W(X)$ - закон распределения случайного сигнала.

Количество информации, которая содержится в выборке случайных величин X_1, X_2, …, X_n с вероятностями появления $P(X_1)$, $P(X_2)$, …, $P(X_n)$, согласно формуле К. Шеннона имеет вид:

$$I_1 = -n\sum_{i=1}^{n} P(X_i)\log_2 P(X_i) = -n\sum_{i=1}^{n} W(X_i)\Delta X \log_2\left[W(X_i)\Delta X\right].$$

Будем считать, что реализовано значения X_i, если *X* удовлетворяет неравенство

$$X_i - \Delta X/2 \le X \le X_i + \Delta X/2.$$

Жозеф Луи Лагранж (Joseph Louis Lagrange, 1736 - 1813), французский математик и механик. Автор классического трактата "Аналитическая механика", который расширил основы статики и механики, установив общую формулу, известную также как принцип возможных перемещений. Формулу конечных приростов и несколько другие теоремы названы его именем. Лагранж сделал важный вклад во много областей математики, включая вариационное исчисление, теорию дифференциальных уравнений, решение задач на отыскание максимумов и минимумов, теорию чисел (теорема Лагранжа), алгебру и теорию вероятностей.

Вероятность того, что это будет X_i, равняется

$$P(X_i) = \int_{U_i-0.5\Delta U}^{U_i+0.5\Delta U} W(X)dX \approx W(X_i)\Delta X, \quad (2.75)$$

где $W(X)$ - закон распределения случайного сигнала.

Количество информации, которая содержится в выборке случайных величин X_1, X_2, …, X_n с вероятностями появления $P(X_1)$, $P(X_2)$, …, $P(X_n)$, согласно формуле К. Шеннона имеет вид:

$$I_1 = -n\sum_{i=1}^{n} P(X_i)\log_2 P(X_i) =$$

$$= -n\sum_{i=1}^{n} W(X_i)\Delta X \log_2\left[W(X_i)\Delta X\right].$$

Количество информации, которая приходится в среднем на один символ, запишем в виде

$$i_1 = -\sum_{i=1}^{n} W(X_i)\log_2\left[W(X_i)\Delta X\right] - \log_2(\Delta X),$$

причем $\sum_{i=1}^{n} W(X_i)\Delta X = 1$. Заменив сумму интегралом, получим

$$i_1 = -\int_{-\infty}^{\infty} W(X)\lg_2 W(X)dX - \lg_2(\Delta X). \quad (2.76)$$

Первый член в этом выражении имеет конечное значение и зависит только от закона распределения сигнала, который несет информацию, второй зависит от выбора интервала квантования ΔX. Если считать, что при сравнении количества информации разных сигналов интервал квантования для всех будет тот же, то для измерения информации можно воспользоваться формулой

$$i = -\int_{-\infty}^{\infty} W(X)\log_2\left(W|X|\right)dX. \quad (2.77)$$

Среднее количество информации на один отсчет равняется математическому ожиданию двоичного логарифма закона распределения для непрерывных сигналов.

Пример. В каком источнике непрерывных сообщений содержится максимальное количество информации?

Рассмотрим два вида ограничений на сигналы:

1) сигналы имеют конечную мощность (дисперсию);

2) сигналы имеют ограниченную амплитуду ($\pm A$).

Предположим, что случайный сигнал с нулевым математическим ожиданием (постоянная не несет информации) имеет конечную дисперсию. Тогда имеем

$$D = \int_{-\infty}^{\infty} x^2 W(x)\,dx, \quad \int_{-\infty}^{\infty} W(x)\,dx = 1. \tag{2.78}$$

Это означает, что на вид неизвестной функции $W(x)$ накладываются ограничение вида (2.78). Необходимо определить, которая из функций $W(x)$ максимизирует функционал

$$i = -\int_{-\infty}^{\infty} \ln\left[W(x)\right] W(x)\,dx. \tag{2.79}$$

Это задача вариационного исчисления с ограничениями (2.78). Методика ее решения известная. Необходимо сформировать функционал Лагранжа путем присоединения к функционалу (2.79) ограничений (2.78) с помощью множителей λ_1 и λ_2:

$$L = -\int_{-\infty}^{\infty} \ln\left[W(x)\right] W(x)\,dx + \lambda_1\left(D - \int_{-\infty}^{\infty} x^2 W(x)\,dx\right) + \lambda_2\left(1 - \int_{-\infty}^{\infty} W(x)\,dx\right).$$

Функционал Лагранжа в рассмотренном случае сводится к виду

$$L = \left(\lambda_2 + \lambda_1 D\right) - \int_{-\infty}^{\infty} \left[\lambda_2 + \lambda_1 x^2 + \ln W(x)\right] W(x)\,dx. \tag{2.80}$$

Условие максимума - равенство нулю первой вариации функционала (2.80):

$$\delta L = \int_{-\infty}^{\infty} \left[\left(\lambda_2 + \lambda_1 x^2 + \ln W(x)\right)\delta W(x) - W(x)\,\delta\left(\ln W(x)\right)\right] dx = 0.$$

Поскольку вариацию $\delta W(x)$ можно подать как $W(x)\delta[\ln(x)] = \delta W(x)$, то условие экстремума запишется как уравнение относительно неизвестной функции $W(x)$:

$$\ln W(x) + \lambda_2 + \lambda_1 x^2 - 1 = 0. \tag{2.81}$$

Решив это уравнение, получим

$$W(x) = \exp\left(-\lambda_1 x^2 - \lambda_2 + 1\right). \tag{2.82}$$

Для определения множителей Лагранжа используем ограничение (2.78). Первое ограничение дает возможность записать уравнение

$$\int_{-\infty}^{\infty} W(x)\,dx = 1 = \int_{-\infty}^{\infty} e^{-(\lambda_2 - 1)} e^{-\lambda_1 x^2}\,dx.$$

Поскольку $\int_{-\infty}^{\infty} e^{-\lambda_1 x^2} dx = \sqrt{\pi/\lambda_1}$, то связь между λ_1 и λ_2 имеет вид $e^{-(\lambda_2 - 1)} = \sqrt{\lambda_1/\pi}$. Второе ограничение запишется как интеграл

$$D = \sqrt{\frac{\lambda_1}{\pi}} \int_{-\infty}^{\infty} x^2 e^{-\lambda_1 x^2} dx = \frac{1}{2\lambda_1}.$$

Подставив в (2.80) значение для λ_1 и λ_2, получим

$$W(x) = \frac{1}{\sqrt{2\pi D}} e^{-\frac{x^2}{2D}},$$

т.е. максимальное количество информации содержат непрерывные сигналы с нормальным (гауссовским) законом распределения вероятностей.

Вторая задача решается при одном ограничении $\int_{-a}^{a} W(x)\,dx = 1$. Функционал Лагранжа

$$L = -\int_{-a}^{a} \left[\ln W(x)\right] W(x)\,dx + \lambda_1 \left[1 - \int_{-a}^{a} W(x)\,dx\right],$$

а его первая вариация

$$\delta L = -\int_{-a}^{a} \left[\ln W(x) + \lambda_1 - 1\right] \delta W dx = 0.$$

Из уравнения $\ln W(x) + \lambda_1 - 1 = 0$ с учетом ограничения получаем, $W(x) = e^{-(\lambda_1 - 1)} = 1/2a$, т.е. среди ограниченных за амплитудой сигналов максимальное количество информации содержат сигналы с равномерным законом распределения вероятностей.

2.6. Связь информации с параметрами сигналов

Источники информации имеют физическую природу и различаются видом формируемых сообщений, энергетической активностью, вероятностными характеристиками и т.п.

Для анализа интересными являются не только характеристики определенных сообщений, а и потоки сообщений как специфический случайный процесс. В информационных системах информация с носителей разной физической природы (голос, изображение, символы на бумаге, ленте, вибрации и т.п.) преобразуется к универсальному виду и фиксируется на универсальных носителях.

В качестве универсального носителя информации используется электрический сигнал (или материалы, которые имеют электромагнитные свойства, дающие возможность просто снимать с них информацию в виде электрических сигналов).

Электрические сигналы являются носителями информации, а материалы - носители информации - выполняют функции ее хранения.

Как правило, первичные информационные сообщения - язык, музыка, изображение, значение параметров окружающей среды и т.д. - представляют собой функции времени $X(t)$ или других аргументов $X(x,y,z)$ неэлектрической природы (акустическое давление, температура, распределение яркости на некоторой плоскости и т.п.). С целью передачи информации к потребителю эти сообщения обычно превращаются в электрический сигнал, изменения которого во времени $X(t)$ отображают переданную информацию. Такие сообщения называются *непрерывными*, или *аналоговыми*, *сообщениями* (*сигналами*), и для них выполняются условия

$$X(t) \in (X_{\min}, X_{\max}),\ t \in (0, t),$$

т.е. как значение функции, так и значение аргумента для таких сообщений непрерывны или определены для любого значения из интервала, непрерывного как за X, так и за t .

Пример. Преобразование признаков любой физической природы в электрический сигнал можно рассмотреть на примере работы микрофонной цепи телефонного аппарата. Звуковая энергия человека, который говорит, в виде переменного давления согласно информационному сообщению $c(t)$ влияет на микрофон, который содержит внутри угольный порошок. Из-за этого изменяется электрическое сопротивление микрофона, вследствие чего ток $I(t)$ повторяет звуковые колебания функции времени $c(t)$.

Такие преобразования могут осуществляться с использованием цепи переменного тока путем влияния на индуктивность, или емкость, колебательного контура, в результате чего параметры колебательного процесса превратят закономерности информационного процесса в электрический ток.

Если информационное сообщение является некоторой функцией $c(t)$, то электрический сигнал будет иметь вид

$$X(t) = F_c[c(t)],$$

где $F_c[c(t)]$ - оператор преобразования,

Жан Батист Жозеф Фурье (Jean Baptiste Joseph Fourier, 1768 - 1830), французский математик и физик. Основная работа Фурье - "Аналитическая теория теплоты" (1822), где изложена математическая теория теплопроводности. Эта теория стала основанием современных методов математической физики, которые касаются интегрирования уравнений в частных производных при заданных предельных условиях. Метод Фурье, который заключается в представлении функций в виде тригонометрических рядов (рядов Фурье), нашел широкое применение в разных разделах физики и математики. Кроме этого, Фурье построил первую математическую теорию теплового излучения.

Фридрих Вильгельм Бессель (Friedrich Wilhelm Bessel, 1784 - 1846),
немецкий астроном, геодезист, математик, иностранный почетный член Петербургской академии наук (1814). Создал теорию и методы учета инструментальных и личных ошибок в астрономических наблюдениях. Одним из первых он измерил звездный параллакс (1838). В геодезии известные его работы по определению длины секундного маятника. В математике исследовал функции, которые нашли широкое применение в физике, астрономии, технике. Позднее их стали называть его именем - функциями Бесселя.

который должен повторять закон изменения информационного сообщения $c(t)$ в изменении своих параметров, которыми на этом этапе преобразования являются мгновенное значение электрического тока или напряжения, фазы и т.п.

На практике добиваются практически абсолютного сходства функций информационного сообщения $c(t)$ и функции электрического сигнала $X(t)$ при любой сложности их отображения (или с точностью до масштабного множителя). Поэтому их математические модели должны быть одинаковыми.

С целью упрощения анализа сигналов, которые отображают информационное сообщение $c(t)$ произвольной сложности, подают в виде суммы элементарных колебаний $\eta(t)$, которые называются *базисными функциями*

$$c(t) = \sum_{k=0}^{\infty} C_k \eta_k(t), \qquad (2.83)$$

где C_k - коэффициенты.

Как базисные могут использоваться известные системы функций Фурье, Бесселя, Лежандра, Чебышева, Уолша и др.

Классической суммой в этом понимании есть ряд Фурье, в котором базисными есть гармонические колебания, а также базисные функции, имеющие вид $\sin x/x$, расписанные В. А. Котельниковым.

Если информационное сообщение является некоторой функцией $c(t)$, то электрический сигнал будет иметь вид

$$X(t) = F_c[c(t)],$$

где $F_c[c(t)]$ - оператор преобразования, который должен повторять закон изменения информационного сообщения $c(t)$ в изменении своих параметров, которыми на этом этапе преобразования есть наиболее частое мгновенное значение электрического тока или напряжения, фазы и т.п.

На практике добиваются практически абсолютного сходства функций информационного сообщения $c(t)$ и функции электрического сигнала $X(t)$ при любой сложности их отображения (или с точностью до масштабного множителя).

Поэтому их математические модели должны быть одинаковыми.

С целью упрощения анализа сигналов, которые отображают информационное сообщение $c(t)$ произвольной сложности, подают в виде суммы элементарных колебаний $\eta(t)$, которые называются *базисными функциями:*

$$c(t) = \sum_{k=0}^{\infty} C_k \eta_k(t), \qquad (2.83)$$

где C_k - коэффициенты.

Как базисные функции могут использоваться такие известные системы функций, как функции Фурье, Бесселя, Лежандра, Чебышева, Уолша и т.д.

Классической суммой в этом понимании есть ряд Фурье, в котором базисными есть гармонические колебания, а также базисные функции, имеющие вид sin*x*/*x*, расписанные В.А. Котельниковым.

Преобразование Фурье дает возможность перевести информационное сообщение произвольной формы в совокупность элементарных гармонических колебаний. При этом анализ преобразований сигнала сводится к анализу изменений параметров гармонических колебаний амплитуды, фазы и частоты (*элементарных* «*кирпичиков*» *сложного информационного объекта*) *как базисных функций и их весовых коэффициентов.*

Преобразование Фурье переводит анализ сообщений и сигналов в частотную область. Для исследования сигналов во временной области применяется *теорема В.А.Котельникова* (*теорема отсчетов*).

Если функция $X(t)$ *не содержит частот выше* F_m*, то она полностью определяется последовательностью своих значений в моменты времени, которые отличаются друг от друга на*

$$\Delta t = 1/(2F_m).$$

Согласно теореме В.А. Котельникова можно подавать сигналы как функции от любого параметра, не только от времени.

Таким образом, математические модели сообщений и видеосигналов имеют одинаковую функциональную структуру и параметры и пригодные для анализа информационных сообщений.

Адриан Мари Лежандр (Adrien-Marie Legendre, 1752 - 1833), французский математик член Парижской академии наук (1783). Лежандр обосновал и развил теорию геодезических измерений и первым открыл и применил в вычислениях метод наименьших квадратов. В области математического анализа ввел так называемые многочлены Лежандра, преобразование Лежандра, а также исследовал интегралы Эйлера I і II рода. Лежандр доказал сведение эллиптических интегралов к каноническим формам, нашел соответствующие разложения в ряды, составил таблицы их значений.

Рассмотренные видеосигналы $X(t)$ не являются основными носителями информации, поскольку они переносят ее по внутренним цепям.

Носителями информации на далекие расстояния есть совокупность электрических или электромагнитных гармонических колебаний (*в том числе оптического диапазона, где генераторами являются лазеры*). *Частоту основного из таких колебаний называют несущей, а само колебание - несущим (таким, что несет информацию).*

Чтобы данное колебание было несущим, необходимо выполнить два основных условия относительно функций параметров такого сигнала:

1) среда распространения сигналов должна хорошо пропускать колебания с несущей частотой $\omega_н$;

2) частота несущего колебания должна быть намного большей, чем верхняя частота Ω_m в спектре переданного сообщения: $\omega_н \square \Omega_m$, $\Omega_m = 2\pi F_m$ - верхняя частота в спектре сообщения (видеосигнала).

Второе условие вытекает из требования, чтобы за один период несущего колебания информационный параметр почти не изменился, иначе возникнут искажения.

Несущий сигнал можно представить в виде колебания

$$S(t) = A(t)\cos[\omega_н t + \theta(t)] = A(t)\cos\psi(t), \qquad (2.84)$$

в котором амплитуда A(t) или фаза θ(t) изменяется по закону переданного информационного сообщения.

Тем не менее в рассмотренных только что случаях не учитываются корреляционные связи между сообщениями, которые существуют в языках общения людей. Корреляция (взаимосвязь) существует не только между соседними сообщениями, она охватывает параметры этих сообщений, которые приводят к информационной чрезмерности, оцениваемой коэффициентом R.

Исследования показали, что в украинском языке при учете взаимосвязи только между соседними сообщениями (буквами алфавита) $H_1(x) = 4{,}05$ бит/сообщение. При $v = 2$ значение $H_2(x) = 3{,}52$ бит/сообщение. При $v = 3$ значение $H_3(x) = 2{,}97$ бит/сообщение. А при независимых сообщениях $H_0(x) = 5$ бит/сообщение. Поэтому $R_1(x) \approx 19\,\%$; $R_2(x) \approx 30\,\%$; $R_3(x) \approx 41\,\%$. Наличие естественной чрезмерности языка повышает помехоустойчивость сообщений при их передаче, но неэкономно тратит канальные ресурсы, объемы машинной памяти, время на анализ при обработке и т.п.

Исследование источников информации с корреляционными сообщениями удобно проводить с использованием аппарата цепей Маркова. Марковские источники характеризуются состоянием и правилами перехода из одного состояния в другое. Для них характерно, что вероятность любого состояния системы в будущем зависит только от ее состояния в данный момент и не зависит от того, каким образом система пришла в это состояние, т.е. не зависит от предыстории.

Основные выводы

Существуют три подхода к анализу информации: синтаксический, семантический и прагматический.

На синтаксическом уровне учитываются тип носителя и способ представления информации, скорость передачи и обработки, размеры кодов представленной информации, надежность и точность преобразования этих кодов и т.п.

Семантический аспект допускает учет содержания информации.

Для информационного взаимодействия чаще всего используется часть информации, которая содержит определенный (из тех или других соображений допустимый) объем данных. Такие части данных называют информационными объектами, или сообщениями.

Информация является характеристикой не сообщения, а соотношением между источником информации (объектом исследования), сообщением и его потребителем.

Информационные объекты - предметы, процессы, явления материальной или нематериальной природы, которые рассматриваются с точки зрения их информационных свойств.

Источник информации или сообщение - это физический объект, система или явление, которое формирует переданное сообщение. Само сообщение - это значение или изменение некоторой физической величины, которая отражает состояние информационного объекта (системы или явления).

Пути и процессы, которые обеспечивают передачу сообщений от источника информации к ее потребителю, называются информационно-коммуникационными системами.

Степень изменения неопределенности ситуации положено в основу количественной меры информации. При введении количественной меры информации принято не учитывать содержание сообщений (семантику), а ограничиваться только формальными признаками, важными с точки зрения передачи сообщений по каналам связи.

Часто используют простой способ определения количества информации, который можно назвать объемным. Он базируется на подсчете количества символов в сообщении, т.е. связан с его длиной и не учитывает содержание.

В теории информации под количеством информации понимают меру уменьшения неопределенности знания. Нахождение такой меры требует оценивания и учета количества переданной информации. В теории информации количеством информации называют числовую характеристику сигнала, которая не зависит от его формы и содержания и характеризует неопределенность, которая исчезает после получения сообщения в виде данного сигнала.

Для количественного оценивания информации довольно часто применяют синтаксическую, семантическую и прагматическую меры информации.

Количество информации на синтаксическом уровне определяется через понятие энтропии системы.

Энтропия определяет меру неопределенности всего множества сообщений на входе системы и вычисляется как среднее количество собственной информации во всех сообщениях.

Энтропия источника и количество информации тем больше, чем больше размер алфавита источника.

Энтропия источника зависит от статистических свойств сообщений. Энтропия максимальная, если сообщения источника равновероятные и статистически независимые.

Энтропия источника, который вырабатывает равновероятные сообщения, всегда меньше максимально допустимой.

При определении энтропии и количества информации в сообщениях, элементы которых статистически связаны, нельзя ограничиваться только безусловными вероятностями - необходимо учитывать также условные вероятности появления отдельных сообщений.

Общая энтропия двух сообщений равняется сумме безусловной энтропии одного из сообщений и условной энтропии второго.

При наличии связи между элементарными сообщениями энтропия источника снижается.

Универсальный носитель информации - электрический сигнал (или материалы, которые имеют электромагнитные свойства, и дают возможность просто снимать с них информацию в виде электрических сигналов).

С целью упрощения анализа информации, заложенной в параметрах сообщений, сигналы, которые отображают информационное сообщение произвольной сложности, подают в виде суммы элементарных колебаний (базисных функций). Как базисные функции можно использовать известные системные функции Фурье, Бесселя, Лежандра, Чебышева, Уолша и др.

Преобразование Фурье переводит анализ сообщений и сигналов в частотную область. Для исследования сигналов в часовой области применяется теорема В. А. Котельникова (теорема отсчетов).

Вопросы для самоконтроля

1. С проявлениями каких функций информации чаще всего мы сталкиваемся?

2. Насколько актуальны в исследовании информации ее динамические свойства?

3. В чем заключается отличие в понятиях «данные» и «информация»?

4. Приведите примеры сложных информационных объектов.

5. Запишите формулы для вычисления количества информации, которое содержит сообщение.

6. Раскройте свойство симметрии для количества информации.

7. Раскройте свойство аддитивности для количества информации.

8. Запишите формулу для среднего количества информации в сообщении.

9. Какие меры информации используют для ее количественной оценки?

10. Назовите единицы измерения информации.

11. Приведите классификацию мер информации и их свойств.

12. Дайте определение энтропии.

13. Выведите формулу для определения общей энтропии двух источников X и Y.

14. Назовите основные свойства энтропии сложных сообщений.

15. Определите границы, в которых может изменяться условная энтропия.

16. Как изменяется энтропия при наличии статистических связей между элементарными сообщениями?

17. Какие параметры сигнала могут быть функционально зависимыми от информационного сообщения?

The main conclusions

There are three approaches to the analysis of the information: syntactical, semantic and pragmatic

At a syntactical level the type of the carrier and a way of representation of the information, a transmission rate and processing, the sizes of codes of representation of the information, reliability and accuracy of transformation of these codes and so on are taken into account.

The semantic aspect assumes taking into account a content of the information.

The particle of information which contains certain, for these or those reasons, admissible volume of data is oftener used for informational interaction It is accepted to name such particles of data as informational objects or messages.

The information is not the characteristic of messages, but it is a correlation between the source of information (the object of research), the message and its consumer.

Informational objects are subjects, processes, phenomena of material or nonmaterial property that are considered from the point of view of their informational properties.

The Source of information or message is a physical object, system or phenomenon that forms the transmitted message. The message in itself is a value or change of some physical quantity that reflect a state of informational object (system or the phenomenon).

Ways and processes which provide message transmission from a source of information to its consumer are called information-communication systems.

The degree of change of uncertainty of a situation is put in a basis of quantitative measure of the information. While leading of quantitative measure it is not

accepted to take into account a content of messages (semantics), but to be limited only to formal characteristics, important from the point of view of message transmission through communication channels.

A simple and rough way of determining the quantity of information, that can be named a volume way is often used. It is based on counting the quantity of symbols in the message, in other words it is connected with its length also does not consider content.

In information theory the quantity of information is a measure of decrease of uncertainty of knowledge. Finding of such measure demands estimation and registration of quantity of the transmitted information. In information theory the quantity of information is the numerical characteristic of a signal that does not depend on its form and content and characterizes uncertainty that disappears after obtaining the message in the form of the given signal.

Syntactic, semantic and pragmatic measures of the information are often used for a quantitative estimation of the information. The quantity of the information at a syntactical level is determined as a notion of entropy of system.

Entropy defines a measure of uncertainty of a great number of messages on an entry of system and it is calculated as average quantity of own information in all the messages.

The more is the size of the alphabet of a source the more is entropy of a source and the quantity of the information.

Entropy of a source depends on statistical qualities of messages. Entropy is maximal if the messages of a source are equally possible and statistically independent.

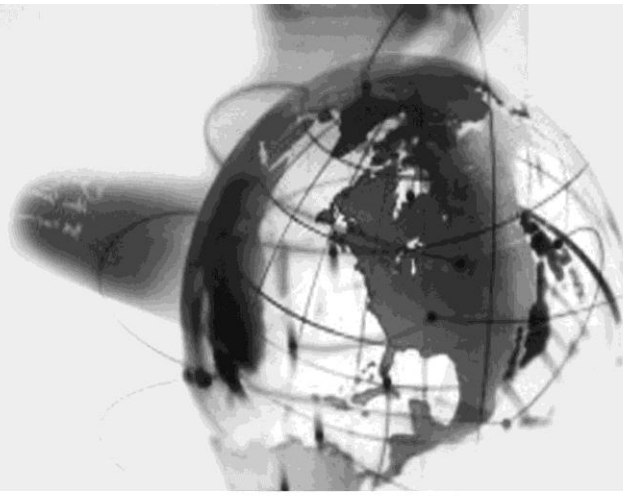
The common entropy of two messages is equal to the sum of unconditional entropy of one of them and conditional entropy of the other.

Signals that display an information message of arbitrary complexity are represented in the form of the sum of elementary oscillations (basic functions) with the purpose of simplification of the analysis of information put in parameters of messages. Many famous systems of functions of such scientists as Furie, Bessel, Lezhandr, Chebishev, Walsh and so on can be used as basic functions.

Furie's transformations convert the analysis of messages and signals in area of frequency. Theorem of V. A. Kotelnikov (sampling theorem) is used for research of signals in time area.

Ключевые слова

Русский	**Английский**
информационное сообщение	information message
источник информации	information source
количество информации	amount of information
мера информации	measure information



ИНФОРМАЦИОННЫЕ СИГНАЛЫ И ИХ МАТЕМАТИЧЕСКИЕ МОДЕЛИ

3

3.1. Виды информационных сигналов и их математические модели
3.2. Случайные сигналы и помехи
3.3. Численные характеристики сигналов и помех
3.4. Математические модели сигналов с ограниченным спектром
3.5. Дискретные сигналы

3.1. Виды информационных сигналов и их математические модели

Согласно с общепринятой терминологией ***информационным сигналом*** (или сигналом) называют *процесс, который характеризует изменение во времени физического состояния некоторого объекта и используется для отображения, регистрации, передачи, принятия и обработки сообщений.*

Информационные сигналы могут иметь разную физическую природу. Широко известны электрические сигналы (ток или напряжение), конкретную реализацию которых можно наблюдать на экране осциллографа или в других устройствах от изображения сигналов; акустические сигналы, воспринимаемые органами слуха, и т.п.

Знание математических моделей сигналов дает возможность сравнивать их между собой, устанавливать тождественность и разногласия, а в конечном итоге - классифицировать.

Рассмотрим ряд критериев (признаков) классификации сигналов:

Критерий пространственно-временного представления сигналов. Любые сигналы существуют в пространстве и во времени. При этом как в пространстве (по величине), так и во времени сигналы могут иметь или непрерывные, или дискретные значения.

Если в качестве классификационного признака взять *характер изменения сигнала по величине и во времени, то возможны такие четыре класса сигналов (рис. 3.1):*

сигналы произвольные по величине и непрерывные во времени;
сигналы произвольные по величине и дискретные во времени;
сигналы квантованные по величине и непрерывные во времени;
сигналы квантованные по величине и дискретные во времени.

Сигналы первого класса (рис. 3.1, *а*) называют *аналоговыми*, или *непрерывными.* Общепринятыми являются такие обозначения математических моделей этих сигналов: $s(t)$, $x(t)$, $u(t)$,... .

Сигналы второго класса (рис. 3.1, *б)* называют *дискретными*, или *дискретизированными*. Термин «дискретный» в этом случае характеризует не сам сигнал, а способ представления его в часовой области. Математическую модель дискретного сигнала в общем случае обозначают как*:*

$$s(t_n),\ x(t_n),\ u(t_n),\ n=0,1,\ \ldots.$$

Временной интервал $T_n = t_{n+1} - t_n$, $n = 0,1,\ \ldots$ называют *n-м* шагом временной дискретизации.

Если шаг временной дискретизации T_n не зависит от *n*, т. е. $T_n = T$, $n = 0,1,\ldots,$ то это значит, что речь идет об эквидистантной во времени дискретизации сигнала. Интервал *T* называют *периодом дискретизации*. Обозначим *математическую модель дискретного сигнала*

$$s(nT), x(nT),\ u(nT),\ n = 0,1,\ldots,$$

или, отбрасывая для упрощения T, в виде $s(n), x(n), u(n), n = 0,1,\ldots.$

Если нет предостережений относительно другого способа часовой дискретизации, то под дискретными сигналами понимают последнюю форму их математических моделей.

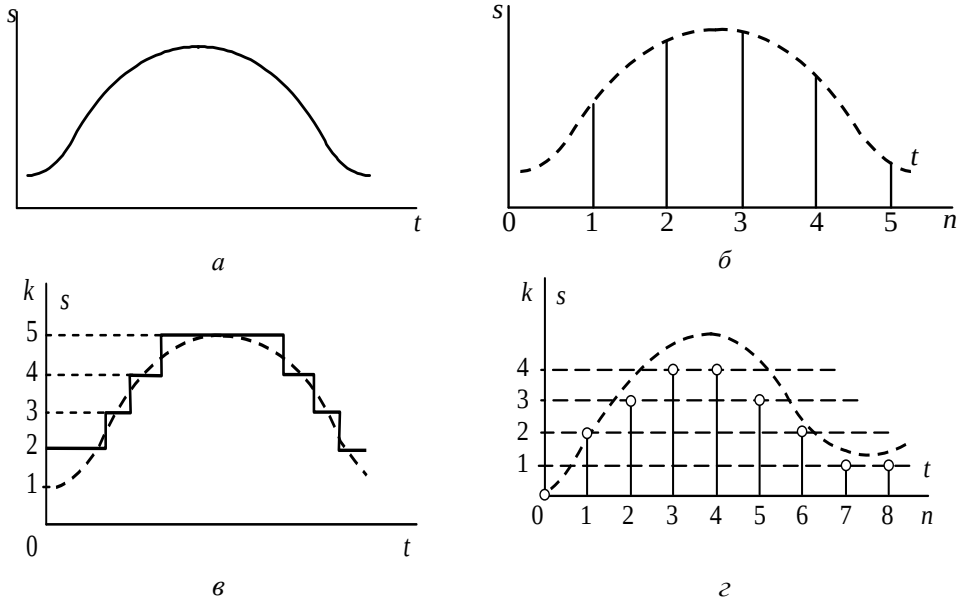


Рис. 3.1. Основные пространственно-временные модели сигналов

Сигналы третьего класса (рис. 3.1, *в*) называют *квантованными*. Для этого класса сигналов осуществляется дискретизация по уровню, т.е. *квантованный сигнал может приобретать только дискретные значения. Математическую модель квантованного сигнала можно обозначить как* $\hat{s}(t)$, *при этом время t непрерывно (в общем случае в интервале от минус бесконечности к плюс бесконечности), а величина* $\hat{s}$ может приобретать одно из множества значений s_k, т.е. $\hat{s} \in [s_0,\ s_1,\ ...,\ s_{m-1}]$.

Шаг квантования сигнала по уровню, как правило, выбирается постоянным.

Сигналы четвертого класса (рис. 3.1, *г*) называют *цифровыми*. *Такие сигналы образуются из аналоговых сигналов в результате их дискретизации во времени и квантования по величине*.

Критерий предвиденья сигналов. Если в качестве классификационного признака взять *предвиденье мгновенного значения сигнала в любой момент*

времени, то всю совокупность сигналов можно разделить на такие два класса: *детерминированные сигналы и случайные сигналы*.

Детерминированными *называются сигналы, значения которых могут быть вычислены в любой момент времени, т.е. они предвидятся с вероятностью, которая равняется единице.* Самым простым примером математической модели детерминированного сигнала может быть гармоническое колебание

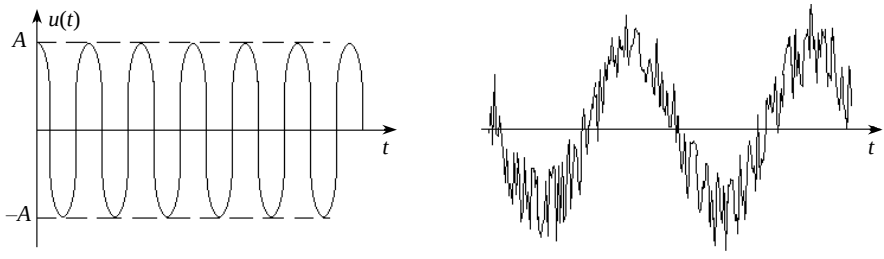
$$u(t) = A\cos(\omega_0 t + \varphi_0), \tag{3.1}$$

где амплитуда A, угловая частота ω_0 и начальная фаза φ_0 колебаний заданы.

Случайными *называются сигналы, значение которых в любой момент времени непредсказуемы, т.е. в заданный момент времени t их невозможно определить с вероятностью, которая равняется единице.*

Примерами случайных сигналов могут быть акустические колебания, воспроизводимые акустической аппаратурой; электромагнитные помехи, создаваемые атмосферными явлениями, и тому подобное.

Осциллограмма детерминированного сигнала (3.1) с начальной фазой $\varphi_0 = 0$ изображена на рис. 3.2, *а*, а случайного - на рис. 3.2, *б*.



а *б*

Рис. 3.2. Осциллограммы типичных сигналов:
а - детерминированный; *б* - случайный

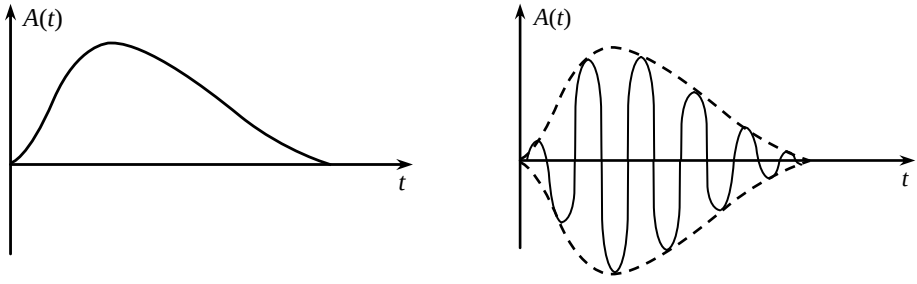
Все сигналы, несущие информацию, являются случайными, поскольку детерминированный (полностью известный) сигнал информации не содержит. Он может быть полностью воспроизведен в месте принятия без передачи по каналу связи.

Между детерминированными и случайными сигналами нет непреодолимой границы. Очень часто в условиях, когда уровень помех значительно меньше уровня полезного сигнала с известной формой, более простая детерминированная модель оказывается вполне адекватной поставленному заданию.

Критерий области существования сигналов. Если за классификационный признак взять *длину часового интервала, в пределах которого существу-*

ет сигнал, то можно выделить такие классы сигналов: бесконечные во времени сигналы и *импульсные сигналы*, т.е. сигналы, которые существуют в пределах конечного интервала времени.

Общая форма записи бесконечных во времени сигналов $s(t)$, $t \in (-\infty, \infty)$, а импульсных сигналов $s(t)$, $t \in [a,b]$, где a *и* b границы временного интервала, в котором существует сигнал.



а *б*

Рис. 3.3. Импульсные сигналы

В электро- и радиотехнике широко используется понятие видеоимпульсов (рис. 3.3, *а*) и радиоимпульсов (рис. 3.3, *б*). Их характерные особенности легко установить по временным диаграммам. Отличие между этими двумя основными импульсными сигналами заключается в следующем. Если $A(t)$ видеоимпульс, то соответствующий ему радиоимпульс

$$u_{\text{р}}(t) = A(t)\cos(\omega_0\, t + \varphi_0), \tag{3.2}$$

причем сигнал $A(t)$ называется *огибающей* радиоимпульса $u_p(t)$, а функция $\cos(\omega_0 t + \varphi_0)$ его *высокочастотным заполнением.*

Из сравнения выражений (3.1) и (3.2) выплывает, что математическую модель радиоимпульса $u_p(t)$ можно получить из модели гармонического колебания $u_p(t)$, заданного формулой (3.1), если в ней выполнить замену амплитуды A (постоянной величины) на функцию времени $A(t)$, что описывает видеоимпульс.

Критерий размерности сигналов. Допустим, что наблюдается сигнал в виде напряжения $u_p(t)$ на зажимах любого элемента электрического цепи или тока $i(t)$, что протекает в заданной ветке цепи. Такой сигнал $u_p(t)$ или $i(t)$, что описывается одной функцией времени, называют *одномерным сигналом*. Так называемые *многомерные сигналы* вида $\vec{V}(t) = \{v_1(t), v_2(t), ..., v_N(t)\}$ со-

стоят из некоторого множества одномерных сигналов.

Целое число N называют *размерностью* сигнала. В качестве примера многомерного сигнала можно привести систему напряжений на зажимах многополюсника.

Критерий динамики сигналов. За характером изменения во времени различают *статические* и *динамические сигналы. В статической модели сигнала нет часового параметра. Такие сигналы используются прежде всего для хранения информации в виде цифровых кодов, например в ячейках памяти цифровых вычислительных машин, в программируемых логических матрицах и т.п. Динамические сигналы зависят от времени. Их математические модели содержат часовой аргумент.*

Критерий вещественности сигналов. Часовые функции, с помощью которых задаются модели сигналов, могут приобретать как вещественные, так и комплексные значения. Выбор той или другой модели сигнала предопределяется лишь простотой математического анализа. За так называемым *критерием вещественности* все сигналы можно разделить на два класса: *вещественные и комплексные сигналы.*

Все рассмотренные ранее модели сигналов принадлежат к классу вещественных. Математическую модель комплексных сигналов можно подать в общем виде

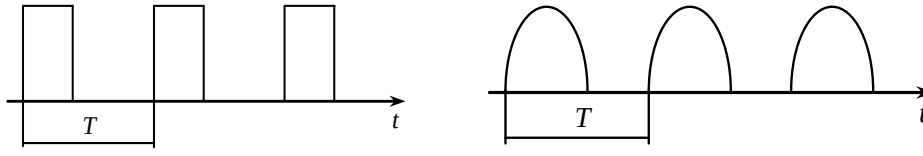
$$\dot{z}(t) = x(t) + jy(t), \tag{3.3}$$

где $x(t)$ и $y(t)$ - действительные сигналы; $j = \sqrt{-1}$ - мнимая единица.

Частным случаем модели (3.3) является так называемый комплексно-экспоненциальный сигнал $\dot{x}(t) = e^{j\omega t} = \cos\omega t + j\sin\omega t,$ применение которого продуктивно при решении задач спектрального анализа.

Критерий повторяемости сигналов. *Периодичность* также может быть классификационным признаком сигналов. За этим критерием различают два основных класса сигналов: *периодические* и *апериодические* (*одиночные*) *сигналы.*

Общая форма математической модели периодического сигнала имеет вид $x(t) = x(t + nT),\ n = \pm 1, \pm 2, ...,$ где T - период сигнала. Некоторые периодические сигналы изображены на рис. 3.4.



а *б*

Рис. 3.4. Примеры периодических сигналов:
а - прямоугольные; *б* - срезанные косинусные

Апериодический (импульсный, или одиночный) сигнал $x(t)$ является частным случаем периодического сигнала и получается из него, когда период следования импульсов T устремляется к бесконечности, т.е.

$$x(t)=\lim_{t\to\infty} x(t+nT).$$

Вся совокупность рассмотренных видов сигналов и их классификационных признаков (критериев) сведена в таблицу

Классификационные признаки сигналов	Типы сигналов
Характер изменения по величине и во времени	Аналоговые
	Дискретные
	Квантованные
	Цифровые
Предвидение мгновенных значений	Детерминированные
	Случайные
Длина интервала существования	Импульсные
	Бесконечные
Размерность	Одномерные
	Многомерные
Динамичность	Статические
	Динамические
Действительность	Действительные
	Комплексные
Повторяемость	Периодические
	Одиночные

Рассмотрим математические модели самых простейших типовых сигналов.

Гармонические сигналы. Гармонические сигналы называют еще тригонометрическими сигналами. Математическая модель таких сигналов определяется формулой (3.1), а типичная осциллограмма изображена на рис. 3.2, *а*.

Комплексно-экспоненциальные сигналы. Математическая модель комплексно-экспоненциальных сигналов имеет вид

$$\dot{z}(t)=Ae^{j(\omega_0 t+\varphi_0)}. \qquad (3.4)$$

Воспользовавшись формулой Эйлера, представим модель (3.4) следующим образом:

$$\dot{z}(t)=A\cos(\omega_0 t+\varphi_0)+jA\sin(\omega_0 t+\varphi_0). \qquad (3.5)$$

Слагаемые в правой части уравнения (3.5) - соответственно вещественная и мнимая составляющие комплексно-экспоненциального сигнала (3.4).

Представим выражение (3.5) в несколько иной форме:

$$\dot{z}(t) = x(t) + jy(t),$$

где $x(t)$ и $y(t)$ - действительные сигналы; $j = \sqrt{-1}$ - мнимая единица.

На рис. 3.5 изображена структурная схема устройства физического моделирования пары действительных сигналов $x(t)$ и $y(t)$, которые еще называются *квадратурными сигналами:*

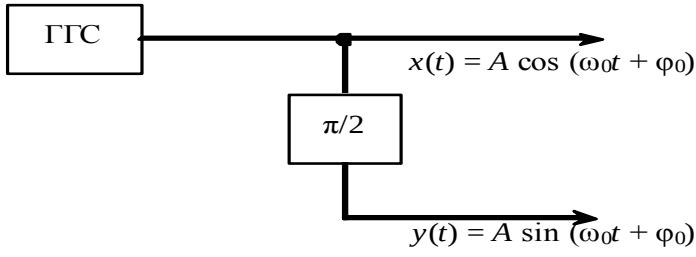


Рис. 3.5. Схема формирования квадратурных сигналов:
ГГС - генератор гармонических сигналов; π/2 - фазовращатель

Комплексность сигнала $\dot{z}(t)$, составленного из вещественных сигналов $x(t)$ и $y(t)$, можно особенным способом организовать при программной или аппаратурной реализации соответствующих алгоритмов преобразования сигнала $\dot{z}(t)$.

Прямоугольные видеоимпульсы. Математическая модель одиночного прямоугольного видеоимпульса, симметрично расположенного относительно начала отсчета времени (рис. 3.6), задается соотношением

$$s(t) = \begin{cases} U,\ t \in [-\tau/2, \tau/2], \\ 0,\ \ t \notin [-\tau/2, \tau/2], \end{cases}$$

где U - амплитуда; τ - длительность прямоугольного видеоимпульса.

В общем случае видеоимпульс может быть смещен относительно начала отсчета времени вправо (*задержанный сигнал*; рис. 3.7, *а*) или влево (*опережающий сигнал;* рис. 3.7*, б).*

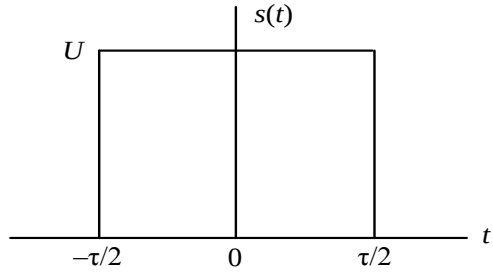


Рис. 3.6. Прямоугольный видеоимпульс

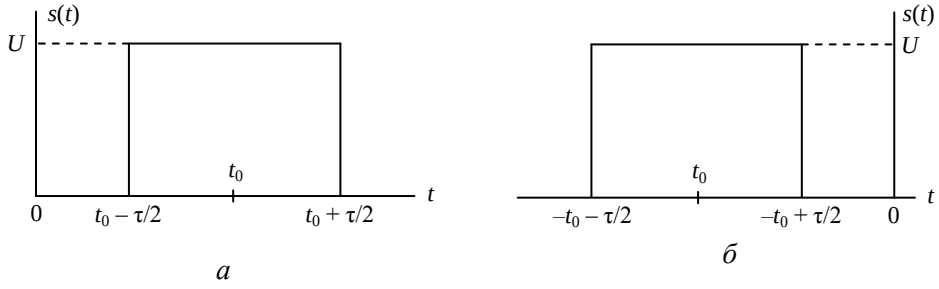


Рис. 3.7. Смещенные по времени прямоугольные видеоимпульсы

Математические модели смещенных во времени прямоугольных видеоимпульсов можно записать в виде систем уравнений:

$$s(t)=\begin{cases}U,\ t\in[t_0-\tau/2,\ t_0+\tau/2]\\ 0,\ \ t\notin[\cdot]\end{cases}$$ - для задержанного сигнала;

$$s(t)=\begin{cases}U,\ t\in[-t_0-\tau/2,-t_0+\tau/2]\\ 0,\ \ t\notin[\cdot]\end{cases}$$ - для опережающего сигнала.

В общем виде запись $s(t-t_0)$ относится к модели задержанного сигнала, а $s(t+t_0)$ - к модели опережающего сигнала, причем t_0 означает интервал временного смещения сигнала.

Треугольные видеоимпульсы. Математическая модель одиночного треугольного видеоимпульса, симметрично расположенного относительно начала отсчета времени (рис. 3.8), задается системой равенств:

$$s(t)=\begin{cases} U\left(1+\dfrac{t}{\tau/2}\right), & t\in[-\tau/2,0]; \\ U\left(1-\dfrac{t}{\tau/2}\right), & t\in[0,\tau/2]; \\ 0, & t\notin[-\tau/2,\tau/2]. \end{cases}$$

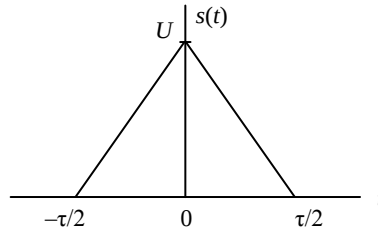


Рис. 3.8. Треугольный видеоимпульс

Треугольный видеоимпульс в общем случае также может быть смещенным относительно начала отсчета времени (рис. 3.9).

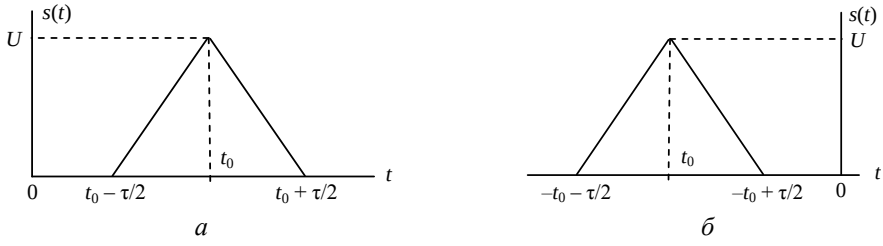


Рис. 3.9. Смещенные во времени треугольные видеоимпульсы

Математическая модель треугольного сигнала, задержанного на промежуток времени t_0 (рис. 3.9, *а*), имеет вид

$$s(t-t_0)=\begin{cases} U\left(1+\dfrac{t-t_0}{\tau/2}\right), & t\in[t_0-\tau/2,t_0]; \\ U\left(1-\dfrac{t-t_0}{\tau/2}\right), & t\in[t_0,t_0+\tau/2]; \\ 0, & t\notin[t_0-\tau/2,t_0+\tau/2]. \end{cases}$$

Для опережающего треугольного видеоимпульса (рис. 3.9, *б)*

$$s(t+t_0)=\begin{cases} U\left(1+\dfrac{t+t_0}{\tau/2}\right), & t\in\left[-t_0-\tau/2,\ -t_0\right];\\ U\left(1-\dfrac{t+t_0}{\tau/2}\right), & t\in\left[-t_0,-t_0+\tau/2\right];\\ 0, & t\notin[-t_0-\tau/2,\ -t_0+\tau/2]. \end{cases}$$

Ступенчатые сигналы. Ступенчатые сигналы (рис. 3.10) еще называют *функциями включения*, *сигма-функциями* или *функциями Хевисайда.*

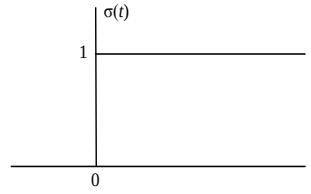


Рис. 3.10. Ступенчатый сигнал

Математическая модель ступенчатого сигнала аналитически описывается системой равенств

$$\sigma(t)=\begin{cases} 1, & t\ge 0;\\ 0, & t<0. \end{cases} \qquad (3.6)$$

В общем случае ступенчатая функция может быть смещена относительно начала отсчета времени (рис. 3.11) на величину t_0 (задержки или опережения).

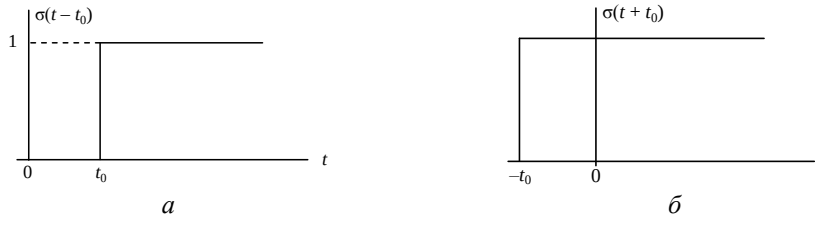


Рис. 3.11. Смещенные ступенчатые сигналы:
а - задержанный; *б* - опережающий

Математическая модель ступенчатого сигнала, задержанного на промежуток времени t_0, имеет вид

$$\sigma(t-t_0)=\begin{cases}1, & t\ge t_0;\\ 0, & t<t_0.\end{cases}$$

Математическая модель опережающей на промежуток времени t_0 ступенчатой функции определяется системой

$$\sigma(t+t_0)=\begin{cases}1, & t\ge -t_0;\\ 0, & t<-t_0.\end{cases}$$

Графики, приведенные на рис. 3.11, отображают характер сдвига задержанного и опережающего сигналов относительно начала отсчета времени.

Дельта-функция. Рассмотрим импульсный сигнал $v(t;\tau)$ прямоугольной формы, основа которого равняется τ, а высота h - величина, обратная τ: $h=1/\tau$. Очевидно, что при любом τ площадь Π_v, ограниченная таким сигналом, равняется единице: $\Pi_V=\tau h=1$. Графики сигналов $v(t;\tau)$, симметрично размещенных относительно начала отсчета времени, для двух значений τ изображены на рис. 3.12.

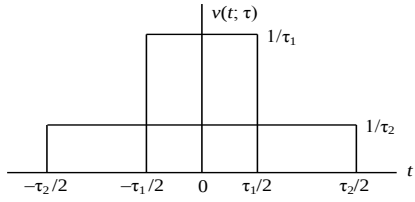


Рис. 3.12. Изображение функций $v(t;\tau)$

Этот импульс характерен тем, что при любом выборе параметра τ его площадь равняется единице:

$$\Pi_v=\int_{-\infty}^{\infty} v\,dt \equiv 1. \qquad (3.7)$$

Пусть теперь величина τ стремится к нулю. Импульс, сокращаясь по длительности, сохраняет свою площадь, поэтому его высота должна неограниченно возрастать. Предел такой функции при $\tau\to 0$ называется *дельта-функцией*, или *функцией Дирака*

$$\delta(t)=\lim_{\tau\to 0} v(t;\tau). \qquad (3.8)$$

Дельта-функция равняется нулю везде, за исключением точки $t=0$, в которой она приобретает бесконечное значение, т.е.

$$\delta(t)=\begin{cases}\infty, & t=0;\\ 0, & t\neq 0.\end{cases} \tag{3.9}$$

Относительно дельта-функции, заданной выражением (3.9), говорят, что она сосредоточена в точке $t=0$. Однако, как следует из соотношений (3.7) и (3.8), площадь, ограниченная дельта-функцией, конечна

$$\int_{-\infty}^{\infty}\delta(t)dt=1. \tag{3.10}$$

Выражение (3.10) определяет *условие нормирования дельта-функции*. Для сокращения вместо дельта-функции записывают иногда δ-функция. Также как и ступенчатая, дельта-функция может быть смещена на некоторый промежуток времени t_0 (рис. 3.13).

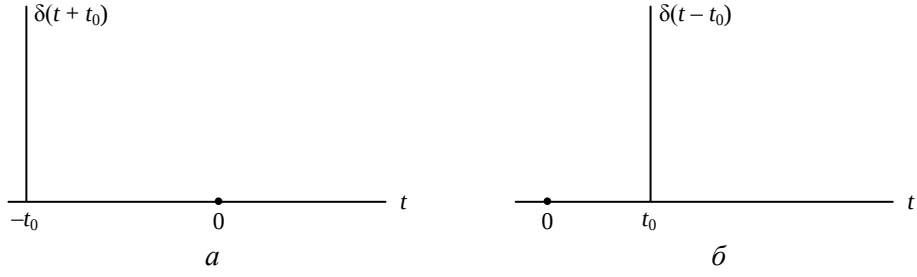


Рис. 3.13. Изображения дельта-функций:
а - задержанная; *б* - опережающая

Смещенную справа на оси времени дельта-функцию называют задержанной дельта-функцией. Ее модель записывают системой равенств

$$\delta(t-t_0)=\begin{cases}\infty, & t=t_0;\\ 0, & t\neq t_0.\end{cases}$$

В случае смещения дельта-функции влево относительно начала отсчета времени ее называют опережающей дельта-функцией. Ее модель такова:

$$\delta(t+t_0)=\begin{cases}\infty, & t=-t_0;\\ 0, & t\neq -t_0.\end{cases}$$

Между дельта-функцией и ступенчатой функцией существует взаимо одно-

значное соответствие. Это соответствие можно легко установить с помощью вспомогательной функции, *математическая* модель которой задается системой равенств

$$u(t;\tau)=\begin{cases}0, & t<-\tau/2;\\ \dfrac{t}{\tau}+\dfrac{1}{2}, & -\tau/2\le t\le\tau/2;\\ 1, & t>\tau/2.\end{cases} \qquad (3.11)$$

График этой функции достаточно простой (рис. 3.14).

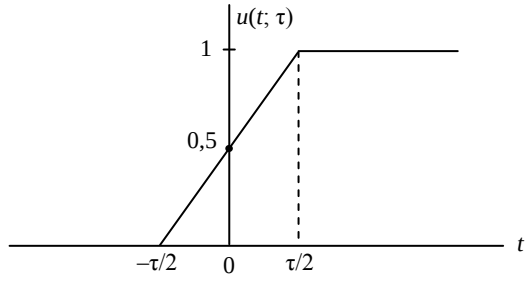


Рис. 3.14. Функция $u(t;\tau)$

Из соотношения (3.11) и рис. 3.14 непосредственно вытекает, что, во-первых, при $\tau\to 0$ функция $u(t;\tau)$ переходит в ступенчатую функцию

$$\lim_{\tau\to 0}u(t;\tau)=\sigma(t), \qquad (3.12)$$

и, во-вторых, производная по времени от функции $u(t;\tau)$ совпадает с функцией $v(t;\tau)$, заданной соотношениям *(3.11)*, т.е. $u'(t;\tau)=v(t;\tau)$, а при $\tau\to 0$

$$\lim_{\tau\to 0}u'(t;\tau)=\delta(t). \qquad (3.13)$$

Из предельных соотношений (3.12) и (3.13) непосредственно следует, что дельта-функция есть производная от ступенчатой функции

$$\delta(t)=\sigma'(t).$$

В равной степени правильно и обратное соотношение: *ступенчатая функция* есть *интеграл от дельта-функции*

$$\sigma(t)=\int_{-\infty}^{t}\delta(\tau)\,d\tau.$$

Если некоторую непрерывную функцию $x(t)$ умножить на дельта-функцию $\delta(t-t_0)$ и произведение проинтегрировать по времени t, то результат будет равняться значению непрерывной функции в той точке t_0, где сосредоточен δ-импульс. Действительно, в соотношении $\int\limits_{-\infty}^{\infty} x(t)\,\delta(t-t_0)\,dt$ подинтегральное выражение отличается от нуля лишь для $t=t_0$. В этой точке функция $x(t)$ приобретает конкретное значение $x(t_0)$, которое можно вынести из-под знака интеграла. Часть интеграла, которая осталась, согласно условию нормировки дельта-функции (3.10), равна единице. Таким образом, получаем выражение

$$x(t_0)=\int\limits_{-\infty}^{\infty} x(t)\,\delta(t-t_0)\,dt.$$

В этом проявляется *фильтрующее свойство дельта-функции*, которое можно использовать в устройствах измерения мгновенных значений некоторого сигнала $x(t)$. Структурная схема измерителя (рис. 3.15) состоит из двух звеньев: умножителя и интегратора.

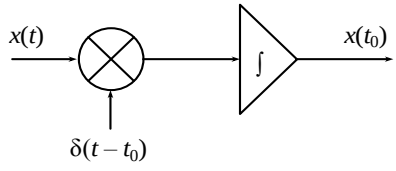


Рис. 3.15. Измеритель мгновенных значений сигнала

Значение $x(t_0)$ будет измерено тем точнее, чем короче реальный сигнал, который приближенно подает дельта-функция.

Сигнум-функция. Функция $\operatorname{sign}(t)$, математическая модель которой представляется системой равенств

$$\operatorname{sign}(t)=\begin{cases}1, & t>0;\\ 0, & t=0;\\ -1, & t<0,\end{cases}$$

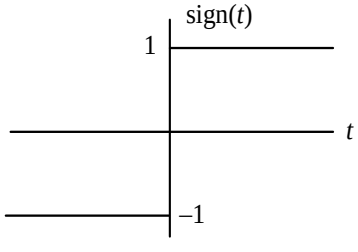


Рис. 3.16. Сигнум-функция

называется *сигнум-функцией* (рис. 3.16).

Ступенчатая функция $\sigma(t)$ и сигнум-функция $\operatorname{sign}(t)$ связаны такими очевидными соотношениями

$$\sigma(t)=\frac{1}{2}\left(1+\operatorname{sign}(t)\right);\ \operatorname{sign}(t)=2\sigma(t)-1.$$

Леонард Эйлер (Leonhard Euler, 1707-1783), выдающийся математик, который сделал огромный вклад в развитие математики, а также механики, физики, астрономии и ряда прикладных наук. Эйлер - автор свыше 800 работ по математическому анализу, дифференциальной геометрии, теории чисел, приближенных вычислений, небесной механики, математической физики, оптики, баллистики, кораблестроения, теории музыки и многих других, которые имели значительное влияние на развитие науки. В течение 1731 - 1741 гг. и начиная с 1766 г. был академиком Петербургской академии наук.

Энергетические характеристики сигнала. Известные из физики понятия мощности и энергии электрических сигналов можно легко обобщить и перенести на соответствующие понятия мощности и энергии произвольных сигналов. Рассмотрим простую электрическую цепь (рис. 3.17), которая состоит из генератора напряжения $u(t)$, к клеммам которого подключено резистор с сопротивлением R. В цепи протекает ток $i(t)$, значения которого определяется по закону Ома:

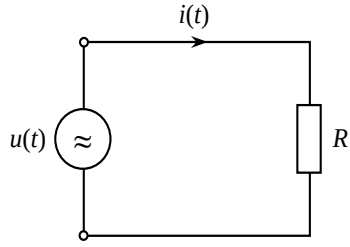


Рис. 3.17. К определению мощности сигнала

$$i(t)=\frac{u(t)}{R}.$$

Мгновенное значение мощности генератора $p(t)$ равняется произведению напряжения $u(t)$ и тока $i(t)$:

$$p(t)=u(t)\,i(t).$$

Энергия E, отдаваемая генератором в сеть (выделяемая на резисторе R), в интервале времени $[a, b]$ определяется соотношением

$$E=\int_a^b p(t)dt=\int_a^b u(t)\,i(t)dt. \qquad (3.14)$$

В частном случае, когда $R=1$, ток $i(t)$ в цепи равняется напряжению генератора $u(t)$, а значение энергии E_u вычисляется по формуле

$$E_u=\int_a^b u^2(t)dt.$$

Обобщим понятие мощности и энергии в электрической цепи, в которой протекает ток $i(t)$ под воздействием напряжения $u(t)$ и перенесем их на соответствующие понятия мощности и энергии произвольных сигналов $u(t)$ и $v(t)$. Итак, пусть $u(t)$ и $v(t)$ - некоторые сигналы (точнее, математические модели сигналов). По аналогии с соотношением (3.14) произведение

$$p_{u,v}(t) = \frac{1}{2}u(t)v(t)$$

будем называть *взаимной мощностью сигналов,* а интеграл

$$E_{u,v} = \int_{-\infty}^{\infty} u(t)v(t)dt \tag{3.15}$$

- *взаимной энергией* этих *сигналов*.

Тогда полная энергия сигнала при $v(t)$, тождественно равном $u(t)$, определится выражением

$$E_u = \int_{-\infty}^{\infty} u^2(t)dt. \tag{3.16}$$

В математике величина $\int_{-\infty}^{\infty} u(t)v(t)dt$ называется *скалярным произведением функций* $u(t)$ *и* $v(t)$ и обозначается как

$$(u,v) = \int_{-\infty}^{\infty} u(t)v(t)dt. \tag{3.17}$$

Из сравнения выражений (3.17) и (3.15) вытекает, что *скалярное произведение двух сигналов* (*функций*) *это не что иное, как взаимная энергия этих сигналов:*

$$(u,v) = E_{u,v}. \tag{3.18}$$

Как частный случай из формулы (3.18) получаем

$$(u, u) = E_u.$$

Таким образом, *полная энергия сигнала определяется как скалярное произведение сигнала с самим собой.*

Перенесем понятие взаимной и полной энергии на комплексные сигналы. Если попробуем непосредственно по формуле (3.16) вычислить полную энергию комплексного сигнала, т.е. воспользуемся выражением

$$E_{\dot{u}} = E_{\dot{u}} = = \int_{-\infty}^{\infty} \dot{u}^2(t)\,dt,$$

то придем к тому, что энергия такого сигнала также окажется комплексной (поскольку квадрат комплексной функции есть функция комплексная), а это недопустимо. Поэтому взаимную энергию двух комплексных сигналов $\dot{u}(t)$ и $\dot{v}(t)$ зададим соотношением

$$E_{\dot{u},\dot{v}} = (\dot{u},\dot{v}^*) = \int_{-\infty}^{\infty} \dot{u}(t)\,\dot{v}^*(t)\,dt, \qquad (3.19)$$

где $v^*(t)$ - сигнал, комплексно-сопряженный с сигналом $v(t)$.

Из соотношения (3.19) как частный случай получим полную энергию комплексного сигнала $\dot{u}(t)$:

$$E_{\dot{u}} = (\dot{u},\dot{u}^*). \qquad (3.20)$$

Произведение комплексной и комплексно-сопряженной с ней величины (функции) равняется квадрату модуля этой величины: $\dot{u}\dot{u}^* = |\dot{u}|^2$. Поскольку квадрат модуля комплексной функции - функция вещественная, полная энергия комплексного сигнала становится вещественной величиной, т.е.

$$E_{\dot{u}} = \int_{-\infty}^{\infty} |\dot{u}(t)|^2\,dt.$$

Корень квадратный из полной энергии сигнала (функции) называют *нормой сигнала (функции)*

$$\|\dot{u}\| = \sqrt{\int_{-\infty}^{\infty} |\dot{u}(t)|^2\,dt}.$$

Функцию $u(t)$, для которой выполняется условие $\|u\|^2 = 1$, называют *нормированной*.

Таким образом, *нормированными называются функции, полная энергия которых равняется единице*.

Ортогональные сигналы и ортогональные базисы. Два сигнала $u(t)$ и $v(t)$ называются *ортогональными*, если их скалярное произведение, а следовательно, и взаимная энергия, равняются нулю

$$(u,v)=\int_{-\infty}^{\infty} u(t)v(t)dt=0.$$

Бесконечная система действительных функций (сигналов)

$$\{\varphi_k(t)\}=\{\varphi_0(t),\ \varphi_1(t),\ ...,\ \varphi_n(t),\ ...\}$$

называется *ортогональной*, если скалярное произведение двух разных сигналов, а, следовательно, и их взаимная энергия, равняется нулю:

$$(\varphi_n,\ \varphi_m)=\int_{-\infty}^{\infty}\varphi_n(t)\varphi_m(t)dt=0 \qquad \text{при} \qquad n\neq m.$$

Предполагается, что энергия каждого сигнала из системы не равна нулю, т. е.

$$\int_{-\infty}^{\infty}\varphi^2(t)\,dt\neq 0.$$

Это, в частности, значит, что ни одна из функций системы не равняется тождественно нулю.

Бесконечную систему функций $\{\varphi_k(t)\}$, попарно ортогональных друг другу, и таких, которые имеют единичные нормы

$$(\varphi_n,\ \varphi_m)=\begin{cases}1, & n=m;\\ 0, & n\neq m,\end{cases} \tag{3.21}$$

называют *системой ортонормированных* функций, или *ортонормированным базисом.*

Совсем не обязательно, чтобы ортонормированность базисных функций обеспечивалась лишь на бесконечном интервале времени. Существуют многочисленные базисы, в которых системы функций $\{\varphi_k(t)\}$ ортонормированы на конечном интервале времени.

Пример. Рассмотрим один из наиболее важных и распространенных базисов, что образуется ортонормированной системой гармонических функций. На интервале времени $[-T/2, T/2]$ система тригонометрических функций с

кратными частотами, дополненная постоянным во времени сигналом φ_0:

$$\varphi_0(t) = 1/\sqrt{T};$$
$$\varphi_1(t) = \sqrt{2/T}\sin 2\pi t/T;$$
$$\varphi_2(t) = \sqrt{2/T}\cos 2\pi t/T;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\varphi_{2k-1}(t) = \sqrt{2/T}\sin 2\pi kt/T;$$
$$\varphi_{2k}(t) = \sqrt{2/T}\cos 2\pi kt/T, \ldots,$$

образует ортонормированный базис.

В ортогональности этих функций легко убедиться геометрическим построением (рис. 3.18) на примере функций φ_1 и φ_2.

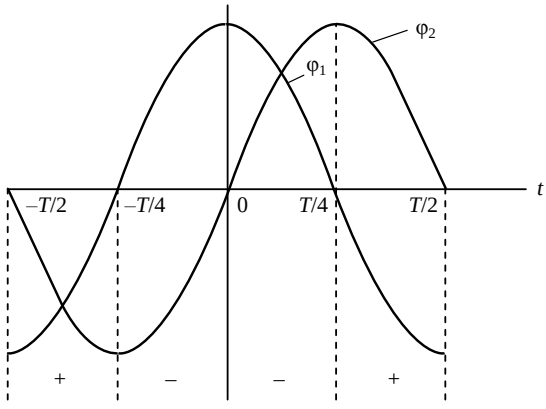


Рис. 3.18. Гармонические функции

Определим энергию базисной функции $\varphi_k(t)$ на интервале ортогональности:

$$E_k = \int_{-T/2}^{T/2} \varphi_k^2(t)\,dt. \qquad (3.22)$$

Очевидно, что энергия, а следовательно, и норма функции $\varphi_0(t)$ равняется единице.

При вычислении энергии функций $\varphi_k(t)$, $k \geq 1$ по формуле (3.22) примем во внимание, что

$$\int \sin^2(pt)\,dt = \frac{t}{2} - \frac{\sin(2pt)}{4p},$$

а

$$\int \cos^2(pt)\,dt = \frac{t}{2} + \ + \frac{\sin(2pt)}{4p}.$$

Пусть

$$\varphi_k(t) = \sqrt{\frac{2}{T}}\cos\frac{2\pi kt}{T},$$

тогда для этой функции

$$E_{\varphi_k} = \int_{-T/2}^{T/2} \varphi_k^2(t)\,dt = \frac{2}{T}\left(\frac{t}{2} + \frac{\sin 4\pi kt/T}{8\pi k/T}\right)\Bigg|_{-T/2}^{T/2}. \quad (3.23)$$

Подставив границы интеграции в правой части выражения (3.23), получим, что энергия (так же, как и норма) гармонических базисных функций (как синусных, так и косинусных) равняется единице.

Рассмотренный базис гармонических функций является одной из многих систем ортонормированных функций.

Выбор той или другой системы ортонормированных функций (*базиса*) *зависит как от свойств сигнала, так и от конкретных условий применения базиса.*

Общие рекомендации относительно выбора базиса можно сформулировать следующим образом:

1. *Базисные функции должны быть по мере сил максимально согласованными за формой с фильтруемым сигналом.*

Например, если априори известно, что некоторый сложный сигнал содержит гармоническую составляющую, информацию о параметрах которой получаем на основании разложения сигнала в обобщенный ряд Фурье, то как базисные функции целесообразно использовать именно систему ортонормиро-

Оливер Хевисайд (Oliver Heaviside, 1850-1925), английский ученый самоучка, инженер, математик и физик. Создал теорию передачи сигналов на далекие расстояния. Впервые применил комплексные числа для изучения электрических цепей, разработал технику применения преобразования Лапласа для решения дифференциальных уравнений, сформулировал уравнение Максвелла в терминах электрической и магнитной сил и потока, а также независимо от других математиков создал векторный анализ. Впервые разработал операции неисчисления, которые широко применяются в физике и других науках.

ванных гармонических функций, согласованных по форме с формой фильтруемого сигнала.

2. *Нужно стремиться к выбору максимально простого с точки зрения аппаратной или программной реализации базиса*.

Эти рекомендации не всегда удается выполнить в полном объеме, поскольку их реализация может приводить к конфликтным ситуациям. В связи с этим иногда возникает необходимость поиска разумного компромисса между согласованностью формы базисных функций с формой фильтруемого сигнала и простотой программной или аппаратной реализации базиса.

Обобщенные ряды Фурье. Из математики известно, что любой сигнал $s(t)$ с конечной энергией, т.е. такой, для которого выполняется условие

$$\int_{-\infty}^{\infty} s^2(t)\,dt < \infty,$$

можно подать в виде ряда

$$s(t) = \sum_{k=0}^{\infty} c_k \varphi_k(t), \qquad (3.24)$$

где c_k - коэффициенты разложения, называемые *спектром сигнала*; φ_k - система ортонормированных вещественных функций (*базис*).

Представление (*3.24*) называется *обобщенным рядом Фурье сигнала* $s(t)$ *в выбранном базисе* $\{\varphi_k\}$.

К разложению (3.24) некоторого сигнала $s(t)$ следует относиться как к такому, который вводится аксиоматически, т.е. можно пытаться аппроксимировать сигнал $s(t)$ бесконечной суммой произведений предварительно выбранных действительных функций $\varphi_k(t)$ (которые совсем не обязательно должны быть ортонормированными) и пока не известных коэффициентов разложения c_k.

Задача заключается в том, чтобы выбрать такую систему функций $\{\varphi_k\}$, *которая бы, во-первых, обеспечивала простоту вычисления коэффициентов разложения* $\{c_k\}$ *и, во-вторых, минимизировала погрешность аппроксимации сигнала* $s(t)$ *конечномерным рядом Фурье.*

В дальнейшем будет показано, что если в качестве системы $\{\varphi_k\}$ *взять совокупность ортонормированных функций, то обозначенная задача может быть успешно решена.*

Следовательно, пусть в формуле (3.24) $\{\varphi_k\}$ бесконечная система вещественных функций, ортонормированных на интервале времени $[-\infty, \infty]$, удовлетворяющих условию (3.21), т.е.

$$\int_{-\infty}^{\infty} \varphi_k(t)\varphi_n(t)\,dt = \begin{cases} 1, & n = k; \\ 0, & n \neq k. \end{cases} \qquad (3.25)$$

Умножим обе части разложения (3.24) на $\varphi_n(t)$ и проинтегрируем в пределах области ортогональности

$$\int_{-\infty}^{\infty} s(t)\varphi_n(t)\,dt = \sum_{k=0}^{\infty} c_k \int_{-\infty}^{\infty} \varphi_k(t)\varphi_n(t)\,dt. \qquad (3.26)$$

Согласно ограничению (3.25) интеграл в правой части выражения (3.26) отличен от нуля и равен единице только при $k = n$. Следовательно,

$$c_n = \int_{-\infty}^{\infty} s(t)\ \varphi_n(t)\ dt = (s, \varphi_n). \qquad (3.27)$$

Соотношение (3.27) является фундаментальным в теории рядов Фурье и определяет алгоритм вычисления коэффициентов разложения сигнала $s(t)$ для заданного ортонормированного базиса, а именно: *коэффициенты разложения обобщенного ряда Фурье* (*спектр*) *временной функции (сигнала)* $s(t)$ *определяются скалярным произведением этого сигнала с соответствующими базисными функциями*

$$c_k = (s(t),\ \varphi_k(t)),\ k = 0, 1, \ldots \qquad (3.28)$$

Возможность представления сигналов посредством обобщенных рядов Фурье является фактом принципиального значения. Их достоинство и удобство применения таких рядов при анализе сигналов состоит в том, что вместо того, чтобы изучать функциональную зависимость в континуальном (несчетном) множестве точек, достаточно характеризировать эти сигналы счетной (бесконечной) системой коэффициентов обобщенного ряда Фурье c_k.

Поль Адриен Морис Дирак (Paul Adrien Maurice Dirac, 1902-1984), английский физик, один из творцов квантовой механики, лауреат Нобелевской премии по физике 1933 г. (вместе с Ервином Шредингером). Работы Дирака посвящены квантовой механике, квантовой электродинамике, теории поля, теории элементарных частиц, статистической физике. В 1926 - 1927 гг. разработал математический аппарат квантовой механики - теорию преобразований, ввел так называемую дельта-функцию. В 1927 г. применил принципы квантовой механики к электромагнитному полю и построил модель квантовой электродинамики.

Отметим еще раз, что *спектр сигнала всегда определяется относительно конкретного базиса. С изменением базиса изменяется и спектр, хотя сигнал остается неизменным.*

Для комплексного сигнала $s(t)$ его разложение в ряд Фурье записывается в виде

$$\dot{s}(t) = \sum_{k=0}^{\infty} \dot{c}_k \varphi_k(t), \tag{3.29}$$

т.е. поскольку система базисных функций $\{\varphi_k\}$ есть система вещественных функций, то для комплексного сигнала $s(t)$ *его спектр* $\{\dot{c}_k\}$ также оказывается комплексным и определяется соотношением

$$\dot{c}_k = \left(\dot{s}(t), \varphi_k(t)\right). \tag{3.30}$$

Вычислим энергию в общем случае комплексного сигнала, представленного в форме обобщенного ряда Фурье (3.29). Воспользовавшись соотношениями (3.29) и (3.30), на основании выражения (3.20) получим

$$E_{\dot{s}} = \left(\dot{s}(t), \dot{s}^*(t)\right) = \int_{-\infty}^{\infty} \dot{s}(t)\dot{s}^*(t)dt = \int_{-\infty}^{\infty} \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \left(\dot{c}_k \dot{c}_n^*\right)\varphi_k(t)\varphi_n(t)dt.$$

Поменяв местами операции суммирования и интегрирования (в силу их линейности), получим

$$E_{\dot{s}} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\left(c_k c_n^*\right)\int_{-\infty}^{\infty} \varphi_k(t)\varphi_n(t)\,dt.$$

Поскольку базисные функции ортонормированны, то в сумме правой части последнего выражения будет отличаться от нуля только слагаемое при $n = k$. В итоге приходим к очень важному результату

$$E_{\dot{s}} = \sum_{k=0}^{\infty} \dot{c}_k \dot{c}_k^* = \sum_{k=0}^{\infty} c_k^2,$$

где $c_k = |\dot{c}_k|$ - *модуль k-го коэффициента разложения обобщенного ряда Фурье.*

Смысл полученного выражения состоит в том, что *энергия сигнала равна сумме энергий всех компонентов* (*гармоник*), *из которых состоит обобщенный ряд Фурье.*

Другими словами, *энергия сигнала равна сумме энергий спектральных составляющих сигнала, а квадрат модуля коэффициентов обобщенного ряда Фурье численно равен той доле энергии сигнала, которая содержится в соответствующей спектральной составляющей сигнала.*

Докажем оптимальность разложения сигнала по ортонормированному базису. Для вещественного сигнала $s(t)$ *введем конечномерную аппроксимацию*

$$\tilde{s}(t) = \sum_{n=0}^{N} c_n \varphi_n(t)$$

с неизвестными коэффициентами c_n и потребуем, чтобы эти коэффициенты были выбраны из условия минимальности энергии ошибки аппроксимации, которое иначе называется *условием минимальности квадратической ошибки* аппроксимации

$$\mu = \|s - \tilde{s}\|^2 = \int_{-\infty}^{\infty} \left[s(t) - \sum_{n=0}^{N} c_n \varphi_n(t) \right]^2 dt = \min.$$

Необходимое условие минимума заключается в том, что коэффициенты c_k должны удовлетворять системе линейных уравнений

$$\frac{\partial \mu}{\partial c_k} = 0,\ k = \overline{0, N}. \qquad (3.31)$$

В развернутом виде погрешность аппроксимации

$$\mu = \int_{-\infty}^{\infty} \left(s^2 - 2s \sum_{k=0}^{N} c_k \varphi_k + \sum_{k=0}^{N} \sum_{n=0}^{N} c_k c_n \varphi_k \varphi_n \right) dt. \qquad (3.32)$$

Для простоты записи аргумент t базисных функций φ в выражении (3.32) опущен. Рассмотрим последнее слагаемое в выражении (3.32)

$$A_{kn} = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{N} \sum_{n=0}^{N} c_k c_n \varphi_k \varphi_n \right) dt.$$

Согласно свойству линейности операции интегрирования и суммирования можно поменять местами, а слагаемое A_{kn} записать в виде

$$A_{kn} = \sum_{k=0}^{N} \sum_{n=0}^{N} c_k c_n \int_{-\infty}^{\infty} \varphi_k \varphi_n dt.$$

Якоб Бернулли (Jakob I (James) Bernoulli, 1655-1705), Швейцарский математик, брат Йогана Бернулли, старший из знаменитой династии ученых. Якобу Бернулли принадлежат значительные достижения в теории рядов, дифференциальному исчислении, вариационному исчислении, теории вероятностей и теории чисел, где его именем названы числа с некоторыми важными свойствами. Якобу Бернулли принадлежат также работы по физике, арифметике, алгебре и геометрии.

Поскольку базис $\{\varphi_k\}$ ортонормирован, то, воспользовавшись системой (3.25), получим $A_{kn} = \sum_{k=0}^{N} c_k^2$. Тогда соотношение (3.32) можно представить в виде

$$\mu = \int_{-\infty}^{\infty} \left(s^2 - 2s \sum_{k=0}^{N} c_k \varphi_k \right) dt + \sum_{k=0}^{N} c_k^2 . \qquad (3.33)$$

Принимая во внимание, что в формуле (3.33) интеграл $\int_{-\infty}^{\infty} s^2(t)dt$ не зависит от коэффициента c_k, после дифференцирования правой части выражения (3.33) по c_k получим

$$-2\int_{-\infty}^{\infty} s(t)\varphi_k(t)dt + 2c_k = 0,$$

а это непосредственно приводит к выводу о том, что равенства (3.31) будут выполняться при следующем выборе коэффициентов разложения:

$$c_k = \int_{-\infty}^{\infty} s(t)\varphi_k(t)dt = (s, \varphi_k). \qquad (3.34)$$

Это полностью совпадает с выражением (3.27) для коэффициентов обобщенного ряда Фурье.

Более тщательный анализ, когда рассматривается не только первая, но и вторая производная энергии ошибки, показывает, что *ряд Фурье обеспечивает не просто экстремум, а именно минимум ошибки аппроксимации. Эта ошибка тем меньше, чем больше N. В пределе, когда N стремится к бесконечности, ошибка становится равной нулю и разложение типа* (*3.24*) *точно описывает сигнал* $s(t)$.

В заключение рассмотрим структурную схему устройства (рис. 3.19) для экспериментального определения коэффициентов разложения произвольного сигнала $s(t)$ в обобщенный ряд Фурье (3.24) по заданной системе ортонормированных функций $\varphi_k(t)$, $k = 0, 1, ...$

Основные элементы устройства это генераторы базисных функций (ГФ) $(\varphi_k(t), k = 0, 1, ...)$, по которым проводится разложение сигнала $s(t)$ в обобщенный ряд Фурье (осуществляется вычисление коэффициентов разложения, совокупность которых образует спектр c_k, $k = 0,1, ...$). Анализируемый сигнал $s(t)$ одновременно подается на первые входы все множительных звеньев, на

вторые входы которых подводятся базисные функции $\varphi_k(t)$, $k = 0,1,...$ С выхода умножителей сигналы поступают на интеграторы, откликом которых являются коэффициенты c_k.

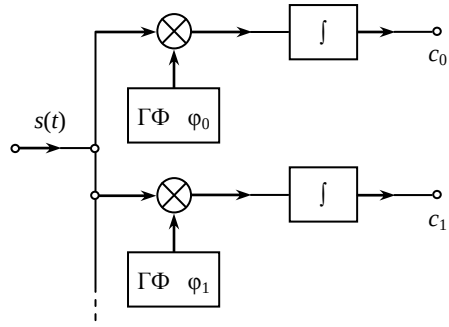


Рис. 3.19. Структурная схема устройства спектрального анализа сигналов

При таком методе обработки сигнала в конце промежутка времени интегрирования (t_1, t_2) на выходе каждого интегратора возникает неизменный во времени сигнал, значение которого согласно с формулой (3.33) в точности равно тому или иному коэффициенту разложения обобщенного ряда Фурье.

Следовательно, работоспособность системы в целом будет зависеть от того, насколько точно удается воссоздать базисные функции, а также от совершенства функционирования умножителей и интеграторов.

Совсем не обязательно вычислять спектр сигнала аппаратным способом. Алгоритм определения коэффициентов разложения обобщенного ряда Фурье можно реализовать и на программном уровне.

Анализируя устройство (см. рис. 3.19), убеждаемся в том, что *вся информация, заключенная в сигнале* $s(t)$, *может быть представлена в виде хотя и бесконечной, но все же счетной совокупности чисел* c_k, $k = 0,1,...$, *образующих спектр сигнала.*

Этот вывод вполне очевиден, поскольку является следствием представления сигнала в виде обобщенного ряда Фурье (см. формулу (3.24)).

Итак, *любая функция* (или *сигнал*), *аналитически описанная во временном пространстве* (*такую функцию называют оригиналом*), *может быть представлена счетной совокупностью комплексных амплитуд* $\dot{c}_k$, $k = 0,1,...$, *образующих спектр в частотном пространстве (совокупность* $\dot{c}_k$ *называют изображением). Между оригиналом и его изображением существует взаимно однозначное соответствие, т.е., зная оригинал, можно найти изображение и наоборот. Это указывает на дуальность процесса в смысле равнозначности его представления или в виде функции времени (в пространстве оригиналов), или в виде совокупности элементарных частотных гармоник (в пространстве изображений).*

3.2. Случайные сигналы и помехи

Главное отличие между случайными и детерминированными сигналами заключается в том, что после наблюдения их на конечном отрезке времени $T_{сп}$ невозможно предусмотреть их будущее. Все случайные сигналы и помехи являются непредсказуемыми. Таким образом, для случайных сигналов невозможно найти математическую формулу, по которой можно было бы рассчитать их мгновенные значения. Случайные сигналы и помехи принадлежат к тем явлениям природы, изучением основных закономерностей которых занимается теория вероятностей [3].

Одна из задач, которая решается в теории вероятностей, - нахождение таких *характеристик случайных явлений*, которые были бы *неслучайными и давали возможность проводить математические вычисления характеристик случайных явлений. Исследования выполняются статистическими методами, для которых характерным является принципиальный отказ от определения результатов каждой отдельной попытки и переход к рассмотрению массовых попыток, т.е. попыток, которые осуществляются многократно в таких же условиях. Определяемые при этом характеристики называются статистическими.*

Все случайные явления, которые изучаются в теории вероятностей, можно разделить на три типа: случайные события; случайные величины; случайные процессы. Каждый из этих типов случайных явлений имеет свои особенности и характеристики.

Реальные случайные сигналы и помехи, как и детерминированные, могут быть простыми и сложными, аналоговыми, дискретными и цифровыми. Вообще же все они - дискретные или непрерывные функции времени, причем в зависимости от конкретных условий их можно рассматривать и как случайные величины, и как случайные события. Для их математического описания, то есть выбора математической модели сигнала, необходимо решить две задачи:

1) установить, к какому типу случайных явлений отнести случайный сигнал (помеху) в конкретной ситуации;

2) определить необходимые статистические характеристики.

Напомним важнейшие понятия теории вероятностей, необходимые для выбора математической модели случайных сигналов и помех.

Случайные события. *Случайным называют событие, которое в результате попытки может наступить или не наступить.* Это и передача текста без ошибок, и работа канала связи без повреждений не менее чем $T_{сп}$ часов, и превышение помехой данного значения и тому подобное. Обозначаются случайные события начальными большими буквами латинской азбуки *A, B, C.*

Числовыми характеристиками возможности наступления какого-то события *A* в тех или других условиях попытки есть частота появления события и ее вероятность.

Частота наступления события A в данной серии попыток - это отношение количества попыток m, в которых наступило событие A, к общему количеству попыток n: $\nu = m / n$.

Вероятность события

$$P(A) = \lim_{n\to\infty} \nu = \lim_{n\to\infty} m / n, \qquad (3.35)$$

т.е. вероятностью случайного события $P(A)$ является частота его наступления при неограниченном увеличении количества независимых однородных попыток. Это доказал *швейцарский математик Я.Бернулли*. С достаточной для практических расчетов точностью можно считать: если количество попыток, в которых наступило событие A, большее 20, то частота случайного события совпадает с его вероятностью $P(A)$. Так, если из 50 принятых знаков два ошибочных, то частота ошибок $\nu = 2/50 = 4\cdot 10^{-2}$, если зафиксировано 40 ошибочных знаков на 1 тыс. принятых, то можно считать, что вероятность ошибки $P(A) = 40/100 = 4\cdot 10^{-2}$.

Для характеристики зависимых случайных событий A *и* B вводится условная вероятность $P(A/B)$, что обозначает вероятность события A при условии, что событие B уже наступило.

Случайные величины. *Величина, значение которой изменяется от попытки к попытке случайным образом, называется случайной.* Для такой величины невозможно вероятно предусмотреть, какое значение она приобретет в конкретных условиях попытки. Количество ошибок в тексте, количество занятых каналов многоканальной связи, мощность сигнала на выходе линии связи, значения помехи в канале - это все примеры случайных величин. Можно даже сказать так: реалии мира таковы, что любая физическая величина является случайной. Обозначаются случайные величины большими буквами латинской азбуки X, Y, Z, а значения, которые они приобретают, - соответствующими малыми буквами x, y, z.

Александр Яковлевич Хинчин (1894-1959), русский математик, член-корреспондент АН СССР (1939). Закончил Московский университет (1916), с 1922 г. - профессор там же. Первые работы касаются теории функций действительной переменной. Перенес методы метрической теории функций в теорию чисел и теорию вероятностей. Является одним из творцов советской школы теории вероятностей (получил важные результаты в области предельных теорем, открыл закон повторного логарифма, дал определение случайного стационарного процесса и заложил основы теории таких процессов).

Случайные величины разделяются на дискретные и непрерывные.

Дискретная случайная величина *X* может приобретать только конечное множество значений $x_1, x_2, \ldots, x$. *Непрерывная* случайная величина *X* может приобретать любые значения из некоторого интервала, даже бесконечного.

Для математического описания случайных величин введены следующие неслучайные статистические характеристики.

Функция распределения вероятности

$$F(x) = P(X \le x) \tag{3.36}$$

показывает вероятность того, что случайная величина *X* не превышает конкретного значения *x*. Если случайная величина *X* является дискретной, то $F(x)$ - дискретная функция. Если *X* - непрерывная случайная величина, то $F(x)$ -монотонно возрастающая функция, значение которой изменяется в интервале $0 \le F(x) \le 1$, при этом $F(-\infty) = 0$ и $F(\infty) = 1$. Функция распределения $F(x)$ - величина безразмерная.

Плотность распределения вероятности p(x), или плотность вероятности, рассчитывается как производная от функции деления

$$p(x) = dF(x)/dx.$$

Физически $p(x)$ является вероятностью того, что случайная величина попадает в малый интервал dx в окрестности точки *x*. Взаимозависимость между $F(x)$ и $p(x)$ определяется формулой

$$F(x) = \int_{-\infty}^{\infty} p(x)dx. \tag{3.37}$$

Единица измерения $p(x)$ - обратная к единице измерения случайной величины.

Математическое ожидание

$$M(X) = \int_{-\infty}^{\infty} xp(x)dx = \sum_{i=1}^{n} x_i P(x_i), \tag{3.38}$$

где $p(x)$ - плотность вероятности; $P(x)$ - вероятность значения x_i случайной величины.

По своей сути *математическое ожидание является средним значением случайной величины.* Если *X* - случайное напряжение или ток, то *M(X)* - *постоянная составляющая* напряжения или тока. В выражении (3.38) интегрирование используется при вычислении математического ожидания непрерывной

случайной величины, суммирование - при вычислении математического ожидания дискретной случайной величины. Единица измерения $M(X)$ совпадает с единицей измерения случайной величины.

Дисперсия $D(X)$ *количественно характеризует меру разброса результатов отдельных попыток относительно среднего значения.* Дисперсия рассчитывается как математическое ожидание квадрата отклонения случайной величины от его математического ожидания:

$$D(X) = M\left[X - M(X)\right]^2 = M\left(X^2\right) - M^2(X). \quad (3.39)$$

Единица измерения $D(X)$ - квадрат единицы измерения случайной величины.

Величина $\sigma = \sqrt{D(X)}$, т.е. квадратный корень из дисперсии, называется *средним квадратическим отклонением*. По физической сути это то, что в электротехнике называют *эффективным значением*.

Случайные процессы. *Случайным процессом* $X(t)$ называется функция, *значение которой при любом значении аргумента t является случайной величиной.* Из этого определения вытекает, что когда осуществлять наблюдение изменения во времени любой случайной величины X, то результатом такого наблюдения и будет случайный процесс $X(t)$. Напряжение шума на выходе линии связи, температура воздуха, ток через микрофон при разговоре и т. п., если наблюдать за изменениями мгновенных значений перечисленных физических величин во времени, являются примерами случайных процессов.

Результаты отдельных наблюдений, которые проходят в одних и тех же условиях, дают каждый раз разные функции $x_k(t)$ - разные экземпляры, или реализации, случайного процесса. Совокупность $\{x_k(t)\}$ всех возможных реализаций данного случайного процесса называется *ансамблем*. Таким ансамблем может быть набор сигналов $x_1(t), \ldots, x_k(t)$, которые наблюдаются в то же время на выходах разных каналов системы многоканальной связи (рис. 3.20).

Поль Адриен Карл Фридрих Гаусс (Carl Friedrich Gauss) (1777-1855), немецкий математик, астроном и физик. Учился (1795 - 1798) в Геттингенском университете, с 1807 г. - профессор этого университета и директор астрономической обсерватории. Характерными особенностями исследований Гаусса являются чрезвычайная разносторонность и органическая связь в них между теоретической и прикладной математикой. Работы Гаусса имели большое влияние на все дальнейшее развитие высшей алгебры теории чисел и многих других наук.

Вовсе не обязательно, чтобы реализации случайного процесса были сложными функциями, как это приведено на рис. 3.20. Гармонический сигнал, у которого хотя бы один из параметров U_m, ω, φ - случайная величина, является также случайным процессом.

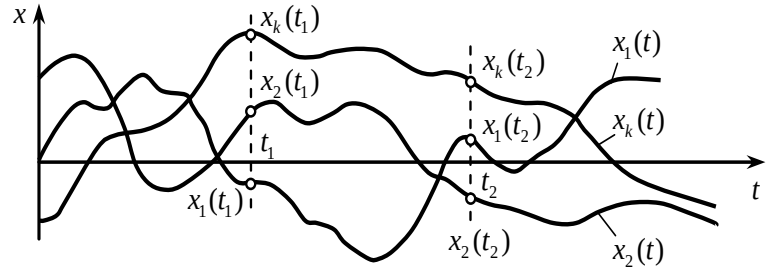


Рис. 3.20. Реализация случайного процесса $X(t)$

Случайные процессы бывают разных типов: *нестационарные, стационарные, квазистационарные, эргодичные*. Но в технике большинство случайных сигналов и помех принадлежат к стационарным эргодичным случайным процессам. Случайный процесс является *стационарным,* если такие его характеристики, как функция распределения $F(x)$, плотность вероятности $p(x)$, математическое ожидание $M(X)$, дисперсия $D(X)$, не зависят от времени. Определяются они так же, как и для случайной величины.

Стационарный случайный процесс называется *эргодичным*, если для него усреднение по времени одной реализации дает те же результаты, что и статистическое усреднение по всем реализациям. Физически это значит, что все реализации похожи одна на другую, поэтому измерение и расчеты характеристик такого случайного процесса можно проводить по одной (любой) реализации, что значительно проще.

Нестационарными являются случайные процессы, для которых не выполняются приведенные только что условия стационарности. Случайные процессы, характеристики которых приближаются к стационарным, получили название *квазистационарных.*

Кроме четырех отмеченных характеристик - функции распределения $F(x)$, плотности вероятности $p(x)$, математического ожидания $M(X)$, дисперсии $D(X)$ - для описания случайного процесса используются еще две: *функция корреляции* $K_X(\tau)$ и *спектральная плотность мощности* $G_X(f)$ или $G_X(\omega)$.

Функция корреляции $K_X(\tau)$ характеризует меру взаимосвязи между значениями случайного процесса в разные моменты времени t и $t+\tau$. Для эргодичных слу-

чайных процессов $K_X(\tau)$ вычисляется усреднением по времени произведения $x_k(t)$ и $x_k(t+\tau)$:

$$K_X(\tau) = \int_0^{T_{сп}} x_k(t)\, x_k(t+\tau)\, dt \qquad (3.40)$$

где $T_{сп}$ - время наблюдения (или длительность) реализации случайного процесса $X(t)$. Единица измерения $K_X(\tau)$ совпадает с единицей измерения дисперсии.

Спектральная плотность мощности $G_X(f)$ или $G_X(\omega)$ представляет распределение мощности случайного процесса за частотами и на любой частоте определяется как отношение

$$G_X(f) = \lim_{\Delta f \to 0} \Delta P / \Delta f; \qquad (3.41)$$

$$G_X(\omega) = 2\pi \lim_{\Delta\omega \to 0} \Delta P / \Delta\omega, \qquad (3.42)$$

где ΔP - мощность случайного процесса, которая попадает в полосу частот Δf или $\Delta\omega$. $K_X(\tau)$ измеряется в ваттах на герц (Вт/Гц), а $G_X(\omega)$ - в ватт-секундах на радиан (Вт·с/рад).

Для эргодичного случайного процесса функция корреляции $K_X(\tau)$ и спектральная плотность мощности $K_X(\tau)$ связаны между собой согласно с теоремой Хинчина - Винера интегральными преобразованиями Фурье. Поскольку по физической сущности и функция корреляции, и спектральная плотность мощности - всегда вещественные парные функции, интегральные преобразование Фурье для них можно записать в виде

$$K_X(\tau) = \int_0^{\infty} G_X(f) \cos 2\pi f \tau\, df\,, \qquad (3.43)$$

$$G_X(f) = 4\int_0^{\infty} K_X(\tau) \cos 2\pi f \tau\, d\tau\,. \qquad (3.44)$$

Норберт Винер (Norbert Wiener, 1894-1964), американский ученый - выдающийся математик и философ, основатель кибернетики и теории искусственного интеллекта. Заинтересовавшись автоматическими расчетами и теорией обратной связи, сформулировал в своей фундаментальной работе "Кибернетика" (1948) положение этой науки, предметом которой стали управления, связь и обработка информации в технике, живых организмах и человеческом обществе. Автор работ по математическому анализу, теории вероятностей, электрических сетей и вычислительной техники.

Эти соотношения широко используются в расчетах характеристик случайных процессов.

Флуктуационный шум наиболее характерен для большинства телекоммуникационных каналов. Для количественных расчетов воздействия флуктуационного шума на сигнал необходимо знать его основные статистические характеристики. Поскольку флуктуационный шум создается как сумма большого количества независимых колебаний, то он согласно центральной предельной теореме, доказанной в 1901 г. академиком А. И. Ляпуновым, является *стационарным эргодичным* случайным процессом с *гауссовым* (нормальным) распределением вероятности.

Плотность вероятности гауссового процесса описывается формулой

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right], \qquad (3.45)$$

в которую входят два числовых параметра m и σ^2, математического ожидания и дисперсии соответственно

$$m = M(X); \qquad \sigma^2 = D(x).$$

График плотности вероятности $p(x)$ является звоноподобной кривой с единственным максимумом в точке $x = m$ (рис. 3.21, *а*). На графике привлекает внимание то, что с уменьшением σ кривая все более локализуется вокруг точки $x = m$. *Для* флуктуационного шума $M(X) = 0$.

Функция распределения вероятности гауссового случайного процесса согласно формуле (3.37)

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right] dx \qquad (3.46)$$

и после введения новой переменной $y = (x-m)/\sigma$ сводится к виду

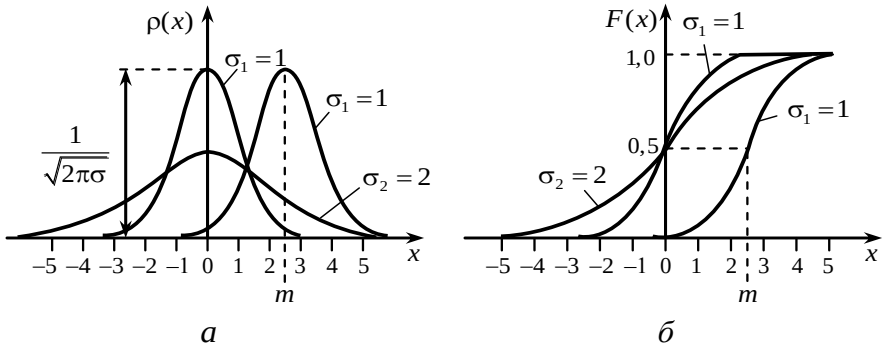
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-m)/\sigma} \exp\left(-y^2/2\right) dy = 0{,}5 + \Phi_0\left[(x-m)/\sigma\right] \qquad (3.47)$$

где $\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-y^2/2\right) dy.$

График функции $F(x)$ (рис. 3.21, *б*) имеет вид монотонной растущей кривой от нуля до единицы. Функция $\Phi_0(z)$, которая входит в выражение

(3.47), называется *интегралом вероятности*, и она табулирована в математических справочниках.

Спектральная плотность мощности флуктуационного шума зависит от физической природы его образования, а также от точки, где он наблюдается. Как правило, спектральная плотность мощности $G_X(f)$ флуктуационного шума равномерна от нуля до $10^{12}-10^{13}$ Гц, т.е. можно считать, что $G_X(f)=N_0$, для $0\le f\le\infty$.



а *б*

Рис. 3.21. Распределение вероятности Гаусса:
а – плотность вероятности; *6* - функция распределения

В этом случае шум называют *белым.* Это название присвоено по аналогии с белым светом, который имеет все частотные компоненты. Если спектральная плотность мощности шума равномерна только в ограниченной полосе частот, например сигнала, то шум называют *квазибелым.*

3.3. Числовые характеристики сигналов и помех

Энергетические характеристики. Основными энергетическими характеристиками действительного сигнала $s(t)$ является его мощностью и энергия. Если $s(t)$ - напряжение $u(t)$ или ток $i(t)$, то *мгновенная мощность*, которая выделяется на сопротивлении R, определяется через квадрат мгновенного значения

$$p(t)=u^2(t)R=i^2(t)/R.$$

Измеряется мгновенная мощность в ваттах (Вт). В теории сигналов, как правило, принимают для расчетов R = 1 Ом (кроме некоторых случаев), и тогда в общем виде

$$p(t) = s^2(t). \tag{3.48}$$

Принятие такого условия связано с тем, что во многих задачах теории сигналов используют в расчетах не конкретные значения мощности, а отношение мощности сигнала к мощности помехи. При расчетах отношения сопротивление R сокращается, и для упрощения расчетов его считают единичным. Чтобы отличить расчеты мощности при таких условиях $(R = 1\ \text{Ом})$ от мощности на каком-то сопротивлении $R \neq 1$ Ом, в формуле (3.41) и других, куда входит мощность, за единицу мощности берут вольт в квадрате (В^2), а не ватт (Вт).

Энергия сигнала на интервале (t_1, t_2) определяется как интеграл его мгновенной мощности

$$E_{\text{с}} = \int_{t_1}^{t_2} p(t)\,dt = \int_{t_1}^{t_2} s^2(t)\,dt\,. \tag{3.49}$$

Энергию сигнала $s(t)$ можно вычислить также по его спектральной плотности $S(j\omega)$ или $S(f)$ по формуле Релея

$$E_{\text{с}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|S(j\omega)\right|^2 d\omega = \int_{0}^{\infty} S^2(f)\,df\ . \tag{3.50}$$

Величину $\left|S(j\omega)\right|^2$ называют *спектральной плотностью энергии*, или *энергетическим спектром сигнала. Из формулы (3.50) вытекает,* что

$$S(f) = \sqrt{2}S(\omega), \tag{3.51}$$

поскольку $S(\omega)$ - двухсторонняя $(-\infty < \omega < \infty)$, а $S(f)$ - односторонняя $(0 \le f \le \infty)$ спектральная плотность мощности.

Отношение

$$E_{\text{с}}/(t_1 - t_2) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} S^2(t)\,dt \tag{3.52}$$

определяет среднюю мощность $P_c = \overline{s_2(t)}$ на интервале (t_1, t_2).

Расчеты средней мощности по спектру. Средняя мощность периодического сигнала, которая рассчитывается на всей оси времени $\left(-\infty < t < \infty\right)$, совпадает со средней мощностью за период. Для гармонического сигнала $u(t) = U_m \cos(\omega t + \varphi_0)$ согласно с соотношением (3.50) средняя мощность (на R = 1 Ом) подается в виде

$$\overline{u^2(t)} = P_u = \left(U_m^2/T\right)\int_0^T \cos^2\left(\omega t + \varphi_0\right) dt = U_m^2/2 \quad (3.53)$$

и не зависит ни от частоты, ни от начальной фазы.

Поскольку периодический сигнал $S(t)$ можно представить в виде тригонометрического ряда Фурье, а интеграл суммы равняется сумме интегралов, то *полная средняя мощность периодического сигнала равняется сумме средних мощностей, которые выделяются отдельно постоянной составляющей* $a_0/2$ *и гармониками с амплитудами* A_{m1}, A_{m2}, *... при этом она не зависит от частот и фаз отдельных гармоник.*

Для случайных сигналов (помех) среднюю мощность можно вычислить по спектральной плотности мощности $G_x(f)$ или $G_x(\omega)$. Поскольку функции $G_x(f)$ и $G_x(\omega)$ показывают деление мощности за частотами (см. формулу (3.52)), то средняя мощность определяется интегралом

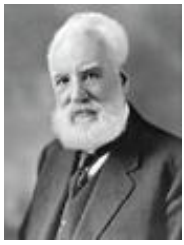
$$P_x = \int_0^\infty G_x(f)\,df = 2\int_0^\infty G_x(\omega)\,d\omega. \quad (3.54)$$

Заметим, что в выражении (3.54), как и в формуле (3.50), $G_x(f) = 2G_x(\omega)$. Это условие взято из тех соображений, что $G_x(\omega)$ - двусторонняя $-\infty < \omega < \infty$, а $G_x(f)$ - односторонняя $0 \le f \le \infty$ спектральная плотность мощности.

Чтобы найти, например, мощность случайного сигнала (помехи) в некоторой полосе частот от f_1 до f_2, необходимо осуществить интегрирование согласно выражению (3.54) в этой полосе

$$P_{x1,2} = \int_{f_1}^{f_2} G_x(f)\,df.$$

Александр Михайлович Ляпунов (1857—1918), русский математик и механик, академик Петербургской академии наук. Ученик П. Л. Чебышева. Основные работы посвящены теории устойчивости равновесия и движения механических систем, теории фигур равновесия жидкости, которая равномерно оборачивается, и математической физике. Важнейшим достижением являются создание современной теории устойчивости равновесия и движения механических систем, определенных конечным количеством параметров. Получил ряд весомых результатов в теории линейных и нелинейных дифференциальных уравнений.



Александр Грехем Белл (Alexander Graham Bell, 1847-1922), американский физик шотландского происхождения, творец телефонного аппарата с металлической мембраной, один из изобретателей телефона. В 1865 г. задумал передать язык электрическими волнами. Воплощение идеи заняло следующие 10 лет. В 1876 г. первое сообщение было успешно передано по проводам. В 1877 г. была образована телефонная компания Bell. Выполнял работы по использованию в телекоммуникациях светового луча - направление, которое со временем привело к созданию волоконно-оптических технологий.

Пример. Найти среднюю мощность $P_{ш}$ белого шума со спектральной плотностью мощности $G_x(f) = N_0 = 10^{-6}$ Вт/Гц в полосе Δf =3100 Гц.

Среднюю мощность шума в полосе Δf находим согласно выражению (3.54), если границы интеграции берем от f_1 до $f_1 + \Delta f$:

$$P_{ø} = \int_{f_1}^{f_1+\Delta f} N_0 df = N_0 \Delta f = 10^{-6} \cdot 3100 = 3{,}1 \cdot 10^{-3} \text{ (Вт)}.$$

Уровни сигналов (помех). Под *уровнем* понимают отношение значения мощности P_x или напряжения U_x в некоторой точке *x* электрической цепи к выбранному для сравнения значению мощности P_0 или напряжения U_0 Поскольку значения мощности и напряжения могут изменяться в достаточно больших (сотни и тысячи раз) границах, то для измерения уровней введена логарифмическая единица уровня *децибел* (дБ), который равняется $10\lg(P_x / P_0)$ по мощности и $20\lg(U_x / U_0)$ по напряжению. Например, в технике связи за абсолютный нулевой уровень взята мощность P_0 = 1 мВт на сопротивлении *R* = 600 Ом. Тогда

$$U_0 = \sqrt{P_0 R} = 0{,}7748 \approx 0{,}775 \text{ (В)}.$$

Децибелы, определенные относительно мощности P_0 = 1 мВт, называются *децибелами относительно 1 мВт и сокращенно обозначаются дБп или дБ(мВт)*.

В случае использования логарифмической единицы измерения уровней такая характеристика качества, как отношение сигнал/помеха, будет равняться разнице уровней сигнала $L_с$ и помехи $L_з$, поскольку

$$\rho = 10\lg(P_с / P_з) = -10\lg(P_з / P_0) = L_0 - L_з = 10\lg(P_с / P_0)$$

Динамический диапазон и коэффициент амплитуды. *Динамический диапазон* D_C, *дБ, сигнала* $s(t)$ *характеризует границы изменения мгновенной мощности и определяется выражением*

$$D_C = 10\lg(p_{\max}/p_{\min}), \tag{3.55}$$

где $p_{\max}$, $p_{\min}$ - соответственно максимальное и минимальное значения мгновенной мощности, определенные любым способом. Например, минимальную мощность, если ее тяжело найти, считают такой, которая равняется мощности помехи или средней квадратичной погрешности.

Коэффициентом амплитуды сигнала K_A называется отношение его максимальной мощности к средней. В логарифмических единицах, дБ, имеем

$$K_A^{\,2} = 10\lg(p_{\max}/P_x). \tag{3.56}$$

В некоторых случаях динамический диапазон и коэффициент амплитуды определяются не в логарифмических, а в абсолютных единицах («в разах»).

Длительность и ширина спектра сигнала (**помехи**). Под *длительностью сигнала* понимают интервал времени его существования. Вычисляется длительность сигнала как разница между временем окончания сигнала $t_к$ и временем его начала $t_п$

$$T_s = t_к - t_п.$$

Ширина спектра - это интервал частот, который занимает спектр. Вычисляется ширина спектра как разница между максимальной $f_{\max}$ и минимальной $f_{\min}$ частотой спектра

$$F_s = f_{\max} - f_{\min}.$$

Расчеты длительности сигнала (помехи) и ширины спектра не вызывают осложнений, если этот сигнал (помеха) имеет четко определенное начало или конец, а его спектр - граничные частоты. Но с преобразования Фурье вытекает, что когда сигнал имеет конечную длительность, то спектр его нескончаемый, и наоборот. Поэтому практически необходимо договориться об определении длительности и ширины сигнала (помехи).

На практике используются разные методы нахождения $T_с$ и F_c, выбор которых зависит от назначения сигнала, его формы и структуры. Наиболее применяемые следующие методы определения $T_с$ и F_c:

1. *Отсчет на заданном уровне от максимального.* Обычно длительность

импульсного сигнала $s(t)$ и ширину его спектра $S(f)$ определяют на уровне $1/\sqrt{2}$ от максимального значения из этих величин. Однако для расчетов можно выбрать и любое другое значение, например 5 % от максимального, как это показано на рис. 3.22. В этом - неопределенность метода.

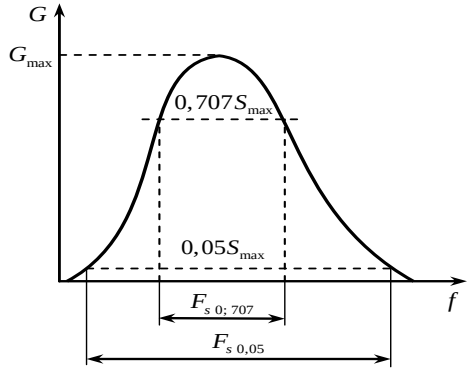


Рис. 3.22. Определение ширины спектра на заданном уровне

2. *Энергетический метод*. За длительность сигнала (ширину спектра) берут такой интервал времени (частот), в который попадает заданная часть энергии сигнала, например 0,9 или 0,95.

3. *Замена реального сигнала (спектра) равновеликим прямоугольным*. Такую процедуру, чаще всего применяемую для вычисления спектральной плотности мощности сигнала или помехи, показывает наглядно рис. 3.23, где изображена спектральная плотность мощности помехи $G_{\text{п}}(f)$. Площади прямоугольника и фигуры, ограниченной кривой $G_{\text{п}}(f)$ и осями координат, одинаковые, т.е. прямоугольник и эта фигура равновелики.

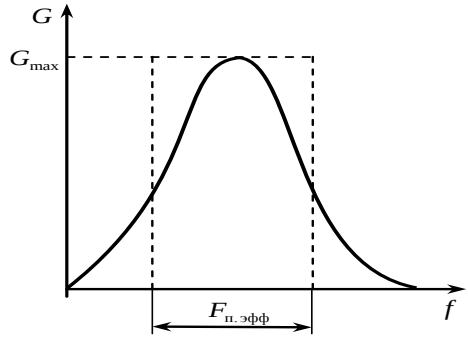


Рис. 3.23. Определение эффективной ширины спектра

Из рис. 3.23 вытекает, что ширина спектра, которую называют *эффективной*, определяется как

$$F_{\text{п.эфф}} = \frac{1}{G_{\max}} \int_0^{\infty} G_{\text{п}}(f)\,df.$$

Числовые характеристики сигналов и помех широко используются в телекоммуникационных системах. По энергетическим характеристикам определяется необходимое отношение сигнал/помеха, по ширине спектра сигнала устанавливается полоса пропускания канала связи, необходимая для неискаженной передачи. Для непрерывных первичных сигналов ширина спектра определяется, как правило, экспериментально. При определении ширины спектра импульсных сигналов можно воспользоваться одним из важнейших положений теории сигналов и спектров: если $F_{\text{с}}$ *означает ширину спектра некоторого сигнала длительностью* $T_{\text{с}}$, *то всегда выполняется соотношение*

$$T_{\text{с}} F_{\text{с}} \approx \mu, \qquad (3.57)$$

где μ - постоянная величина, близкая единице ($\mu \approx 1$) для видеоимпульсов и двум ($\mu \approx 2$) для радиоимпульсов.

Суть этого соотношения в том, что *ширина спектра сигнала обратно пропорциональна его длительности.*

3.4. Математические модели сигналов с ограниченным спектром

Все рассмотренные ранее сигналы принадлежат к таким, которые теоретически имеют бесконечно широкий спектр. Это значит, что при попытке восстановления исходного сигнала методом суммирования его гармоник необходимо учитывать бесконечное множество спектральных компонентов. Потеря каждого из них, а тем более некоторого их подмножества, сопровождается искажением формы сигнала. Искажение будет тем большим, чем большее количество гармоник утрачено при восстановлении сигнала по его спектральным составляющим.

С физической точки зрения процедура восстановления сигнала, которая основывается на учете всех спектральных составляющих из бесконечно широкого спектра, неисполнимая. Не следует забывать также о том, что вклад, сделанный спектральными компонентами при $\omega \to \infty$, становится ничтожно малым по сравнению с самими сигналами, энергия которых конечна. Кроме того, любое реальное устройство, предназначенное для передачи и обработки сигналов, имеет конечную ширину полосы пропускания. Наиболее характерно это для устройств типа частотных фильтров.

Идеальный низкочастотный сигнал. Рассмотрим *особенный класс сигналов, спектральная плотность которых отличается от нуля лишь в пределах некоторого интервала частот конечной длины.*

Пример таких сигналов - радиоимпульс с линейной частотной модуляцией при значении базы B, которая стремится к бесконечности и имеет конечную ширину спектра.

Пусть D - частотный интервал, в пределах которого спектральная плотность $\dot{U}(\omega)$ некоторого сигнала $u(t)$ не равняется нулю, т.е. $\dot{U}(\omega) \neq 0$, если $\omega \notin D$. В общем виде математическая модель сигнала с ограниченным спектром определяется формулой обратного преобразования Фурье

$$u(t) = \frac{1}{2\pi} \int_D \dot{U}(\omega) e^{j\omega t} d\omega.$$

В зависимости от выбора интервала D и функции $\dot{U}(\omega)$ можно получить самые разнообразные сигналы с ограниченным спектром.

Рассмотрим колебание, спектральная плотность которого постоянная и приобретает действительное значение в пределах частотного интервала, ограниченного некоторой верхней частотой $\omega_в$. Вне этого интервала спек-

тральная плотность превращается в ноль:

$$\dot{U}(t)=\begin{cases}U_0, & \omega\in[-\omega_в,\omega_в];\\ 0, & \omega\notin[\cdot].\end{cases} \tag{3.58}$$

График спектральной плотности (3.58) изображен на рис. 3.24. Мгновенное значение этого сигнала находим за формулой обратного преобразования Фурье.

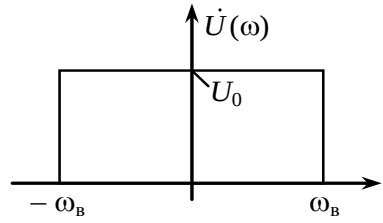


Рис. 3.24. Спектральная плотность ИНС

После интегрирования и элементарных преобразований получим:

$$u(t)=\frac{U_0\omega_в}{\pi}\cdot\frac{\sin\omega_в t}{\omega_в t}. \tag{3.59}$$

Это колебание называется *идеальным низкочастотным сигналом* (ИНС), график которого, построенный по формуле (3.59), имеет вид осциллирующей кривой, парной относительно начала отсчета времени (рис. 3.25). С увеличением верхней предельной частоты $\omega_в$ растут как значение центрального максимума, так и частота осцилляций.

В пределе при $\omega_в$, стремящейся к бесконечности, сигнал $u(t)$ переходит в дельта-функцию, т.е. $\lim\limits_{\omega_в\to\infty}u(t)=\delta(t)$.

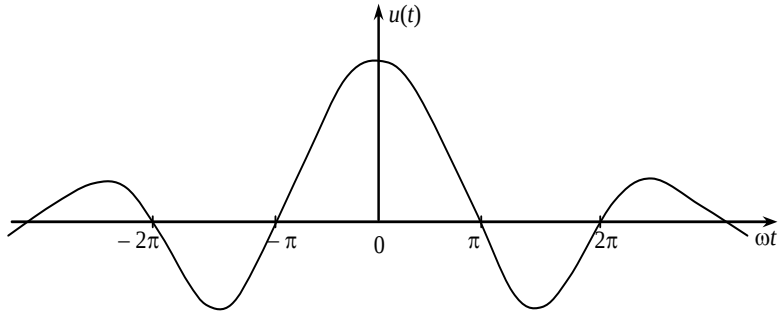


Рис. 3.25. ИНС

ИНС можно получить, подав на вход идеального фильтра нижних частот

(ФНЧ) сигнал $s(t)$ с равномерной на всей оси частот спектральной плотностью (рис. 3.26).

Как известно, равномерную в бесконечном интервале частот спектральную плотность имеет сигнал $s(t)$ типа дельта-функции $\delta(t)$ (рис. 3.27).

Рис. 3.26. Схема моделирования ИНС

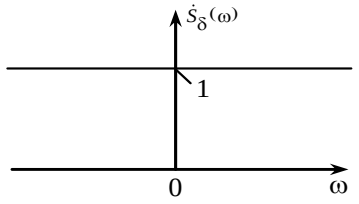


Рис. 3.27. Спектральная плотность дельта-функции

Для того чтобы обеспечить на выходе фильтра (см. рис. 3.26) формирование сигнала $\tilde{u}(t)$ с равномерной в интервале частот $\omega \in [0, \omega_{\text{â}}]$ (для физически реализуемого фильтра), частотная передаточная функция фильтра $K(\omega)$ должна быть такой, как на рис. 3.28.

Очевидно, что идеальный ФНЧ (см. рис. 3.26) в случае подачи на его вход дельта-функции вырезает из ее спектра (см. рис. 3.27) участок частот от 0 к $\omega_{\text{в}}$. Сигналу $\tilde{u}(t)$ на выходе фильтра будет соответствовать спектральная плотность

$$\tilde{U}(\omega) = \begin{cases} U_0, & \omega \in [0, \omega_{\text{â}}]; \\ 0, & \omega \notin [\cdot], \end{cases}$$

отличающаяся от плотности (3.58) отсутствием компонент на отрицательных частотах.

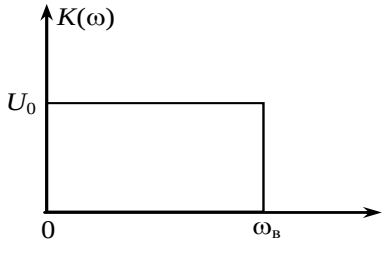


Рис. 3.28. Частотная передаточная функция идеального ФНЧ

Рис. 3.29. Отклик идеального ФНЧ на входящий сигнал типа дельта-функции

Заметим, что физически реализуемый ФНЧ формирует отклик (исходный сигнал) не раньше, чем с момента появления входного сигнала. Если короткий

входной сигнал прямоугольной формы (аналог дельта-импульса) появляется в момент времени $t = 0$, то реакция ФНЧ на такой сигнал будет иметь вид функции $\tilde{U}(t)$, изображенной на рис. 3.29.

Максимальное значение отклика U_m в момент времени $t = 0$ определяется соотношением

$$U_m = \frac{U_0\,\omega_{\text{â}}}{\pi}.$$

Как еще одну модель сигнала с ограниченным спектром рассмотрим идеальный полосовой сигнал.

Идеальный полосовой сигнал. Построение математической модели полосового сигнала опирается на предположение, что его спектр ограничен полосой частот шириной П = 2 Δω с центром на частотах $\pm\omega_0$. Если в пределах этой полосы спектральная плотность сигнала постоянная (рис. 3.30):

$$\dot{U}(\omega) = \begin{cases} U_0, & \left[\begin{array}{c} -\omega_0 - \Delta\omega < \omega < -\omega_0 + \Delta\omega \\ \omega_0 - \Delta\omega < \omega < \omega_0 + \Delta\omega \end{array}\right]; \\ 0 - & \text{вне полосы пропускания,} \end{cases}$$

то по аналогии из ИНС его называют *идеальным полосовым сигналом (ИПС)*.

Мгновенные значения ИПС можно найти по формуле обратного преобразования Фурье. Как следует из рис. 3.30, спектральная плотность ИПС является парной функцией относительно начала оси частот и в общем случае

$$u(t) = \frac{1}{\pi}\int_0^{\infty} \dot{U}(\omega)\cos\omega t\, d\omega.$$

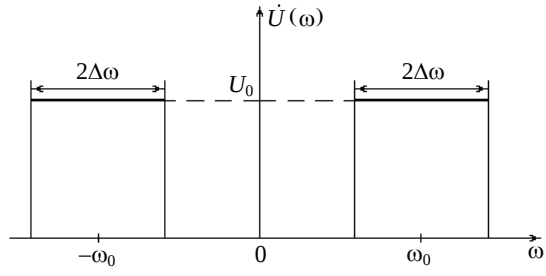


Рис. 3.30. Спектральная плотность ИПС

Поскольку $\dot{U}(\omega)$ задано как действительную функцию, которая равняется U_0 в пределах полосы частот $\omega \in (\omega_0 - \Delta\omega, \omega_0 + \Delta\omega)$, то последний интеграл можно записать в виде

$$u(t) = \frac{U_0}{\pi} \int_{\omega_0 - \Delta\omega}^{\omega_0 + \Delta\omega} \cos \omega t \, d\omega.$$

После интегрирования и элементарного превращения получим

$$u(t) = \frac{2U_0 \Delta\omega}{\pi} \cdot \frac{\sin \Delta\omega t}{\Delta\omega t} \cos \omega_0 t.$$

График на рис. 3.31 показывает наглядно структуру ИПС. Функция $\sin(\Delta\omega t)/(\Delta\omega t)$ с точностью до масштабного коэффициента $2U_0\Delta\omega/\pi$ описывает закон изменения огибающей ИПС.

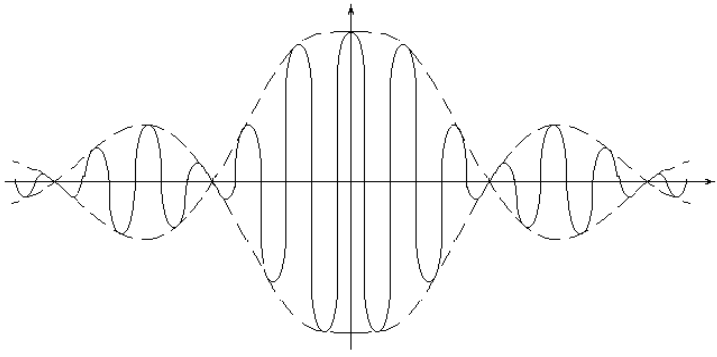


Рис. 3.31. Реализация ИПС

Способ образования ИПС вполне очевиден: на вход идеального полосового фильтра (рис. 3.32), пропускающего колебание с частотами в пределах полосы $(\omega_0 - \Delta\omega, \omega_0 + \Delta\omega)$, нужно подать широкополосное влияние типа дельта-импульс.

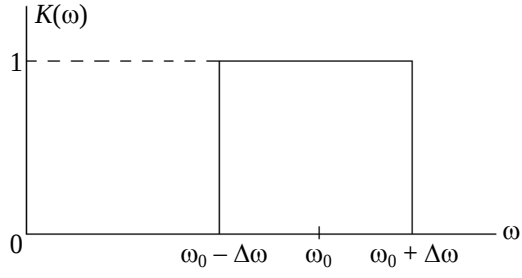


Рис. 3.32. Частотная характеристика идеального полосового фильтра

На выходе идеального полосового фильтра будет наблюдаться сигнал, осциллограмма которого соответствует правой половине графика, приведенно-

го на рис. 3.31. Это выплывает из условия физической реализуемости сигнала: отзыв цепи (системы и тому подобное) начинает формироваться не раньше за появление влияния на входе.

Базис Котельникова. Как известно, базис в общем случае образует бесконечную совокупность ортогональных функций, норма каждой из которых равняется единице.

Напомним, что *ортогональными* называются такие сигналы $u(t)$ и $v(t)$, скалярное произведение которых равно нулю:

$$(u, v) = \int_{-\infty}^{\infty} u(t)v(t)dt = 0.$$

Если к тому же энергия каждого сигнала равна единице, то сигналы $u(t)$ и $v(t)$ называются *ортонормированными.* Для сигнала $u(t)$ его энергия E_u определяется соотношением

$$E_u = (u, u) = \int_{-\infty}^{\infty} u^2(t)\, dt.$$

В соответствии с обобщенной формулой Релея сигналы $u(t)$ и $v(t)$ будут ортогональными, если выполняется условие

$$(u,v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \dot{U}(\omega)\dot{V}^*(\omega)\, d\omega = 0. \qquad (3.60)$$

Ограничения, которые накладываются на полосу частот сигнала, дают возможность находить интересные и важные классы ортогональных сигналов. В качестве самого простого примера ортогональных сигналов можно привести такую пару полосовых сигналов, спектры которых не пересекаются. Равенство нулю скалярного произведения этих сигналов непосредственно вытекает из формулы (3.60).

Менее очевидный способ ортогонализации сигналов с ограниченным спектром заключается в их сдвиге во времени.

Рассмотрим два ИНС $u(t)$ и $v(t)$. Оба они имеют одинаковые параметры U_0 и $\omega_в$, но отличаются тем, что сигнал $v(t)$ запаздывает относительно сигнала $u(t)$ на время t_0, т.е.

$$v(t) = u(t - t_0). \qquad (3.61)$$

Спектральная плотность сигнала $v(t)$ определяется соотношением

$$\dot{V}(\omega)=\dot{U}(\omega)\,e^{-j\omega t_0}$$

или согласно выражению (3.58)

$$\dot{V}(\omega)=U_0\,e^{-j\omega t_0},\ \omega\in[-\omega_в,\omega_в]. \tag{3.62}$$

Допустим, что спектральная плотность сигнала $u(t)$ задана соотношением (3.58). Подставив выражения (3.58) и (3.62) в (3.60), получим формулу для скалярного произведения этих сигналов

$$(u,v)=\frac{U_0^2}{2\pi}\int\limits_{-\omega_в}^{\omega_в} e^{j\omega t_0}d\,\omega. \tag{3.63}$$

В результате интегрирования в формуле (3.63) получим

$$(u,v)=\frac{U_0^2\omega_в}{\pi}\,\frac{\sin\omega_в\,t_0}{\omega_в\,t_0}. \tag{3.64}$$

Из соотношения (3.64) вытекает, что два одинаковых по форме ИНС оказываются ортогональными, если сдвиг между ними во времени удовлетворяет условию

$$\omega_в\,t_0=k\pi,\quad k=\pm1,\pm2,... \tag{3.65}$$

Минимально возможен сдвиг Δt (назовем его *шагом временного сдвига), который приводит к ортогонализации, имеем при* $k=\pm1$, *т. е.:*

$$\Delta t=\pm\frac{\pi}{\omega_в}=\pm\frac{1}{2F_в}, \tag{3.66}$$

где $F_в$ - верхняя предельная частота среза (в герцах) идеального ФНЧ, что отвечает верхней гармонике колебаний в ИНС.

В соотношении (3.66) взято во внимание, что $\omega_в=2\pi F_в$.

Принципиально важно, что *условием* (3.65) *удается не только добиться ортогонализации двух ИНС, но и построить ортогональный базис для сигналов, в спектре которых отсутствуют частоты, выше чем* $\omega_в$.

Покажем это на примере ИНС. Воспользовавшись соотношениями (3.61), (3.65) и (3.66), образуем совокупность сигналов

$$v_k(t)=u(t-k\Delta t)=u\left(t-k\frac{\pi}{\omega_в}\right) \tag{3.67}$$

Сигналам $v_k(t)$ отвечает спектральная плотность

$$\dot{V}_k(\omega) = U_0\, e^{-j\omega k\frac{\pi}{\omega_в}} \qquad \omega \in [-\omega_в, \omega_в]. \tag{3.68}$$

При $k = 0$ спектральная плотность $\dot{V}_0(\omega)$ сигнала $v_0(t)$ совпадает со спектральной плотностью (3.58) ИНС $u(t)$, временная функция которого задана соотношением (3.59). Покажем, что совокупность $v_k(t)$ сигналов (3.68) образует исчислимое множество ортогональных функций. С этой целью нам достаточно убедиться в том, что взаимная энергия E_l сигналов $v_k(t)$ и $v_{k+l}(t)$, разнесенных на l интервалов времени Δt, равняется нулю. Действительно, взаимная энергия E_l двух сигналов численно равняется скалярному произведению этих сигналов $E_l = (v_k, v_{k+l})$.

В свою очередь скалярное произведение можно определить обобщенной формулой Релея

$$E_l = \frac{1}{2\pi}\int_{-\infty}^{\infty} \dot{V}_k(\omega)\dot{V}^*_{k+l}(\omega)\, d\omega.$$

Это равенство с учетом выражения (3.68) сводится к виду

$$E_l = \frac{U_0^2}{2\pi}\int_{-\omega_в}^{\omega_в} e^{j\omega l\frac{\pi}{\omega_в}} d\omega = \frac{U_0^2}{\pi}\int_0^{\omega_в} \cos\left(\omega l\frac{\pi}{\omega_в}\right) d\omega. \tag{3.69}$$

Вычисляя интеграл (3.69), получим

$$E_l = \frac{U_0^2\,\omega_в}{\pi}\,\frac{\sin(l\pi)}{l\pi},\ l = 0, 1, \ldots \tag{3.70}$$

Следовательно

$$E_l = \begin{cases} \dfrac{U_0^2\,\omega_в}{\pi}, & l = 0; \\ 0, & l \neq 0. \end{cases} \tag{3.71}$$

Из системы (3.71) вытекает, что взаимная энергия E_l сигналов $v_k(t)$, заданных соотношением (3.67) и разнесенных на l интервалов Δt, определяющихся по формуле (3.66), равняется нулю, т.е. функции $v_k(t)$ образуют ансамбль ортогональных функций.

К аналитической форме ансамбля ортогональных функций $v_k(t)$ перей-

дем, воспользовавшись выражениями (3.58) и (3.67). Получим

$$v_k(t) = \frac{U_0\omega_в}{\pi}\frac{\sin\omega_в(t - k\pi/\omega_в)}{\omega_в(t - k\pi/\omega_в)}, \quad k = 0, \pm 1, \pm 2, \ ... \qquad (3.72)$$

Таким образом, ансамбль ортогональных функций $v_k(t)$ образован за счет часового сдвига идеального низкочастотного сигнала (3.59) на k интервалов $\Delta t = \pi/\omega_в$. Графики сигналов (3.72) для $k = 0$ и $k = 2$ приведены на рис. 3.33.

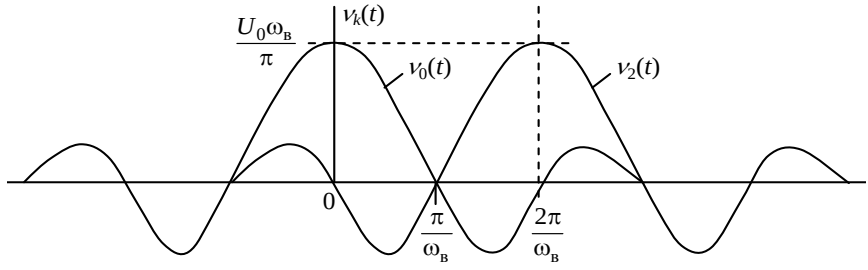


Рис. 3.33. Графики ортогональных ИНС

Построим на основании функций $v_k(t)$, заданных соотношениям (3.72), систему ортонормированных функций (базис), которые обозначим как $\varphi_k(t)$. Для этого необходимо требовать, чтобы энергия каждой функции $v_k(t)$ равнялась единице. Согласно соотношению (3.71) энергия E каждого сигнала составляет

$$E_0 = \frac{U_0^2\,\omega_в}{\pi}. \qquad (3.73)$$

Приравняв правую часть (3.73) к единице, получим, что система ортогональных функций $v_k(t)$ становится нормируемой при условии

$$U_0 = \sqrt{\frac{\pi}{\omega_в}}. \qquad (3.74)$$

Подставив значение (3.74) в формулу (3.72), приходим к системе ортонормированных функций:

$$\varphi_k(t) = \sqrt{\frac{\omega_в}{\pi}} \cdot \sin\omega_в\left(t - \frac{k\pi}{\omega_в}\right) \Big/ \omega_в\left(t - \frac{k\pi}{\omega_в}\right), \quad k = 0, \pm 1, \pm 2, ... \qquad (3.75)$$

которая образует так называемый *базис Котельникова* в пространстве ИНС с

частотами, ограниченными сверху величиной $\omega_в$. Отдельная функция $\varphi_k(t)$ называется *k-й отсчетной функцией*.

Таким образом, *базис Котельникова является совокупностью ортонормированных функций* $\varphi_k(t)$, *образованных из идеального низкочастотного сигнала* (*с верхней предельной частотой* $\omega_в$ *и спектральной плотностью в полосе частот* $[-\omega_в, \omega_в]$, *что равняется* $\sqrt{\pi/\omega_в}$ *) за счет его часового сдвига на промежуток времени* $t_k = k\Delta t = k\dfrac{\pi}{\omega},\ k = 0, \pm 1, \pm 2, ...$

Теорема Котельникова (**теорема отсчетов**). Теорема, которую доказал В. О. Котельников в 1933 г., является одним из фундаментальных положений теоретической радиотехники. *Теорема устанавливает возможность как угодно точного восстановления сигнала с ограниченным спектром по его дискретным значениям, взятым через равные промежутки времени.*

Пусть $s(t)$ - произвольный сигнал, спектральная плотность которого отличается от нуля лишь в интервале частот $-\omega_в < \omega < \omega_в$. *Его можно разложить в обобщенный ряд Фурье по базису Котельникова, т.е. подать в виде*

$$s(t) = \sum_{k=-\infty}^{\infty} c_k \varphi_k(t). \qquad (3.76)$$

Коэффициенты c_k ряда (3.76) являются скалярными произведениями сигнала $s(t)$ и *k*-й отсчетной функции $c_k = (s, \varphi_k)$. Удобный способ вычисления этих коэффициентов заключается в применении обобщенной формулы Релея

$$c_k = \frac{1}{2\pi}\int_{-\infty}^{\infty} \dot{S}(\omega)\dot{\Phi}_k^*(\omega)\, d\omega. \qquad (3.77)$$

Здесь $\dot{S}(\omega)$ - спектральная плотность сигнала $s(t)$, $\dot{\Phi}_k(t)$ - спектральная плотность *k*-й отсчетной функции базиса Котельникова $\varphi_k(t)$, т.е. $\dot{\Phi}_k(\omega) = F\{\varphi_k(t)\}$, где F - оператор прямого преобразования Фурье.

Приступим к вычислению спектральной плотности $\dot{\Phi}_k(\omega)$. Заметим, что из сравнения выражений (3.74) и (3.77) вытекает

$$\varphi_k(t) = \sqrt{\frac{\pi}{\omega_в}} v_k(t)\Big|_{U_0=1}. \qquad (3.78)$$

Таким образом, согласно равенству (3.78) базисная функция Котельникова $\varphi_k(t)$ с точностью до коэффициента $\sqrt{\pi/\omega_k}$ совпадает с выражением

(3.75) для функции $v_k(t)$, спектральная плотность которой задана соотношениям (3.67). Следовательно

$$\dot{\Phi}_k(\omega) = \sqrt{\frac{\pi}{\omega_в}} e^{-j\omega k\frac{\pi}{\omega_в}}, \tag{3.79}$$

причем в формуле (3.79) учтено условие выражения (3.78), в соответствии с которым $U_0 \equiv 1$. Подставив значение (3.79) в (3.77), получим

$$c_k = \sqrt{\frac{\pi}{\omega_в}} \left\{ \frac{1}{2\pi} \int_{-\omega_в}^{\omega_в} \dot{S}(\omega) e^{j\omega\frac{k\pi}{\omega_в}} d\omega \right\}. \tag{3.80}$$

Выражение в фигурных скобках правой части формулы (3.80) есть не что иное, как мгновенное значение сигнала $s(t_k) = s_k$ в k-й отсчетной точке: $t_k = \frac{k\pi}{\omega_в} = \frac{k}{2F_в}$. Таким образом, $c_k = \sqrt{\frac{\pi}{\omega_в}} s_k$, откуда вытекает окончательная форма ряда Котельникова

$$s(t) = \sum_{k=-\infty}^{\infty} s_k \sin\omega_в\left(t - \frac{k\pi}{\omega_в}\right) \Big/ \omega_в\left(t - \frac{k\pi}{\omega_в}\right). \tag{3.81}$$

Формула (3.81) является содержанием теоремы Котельникова (теоремы отсчетов): *произвольный сигнал, спектр которого не содержит частот, выше* $F_в$, *можно представить последовательностью дискретных отсчетов этого сигнала, взятых через одинаковые промежутки времени* $1/(2F_в)$.

Обозначив рассмотренный промежуток времени отбора дискретных отсчетов сигнала (назовем его периодом дискретизации) T, т.е. взяв

$$T = \frac{1}{2} F_в \tag{3.82}$$

и введя понятие частоты дискретизации $F = 1/T$, можно сформулировать ***теорему Котельникова*** таким образом: *для неискаженного представления сигнала с ограниченным спектром последовательностью его дискретных отсчетов частота дискретизации F должна равняться удвоенной верхней частоте спектра сигнала* $F_в$, *т.е.* $F = 2F_в$.

Особенность теоремы Котельникова заключается в ее конструктивном характере. Эта теорема не только указывает на возможность разложения сиг-

нала в соответствующий ряд, но и определяет способ восстановления непрерывного сигнала, заданного своими отсчетными значениями (рис. 3.34).

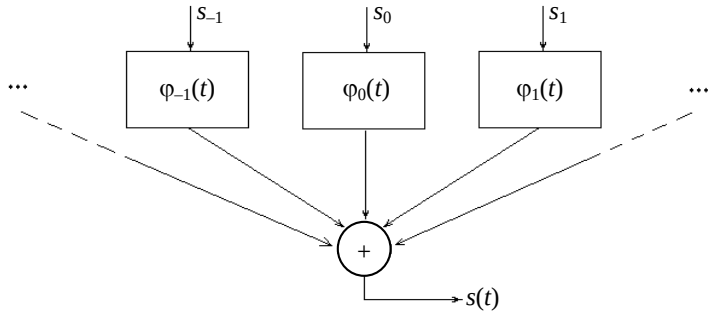


Рис. 3.34. Алгоритм синтеза сигнала

Пусть имеем совокупность генераторов, которые создают на своих выходных зажимах отсчетные функции

$$\varphi_k(t) = \sin\omega_в\left(t - \frac{k\pi}{\omega_в}\right) \Big/ \omega_в\left(t - \frac{k\pi}{\omega_в}\right). \qquad (3.83)$$

Генераторы являются управляемыми - амплитуда их сигналов пропорциональна отсчетным значениям s_k. Если теперь объединить колебание на выходах, подав их на сумматор, то на выходе сумматора согласно формуле (3.81) появится мгновенное значение синтезированного сигнала

$$s(t) = \sum_{k=-\infty}^{\infty} s_k \varphi_k(t). \qquad (3.84)$$

Физически восстановление сигнала с ограниченным спектром из последовательности его дискретных отсчетов реализуется с помощью идеального ФНЧ, частота среза которого выбирается равной верхней частоте спектра сигнала.

Допустим, что дискретные отсчеты сигнала формируются, как изображено на рис. 3.35. Последовательность s_k образуется как результат умножения сигнала $s(t)$ и задержанной на k периодов дискретизации дельта-функции, т.е. $s_k = s(t)\delta(t - kT)$, причем T выбирается из условия (3.82).

Реакцией идеального ФНЧ на входное воздействие типа дельта-импульс $\delta(t - kT)$ является функция вида (3.83), или отсчетная функция ряда Котельникова. С учетом того, что ФНЧ является интегрирующим звеном (сумматором), нетрудно понять, что этот ФНЧ именно и реализует формулу (3.84), т.е. восстанавливает сигнал $s(t)$.

Таким образом, *для восстановления сигнала* $s(t)$, имеющего ограниченный спектр $F_в$, *по заданной последовательности его дискретных отсчетов* s_k, *взятых с частотой дискретизации* $F = 2F_в$, *необходимо пропустить эту последовательность через идеальный ФНЧ, частота среза которого выбирается такой, которая равняется верхней частоте спектра сигнала* $F_в$.

Аппроксимация прямоугольного сигнала рядом Котельникова. Ряд Котельникова часто используют для приближенного описания сигналов с неограниченным спектром, значительная часть энергии которого сосредоточена в низкочастотной области.

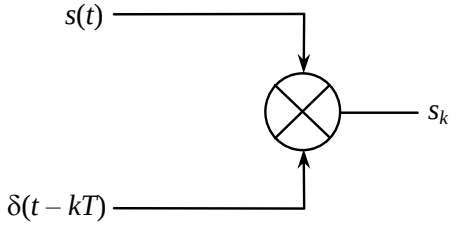


Рис. 3.35. Формирование дискретных отсчетов сигнала

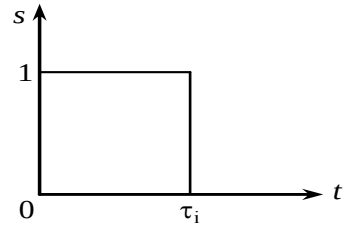


Рис. 3.36. Прямоугольный видеоимпульс

Пример. Рассмотрим прямоугольный видеоимпульс. Прямоугольный видеоимпульс не принадлежит к сигналам с ограниченным спектром, однако модуль его спектральной плотности достаточно быстро (по закону $1/\omega$) уменьшается с ростом частоты.

Чтобы упростить представления прямоугольного видеоимпульсу рядом Котельникова, разместим импульс на оси времени так, как это показано на рис. 3.36. Математическая модель $s(t)$ такого сигнала достаточно простая

$$s(t) = \begin{cases} 1, & t \in [0, \tau_з]; \\ 0, & t \notin [0, \tau_з]. \end{cases} \tag{3.85}$$

Поскольку согласно равенству (3.85) в области отрицательных значений времени сигнал отсутствует, представим ряд Котельникова такого сигнала (см. формулу (3.81)) выражением

$$\tilde{s}_n(t) = \sum_{k=0}^{n-1} s_k \sin \omega_в \left(t - \frac{k\pi}{\omega_в} \right) \Big/ \omega_в \left(t - \frac{k\pi}{\omega_в} \right), \tag{3.86}$$

в котором n - количество отсчетов s_k прямоугольного видео-импульса, которое эквидистантно подбирается на всем интервале его существования от нуля

до τ_u.

Поскольку по условию (3.85) все отсчеты s_k импульса в пределах его длительности τ_u равняются единице, ряду (3.86) можно придать вид

$$\tilde{s}_n(t) = \sum_{k=0}^{n-1} \sin\omega_в\left(t - k\pi/\omega_в\right) \Big/ \omega_в\left(t - k\pi/\omega_в\right), \qquad (3.87)$$

где осталось еще неопределенным значение верхней частоты $\omega_{â}$ сигнала $\tilde{s}(t)$, что аппроксимирует функцию $s(t)$.

Пусть задано количество n отсчетов прямоугольного видеоимпульса, которые эквидистантно подбираются по всей области существования $t \in (0, \tau_з)$. Очевидно, что период дискретизации

$$T = \frac{\tau_з}{n-1}. \qquad (3.88)$$

Это значит, что когда $n = 2$, период дискретизации равняется длительности импульса τ_u, а два отсчета сигнала s_0 и s_1 берутся соответственно в начале и в конце импульса. При $n = 3$ отсчета s_0, s_1 и s_2 берутся с интервалом $T = \tau_u / 2$ соответственно в начале, середине и конце импульса. И, наконец, при $n = 5$ отсчет s_0 берется в самом начале импульса, а все следующие - через четверть длительности импульса, т.е. $T = \tau_u / 4$ и т.д.

В соответствии с выражением (3.66) ИНС ортогонализируется при сдвиге на промежуток времени $\Delta t = T$, если его верхняя частота $\omega_{â}$ удовлетворяет условию

$$\omega_в = \pi / T. \qquad (3.89)$$

Подставив выражение (3.88) в формулу (3.89), получим значение верхней частоты $\omega_в$ спектра сигнала $\tilde{s}(t)$ при условии, что прямоугольный видеоимпульс подается последовательностью n эквидистантно размещенных по всей длительности τ_u отсчетов импульса $s(t)$, т.е.

$$\omega_в = \pi(n-1)/\tau_i. \qquad (3.90)$$

На основании соотношений (3.87) и (3.90) приходим к окончательной формуле ряда Котельникова, который аппроксимирует прямоугольный видеоимпульс

$$\tilde{s}_n(t) = \sum_{k=0}^{n-1} \sin\left(\pi(n-1)/\tau_i\right)\left(t - k\tau_i/(n-1)\right) \Big/ \left(\pi(n-1)/\tau_i\right)\left(t - k\tau_i/(n-1)\right). \qquad (3.91)$$

Т.е. если представить прямоугольный видеоимпульс всего двумя его отсчетами, взятыми в начале и в конце импульса, то это значит, что в спектре этого импульса будут учтены составляющие, ограниченные частотой $\omega_в = \pi / \tau_i$. По формуле (3.91) находим приближенное выражение математической модели сигнала аппроксимации

$$\tilde{s}_2(t) = \frac{\sin \pi t/\tau_i}{\pi t/\tau_i} + \frac{\sin \pi/\tau_i\,(t-\tau_i)}{\pi/\tau_i\,(t-\tau_i)}.$$

А если представить этот импульс тремя равноотдаленными отсчетами (то есть взять $n=3$), то, как следует из соотношения (3.90), в спектре сигнала будут учтены все частоты, вплоть до $\omega_в = 2\pi/\tau_i$, и поэтому:

$$\tilde{s}_3(t) = \sin\frac{2\pi t}{\tau_i}\Big/\frac{2\pi t}{\tau_i} + \sin\frac{2\pi}{\tau_i}\left(t-\frac{\tau_i}{2}\right)\Big/\frac{2\pi}{\tau_i}\left(t-\frac{\tau_i}{2}\right) + \sin\frac{2\pi}{\tau_i}(t-\tau_i)\Big/\frac{2\pi}{\tau_i}(t-\tau_i).$$

Соответствующие графики изображены на рис. 3.37.

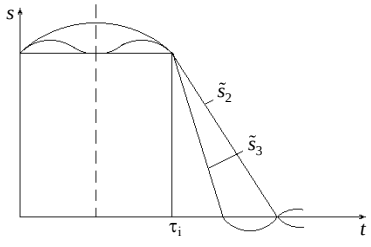


Рис. 3.37. Аппроксимация прямоугольного видеоимпульса рядом Котельникова

Естественно, что с ростом числа n, т.е. с уменьшением часового интервала T между выборочными отсчетами, точность аппроксимации будет повышаться.

Дискретизация полосовых сигналов. Предыдущий анализ по умалчиванию базировался на том предположении, что аналоговые сигналы (импульсы) с неограниченным спектром принадлежат к классу так называемых *низкочастотных сигналов, для которых большая часть энергии сигнала сосредоточивается* в низкочастотной области.

Существует и другой класс сигналов, энергия которых сосредоточивается в высокочастотной области и практически не существует в области низких частот. Назовем такой класс сигналов *высокочастотным.* Из этого класса сигналов широкого применения приобрели так называемые *полосовые сигналы*, примером которых являются *модулируемые сигналы.*

Дадим некоторые рекомендации относительно алгоритмов дискретизации аналоговых полосовых сигналов и их восстановления из последовательности дискретных отсчетов.

Следовательно, пусть имеем некоторый сигнал *s(t)*, спектр которого содержится в интервале $(\omega_0 - \Delta\omega, \omega_0 + \Delta\omega)$, в котором ω_0 - центральная несущая частота, а $\Delta\omega$ - девиация частоты. Будем считать, что вне отмеченного интервала гармоник спектра полосового сигнала нет. Для случая амплитудной модуляции $\Delta\omega$ являет собой верхнюю гармонику $\omega_â$ спектра модулированного сигнала. В случае угловой модуляции $\Delta\omega$ - это непосредственно девиация частоты фазомодулированных (ФМ) или частотномодулированных (ЧМ) сигналов.

Теорема Котельникова и для таких сигналов дает возможность выбирать значение частоты дискретизации F полосового сигнала $s(t)$. Верхней угловой частоте $\omega_в = \omega_0 + \Delta\omega$ сигнала *s(t)* соответствует верхняя циклическая частота $F_в$, связанная с $\omega_â$ выражением $\omega_в = 2\pi F_в$.

Частота дискретизации согласно теореме Котельникова должна удовлетворять условию $F \geq 2F_{в}$.

Для восстановления аналогового полосового сигнала $s(t)$ последовательность дискретных отсчетов $\{s_k\}$ следует подать на вход идеального полосового фильтра с частотами среза: верхней $\omega_{в}$, определенной ранее, и нижней $\omega_{н} = \omega_0 - \Delta\omega$. Отклик такого фильтра и будет являться достаточно приемлемой аппроксимацией исходного полосового сигнала.

3.5. Дискретные сигналы

Напомним основные свойства дискретных сигналов. Значение дискретных сигналов определено не для всех моментов времени, а лишь в исчисляемом множестве точек $(..., t_0, t_1, t_2, ...)$. Поэтому если математическая модель аналогового сигнала $x(t)$ имеет обычные свойства гладкой функции, то дискретный сигнал $x_{д}(t)$ описывается последовательностью $..., x_0, x_1, x_2, ...$ своих отсчетных значений в моменты времени соответственно $..., t_0, t_1, t_2, ...$

Дискретные сигналы приобретают особое значение в последние десятилетия в связи с развитием цифровой техники связи и способов обработки информации на быстродействующих ЭВМ. Наметилась тенденция создания специализированных вычислительных устройств, так называемых *цифровых фильтров*, которые используются для обработки дискретных сигналов.

Дискретные сигналы $x(nT)$, $n = 0, 1, 2, ...$ образуются в результате дискретизации (отбора мгновенных значений) непрерывных (аналоговых) сигналов $x(t)$ в моменты времени t_n, что образуют исчисляемую последовательность $t = nT$, $n = 0, 1, 2, ...$, где T - шаг дискретизации по времени (рис. 3.38). Дискретные сигналы $x(nT)$, $n = 0, 1, 2, ...$ обретают значения из континуального множества во всем интервале значений $(x_{\min}, x_{\max})$ непрерывной функции $x(t)$.

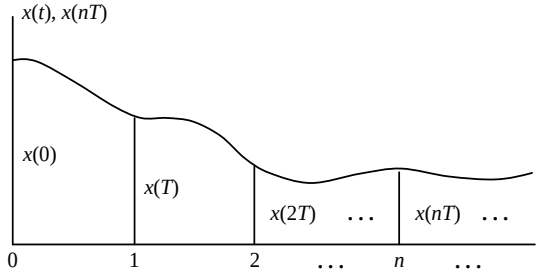


Рис. 3.38. Формирование дискретных сигналов

При спектральном анализе дискретных сигналов удобно применять так называемое *Z-преобразование*, которое относительно дискретных сигналов играет такую же роль, как и интегральное преобразование Фурье для непрерывных сигналов. Основу *Z-преобразования* составляет интегральное преобразование Лапласа.

Интегральное преобразование Лапласа. Наряду с преобразованием Фурье для решения самых разнообразных задач, связанных с изучением сигналов, широко используется еще один вид интегральных преобразований - *преобразование Лапласа*. Различают *одностороннее* и *двустороннее* преобразование Лапласа.

Одностороннее преобразование Лапласа вводится для сигналов $x(t)$, заданных (таких, которые существуют) лишь на положительной полуоси времени.

Пусть $x(t)$ - некоторый сигнал, действительный или комплексный, определенный при $t>0$ и тождественно равный нулю при отрицательном значении времени. Преобразование Лапласа $X(p)$ этого сигнала задается интегралом

$$X(p)=\int_0^{\infty} x(t)\,e^{-pt}dt, \tag{3.92}$$

где p - параметр преобразования, в общем случае - комплексное число $p=\sigma+j\omega$, получившее название *комплексной частоты*.

Для сигналов, определенных на всей оси времени от минус бесконечности до плюс бесконечности, вводят *двустороннее преобразование Лапласа*

$$X(p)=\int_{-\infty}^{\infty} x(t)\,e^{-pt}dt. \tag{3.93}$$

Сигнал $x(t)$ называется *оригиналом*, а функция $X(p)$ - его изображением по Лапласу (для краткости - просто изображением).

Если в формуле (3.92) или (3.93) положим $p=j\omega$, то приходим к преобразованию Фурье. Следовательно, если для сигнала $x(t)$ известно его преобразование Лапласа $X(p)$, то спектральная плотность $\dot{X}(\omega)$ сигнала $x(t)$ определится заменой в изображении $X(p)$ параметра p на $j\omega$,

$$\dot{X}(\omega)=X(p)\big|_{p=j\omega}. \tag{3.94}$$

Пьер Симон Лаплас (Pierre-Simon Laplace, 1749-1827), Основные астрономические работы Лапласа посвящены небесной механике. Этот термин впервые употребил сам Лаплас в названии пятитомной фундаментальной работы "Трактат о небесной механике" (1798-1825). Решил сложные задачи по движению планет и их спутников, в частности Луны. Разработал теорию возмущений траекторий планет, Солнца и Луны, предложил новый способ вычисления орбит, доказал устойчивость Солнечной системы, открыл причины ускорения в движении Луны.

Рассмотрим примеры определения спектральной плотности простейших сигналов по их изображениях по Лапласу.

1. ***Сигнал типа односторонней экспоненты*** $x(t) = U e^{-\alpha t},\ t \geq 0.$

Воспользовавшись соотношениям (3.92), получим

$$X(p) = = U\int_0^\infty e^{-\alpha t} e^{-\alpha p} dt = U\int_0^\infty e^{-(\alpha + p)t} dt.$$

Вычисляя интеграл, получаем

$$X(p) = \frac{U}{\alpha + p}.$$

В соответствии с формулой (3.94) спектральная плотность $\dot{X}(\omega)$ *этого сигнала*

$$\dot{X}(\omega) = \frac{U}{\alpha + j\omega}$$

совпадает из ранее вычисленной путем интегрального преобразования Фурье спектральной плотностью.

2. ***Сигнал типа двусторонней экспоненты*** $x(t) = U e^{-\alpha|t|},\ t \in (-\infty, \infty).$

В соответствии с формулой (3.93) имеем

$$X(p) = U\int_{-\infty}^{\infty} e^{-\alpha|t|}\, e^{-pt} dt = X_1(p) + X_2(p),$$

где

$$X_1(p) = U\int_0^\infty e^{-(\alpha + p)t} dt = \frac{U}{\alpha + p},$$

а

$$X_2(p) = U\int_{-\infty}^{0} e^{(\alpha - p)t} dt = \frac{U}{\alpha - p} e^{(\alpha - p)t}\Big|_{-\infty}^{0} = \frac{U}{\alpha - p}.$$

Таким образом, изображение по Лапласу двусторонней экспоненты определяется выражением

$$X(p) = \frac{2\alpha U}{\alpha^2 - p^2},$$

а на основании соотношения (3.94) ее спектральная плотность

$$\dot{X}(\omega) = \frac{2\alpha U}{\alpha^2 + \omega^2}.$$

Преобразование Лапласа имеет свойства, аналогичные свойствам преобразования Фурье. Приводим важнейшие из них:

$$\sum_k \alpha_k x_k(t) \leftrightarrow \sum_k \alpha_k X(p);$$

$$x(t \pm t_0) \leftrightarrow X(p) e^{\pm p t_0};$$

$$x(\gamma t) \leftrightarrow \frac{1}{\gamma} X(p/\gamma);$$

$$\frac{dx}{dt} \leftrightarrow pX(p); \tag{3.95}$$

$$\int_{-\infty}^{t} x(\tau)\, d\tau \leftrightarrow \frac{1}{p} X(p); \tag{3.96}$$

$$x_1(t) * x_2(t) \leftrightarrow X_1(p) X_2(p).$$

где * - знак оператора свертки.

Множитель p в выражении (3.95) можно назвать *оператором дифференцирования*, а множитель $1/p$ в формуле (3.96) - *оператором интегрирования* в пространстве изображений по Лапласу.

Элементы теории Z-преобразования. Проще всего к *Z-преобразованию* можно прийти в результате дискретизации всех функций времени в преобразовании Лапласа. Напомним, что преобразованием Лапласа аналоговой функции $x(t)$ является функция

$$X(p) = \int_0^{\infty} x(t) e^{-pt} dt. \tag{3.97}$$

Попытаемся вычислить преобразование Лапласа для дискретизированной функции $x(nT)$, формально воспользовавшись соотношением (3.97). В результате временной дискретизации непрерывных функций $x(t)$ и e^{-pt}, которая сводится к замене непрерывного времени t на дискретное nT, приходим соответственно к дискретным функциям $x(nT)$ и e^{-pnT}. Следовательно, подынтегральная функция оказывается отличной от нуля лишь в дискретные моменты времени nT, $n = 0, 1, 2, \ldots$, которые образуют счетное множество. Такая операция дискретизации дает возможность перейти от интеграла к сумме произведений дискретных функций. Обозначив бесконечную сумму таких произведений $\hat{X}(p)$, получим

$$\hat{X}(p) = \sum_{n=0}^{\infty} x(nT) e^{-pnT}. \tag{3.98}$$

Воспользовавшись заменой

$$z = e^{pT} \tag{3.99}$$

и обозначив левую часть формулы (3.98) как $X(z)$, получим окончательно

$$X(z)=\sum_{n=0}^{\infty} x(n)z^{-n}. \qquad (3.100)$$

В выражении (3.100) аргумент T для упрощения последующих преобразований опущен.

Соотношение (3.100) называется *Z-преобразованием последовательности дискретных сигналов* $x(n)$ и часто обозначается как $Z\{x(n)\}$.

Отметим некоторые *свойства Z-преобразования*, которые во многом аналогичные свойствам преобразований Фурье и Лапласа.

1. *Линейность.* Если $\{x_k\}$ и $\{y_k\}$ - две числовые последовательности, отображающие некоторые дискретные сигналы, причем известны соответствующие *Z-преобразования* $X(z)$ и $Y(z)$, то сигналу

$$u(n)=\alpha x(n)+\beta y(n) \qquad (3.101)$$

отвечает преобразование

$$U(z)=\alpha X(z)+\beta Y(z) \qquad (3.102)$$

при любых постоянных α и β.

Доказательство равенства (3.102) проводится подстановкой суммы формулы (3.101) в выражение (3.100).

2. *Z-преобразование смещенного сигнала.* Рассмотрим дискретный сигнал $\{y_n\}$, получающийся из дискретного сигнала $\{x_n\}$ сдвигом последнего на одну позицию в сторону запаздывания, т.е.

$$y(n)=x(n-1).$$

Непосредственное вычисление *Z-преобразования* приводит к такому результату:

$$Y(z)=\sum_{n=0}^{\infty} x(n-1)\,z^{-n}.$$

Преобразуем правую часть последнего выражения к виду

$$z^{-1}\sum_{n=0}^{\infty} x(n-1)\,z^{-(n-1))}.$$

После замены

$$k=n-1$$

имеем

$$z^{-1}\sum_{k=-1}^{\infty} x(k)\,z^{-k}=x(-1)+z^{-1}\sum_{k=0}^{\infty} x(k)\,z^{-k}.$$

Полагая для физически реализуемых сигналов $x(-1)=0,$ получим окончательно

$$Z\left\{x\left(n-1\right)\right\}=z^{-1}X\left(z\right). \tag{3.103}$$

Выражение (3.103) можно обобщить для любого значения задержки *k* последовательности $x\left(n\right),$ а именно

$$Z\{x(n-1)\}=z^{-k}X\left(z\right)z^{-k}. \tag{3.104}$$

Таким образом, *Z-преобразование последовательности, задержанной на k периодов дискретизации, равно произведению Z-преобразования исходной (не задержанной) последовательности на* z^{-k}.

Символ z^{-1} в соотношении (3.103) играет роль *оператора единичной задержки сигнала* на один период дискретизации в Z-области. Формуле (3.103) можно дать графическую интерпретацию (рис. 3.39).

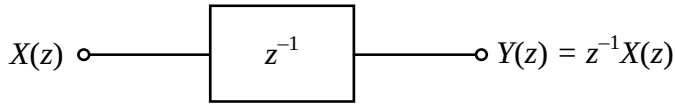


Рис. 3.39. Изображение линии задержки дискретного сигнала на один период дискретизации в *Z*-области

На рис. 3.40 показана интерпретация соотношения (3.104).

На рис. 3.39 и 3.40 обозначено: $X\left(z\right)$- входное влияние, а $Y\left(z\right)$- отзыв цепи, который состоит из последовательно соединенных единичных линий задержки (на один период дискретизации) для Z-области.

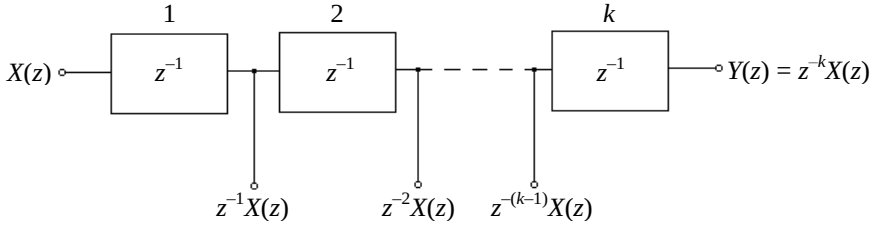


Рис. 3.40. Линия задержки дискретных сигналов на *k* периодов дискретизации

Следовательно, *если исходная последовательность* $y\left(n\right)$ *образована из задержанной на k периодов дискретизации входной последовательности* $x\left(n\right),$ *т.е.*

$$y\left(n\right)=x\left(n-k\right),$$

то Z-изображение отклика цепи $Y(z)$ *связано с Z-изображением входного сигнала* $X(z)$ соотношением

$$Y(z) = z^{-k} X(z),$$

что соответствует выражению (3.104).

3. *Z-преобразование свертки.* Дискретным сигналам $x(n)$ $y(n)$ можно поставить в соответствие дискретную свертку

$$f(k) = \sum_{n=0}^{\infty} x(n) y(k-n) = \sum_{n=0}^{\infty} y(n) x(k-n),$$

которой, как и свертке аналоговых сигналов, соответствует произведение их *Z-изображений (*преобразований), т.е.

$$Z\{x(n) * y(n)\} = X(z) Y(z),$$

где $*$ - знак оператора свертки.

Другими словами, *Z-изображение свертки двух дискретных сигналов равно произведению Z-изображений этих сигналов.*

Спектр дискретных сигналов. Как известно, спектральная плотность $\dot{X}(\omega)$ *аналоговых сигналов* $x(t)$ определяется на основании прямого преобразования Фурье временной функции $x(t)$, т.е.

$$\dot{X}_F(\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t} dt\,. \qquad (3.105)$$

Здесь (и дальше) нижний индекс при $\dot{X}$ соответствует типу изображения (в частности, *F* указывает на Фурье-изображение).

Спектральную плотность аналогового сигнала $x(t)$ можно получить также на основании преобразования Лапласа этого сигнала. Действительно, если известно преобразование Лапласа $X(p)$ сигнала

$$x(t) X(p) = \int_{0}^{\infty} x(t) e^{-pt} dt, \qquad (3.106)$$

то, как это выплывает из сравнения формул (3.105) и (3.106), спектральную плотность $\dot{X}(\omega)$ сигнала $x(t)$ можно определить на основании преобразования Лапласа $X(p)$ заменой параметра p на $j\omega$,

$$\dot{X}_F(\omega) = X(p)\big|_{p=j\omega}.$$

Временная дискретизация функций, входящих в правую часть изображения Лапласа, приводит к *Z*-изображению дискретных сигналов $x(nT)$:

$$X(z) = \sum_{n=0}^{\infty} x(nT)\, z^{-n}, \qquad (3.107)$$

где *z* задано соотношение (3.99).

Полагая, как и при определении спектра сигнала *x(t)* по его преобразова-

нию Лапласа *X(p)*, что параметр *p* равен $j\omega$, приходим к следующему способу вычисления спектральной плотности дискретных сигналов по их *Z*-изображениям:

$$\dot{X}_z(\omega) = X(z)\big|_{z=e^{j\omega T}}. \qquad (3.108)$$

Объединяя формулы (3.107) и (3.108), получим выражение для спектральной плотности последовательности дискретных сигналов

$$\dot{X}_z(\omega) = \sum_{n=0}^{\infty} x(n)\, e^{-j\omega nT} \qquad (3.109)$$

Как видно из соотношений (3.108) и (3.109), спектральная плотность дискретного сигнала является функцией комплексной экспоненты, и согласно формуле Эйлера

$$e^{j\omega T} = \cos\omega T + j\sin\omega T,$$

является периодической функцией. В силу этого спектр дискретного сигнала также становится периодическим.

Пример. Рассмотрим аналоговый сигнал $x(t)$ типа односторонней экспоненты

$$x(t) = e^{-\alpha t},\ t \ge 0. \qquad (3.110)$$

Подвергнув его временной дискретизации с периодом *T*, получим последовательность дискретных сигналов

$$x(nT) = e^{-\alpha nT} = a^n, \qquad (3.111)$$

где $a = e^{-\alpha T} < 1.$

Напомним, что спектральная плотность сигнала (3.110) имеет вид

$$\dot{X}_F(\omega) = \frac{1}{\alpha + j\omega}.$$

Его спектр амплитуд $A(\omega)$ является апериодической функцией

$$A_F(\omega) = \frac{1}{\sqrt{\alpha^2 + \omega^2}}.$$

Вычислим *Z-изображение* сигнала (3.111) по формуле

$$X(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n.$$

Поскольку $a < 1$, а $|z| = 1$, то ряд в правой части последней формулы сворачивается, приобретая вид

$$X(z) = \frac{1}{1 - a/z} = \frac{z}{z - a}. \qquad (3.112)$$

Воспользовавшись соотношениями (3.108) и (3.112), приходим к следующему выражению для спектральной плотности дискретизированной экспоненты:

$$\dot{X}_z(\omega) = \frac{e^{j\omega T}}{e^{j\omega T} - a}. \qquad (3.113)$$

Согласно формуле (3.113) амплитудный спектр $A_z(\omega)$ равен отношению модуля числителя этого выражения к модулю знаменателя. Модуль числителя равняется единице. При определении модуля знаменателя воспользуемся формулой Эйлера для комплексной экспоненты. Тогда

$$\left|e^{j\omega T} - a\right| = \left|(\cos\omega T - a) + j\sin\omega T\right| = \sqrt{(\cos\omega T - a)^2 + \sin^2\omega T}.$$

После элементарных преобразований получим

$$A_z(\omega) = \frac{1}{\sqrt{1 + a^2 - 2a\cos\omega T}}.$$

Графики спектров амплитуд аналоговой $A_F(\omega)$ и дискретизированной $A_z(\omega)$ экспонент изображен соответственно на рис. 3.41.

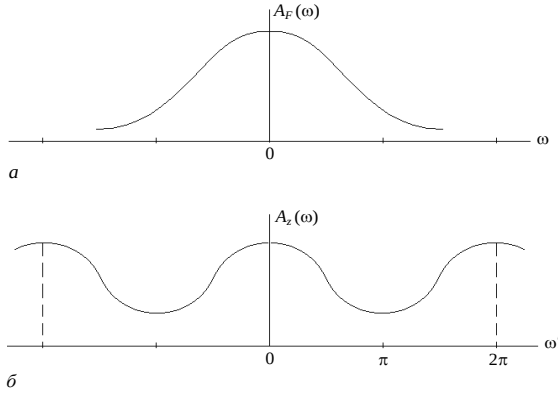


Рис. 3.41. Спектры амплитуд односторонней экспоненты для сигналов:
а - аналогового; *б* - дискретного

Из сопоставления графиков (см. рис. 3.41) видно, что даже в том случае, когда спектр аналогового сигнала апериодический, спектр дискретизированного сигнала становится периодическим, что является следствием дискретизации аналогового сигнала.

Таким образом, можно сформулировать важный вывод: *спектр дискретных сигналов является периодической функцией частоты даже в том случае, когда спектр его аналогового прототипа - апериодическая функция.*

Эффект наложения спектров при дискретизации сигналов. Рассмотрим некоторый гипотетический сигнал, амплитудный спектр которого имеет вид равностороннего треугольника с основой $\pm\omega_в$ (рис. 3.42).

Как известно, при дискретизации любого аналогового сигнала *x(t)* спектр дискретных сигналов *x(n)* становится периодическим, причем интервал периодичности $\omega T = 2\pi$, т.е. модуль и фаза спектральных составляющих сигнала на частотах

$$\omega_k = \omega \pm k\frac{2\pi}{T}, \quad k = 1, 2, \ldots \qquad (3.114)$$

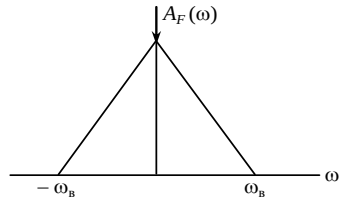


Рис. 3.42. Спектр амплитуд гипотетического сигнала

совпадают по модулю и фазе с гармоникой сигнала на частоте ω.

Действительно, спектральная составляющая дискретного сигнала *x(n)* на некоторой частоте ω определяется общим соотношением (3.108) и является функцией комплексного аргумента

$$e^{j\omega T}. \qquad (3.115)$$

Заменив угловую частоту ω в экспоненте (3.115) значением ω_k, заданным правой частью равенства (3.114), получим $e^{j\omega_k T} = e^{j\left(\omega \pm k\frac{2\pi}{T}\right)T} = e^{j\omega T}e^{\pm j2k\pi}$, т.е. $e^{j\omega_k T} = e^{j\omega T}$, чем и подтверждается периодичность значений спектральных составляющих дискретного спектра с периодом по частоте

$$\Delta_\omega = 2\pi / T. \qquad (3.116)$$

От интервала Δ_ω периодичности спектра дискретных сигналов по угловой частоте (3.116) можно перейти к интервалу Δ_f периодичности спектра по циклической частоте $\Delta_\omega = 2\pi\Delta_f$. Воспользовавшись значениям (3.116), получим

$$\Delta_f = 1/T = F, \qquad (3.117)$$

где F - частота часовой дискретизации сигнала.

Таким образом, согласно формуле (3.117): *спектр дискретизированного сигнала периодический, причем интервал периодичности* Δ_f *совпадает с частотой дискретизации* $F = 1/T$.

Допустим, что форма спектра амплитуд дискретизированного сигнала отвечает форме спектра аналогового сигнала и с учетом периодичности имеет такой вид, как изображено на рис. 3.43.

Графики на рис. 3.43 отражают случай, когда $\omega_â T_1 < \pi$, где $\omega_â$ - максимальная (верхняя) гармоника аналогового сигнала, а T_1 - период часовой дискретизации аналогового сигнала.

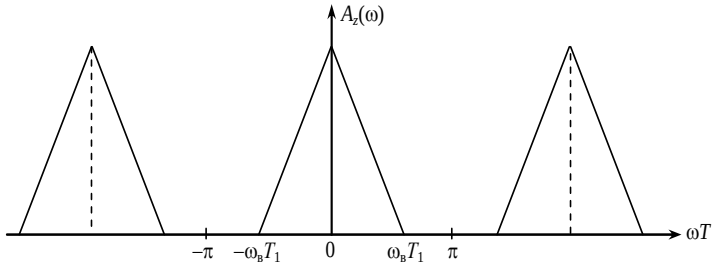


Рис. 3.43. Спектр амплитуд дискретного сигнала

Увеличим теперь период временной дискретизации аналогового сигнала и возьмем его равным значению T, при котором выполняется равенство

$$\omega_в T = \pi, \tag{3.118}$$

в соответствии с которым основы «треугольников» амплитудных спектров будут примыкать друг к другу, как это изображено на рис. 3.44.

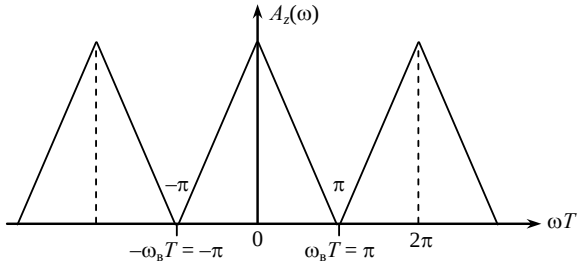


Рис. 3.44. Спектр амплитуд дискретного сигнала,
для которого выполняется условие (3.118)

По мере дальнейшего увеличения периода дискретизации $T_2 > T$ треугольники амплитудного спектра начинают накладываться друг на друга (рис. 3.45). В результате наложения высокочастотных гармоник спектра на низкочастотный амплитудный спектр дискретного сигнала (верхние ломаные прямые) оказывается существенно отличным от спектра аналогового сигнала (см. рис. 3.42).

Причина искажения амплитудного спектра дискретного сигнала (см. рис. 3.45) заключается в том, что частота дискретизации $F_2 = 1/T_2$ взята недостаточно большой, в результате чего высокочастотные составляющие периодического спектра попадают в область низших частот. Такой сдвиг спектральных составляющих из одного диапазона частот в другой называют *наложением спектров.*

Эффект наложения спектров дискретных сигналов можно устранить за счет соответствующего выбора частоты дискретизации аналоговых сигналов.

Выясним условия, при которых наложение спектров в дискретизированных сигналах отсутствует.

Пусть спектр аналогового сигнала ограничен верхней граничной частотой $\omega_в$, как это показано на рис. 3.42. Очевидно, что искажение спектра при дискретизации сигнала не возникает (см. рис. 3.44), если выполняется условие (3.118).

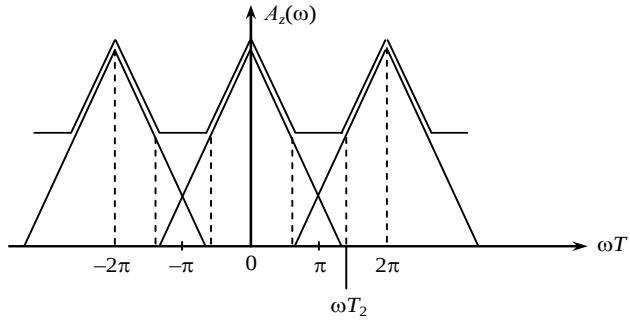


Рис. 3.45. Эффект наложения спектров дискретизированного сигнала

Переходя в соотношении (3.118) от угловой частоты $\omega_в$ к циклической

$$\omega_в = 2\pi F_в$$

($F_в$ - верхняя циклическая гармоника аналогового сигнала) и заменяя период временной дискретизации T частотой дискретизации сигнала, т.е. $F = 1/T$, получаем соотношение

$$F = 2F_в.$$

Это уже известное положение теоремы Котельникова (в зарубежной литературе ее называют теоремой Найквиста, или теоремой отсчетов): *если наивысшая частота в спектре сигнала* $x(t)$ *не превышает* $F_в$, *то функция* $x(t)$ *полностью определяется последовательностью своих значений в моменты времени, отдаленные один от другого не более чем на* $1/2F_в$ *секунд.*

Или другими словами: *любой аналоговый сигнал без искажения можно представить последовательностью его дискретных отсчетов* $x(nT)$ *при условии, что частота дискретизации не менее чем вдвое превышает наивысшую гармонику спектра аналогового сигнала.*

На практике это условие выполнить затруднительно, поскольку спектр аналоговых сигналов чрезвычайно широк. Поэтому при конечной частоте дискретизации спектр дискретизированного сигнала отличается от спектра аналогового сигнала.

Дискретное преобразование Фурье. Как было установлено (см. формулу (3.109)), спектральная плотность $\dot{X}_z(\omega)$ *дискретного сигнала* $x(nT)$ определяется выражением

$$\dot{X}_z(\omega) = \sum_{n=0}^{\infty} x(nT)\ e^{-j\omega nT}, \qquad (3.119)$$

где n - номер дискретного отсчета непрерывной функции; T - период дискретизации непрерывной функции $x(t)$.

Согласно формуле (3.119) спектр дискретного сигнала - сплошной. Но таким он бывает лишь при условии, что объем выборки дискретного сигнала бесконечен. На практике выборка отсчетов сигнала всегда конечномерна. Кроме того, по многим причинам желательно вычислять преобразование Фурье на ЭВМ. Это значит, что конечномерной является не только выборка дискретных отсчетов сигнала, но и соответствующее этой выборке количество гармоник спектра дискретного сигнала.

Допустим, что некоторая непрерывная функция (аналоговый сигнал) $x(t)$ представлена последовательностью N отсчетов этой функции

$$x(n) = x(nT),\ n = \overline{0,\ N-1}, \qquad (3.120)$$

где T - период временной дискретизации аналогового сигнала.

Приведем в соответствие конечновыборочным отсчетам сигнала (3.120) конечновыборочную последовательность спектральных составляющих $\dot{X}(k)$, взяв $k = \overline{0,\ N-1}$. Для вычисления N спектральных составляющих (гармоник спектра) будем действовать таким способом. Сначала заменим в формуле (3.119) угловую частоту ω циклической f

$$\omega = 2\pi f, \qquad (3.121)$$

а затем перейдем от непрерывных частот к дискретным:

$$\begin{cases} \omega \to \omega_k, \\ f \to f_k, \end{cases} \qquad k = \overline{0,\ N-1}. \qquad (3.122)$$

Расставим N спектральных составляющих дискретного сигнала эквидистантно на всем частотном интервале периодичности спектра, который, как показано в предыдущем подразделе, равняется F (рис. 3.46).

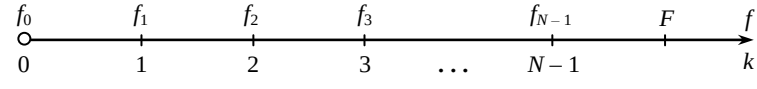


Рис. 3.46. Распределение гармоник спектра по оси частот

Обозначим

$$f_k = k\, f_1,\ k = \overline{0,\, N-1}, \tag{3.123}$$

где

$$f_1 = F/N \tag{3.124}$$

интервал частотной дискретизации спектра.

На основании соотношений (3.121) - (3.124) преобразуем бесконечный ряд (3.119) к конечному, записав его в виде

$$\dot{X}(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}knTF}.$$

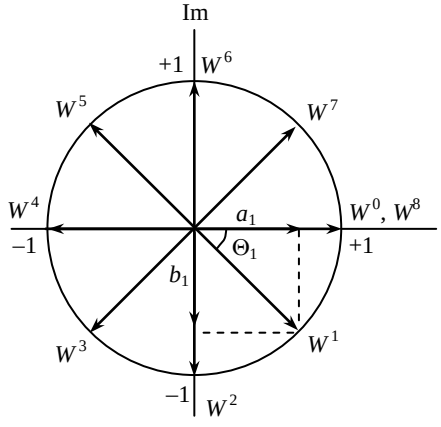


Рис. 3.47. Степени фазового умножителя для восьмиточечного ДПФ

Поскольку $TF = 1$, то имеем

$$\dot{X}(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}.$$

Обозначив

$$W = e^{-j\frac{2\pi}{N}}, \tag{3.125}$$

получим окончательное выражение

$$\dot{X}(k) = \sum_{n=0}^{N-1} x(n) W^{kn},\ k,n = \overline{0,N-1}. \tag{3.126}$$

Соотношение (3.126) имеет название *дискретного преобразования Фурье (ДПФ),* а комплексный множитель W, заданный формулой (3.125), называется *фазовым* (или *поворотным*) *множителем (ФМ).*

Фазовый множитель W является периодической функцией своего аргумента (показателя степени), т.е.

$$W^{mN+k} \equiv W^k. \tag{3.127}$$

В самом деле, подставив значение множителя (3.125) в выражение (3.127), имеем

$$W^{mN+k} = e^{-j\frac{2\pi}{N}(mN+k)} = e^{-j\frac{2\pi}{N}n} e^{-j\frac{2\pi}{N}k}. \tag{3.128}$$

Поскольку для любого целого n

$$e^{-j\frac{2\pi}{N}n} \equiv 1,$$

то из равенства (3.128) вытекает, что

$$W^{mN+k} = e^{-j\frac{2\pi}{N}k} = W^k$$

и, следовательно, тождество (3.127) становится доказанным.

Периодичность ФМ можно достаточно просто проиллюстрировать графически (рис. 3.47) на примере восьмиточечного ($N = 8$) ДПФ. Из периодичности ФМ вытекает периодичность спектра $\dot{X}(k)$ дискретных сигналов $x(n),\ n=\overline{0, N-1},$ т.е. если

$$k = mN + k_0, \tag{3.129}$$

где $m \geq 1,$ а $k_0 < N,$ то

$$\dot{X}(k) \equiv \dot{X}(k_0). \tag{3.130}$$

Действительно, подставив значение k, заданное выражением (3.129), в формулу (3.126), получим

$$\dot{X}(k) = \sum_{n=0}^{N-1} \dot{x}(n)\, e^{-j\frac{2\pi}{N}n(mN+k_0)}. \tag{3.131}$$

Произведя элементарные преобразования в показателе степени экспоненты, имеем

$$e^{-j\frac{2\pi}{N}n(mN+k_0)} = e^{-j\frac{2\pi}{N}nk_0}\ e^{-j2\pi nm}.$$

Поскольку для любого целого m вторая экспонента в правой части последнего выражения равна единице, формула (3.131) приводит к соотношению

$$\dot{X}(k) = \sum_{n=0}^{N-1} \dot{x}(n)\, e^{-j\frac{2\pi}{N}nk_0},$$

т.е.

$$\dot{X}(k) = \dot{X}(k_0),$$

что и требовалось доказать.

Таким образом, *N-выборочной совокупности дискретных отсчетов сигнала, эквидистантно расположенных на оси времени, соответствует N-выборочная совокупность гармоник сигнала, эквидистантно размещенных на оси частот.*

Интервал между соседними гармониками (в герцах) равен

$$f_1 = F / N,$$

где $F = 1/T$ - частота дискретизации, а N - объем выборки сигнала.

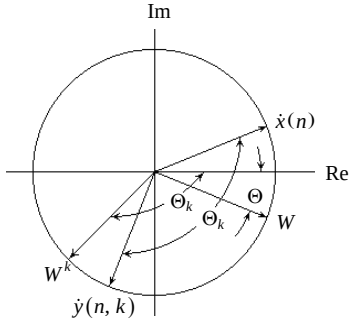


Рис. 3.48. Формирование вектора $\dot{y}(n, k)$

При умножении вектора (в общем случае комплексного) сигнала $\dot{x}(n)$ на k-й степень множителя W в соотношении (3.126) образуется вектор $\dot{y}(n,k) = \dot{x}(n)W^k$, повернутый относительно вектора $\dot{x}(n)$ по часовой стрелке на угол $\theta_k = k\theta$ (рис. 3.48). Этим объясняется название комплексного множителя W как «фазового», или «поворачивающего».

Устройство (*аппаратное или программное*), *реализующее алгоритм ДПФ, называется процессором ДПФ* (рис. 3.49).

Если на вход процессора ДПФ подавать отсчеты сигнала $x(n)$, $n=\overline{0, N-1}$, то на его выходе формируются гармоники $\dot{X}(k)$, $k=\overline{0, N-1}$, что отвечают *N*-выборочной совокупности входных сигналов.

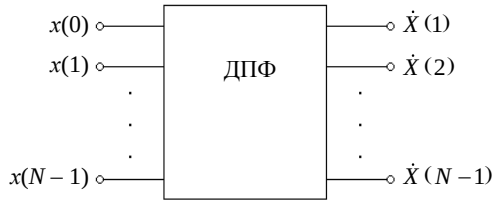


Рис. 3.49. Условное обозначение процессора ДПФ

Развернем формулу (3.119), например, для четырехточечного ДПФ. Имеем

$$\dot{X}(0)=x(0)W^0+x(1)W^0+x(2)W^0+x(3)W^0;$$

$$\dot{X}(1)=x(0)W^0+x(1)W^1+x(2)W^2+x(3)W^3;$$

$$\dot{X}(2)=x(0)W^0+x(1)W^2+x(2)W^4+x(3)W^6;$$

$$\dot{X}(3)=x(0)W^0+x(1)W^3+x(2)W^6+x(3)W^9.$$

Система линейных уравнений (3.132) дает возможность записать соотношение (3.123) в матричной форме

$$W_N\vec{x}_N=\vec{X}_N,$$

в которой W_N - квадратная матрица размерности $N\times N$ весовых коэффициентов преобразования вектора-столбца входных отсчетов сигнала

$$\vec{x}_N=\begin{pmatrix} x(0)\\ x(1)\\ \dots\\ x(N-1)\end{pmatrix}$$ в вектор-столбец $$\vec{X}_N=\begin{pmatrix} \dot{X}(0)\\ \dot{X}(1)\\ \dots\\ \dot{X}(N-1)\end{pmatrix}$$

частотных гармоник дискретного спектра. Гармоники $\dot{X}(k)$, $k=\overline{0, N-1}$ являются компонентами исходного сигнала процессора ДПФ.

Матрицу W_N называют *матрицей преобразования*.

Для $N=4$, т.е. четырехточечного ДПФ, матрица W_4 имеет вид

$$W_4 = \begin{pmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{pmatrix}.$$

а с учетом значений фазовых множителей –

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{pmatrix} \sqrt{a^2+b^2}\,. \qquad (3.132)$$

Из соотношений (3. 126) и (3. 131) следует, что при вычислении гармоник спектра $\dot{X}(k)$, $k=\overline{0,3}$, над входными отсчетами выполняются элементарные операции умножения на величины ± 1 (что тривиально) или на величины $\pm j$ с последующим суммированием результатов перемножения. Отметим, что умножение комплексной величины $\dot{x}(n)$ *на* j означает поворот вектора $x(n)$ против часовой стрелки на угол, равный π/2, тогда как умножение на $-j$ означает поворот вектора $x(n)$ по часовой стрелке на угол π/2. Матрицу (3.132) называют матрицей *преобразования с минимальными фазами.*

Обратное дискретное преобразование Фурье. Обратимся к выражению для ДПФ

$$\dot{X}(k) = \sum_{n=0}^{N-1} \dot{x}(n) W^{nk}. \qquad (3.133)$$

Умножим обе части соотношения (3.133) на W^{-mk} и просуммируем по k от 0 до $N-1$

$$\sum_{k=0}^{N-1} \dot{X}(k) W^{-mk} = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} \dot{x}(n) W^{nk} W^{-mk} = \sum_{k=0}^{N-1} \sum_{n=0}^{N-1} \dot{x}(n) W^{(n-m)k}.$$

Изменив порядок суммирования в правой части последнего выражения, получим

$$\sum_{n=0}^{N-1} \dot{x}(n) \sum_{k=0}^{N-1} W^{(n-m)k}.$$

Внутренняя сумма здесь отлична от ноля и равна N лишь при $n = m$

$$N\,\dot{x}(m) = \sum_{k=0}^{N-1} \dot{X}(k) W^{-mk}.$$

Заменим m на n. В результате такой замены получим окончательно

$$\dot{x}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \dot{X}(k) W^{-nk},\ n = \overline{0,\, N-1}. \qquad (3.134)$$

Выражение (3.134) называется *обратным дискретным преобразованием Фурье*. Его можно представить также в виде

$$\dot{x}(n) = \frac{1}{N} \left(\sum_{k=0}^{N} \dot{X}^{*}(k) W^{nk} \right)^{*}$$

т.е. *с точностью до коэффициента* $1/N$ *обратное ДПФ совпадает с комплексно-сопряженным ДПФ последовательности комплексно-сопряженных гармоник сигнала.*

Соотношения (3.133) и (3.134) образуют пару ДПФ (соответственно прямого и обратного).

Дискретизация периодических сигналов. При исследовании с помощью ЭВМ непрерывный сигнал $\dot{x}(t)$ *на интервале времени* $(0, T)$ заменяется своими отсчетными значениями $(\dot{x}_0, \dot{x}_1, ..., \dot{x}_{N-1})$, взятыми соответственно в моменты времени $(0,\ \Delta,\ 2\Delta,\ ...,\ (N-1)\Delta)$. Полное количество отсчетов

$$N = T / \Delta \qquad (3.135)$$

где Δ - шаг (период) временной дискретизации непрерывного (аналогового) сигнала.

Массив чисел $\{x_n\}$, действительных или комплексных, является той единственной информацией, из которой можно судить о спектральных свойствах сигнала $\dot{x}(t)$. Исследование таких дискретных сигналов можно существенно упростить, если полученную выборку отсчетных значений сигнала повторить бесконечное количество раз влево и вправо по оси абсцисс (рис. 3.50). В результате такой операции приходим к периодическому дискретному сигналу.

На рис. 3.50 черными кружочками обозначена исходная *N-мерная* последовательность дискретных сигналов, а незатемненными кружочками периодическое продолжение последовательности.

Подобрав для сигнала некоторую математическую модель, можно вос-

пользоваться разложением в ряд Фурье и найти соответствующие амплитудные коэффициенты. *Совокупность этих коэффициентов образует спектр дискретного периодического сигнала.*

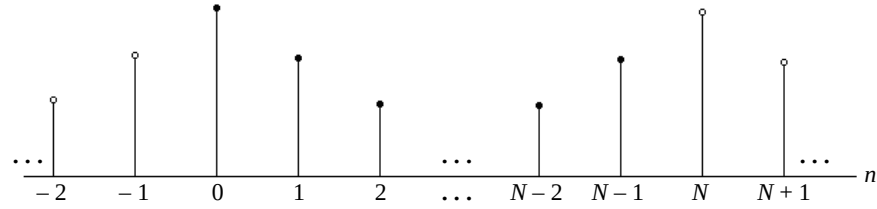


Рис. 3.50. Дискретное представление периодического сигнала

Воспользуемся моделью в виде последовательности дельта-импульсов и поставим в соответствие исходному колебанию $\dot{x}(t)$ его дискретное представление $\dot{x}_д(t)$ на интервале (0, T)

$$\dot{x}_д(t) = \sum_{n=0}^{N-1} \dot{x}_n \delta(t - n\Delta), \tag{3.136}$$

где $\dot{x}_n = \dot{x}(n\Delta)$ - отсчетные значения $\dot{x}(t)$ *в n*-й точке.

Представим дискретную модель (3.136) комплексным рядом Фурье

$$\dot{x}_д(t) = \sum_{k=-\infty}^{\infty} \dot{C}_k e^{j2\pi nt/T}$$

с коэффициентами

$$\dot{C}_k = \frac{1}{T}\int_0^T \dot{x}_д(t) e^{-j2\pi nt/T} dt. \tag{3.137}$$

Подставляя соотношение (3.136) в (3.137), с учетом формулы (3.135) получим

$$\dot{C}_k = \frac{1}{N\Delta}\int_0^{N\Delta} \sum_{n=0}^{N-1} \dot{x}_n \delta(t - n\Delta) e^{-j2\pi kt/T} dt. \tag{3.138}$$

Введем безразмерную переменную

$$\xi = t/\Delta, \qquad dt = \Delta d\xi.$$

Это означает, что необходимо поделить на период дискретизации Δ все переменные, имеющие размерность времени в секундах, а именно: аргумент δ-функции $t - k\Delta$, переменные *t и T* в показателе экспоненты, а также верхнюю границу интегрирования (3.138). В результате такой замены переменных выражение (3.138) приобретет вид

$$\dot{C}_k = \frac{1}{N}\int_0^N \sum_{n=0}^{N-1} \dot{x}_n \delta(\xi - n) e^{-j2\pi k\xi/N} d\xi,,$$

а после перестановки местами операций интеграции и суммирования –

$$\dot{C}_k = \frac{1}{N}\sum_{n=0}^{N-1} \dot{x}_n \int_0^N \delta(\xi - n) e^{-j2\pi k\xi/N} d\xi..$$

Наконец, используя фильтрующие свойства дельта-функции, имеем

$$\dot{C}_k = \frac{1}{N}\sum_{n=0}^{N-1} \dot{x}_n e^{-j\frac{2\pi}{N}nk}. \qquad (3.139)$$

Формула (3.139) определяет последовательность коэффициентов, которые образуют ДПФ рассматриваемого сигнала. С точностью до коэффициента $1/N$ она совпадает с ранее полученной формулой (3.123), но в отличие от нее - математически достаточно корректная. Из соотношения (3.139) четко вытекает, что N-выборочной совокупности отсчетов сигнала $\dot{x}_n$, взятых с интервалом временной дискретизации Δ, соответствует N-выборочная совокупность коэффициентов разложения $\dot{C}_k$, образующие дискретный спектр этого дискретного сигнала.

Если, как и ранее, обозначим через $W = e^{-j\frac{2\pi}{N}}$ фазовый множитель, то выражение (3.139) приобретет более привычный вид

$$\dot{C}_k = \frac{1}{N}\sum_{n=0}^{N-1} \dot{x}_n W^{nk}, \quad k = \overline{0, N-1} \qquad (3.140)$$

прямого ДПФ, которому соответствует *обратное* ДПФ:

$$\dot{x}_n = \sum_{k=0}^{N-1} \dot{C}_k W^{-nk}, \qquad n = \overline{0, N-1}. \qquad (3.141)$$

Взаимно дополняющие друг друга формулы (3.141) и (3.142) являются дискретными аналогами обычной пары преобразований Фурье для дискретных сигналов.

В литературе можно встретить запись пары ДПФ как в виде системы уравнений (3.141) и (3.142), так и в виде системы уравнений (3.133) и (3.134).

Свойства дискретного преобразования Фурье:

1. ДПФ является линейным преобразованием, т.е. сумме взвешенных сигналов отвечает взвешенная сумма их ДПФ:

$$\sum_{m=1}^{M} \alpha_m \left\{\dot{x}_N^{(m)}\right\} \leftrightarrow \sum_{m=1}^{M} \alpha_m \left\{\dot{C}_N^{(m)}\right\},$$

где α_m - весовые коэффициенты *N-мерной m-й* последовательности дискретных сигналов

$$\dot{x}_N^{(m)} = \dot{x}_0^{(m)}, \dot{x}_1^{(m)}, ..., \dot{x}_{N-1}^{(m)},\ m = \overline{1, M},$$

а

$$\dot{C}_N^{(m)} = \dot{C}_0^{(m)}, \dot{C}_1^{(m)}, ..., \dot{C}_{N-1}^{(m)},\ m = \overline{1, M}$$

коэффициент разложения *m*-й последовательности сигналов.

2. Число различных коэффициентов $\dot{C}_0, \dot{C}_1, ..., \dot{C}_{N-1}$, вычисляемых по формуле прямого ДПФ (3.141), равняется числу *N* отсчетов в выборке.

3. Коэффициент $\dot{C}_0$ (постоянная составляющая) является средним значением всех отсчетов:

$$C_0 = \frac{1}{N}\sum_{n=0}^{N-1}\dot{x}_n.$$

4. Если *N* – четное число, то

$$C_{N/2} = \frac{1}{N}\sum_{n=0}^{N-1}\dot{x}_n(-1)^n. \qquad (3.142)$$

Действительно, согласно формуле (4.140)

$$W^k = e^{-j\frac{2\pi}{N}k},$$

т.е. при $k = N/2$:

$$e^{-j\frac{2\pi}{N}\frac{N}{2}} = e^{-j\pi} = -1,$$

что и подтверждает соотношение (3.142).

5. Пусть отсчетные значения x_n - вещественные числа. Тогда коэффициенты ДПФ, номера которых располагаются симметрично относительно $N/2$, образуют комплексно-сопряженные пары, т.е.

6.

$$\dot{C}_{N-k} = \dot{C}_k^*. \qquad (3.143)$$

Действительно, из формулы (3.139) имеем

$$C_{N-k} = \frac{1}{N}\sum_{n=0}^{N-1}x_n e^{-\frac{2\pi}{N}(N-k)n} = \frac{1}{N}\sum_{n=0}^{N-1}x_n e^{+\frac{2\pi}{N}kn} = \dot{C}_k^*,$$

поскольку $e^{-j2\pi n} \equiv 1$ для каждого $n = \overline{0, N-1}$, что и доказывает выражение (3.143).

Поэтому можно считать, что коэффициенты $\dot{C}_{N/2+1}, ..., \dot{C}_{N-1}$ соответствуют отрицательным частотам. При изучении амплитудного спектра сигнала они не применяются, и их можно не вычислять.

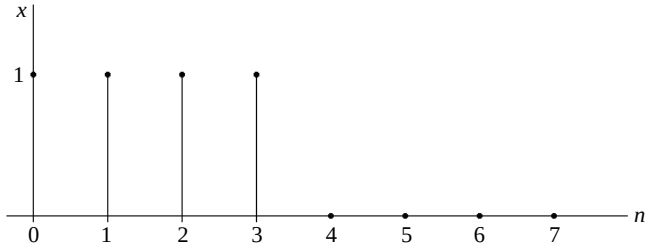


Рис. 3.51. Последовательность дискретных сигналов

Пример расчета коэффициентов ДПФ.

Пусть дискретный сигнал (рис. 3.51) на интервале своей периодичности задан восьмью равноотдаленными отсчетами:

$$\{x_n\} = (1, 1, 1, 1, 0, 0, 0, 0). \tag{3.144}$$

Найдем коэффициенты ДПФ (спектр) этого сигнала. Используя основную формулу (3.139), находим непосредственно $C_0 = 1/2.$

Коэффициенты $\dot{C}_k$ удобно вычислять графически, используя векторное изображение дискретных отсчетов сигнала на комплексной плоскости. Так, для коэффициента

$$\dot{C}_1 = \frac{1}{8}\sum_{n=0}^{3} e^{-j\frac{\pi}{4}n} \tag{3.145}$$

векторная диаграмма компонентов ряда (3.145) изображена на рис. 3.52. Сумму векторов, обозначенных цифрами $0,1,2,3,$ запишем в виде комплексного числа

$$\dot{C}_1 = (a, b),$$

где a - действительная, а b - мнимая части коэффициента $\dot{C}_1$.

Рассматривая выражение (3.145) и рис. 3.52, получаем

$$\dot{C}_1 = \frac{1}{8}\left(1-\left(1+\sqrt{2}\right)\right). \tag{3.146}$$

Для коэффициента $\dot{C}_2$ имеем

$$\dot{C}_2 = \frac{1}{8}\sum_{n=0}^{3} e^{-j\frac{\pi}{4}2n}.$$

В соответствии с диаграммой (рис. 3.53) получаем $\dot{C}_2 = 0.$

Выполняя аналогичные расчеты для других коэффициентов, находим

$$\dot{C}_3 = \frac{1}{8}\left(1, -\left(\sqrt{2}-1\right)\right), \qquad (3.147)$$

$$\dot{C}_4 = 0.$$

Следующие коэффициенты определяем на основании их сопряженности с ранее вычисленными, а именно:

$$\dot{C}_5 = \dot{C}_3^* = \frac{1}{8}\left(1, \sqrt{2}-1\right);$$

$$\dot{C}_6 = \dot{C}_2^* = 0;$$

$$\dot{C}_7 = \dot{C}_1^* = \frac{1}{8}\left(1, 1+\sqrt{2}\right).$$

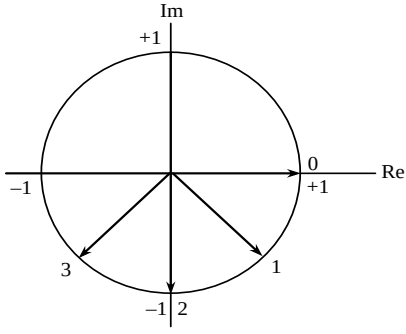


Рис. 3.52. К вычислению коэффициента $\dot{C}_1$

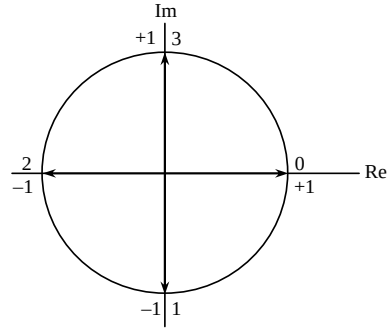


Рис. 3.53. К вычислению коэффициента $\dot{C}_2$

Следовательно, за заданными системой (3.144) значениями дискретного сигнала $\dot{x}(n)$, $n = \overline{0, N-1}$, можно найти постоянную составляющую $C_0 = 1/2$, а также первую и третью гармоники с амплитудами

$$A_1 = \frac{1}{8}\sqrt{4+2\sqrt{2}};\ A_3 = \frac{1}{8}\sqrt{4-2\sqrt{2}}.$$

Коэффициентам $\dot{C}_7$ и $\dot{C}_5$ соответствуют комплексно-спряженные амплитуды соответственно первой и третьей гармоник входного сигнала, который дает возможность записать тригонометрическую форму ряда Фурье для рассмотренного сигнала

$$\dot{x}(t) = 1/2 + 2\left(A_1 \cos\left(2\pi t/T + \varphi_1\right) + A_3 \cos\left(6\pi t/T + \varphi_3\right)\right), \qquad (3.148)$$

в котором начальные фазы, согласно общему соотношению

$$\varphi = \operatorname{arctg}\frac{b}{a},$$

на основании выражений (3.146) и (3.147) равны:

$$\varphi_1 = -\operatorname{arctg}\left(1+\sqrt{2}\right);\ \ \varphi_3 = -\operatorname{arctg}(\sqrt{2}-1).$$

На рис. 3.54 изображен сигнал $\dot{x}(t)$, возобновленный по своим коэффициентам $\dot{C}_k$ согласно формуле (3.148).

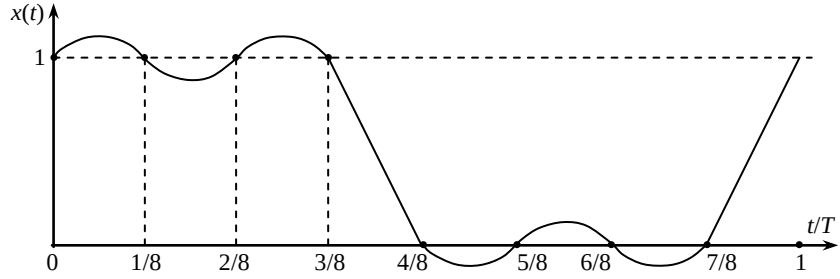


Рис. 3.54. Сигнал, восстановленный по коэффициентам ДПФ

Ряд (3.144) образован в результате дискретизации периодической последовательности прямоугольных видеоимпульсов с амплитудой, которая равняется единице, и относительной длительностью импульса q, которая равняется $q = T/\tau = 2$ (рис. 3.55).

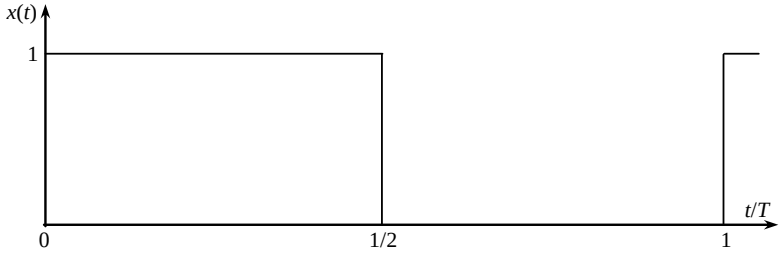


Рис. 3.55. Периодическая последовательность прямоугольных видеоимпульсов

Если вдвое уменьшить период дискретизации Δ, т.е. вдвое увеличить объем выборки *N*, то качество восстановления прямоугольного импульса тригонометрическим рядом Фурье с коэффициентами ДПФ $\dot{C}_0, \dot{C}_1, ..., \dot{C}_{15}$ возрастает, что вытекает из сравнения графиков рис. 3.54 и 3.56.

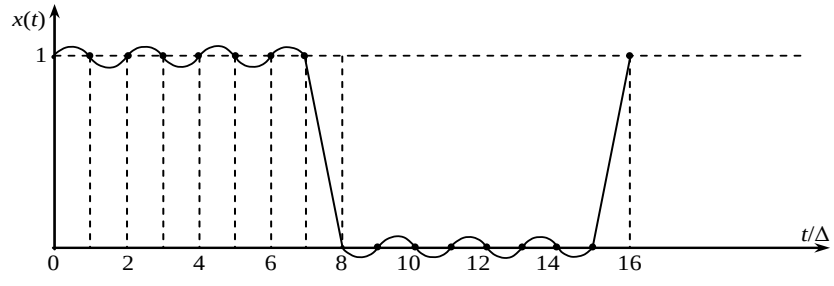


Рис. 3.56. Иллюстрация эффекта удвоения частоты дискретизации аналогового сигнала

Следует особо отметить, что восстановление непрерывного сигнала по формуле (3.148) является не приближенной, а точной операцией, вполне эквивалентной нахождению текущих значений сигнала с ограниченным спектром по его выборкам, которые образуют ряд Котельникова. Однако процедура, использующая ДПФ, в ряде случаев оказывается предпочтительней, поскольку приводит к конечным суммам гармоник, в то время как ряд Котельникова для периодического сигнала принципиально должен содержать бесконечное число членов.

Действительно, пусть наблюдается некоторая периодическая функция

$$x(t) = x(t \pm mT),$$

где T - ее период, а m = 1, 2... Ранее было показано, что располагая парным числом N отсчетов этой функции на всем ее периоде, можно вычислить ее постоянную составляющую и первые $N/2$ гармоники сигнала. Остальные гармоники содержатся в зеркальной отрицательной области частот (с соответствующими комплексно-сопряженными коэффициентами ДПФ) и не оказывают влияния на амплитудный спектр сигнала. Таким образом, при любом четном N число необходимых гармоник ДПФ составляет половину количества отсчетов. Это положение непосредственно вытекает из теоремы Котельникова. Действительно, верхнюю граничную частоту в спектре дискретного сигнала следует (в соответствии с теоремой Котельникова) находить из соотношения

$$f_{\text{в}} = 1/2\Delta,$$

где Δ - период дискретизации периодического сигнала. И поскольку $1/\Delta = F$ - частота дискретизации, которая, в свою очередь, связанная с частотой первой гармоники f_1 соотношением $F = Nf_1$, то непосредственно получим

$$f_{\text{в}} = (N/2)\, f_1,$$

причем

$$f_1 = 1/T,$$

где Т - период повторения периодического сигнала.

Таким образом, *подтверждена целесообразность применения ДПФ при аппроксимации периодических сигналов и его преимущества сравнительно с приближением периодических функций рядами Котельникова.*

Основные выводы

Для теоретического исследования сигналов необходимо построить их математические модели. Математическая модель сигнала являет собой функциональную зависимость, которая адекватно описывает изменение во времени физического состояния некоторого объекта.

Классификация сигналов выполняется на основании существенных признаков соответствующих математических моделей. Различают детерминированные и случайные, действительные и комплексные, одномерные и многомерные, статические и динамические, бесконечные во времени и импульсные, периодические и одиночные, аналоговые и дискретные сигналы. Разновидностью последних являются цифровые сигналы, т.е. сигналы, дискретные не только во времени, но и в пространстве (по величине).

К самым простым типичным элементарным сигналам принадлежат гармонические и комплексно-экспоненциальные сигналы, прямоугольные и треугольные видеоимпульсы, ступенчатые, сигнум- и дельта-функции и др.

Известное в математике понятие скалярного произведения двух функций широко используется в теории спектрального анализа. Скалярным произведением двух функций называется интеграл от произведения этих функций. Границы интеграции определяются областью существования функций.

Скалярное произведение двух сигналов называется взаимной энергией этих сигналов. Полная энергия сигнала определяется скалярным произведением сигнала самого на себя. Другими словами, полная энергия сигнала равняется интегралу от квадрата функции, которая описывает математическую модель сигнала. Корень квадратный из энергии сигнала называется нормой сигнала. Нормируемыми сигналами (функциями) называются сигналы (функции), полная энергия которых равняется единице.

Два сигнала называются ортогональным, если их скалярное произведение равняется нулю.

Бесконечная система действительных функций (сигналов) называется ортогональной, если скалярное произведение двух разных функций (сигналов) равняется нулю. При этом предусматривается, что энергия каждого сигнала из системы не равняется нулю. Это значит, что ни одна из функций, которые рассматриваются, не равняется нулю.

Бесконечную систему функций, попарно ортогональных друг другу и таких, которые имеют единичные нормы, называют системой ортонормированных функций или ортонормированным базисом.

Любой сигнал с конечной энергией можно подать в виде обобщенного ряда Фурье. Обобщенным рядом Фурье называется бесконечная сумма произведений базисных функций и соответствующих им коэффициентов. Совокупность коэффициентов разложения образует спектр сигнала.

Коэффициенты разложения обобщенного ряда Фурье (спектр) часовой

функции (сигнала) определяются скалярным произведением этого сигнала с соответствующими базисными функциями. Такой способ определения коэффициентов разложения обеспечивает минимум квадратичной ошибки аппроксимации сигнала конечномерным рядом Фурье.

Энергия сигнала равняется сумме энергий всех компонентов (гармоник), из которых состоит обобщенный ряд Фурье. Это значит, что энергия сигнала равняется сумме энергий спектральных составляющих, а квадрат модуля коэффициентов обобщенного ряда Фурье численно равняется той частице энергии сигнала, которая содержится в соответствующей составляющей (гармонике) сигнала.

Процесс добывания полезной информации, которая содержится в сигнале, можно представить как аппаратное (или программное) определение числовых значений коэффициентов обобщенного ряда Фурье этого сигнала.

Все случайные сигналы и помехи являются непредсказуемыми. Таким образом, для случайных сигналов невозможно найти математическую формулу, по которой можно было бы рассчитать их мгновенные значения.

Все случайные явления, которые изучаются в теории вероятностей, можно разделить на три типа: случайные события; случайные величины; случайные процессы. Каждый из этих типов случайных явлений имеет свои особенности и характеристики.

Случайным называют такое событие, которое в результате попытки может наступить или не наступить.

Величина, значение которой изменяется от попытки к попытке случайным образом, называется случайной. Для такой величины невозможно предусмотреть, какое значение она приобретет при конкретных условиях попытки.

Функция распределения вероятности показывает вероятность того, что случайная величина не превышает конкретного значения.

Математическое ожидание является средним значением случайной величины.

Дисперсия количественно характеризует меру разброса результатов отдельных попыток относительно среднего значения.

Случайные процессы бывают разных типов: нестационарные, стационарные, квазистационарные, эргодичные. Но в технике большинство случайных сигналов и помех принадлежат к стационарным эргодичным случайным процессам.

Флуктуационный шум наиболее характерен для большинства телекоммуникационных каналов и есть стационарным эргодичным случайным процессом с гауссовым (нормальным) распределением вероятности.

Спектральная плотность мощности флуктуационного шума зависит от физической природы его образования, а также от точки, где он наблюдается.

Основными энергетическими характеристиками действительного сигнала является его мощность и энергия.

Для случайных сигналов (помех) среднюю мощность можно рассчитать

по спектральной плотности мощности.

Динамический диапазон сигнала характеризует границы изменения мгновенной мощности.

Коэффициентом амплитуды сигнала называется отношение его максимальной мощности к средней.

Под длительностью сигнала понимают интервал времени его существования. Вычисляется длительность как разница между временем окончания сигнала и временем его начала.

По энергетическим характеристикам определяется необходимое отношение сигнал/помеха, по ширине спектра сигнала устанавливается полоса пропускания канала связи как необходимая для неискаженной передачи информации.

Теоретически любые сигналы конечной протяжности во времени имеют бесконечно широкий спектр. Реальные технические устройства имеют конечную ширину полосы пропускания. При прохождении сигналов с бесконечно широким спектром через технические устройства происходит искажение спектра, который неминуемо приводит к искажению формы входного сигнала.

Для неискаженного представления сигнала с ограниченным спектром последовательностью его дискретных отсчетов частота дискретизации F должна равняться удвоенной верхней частоте спектра сигнала.

Дискретные сигналы образуются в результате временной дискретизации непрерывных (аналоговых) сигналов. Шаг временной дискретизации выбирается, как правило, постоянным.

Основным математическим аппаратом, который используется при спектральном анализе дискретных сигналов, является аппарат *Z-преобразований*, который играет относительно дискретных сигналов такую же роль, как и интегральное преобразование Фурье для непрерывных сигналов.

Основу *Z*-преобразования составляет интегральное преобразование Лапласа. К *Z*-преобразованию приходят в результате дискретизации всех функций времени в преобразовании Лапласа.

Спектр дискретных сигналов является периодической функцией частоты с интервалом периодичности $\omega T = 2\pi$ даже в том случае, когда спектр его аналогового прототипа апериодическая функция.

Любой сигнал с ограниченным спектром без искажений можно представить последовательностью его дискретных отсчетов при условии, что частота дискретизации не менее чем вдвое превышает наивысшую гармонику спектра аналогового сигнала.

Дискретное преобразование Фурье ставит в соответствие конечномерной выборке дискретных сигналов конечномерный спектр той же размерности, что и объем выборки входных сигналов. Гармоники дискретного спектра эквидистантно размещены на интервале от нуля до частоты дискретизации.

Вопросы для самоконтроля

1. Изобразите основные пространственно-временные модели сигнала.

2. Какие из двух типов сигналов детерминированные или случайные являются носителями сообщений?

3. В какой способ гармоническое колебание можно превратить в пару квадратурных сигналов?

4. Докажите взаимосвязь ступенчатой функции и дельта-функции.

5. В чем заключается условие нормирования и фильтровальное свойство дельта-функции?

6. Запишите общую форму для скалярного произведения двух сигналов.

7. Запишите выражение для коэффициентов разложения обобщенного ряда Фурье.

8. Чему равняется сумма энергий всех гармоник сигнала?

9. Докажите оптимальность по минимуму квадратичной погрешности аппроксимации сигнала конечномерным рядом Фурье.

10. В чем заключается отличие между случайными и детерминированными сигналами?

11. Каким образом формируются математические модели случайных сигналов и помех?

12. Дайте определение случайных событий и их числовых характеристик.

13. Что такое случайные величины и их характеристики?

14. Дайте определение стационарным эргодичным процессам.

15. Запишите основные соотношения для нахождения корреляционной функции и спектральной плотности мощности.

16. Что такое белый шум?

17. Назовите основные числовые характеристики сигналов и помех.

18. Дайте определение мгновенной мощности.

19. Как найти энергетический спектр сигнала?

20. Рассчитайте среднюю мощность по спектру сигнала.

21. Как определяются уровни сигналов и помех?

22. Какие трудности возникают при определении длительности и ширины спектра сигнала или помехи?

23. Что является причиной искажения формы сигнала при его прохождении через реальные технические устройства?

24. Приведите математические модели идеальных низкочастотных и полосовых сигналов.

25. Назовите способы ортогонализации сигналов в часовой и частотной областях.

26. Как выбирается частота дискретизации сигнала с ограниченным спектром для его неискаженного представления последовательностью дис-

кретных отсчетов?

27. Запишите формулу ряда Котельникова и дайте обозначение его компонентов.

28. Каким способом генерируются отсчетные функции базиса Котельникова?

29. Какие меры нужно принять для повышения точности представления прямоугольного видеоимпульса последовательностью его дискретных отсчетов?

30. От каких факторов зависит ошибка аппроксимации сигналов ряда Котельникова?

31. Дайте определение интегрального преобразования Лапласа.

32. Назовите основные свойства преобразования Лапласа.

33. Назовите основные свойства Z-преобразования.

34. По каким причинам спектр дискретных сигналов становится периодическим даже в том случае, когда спектр аналогового прототипа апериодический?

35. В чем заключается «эффект наложения» спектров при дискретизации и какие способы его устранения?

36. В чем заключается подобие, а в чем отличие прямого и обратного ДПФ?

37. Назовите основные свойства ДПФ.

The main conclusions

For theoretical research of signals it is necessary to build their mathematical models. The mathematical model of a signal is the functional dependence that describes a change in time of a physical state of some object adequately.

Classification of signals is realized on the basis of essential characteristics of corresponding mathematical models. Determined and random, real and complex, one-dimensional and multidimensional, static and dynamic, infinite in time and impulsive, periodic and single, analogue and discrete signals are distinguished. A variety of the last are digital signals that are signals, discrete not only in time but also in space (on size).

Harmonic and complex-exponential signals, rectangular and triangular video impulses, step, signum function and delta function and others belong to the simplest typical elementary signals.

A well-known in mathematics notion of a scalar product of two functions is widely used in the theory of the spectral analysis. The scalar product of two func-

tions is the integral from product of these functions. Boundaries of integration are determined by the area of existence of functions.

The scalar product of two signals is the mutual energy of these signals. The complete energy of a signal is determined by a scalar product of a signal on itself. In other words, the complete energy of a signal is equal to integral from a square of function which describes mathematical model of a signal. The square root from energy of a signal is called a norm of a signal. Normalized signals (functions) are the signals (functions), complete energy of which is equal to one.

Two signals are called orthogonal if their scalar product is equal to zero.

Infinite system of real functions (signals) is called orthogonal, if a scalar product of two different functions (signals) is equal to zero. At the same time it is foreseen that the energy of each signal from system is not equal to zero. It means that none of the functions which are considered is equal to zero.

Infinite system of functions, orthogonal in pairs to each other and functions that have single norms is called a system of orthonormal functions or orthonormal basis.

Any signal with finite energy can be given as a generalized row of Furie. Generalized row of Furie is infinite sum of products of basic functions and corresponding coefficients. A collection of coefficients of break-up forms a spectrum of a signal.

Coefficients of break-up of generalized row of Furie (spectrum) of time function (signal) are determined by a scalar product of this signal with corresponding basic functions. Such way of definition of coefficients of break-up provides a minimum of a quadratic error of approximating of a signal by the finite-dimensional row of Furie.

The energy of a signal is equal to the sum of energies of all components (harmonics), that generalized row of Furie consists of them. It means that the energy of a signal is equal to the sum of energies of spectral components, and the square of the modul of coefficients of generalized row of Furie numerically is equal to the particle of energy of a signal which is kept in corresponding constituent (harmonic) of a signal.

Process of getting the useful information which signal contains can be given as hardware (or program) definition of numerical values of coefficients of generalized row of Furie of this signal.

All random signals and interferences are unforeseen. Thus it is impossible to find the mathematical formula for random signals that could be used to calculate their instant values.

All random phenomena that are studied in probability theory can be divided into three types: random events; random sizes; random processes. Each of these types of the random phenomena has its own features and characteristics.

Random event is any fact that can appear or fail to appear as a result of attempt.

The size value of which varies from attempt to attempt in random way is called random. It is impossible to foresee for sure for such size what value it will gain under concrete conditions of attempt.

Function of division of probability shows probability that the random size does not exceed concrete value

The mathematical waiting is average value of a random size.

The dispersion quantitatively characterizes a measure of scattering of results of separate attempts relatively to the average value.

Random processes can be of different types: non-stationary, stationary, quasistationary, ergodic. But in engineering the majority of random signals and interferences are of stationary ergodic random processes

Fluctuation noise is the most typical for the majority of telecommunication channels and it is stationary ergodic random process with gauss (normal) division of probability.

The spectral density of power of fluctuation noise depends on the physical nature of its creation and also on a point where it is observed.

The main power characteristics of the real signal is its power and energy.

The average power can be calculated for random signals (interferences) regarding a spectral density of power.

Regarding a logarithmic unit of measuring of levels such characteristic of quality as the ratio the signal-interference will be equal to a difference of levels of a signal and interference.

The dynamic range of a signal characterizes the boundaries of changes of instant power.

The coefficient of amplitude of a signal is the ratio of its maximal power to average.

The duration of a signal is a time slice of its existence. It is calculated as a difference between the time of ending of a signal and the time of its beginning.

The necessary ratio the signal-interference is determined regarding power characteristics; the bandwidth of communication channel is installed as necessary for undistorted transmission of the information regarding the width of a spectrum of a signal.

Theoretically any signals of finite extension in time have infinitely wide spectrum. Real technical devices have a finite bandwidth. While transmitting the signals with infinitely wide spectrum through technical devices there is a distortion of spectrum which inevitably leads to distortion of forms of an incoming signal.

The frequency of digitization F should be equal to the double upper frequency of a spectrum of a signal for undistorted signal injection with the limited spectrum in sequence of its discrete countings.

Discrete signals are formed as a result of time digitization of continuous (analogue) signals. As a rule, the step of time digitization is selected constant.

The main mathematical device which is used for the spectral analysis of discrete signals is the device of Z-transformations that plays in relation to discrete signals the same role as well as integral transformation of Furie for continuous signals.

The integral transformation of Laplas makes the basis of Z-transformations. As a result of digitization of all functions of time in transformation of Laplas they come to Z-transformation.

The Spectrum of discrete signals is a periodic function of frequency with an interval of periodicity $\omega T = 2\pi$ even in the case when a spectrum of its analogue prototype is aperiodic function.

Any signal with the limited spectrum without distortions can be given as a sequence of its discrete countings on condition that frequency of digitization exceeds not less than twice the highest harmonic of a spectrum of an analogue signal.

Discrete transformation of Furie puts a finite dimensional spectrum of the same dimension, as a sample size of incoming signals in accordance to finite dimensional sample of digital signals. Harmonics of a discrete spectrum are placed equidistantly on an interval from zero to sampling rate.

Ключевые слова

Русский	**Английский**
модель сигнала	model of signal
случайный	casual
сигнал	signal
помехи	hindrances
мощность	power
энергия	energy
спектральная	spectral
плотность	closeness
дискретизация	diskretisation

ОБРАБОТКА ИНФОРМАЦИИ

4

4.1. Аналоговая обработка информации
4.2. Квантование и дискретизация
4.3. Цифровая обработка информации

4.1. Аналоговая обработка информации

В развитии новейших технологий ведущее место принадлежит вопросам обеспечения достоверного приема и передачи информационных потоков по каналам связи. Объекты и процессы, связанные потоками информационных данных различной природы, в частности и электрической, подлежат преобразованию и обработке в информационных системах. В современной радиоэлектронике можно выделить три основных направления развития методов и средств обработки информации (рис. 4.1):

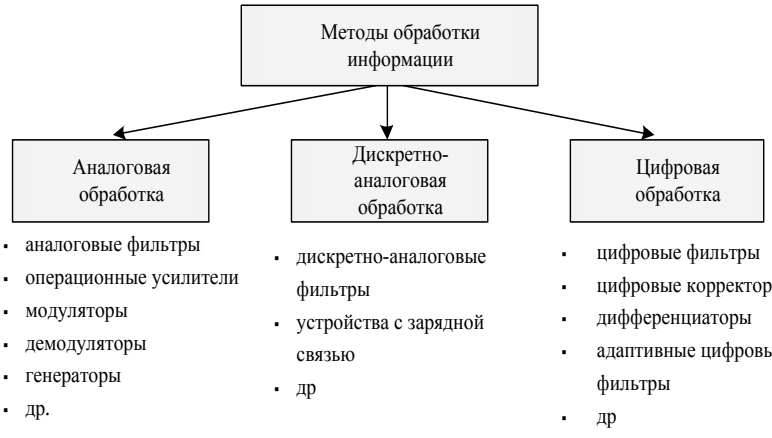


Рис. 4.1. Основные методы и средства обработки информации

аналоговая обработка сигналов с помощью аналоговых фильтров, состоящих из пассивных элементов L, C, R, операционных усилителей и т. д.;

дискретно-аналоговая обработка сигналов с помощью дискретно-аналоговых фильтров на базе вычислительных приборов с обратной связью (ПОС);

цифровая обработка дискретных сигналов с помощью цифровых фильтров на основе цифровых автоматов-вычислителей (процессоров), обрабатывающих ряды числовых последовательностей.

Например, аналоговые и дискретно-аналоговые фильтры на базе пассивных элементов и операционных усилителей имеют меньшую сложность и стоимость, больший частотный диапазон. Однако они не обеспечивают высокой точности и стабильности результатов обработки информации, поскольку имеют неконтролируемые погрешности вследствие нестабильности параметров своих элементов, обусловленной старением последних, изменением температуры, влажности, питающих напряжений, помехами, дрейфом нуля усилителей и другими факторами.

Высокая точность и стабильность результатов присущи цифровой обработке, которая делает невозможными неконтролируемые погрешности. Тем

не менее цифровые фильтры - устройства более сложные, дорогие и ограниченные по частоте - имеют ряд своих, хотя и контролируемых, погрешностей, связанных с дискретизацией и квантованием сигнала, округлением кодов отсчетов и т.д.

Физические системы и их математические модели. Системные операторы. Радиотехнические устройства независимо от своего назначения являются системой или совокупностью физических объектов, между которыми существует определенная связь.

Система рассматривается как взаимодействие некоторого оператора T (рис. 4.2) и входного влияния на систему $U_{вх}(t)$. Оператор T характеризует взаимосвязь в системе. Реакцией системы на влияние есть сигнал $U_{вых}(t)$.

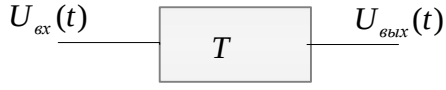


Рис. 4.2. Обобщенная схема типовой системы

Сигнал на выходе $U_{вых}(t) = TU_{вх}(t)$, где $U_{вх}(t)$ - собственная функция оператора T. Для полного описания такой системы отмечают область существования $D_{вх}$ и $D_{вых}$ соответственно для входного и выходного сигналов. Область существования описывает характер входного и выходного сигналов, а также их структуру (аналоговая, дискретная, цифровая).

Математической моделью системы *называют совокупность входного и выходного сигналов, области их существования* $D_{вх}$ и $D_{вых}$, *а также системный оператор взаимосвязи*. Классификацию системы осуществляют на основании свойств заданных математических моделей системы.

Стационарной *называется система* с *реакцией, не зависящей от момента времени, в которой происходит входное влияние.*

Если система стационарная, то $U_{вых}(t \pm t_0) = TU_{вх}(t \pm t_0)$. т.е. T не является функцией времени. Такая система называется системой с *постоянными во времени параметрами.*

Если оператор T является функцией времени $T(t)$, то такая система называется *нестационарной*, или *параметрической.*

Важным свойством для классификации систем является признак *линейности или нелинейности* характеристики системы, которая описывается оператором T (T может быть как линейной, так и нелинейной функцией).

Условия линейности:

$$T\left(U_{вх}(t) + U_{вх_2}(t)\right) = TU_{вх_1}(t) + TU_{вх_2}(t),$$

$$T\left(\alpha U_{вх}(t)\right) = T\alpha U_{вх}(t),$$

где α - произвольное число.

В случае выполнения этого условия система будет *линейной*, а в случае невыполнения - *нелинейной.* Нелинейные системы имеют элементы с нелинейной характеристикой.

Временные и частотные характеристики линейных стационарных систем. Сигнал можно представить как совокупность элементарных единичных функций (импульсов), причем продолжительность импульса стремится к нулю. Такой процесс называется динамическим *представлением сигнала* (см. главу 3):

$$\sum U(\tau)\delta(t).$$

Задача определения выходного сигнала сводится к отысканию функций, являющихся реакцией системы на внешнее влияние (сигнал), которым есть δ-функция. Поскольку совокупность δ-функций - это и есть наш сигнал, то исходное влияние определяется как сумма реакций на все δ-функции, образовывающие сигнал.

Импульсная переходная характеристика *является реакцией на внешнее влияние, которым является* δ*-функция*:

$$h(t) = T\delta(t).$$

Это выражение рассматриваем для стационарной системы. Тогда

$$h(t - t_0) = T\delta(t - t_0).$$

Такая форма записи идеализирована, поскольку реальные системы могут только приближенно создать импульс с единичной площадью и продолжительностью, стремящейся к нулю. Реальный импульс можно считать δ-функцией, если его продолжительность достаточно невелика в сравнении с собственным масштабом времени системы.

На основании рассмотрения динамической системы входной сигнал представим как

$$U_{\text{вх}}(t) = \int_{-\infty}^{\infty} U_{\text{вх}}(\tau)\delta(t - \tau)d\tau.$$

Реакция цепи на такой сигнал

$$U_{\text{вых}}(t) = TU_{\text{вх}}(t) = T\int_{-\infty}^{\infty} U_{\text{вх}}(\tau)\delta(t - \tau)d\tau.$$

Система линейная и стационарная, поэтому оператор системы T можно внести под знак интеграла:

$$U_{\text{вых}}(t) = \int_{-\infty}^{\infty} U_{\text{вх}}(\tau)T\delta(t - \tau)d\tau,$$

Согласно свойствам свертки:

$$U_{\text{вых}}(t) = \int_{-\infty}^{\infty} U_{\text{вх}}(t - \tau)\, h(\tau)\, d\tau.$$

Последние два выражения называют *интегралами Дюамеля*.

Интеграл Дюамеля *дает возможность вычислить реакцию цепи на любое внешнее влияние путем взвешенного суммирования входного сигнала. Весовыми коэффициентами для мгновенного значения сигнала являются значения импульсной характеристики.*

Условия физической реализации импульсной характеристики:

1. Исходный сигнал, соответствующий или являющийся реакцией на входное импульсное влияние, не может появиться до момента возникновения сигнала на входе:

$$h(t) = 0;\ t \le 0.$$

Из этого условия вытекают ограничения, которые накладываются на интеграл Дюамеля: граница интегрирования не $+\infty$, а определенное время T:

$$U_{\text{вих}}(t) = \int_{-\infty}^{T} U_{\text{вх}}(\tau)\, h\,(t-\tau)\, d\tau.$$

Линейная стационарная система обрабатывает поступающий на вход сигнал, выполняя операцию взвешенного суммирования для всех мгновенных значений сигнала, существовавших к началу обработки в интервале $-\infty < \tau \le T$.

2. Импульсная переходная характеристика должна быть стационарной (поскольку система стационарная), т.е. оператор системы не должен зависеть от времени.

Приведенные условия называются *устойчивостью импульсной характеристики*. Т.е. импульсная характеристика соответствует условию полного интегрирования

$$\int_{-\infty}^{\infty} \left|h(t)\right| dt < \infty.$$

Для описания частотных характеристик системы рассмотрим в общем виде линейный электрический фильтр, не содержащий независимых источников. Рассмотрим различные варианты влияния и реакции системы (рис. 4.3).

Раймонд Эдвард Алан Кристофер Пели (Raymond Edward Alan Christopher Paley, 1907—1933), английский математик. В 1930 г. был награжден призом Смита и избран членом колледжа Тринити. Получил образование в Етони. Поступил в колледж Тринити в Кембридже, где проявил себя как лучший студент. Главный вклад в науку - граф Пели (теория графов) и созданная вместе с Норбертом Винером теорема Пели - Винера (гармонический анализ). Сотрудничал с Ентони Зигмундом в области рядов Фурье (неравенство Пели - Зигмунда).

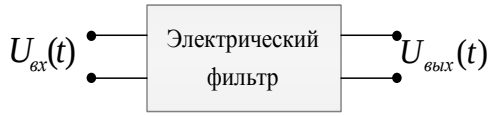


Рис. 4.3. Обобщенный вид электрического фильтра как четырехполюсника

Комплексной функцией *линейной электрической цепи, не содержащей независимых источников энергии, называется отношение комплексных изображений реакции* (*исходной величины*) *к влиянию* (*входной величине*) *в устойчивом режиме*:

$$K(j\omega)=\frac{U_{\text{вых}}(j\omega)e^{j\psi_{\text{вых}}}}{U_{\text{вх}}(j\omega)e^{j\psi_{\text{вх}}}}. \qquad (4.1)$$

Используя комплексные функции, можно оценивать свойства фильтров в частотном диапазоне, т.е. определять реакцию на любое гармоническое влияние:

$$\dot{U}_{\text{вых}}(j\omega)=K(j\omega)\dot{U}_{\text{вх}}(j\omega).$$

Амплитудно-частотная характеристика (***АЧХ***) - *зависимость модуля комплексной функции* $K(j\omega)$ *от частоты. АЧХ отображает изменение соотношения между амплитудами* (*действующими значениями*) выходного *и входного колебания при изменении частоты входного колебания*.

Фазочастотная характеристика (***ФЧХ***) *описывает зависимость аргумента* $\Phi(\omega)$ *комплексной функции от частоты. ФЧХ показывает изменение начальной фазы колебания на выходе цепи относительно начальной фазы колебания на входе при изменении частоты входного колебания*:

$$\Phi(\omega)=\psi_{\text{вых}}-\psi_{\text{вх}}. \qquad (4.2)$$

Действительная частотная характеристика *определяет зависимость действительной части* $\text{Re}[K(j\omega)]$ *комплексной функции от частоты*.

Мнимая частотная характеристика *определяет зависимость от частоты мнимой части* $\text{Im}[K(j\omega)]$ *комплексной функции*.

Комплексная функция объединяет попарно частотные характеристики. В связи с этим комплексную функцию называют *амплитудно-фазовой характеристикой*.

Изменение частоты входного сигнала сопровождается изменением длины и положением вектора $K(j\omega)$ на комплексной плоскости. Конец вектора при изменении частоты от 0 до ∞ описывает траекторию, называемую *частотным годографом*.

Частотный годограф *является амплитудно-фазовой характеристикой цепи*.

На годографе стрелкой указывается направление перемещения вектора, соответствующее увеличению частоты; могут быть обозначены точки, соответствующие характерным значениям частоты. На рис. 4.4 приведен годограф цепи.

Связь между частотными и временными характеристиками цепи. Реакцию цепи на произвольное действие можно рассчитать с помощью ее частотных и временных характеристик. В первом случае используют интеграл Фурье, во втором - интеграл свертки. Частотные и временные характеристики соответствуют разным способам представления свойств цепи: частотному (спектральному) и временному. Они зависят только от конфигурации, состава и параметров элементов цепи и имеют непосредственную связь между собой.

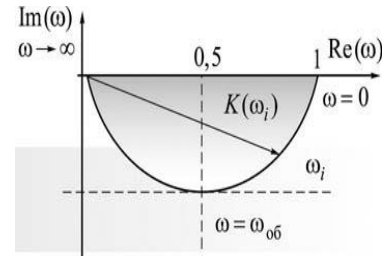


Рис. 4.4. Амплитудно-фазовая характеристика цепи (годограф)

Как известно, импульсная временная характеристика $h(t)$ численно равна реакции цепи на действие δ-функции. Если $K(j\omega)$ - комплексная функция цепи, то с помощью обратного преобразования Фурье находим

$$h(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}K(j\omega)e^{j\omega t}d\omega=F^{-1}[K(j\omega)]. \qquad (4.3)$$

Вместе с тем

$$K(j\omega)=\int_{-\infty}^{+\infty}h(t)e^{-j\omega t}dt=\int_{0}^{+\infty}h(t)e^{-j\omega t}dt=F\left[h(t)\right]. \qquad (4.4)$$

Комплексная функция цепи *равна спектральной плотности ее импульсной временной характеристики, тогда как импульсная характеристика является обратным преобразованием Фурье* (*оригиналом*) *ее комплексной функции*.

Интегрирование в формуле (4.4) осуществляется в пределах от 0 до $+\infty$, поскольку $h(t) = 0$ при $t < 0$.

Частотные и временные характеристики цепи взаимосвязаны. Изменение частотных характеристик всегда служит причиной изменения временных характеристик и наоборот.

В качестве **примера** рассмотрим пропорциональное сжатие частотных характеристик по частоте. Этому соответствует изменение масштаба частоты. Подставив в выражение (4.3) $K(jn\omega)$ вместо $K(j\omega)$, получим

$$h_1(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}K(jn\omega)e^{j\omega t}d\omega.$$

Заменив в этом выражении $n\omega$ на Ω и возвратившись снова к переменной ω, определим

$$h_1(t) = \frac{1}{2\pi n}\int_{-\infty}^{+\infty} K(j\Omega)e^{j\Omega\frac{t}{n}}d\Omega = \frac{1}{2\pi}\int_{-\infty}^{+\infty} K(j\omega)e^{j\omega t}d\omega. \qquad (4.5)$$

Сравнивая полученное выражение с формулой (4.3), определяем, что

$$h_1(t) = \frac{1}{n}h\left(\frac{t}{n}\right), \qquad (4.6)$$

т.е. сжатию частотных характеристик по оси частот соответствует растягивание в столько же раз временной характеристики вдоль оси времени и наоборот. Это полностью согласовывается с выводом о связи между реакцией цепи и шириной соответствующей полосы пропускания: чем уже полоса пропускания, тем медленнее происходят процессы.

Одним из примеров практического использования в радиотехнике частотных свойств электрических цепей, содержащих реактивные элементы, является осуществление с их помощью *частотной избирательности* (*селективности*), *или частотной фильтрации колебаний*. При этом в определенном диапазоне частот, называемом *полосой пропускания*, колебания передаются фильтром с малым затуханием; в других участках частотного диапазона колебания передаются с большим затуханием. Эти участки соответствуют *полосе задержки* (*затухания*).

Электрические фильтры широко применяются в электротехнических и радиотехнических устройствах для распределения электрических колебаний по частоте, т.е. для *частотной селекции сигналов*.

Электрическим фильтром *называют четырехполюсник, через который электрические колебания одних частот проходят с малым затуханием, а других частот - с большим затуханием.*

Полоса пропуска фильтра *- диапазон частот, на границах которого мощность колебаний на выходе цепи уменьшается в 2 раза, а напряжение и ток - в* $\sqrt{2}$ *раз сравнительно с максимальными значениями. Частоты, отвечающие границам полосы пропускания, называются предельными* (*граничными*).

Итак, любая электрическая цепь, в составе которой есть хотя бы один реактивный элемент, имеет свойство цепи, фильтрующей по частоте, *или фильтра.* Тем не менее, эффективность фильтрации различных по составу и структуре цепей далеко не одинакова и определяется их АЧХ.

В зависимости от частотного диапазона, соответствующего полосе пропускания, фильтры разделяют на четыре вида (рис. 4.5).

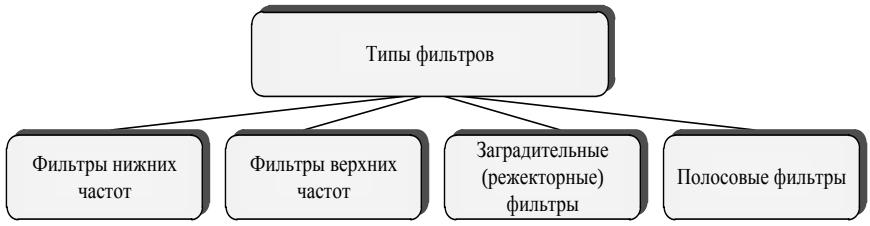


Рис. 4.5. Классификация аналоговых фильтров

Нормированные идеальные АЧХ фильтры приведены на рис. 4.6.

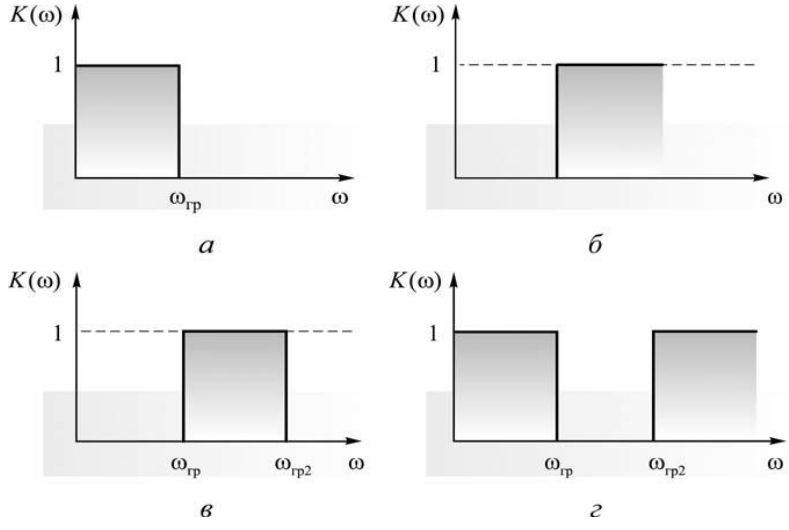


Рис. 4.6. Нормированные идеальные АЧХ фильтров

Фильтр нижних частот (***ФНЧ***) - *электрическая цепь, полоса пропускания которой лежит в диапазоне частот* $\Delta\,\omega_{п} = 0\,...\,\omega_{гр}$, *а полоса задержки* $\Delta\omega_{з} = \omega_{гр}\,...\,\infty$ (см. рис. 4.6, *а*).

Фильтр верхних частот (***ФВЧ***) - *электрическая цепь, полоса пропускания которой лежит в пределах* $\Delta\omega_{п} = \omega_{гр}...\infty$, *а полоса задержки* $\Delta\omega_{з} = 0...\omega_{гр}$ (см. рис. 4.6, *б)*.

Полосовой фильтр (***ПФ***) - *электрическая цепь, полоса пропускания которой лежит в пределах* $\Delta\omega_{п} = \omega_{гр1}...\omega_{гр2}$, *а полоса задержки* $\Delta\omega_{з} = 0...\omega_{гр1}$, $\omega_{гр2}...\infty$ (см. рис. 4.6, *в)*.

Заградительный (режекторный) фильтр (***ЗФ*** или ***РФ***) - *электрическая цепь, полоса пропускания которой лежит в пределах* $\Delta\omega_{п} = 0...\omega_{гр1}$; $\omega_{гр2}...\infty$, *а полоса задержки* $\Delta\omega_{з} = \omega_{гр1} ... \omega_{гр2}$ (см. рис. 4.6, *г)*.

Идеальные выборочные свойства *имеет цепь, АЧХ которой имеет прямоугольную форму, причем в пределах полосы пропускания коэффициент передачи должен быть максимальным, а в полосе задержки - равняться нулю* (см. рис. 4.6).

Далее будут рассмотрены частотные свойства наиболее распространенных на практике типовых фильтров.

Аналоговые фильтры первого порядка. Анализ цепей синусоидального тока показывает, что амплитуды и начальные фазы токов в ветках напряжений на элементах цепей зависят не только от схемы и параметров ее элементов, амплитуды и начальной фазы колебаний источников, действующих в цепи, а и от частоты колебаний. Т.е. характеристики процессов в цепях существенным образом зависят от частоты. Эта зависимость предопределяется частотными свойствами *реактивных элементов системы - индуктивности и емкости.*

Определяя реакции одной и той же цепи на гармонические влияния с одинаковыми амплитудой и начальной фазой, но разной частотой, затем сравнивая их, приходим к выводу: реакцию получают путем умножения влияния на некоторую конкретную для заданной цепи комплексную функцию. Это означает, что реакцию цепи на любое гармоническое влияние можно найти, если известна передающая характеристика (комплексная функция) этой цепи.

Рассмотрим особенности и существенные отличия характеристик некоторых электрических цепей.

Пример. Рассмотрим, как примеры, частотные характеристики простейших *RC*-цепей 1-го порядка (рис. 4.7 и 4.8), нашедших широкое применение в радиоэлектронике, системах связи, автоматического регулирования и т.д. В зависимости от назначения и соотношения параметров элементов, они используются как фильтры нижних и верхних частот, переходные, корректирующие, а также дифференцирующие и интегрирующие сигналы.

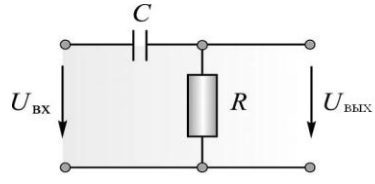


Рис. 4.7. ФВЧ на базе *RC* элементов

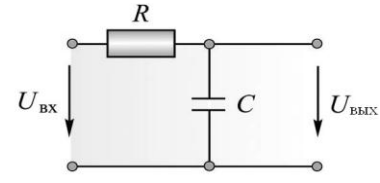


Рис. 4.8. ФНЧ на базе *RC* элементов

Комплексную передающую функцию по напряжению *RC*-схемы (см. рис. 4.7) определим как

$$K_U(j\omega) = U_{\text{вых}}(j\omega)/U_{\text{вх}}(j\omega) = R/(R+1/j\omega C) = \\ = j\omega rC/(1+j\omega rC) = j\omega\tau/(1+j\omega\tau); \\ K_U(j\omega) = \omega\tau/\left(\sqrt{1+(\omega\tau)^2}\right)e^{j(\pi/2-\text{arctg}\omega\tau)}; \\ |K(\omega)| = \omega\tau/\left(\sqrt{1+(\omega\tau)^2}\right);\ \varphi(\omega) = \pi/2 - \text{arctg}\omega\tau. \tag{4.7}$$

Здесь $\tau = RC$ - постоянная времени цепи. Для схемы на рис. 4.8

$$|K(\omega)| = = 1/\left(\sqrt{1+(\omega\tau)}^2\right);\ \varphi(\omega) = -\text{arctg}\omega\tau.$$

Соответствующие характеристики приведены на рис. 4.9 и 4.10.

Пример. Для схем, использующих индуктивность как реактивный элемент, выражения для комплексной передающей функции *RL-цепи* (рис. 4.11) определяются как

$$K_U(j\omega) = U_{\text{вых}}(j\omega)/U_{\text{вх}}(j\omega) = j\omega L/(R+j\omega L) = j\omega\tau/(1+j\omega\tau); \tag{4.8}$$

RL-цепи, приведенной на рис. 4.12, как

$$K_U(j\omega) = R/(R+j\omega L) = 1/(1+j\omega\tau), \tag{4.9}$$

Отличаются эти характеристики лишь значением постоянной времени $\tau = L/R$.

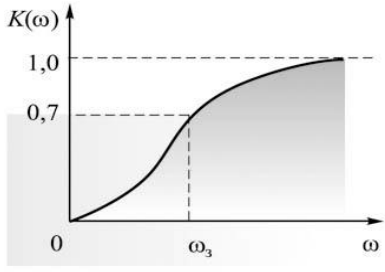


Рис. 4.9. Частотные характеристики ФВЧ

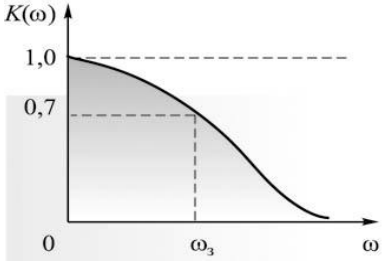


Рис. 4.10. Частотные характеристики ФНЧ

Графики частотных характеристик приведены на рис. 4.9 и 4.10. Частотные характеристики *RL-цепей* (см. рис. 4.11) будут совпадать с

частотными характеристиками соответствующих *RC*-цепей (см. рис. 4.8), а результат (4.9) совпадает с выражением (4.7). Это свидетельствует об общности полученных результатов. Однако существуют качественные отличия электрических цепей первого порядка, которые содержат емкости и индуктивности. На них остановимся ниже.

Анализ частотных свойств. 1. Для схемы, изображенной на рис. 4.7, предельная частота определяется из условия

$$\omega\tau / \left(\sqrt{1+(\omega\tau)^2} \right) = 1/\sqrt{2} \qquad (4.10)$$

и равняется $\omega_{гр} = 1/\tau = 1/R.$. Это ФВЧ (см. рис. 4.9). Полоса пропускания лежит в пределах $\Delta f = \omega_{гр} \ldots \infty.$ Форма АЧХ далека от прямоугольной (идеальной). Однако при технической реализации этого оказывается достаточно. Особенностью фильтра является то, что он не пропускает сигнал с постоянной амплитудой, т.е. через него на выход не проходит постоянный ток. Такая цепь широко используется на практике как переходная. Фильтр пропускает на выход только переменный ток. Сигнал проходит практически беспрепятственно при условии, что минимальная частота колебаний сигнала существенным образом превышает $\omega_{гр}$, т.е. $\omega >> \omega_{гр}$.

Такие фильтры применяются и для импульсных сигналов, продолжительность которых значительно превышает постоянную времени $\tau = CR.$ В случае влияния такого импульсного возрастающего сигнала на *RC-цепь* часть импульса проходит на выход цепи, поскольку происходит заряд емкости. После заряда емкости прохождение тока через резистор должно прекратиться. Дальше происходит разрядка емкости через резистор, источник входного напряжения и нагрузку. Постоянная составляющая такого сигнала будет существенным образом ослабляться и не будет проходить на выход фильтра. Импульс приобретает заостренную вершину. В этом случае выполняется условие $\omega << 1.$

Электрическая цепь имеет комплексный характер частотной характеристики

$$K_U(j\omega) = j\omega\tau / (1 + j\omega\tau)_{\omega\tau<<0} = j\omega\tau.$$

Исходное напряжение оказывается с большой степенью точности пропорциональным производной от входного сигнала:

$$U_{вых}(t) = U_{вх}(t)\, j\omega\tau, \text{ или } U_{вых}(t) = \tau dU_{вх}(t)/dt.$$

Дифференцирующим *называют электрический фильтр, если напряжение на выходе пропорционально производной входного сигнала.*

2. Для фильтра (см. рис. 4.8) предельная частота определится из условия

$$1/\left(\sqrt{1+(\omega\tau)^2} \right) = 1/\sqrt{2}.$$

Ее значение $\omega_{гр}=1/\tau=1/CR$. В этом случае образовывается ФНЧ (см. рис. 4.10). Полоса пропускания лежит в пределах $\Delta f=0...\omega_{гр}$. Форма АЧХ не прямоугольная. Особенность этого фильтра в том, что при условии $\omega\tau>>1$ комплексный коэффициент передачи напряжения близок к величине

$$K_U(j\omega)=1/(1+j\omega\tau)_{/\omega\tau>>1}=1/j\omega\tau.$$

Тогда исходное напряжение оказывается пропорциональным интегралу от $U_{вх}(t)$:

$$U_{вых}(t)=U_{вх}(t)\frac{1}{j\omega\tau},$$

т.е.

$$U_{вых}(t)=\frac{1}{\tau}\int U_{вх}(t)\,dt.$$

Полученный результат свидетельствует о том, что при условиях $\omega\tau>>1$, $\omega>>\omega_{гр}$ такая электрическая цепь может выполнять роль интегрирующего элемента, т.е. накапливать сигнал.

Интегрирующим *называют электрический фильтр, если напряжение на выходе пропорционально интегралу входного сигнала*.

3. Фильтр, схема которого изображена на рис. 4.11, имеет аналогичные параметры. Для этой электрической цепи предельная частота определяется в соответствии с уже известным условием (4.10). Предельная частота такого ФВЧ равна $\omega_{гр}=1/\tau=L/R$.

Фильтр на схеме рис. 4.12 имеет такую же предельную частоту, но является ФНЧ.

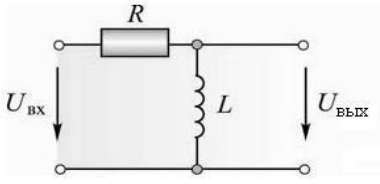


Рис. 4.11. ФВЧ на базе *RL* элементов

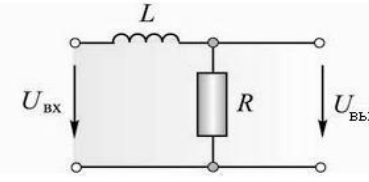


Рис. 4.12. ФНЧ на базе *RL* элементов

4. Более распространены на практике фильтры, построенные по схеме рис. 4.12. Они применяются как фильтрующие элементы в цепях питания различных усилителей, чаще всего вместе с *RC*-фильтрами, которые имеют большую постоянную времени.

Эдвин Хуард Армстронг (Edwin Howard Armstrong, 1890 - 1954), американский радиотехник и радиоизобретатель. Разработал схемы регенеративных (1913), супергетеродинных (1918) и суперрегенеративных (1921) радиоприемников. Показал преимущества частотной модуляции в борьбе с радиопомехами. Разработал до 1933 г. систему вещания в FM-диапазоне. Стал еще в 1920-х годах миллионером, вкладывал все средства в разработку своего детища, а в 1939 г. построил первую FM-радиостанцию.

Фильтры верхних частот, полосовые и заградительные. Характеристические параметры ФВЧ, СФ и ЗФ можно найти методом прямых расчетов, имея комплексную функцию фильтра. Воспользовавшись методом преобразования частоты, можно найти эти параметры по известным параметрам ФНЧ, а расчет свести к расчету ФНЧ, который в этом случае называют низкочастотным *аналогом, или прототипом*.

В общем виде суть метода преобразования частоты сводится к замене операторной частоты p некоторой функцией этой частоты $F(p)$, выбираемой так, чтобы по известным характеристиками ФНЧ получить характеристики фильтров других типов.

1. *Изменение частоты среза ФНЧ.* Изменение частоты среза ФНЧ-прототипа сводится к простому масштабированию частотной оси и выполняется путем функциональной замены p в выражении для функции передачи в операторном виде $p = p'/\omega_з$, где $\omega_з$ - необходимая частота среза ФНЧ.

2. *Преобразование ФНЧ в ФВЧ.* Для преобразования ФНЧ-прототипа в ФВЧ необходима инверсия частотной оси, которая выполняется путем замены оператора p в выражении для функции передачи $p = p'/\omega_з$, где $\omega_з$ - необходимая частота среза ФВЧ; p' - новая операторная частота, предназначенная для синтеза ФВЧ.

3. *Преобразование ФНЧ в СФ.* Для преобразования ФНЧ-прототипа в СФ необходимо более сложная трансформация частотной оси, чем в предыдущих случаях. Так, нулевая и бесконечная частоты должны преобразовываться в бесконечное значение на частотной оси ФНЧ-прототипа (там, где его коэффициент передачи стремится к нулю). Частоты, соответствующие краям необходимой полосы пропускания, должны после преобразования давать значение ± 1, которые равны частоте среза ФНЧ-прототипа. В конце концов, преобразования должны выполняться с помощью дробно-рациональной функции, с целью сохранения дробно-рациональной структуры функции передачи.

Указанным требованиям удовлетворяет такая

замена операторной частоты *p*:

$$p \to Q\frac{\left(p'/\omega_з\right)^2+1}{p'/\omega_з}, \qquad (4.11)$$

где $\omega_з = \sqrt{\omega_1\omega_2}$, $Q = \omega_з/\left(\omega_2-\omega_1\right)$, ω_1 и ω_2 соответственно нижняя и верхняя граница полосы пропускания фильтра.

4. *Преобразование ФНЧ в РФ.* Для преобразования ФНЧ-прототипа в РФ трансформация частотной оси должна быть обратной относительно предыдущего случая. Нулевая и бесконечная частоты должны преобразовываться в нулевое значение на частотной оси ФНЧ-прототипа (там, где коэффициент передачи большой). Частоты, соответствующие краям необходимой полосы задержки, должны после преобразования иметь значения ±1, которые равны частоте среза ФНЧ-прототипа. Кроме этого, некоторое значение частоты в полосе задержки должно преобразовываться в бесконечность (там, где коэффициент передачи ФНЧ-прототипа стремится к нулю). В конце концов, преобразование должно выполняться с помощью дробно-рациональной функции, чтобы сохранить дробно-рациональную структуру функции передачи. Этим требованиям удовлетворяет такая замена *p:*

$$p \to \frac{p'/\omega}{q\left(\left(p'/\omega\right)^2+1\right)},$$

где $\omega_з = \sqrt{\omega_1\omega_2}$, $Q = \omega_з/(\omega_2-\omega_1)$, ω_1 и ω_2 - соответственно нижняя и верхняя границы полосы задержки фильтра.

Фильтры с максимально плоской аппроксимацией АЧХ. ***ФНЧ Баттерворта.*** Идеальная АЧХ фильтра физически не реализуется (теорема Пели - Хинчина) согласно физическим свойствам элементов схемы. Как один из возможных видов аппроксимации характеристики идеального коэффициента передачи $K\left(j\omega\right)$ предложена функция Баттерворта

$$K\left(\omega_н\right) = \frac{1}{1+\omega_н^{2n}}, \qquad (4.12)$$

где $\omega_н = \omega/\omega_з$ - нормированная частота.

Фильтры, использующие такой коэффициент передачи, называются фильтрами с максимально плоской вершиной, или фильтрами Баттерворта.

Число $n = 1, 2, 3, 4, ...$ является порядком фильтра. В полосе пропускания фильтра, где $0 \le \omega_н \le 1$, характеристика должна плавно уменьшаться и на частоте среза ослабления фильтром представлять $10\lg 0{,}5 \approx -3$дБ независимо от порядка системы. Однако видим, что с увеличением порядка фильтра, коэффициент передачи больше приближается к идеальной характеристике (рис. 4.13).

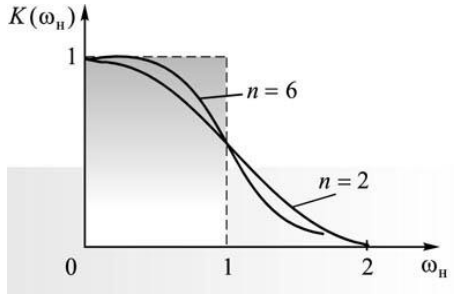


Рис. 4.13. АЧХ фильтров Баттерворта

Для дальнейшего синтеза фильтра необходимо перейти от коэффициента передачи $K(\omega_н)$ к коэффициенту передачи системы в операторном виде

$$K(p_н) = \frac{1}{1+(-1)^n p_н^{2n}}. \qquad (4.13)$$

Отсюда вытекает, что изображенные на комплексной плоскости полюса $p_н = \sigma_н + j\omega_н$ функции $K(p_н)$ соответствуют характеристике Баттерворта *n*-го порядка. Полюса функции являются корнями уравнения $1+(-1)^n p_н^{2n} = 0$ (рис. 4.14).

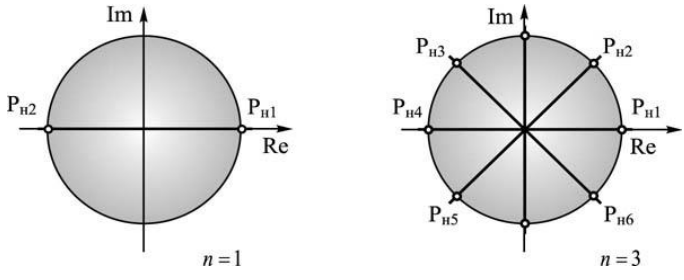


Рис. 4.14. Схематическое изображение комплексной плоскости для размещения корней характеристики фильтра Баттерворта (n = 1, n = 3)

Пример. Рассчитаем аналоговый фильтр Баттерворта 2-го порядка $(n=2)$. Тогда операторный коэффициент передачи будет

$$K(p_н) = \frac{1}{1+(-1)^2 p_н^4}.$$

Важная особенность: если порядок фильтра - нечетное число, то первый корень $p_{н1} = 1$, т.е. показатель степени при e в решении уравнения равен ну-

лю, а если порядок фильтра - число четное, то первый корень начинается с $p_{н1} = e^{j\frac{\pi}{2n}}$:

$$\begin{cases} n \text{ — нечетное}, & p_{н1} = e^{j\frac{k\pi}{2n}}\ k \to 0; \\ n \text{ — четное}, & p_{н1} = e^{j\frac{k\pi}{2n}}\ k \to 1. \end{cases}$$

Найдем корни уравнения, которые входят в знаменатель, и рассмотрим их на комплексной плоскости (см. рис. 4.14).

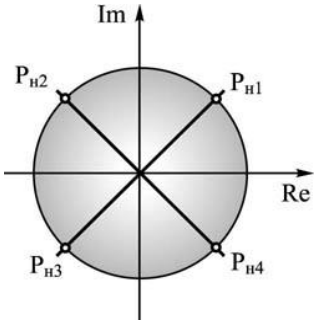


Рис. 4.15. Полюса ПХ фильтра второго порядка

Приравняем знаменатель коэффициента передачи нулю и найдем корни уравнения $1 + p_н^4 = 0, p_н^4 = -1$. Корни уравнения фильтра Баттерворта 2-го порядка

$$p_{н1} = e^{j\frac{1\pi}{2\cdot 2}} = \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}};$$

$$p_{н2} = e^{j\frac{3\pi}{2\cdot 2}} = -\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}};$$

$$p_{н3} = e^{j\frac{5\pi}{2\cdot 2}} = -\frac{1}{\sqrt{2}} - j\frac{1}{\sqrt{2}};$$

$$p_{н4} = e^{j\frac{7\pi}{2\cdot 2}} = \frac{1}{\sqrt{2}} - j\frac{1}{\sqrt{2}}.$$

Рассматривая комплексную плоскость, видим, что полюса $p_{н1}$—$p_{н4}$ размещаются симметрично на комплексном круге (рис. 4.15). Тогда коэффициент передачи можно записать в канонической форме, которая включает все полюса:

$$K(p_н) = \frac{1}{(p_н - p_{н1})(p_н - p_{н2})(p_н - p_{н3})(p_н - p_{н4})}.$$

Тем не менее, ***указанное решение нецелесообразно***, поскольку согласно теореме Пели - Хинчина этот коэффициент передачи физически не реализуем. *Для физической реализации ФНЧ берутся полюса левой полуплоскости.* Таким образом, для ФНЧ решением уравнения будут корни $p_{н2}$ и $p_{н3}$:

$$p_{н2} = -\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}},\ p_{н3} = -\frac{1}{\sqrt{2}} - j\frac{1}{\sqrt{2}}.$$

Найдем операторный коэффициент передачи $K(p_н)$ физически реализуемого фильтра Баттерворта 2-го порядка:

$$K\left(p_{\text{н}}\right)=\frac{1}{\left(p_{\text{н}}-p_2\right)\left(p_{\text{н}}-p_{\text{н}3}\right)}=$$

$$=\frac{1}{\left(p_{\text{н}}-\left(-\frac{1}{\sqrt{2}}+j\frac{1}{\sqrt{2}}\right)\right)\left(p_{\text{н}}-\left(-\frac{1}{\sqrt{2}}-j\frac{1}{\sqrt{2}}\right)\right)}=\ldots=\frac{1}{p_{\text{н}}^2+\sqrt{2}p_{\text{н}}+1}. \quad (4.14)$$

Перейдем от нормированной лапласовской частоты к ненормированной. При этом полюса нужно помножить на частоту среза $\omega_{\text{з}}$:

$$p_2=p_{\text{н}4}\omega_{\text{з}}=\left(-\frac{1}{\sqrt{2}}+j\frac{1}{\sqrt{2}}\right)\omega_{\text{з}},$$

$$p_3=p_{\text{н}3}\omega_{\text{з}}=\left(-\frac{1}{\sqrt{2}}-j\frac{1}{\sqrt{2}}\right)\omega_{\text{з}}.$$

Коэффициент передачи ФНЧ Баттерворта 2-го порядка в операторной форме имеет вид:

$$K(p)=\frac{1}{\left(\frac{p}{\omega_{\text{з}}}\right)^2+\sqrt{2}\left(\frac{p}{\omega_{\text{з}}}\right)+1}=\ldots=\frac{\omega_{\text{з}}^2}{p^2+\sqrt{2}\omega_{\text{з}}p+\omega_{\text{з}}}. \quad (4.15)$$

Для построения АЧХ перейдем от операторного коэффициента передачи к частотному $K\left(j\omega\right)$:

$$K\left(j\omega\right)=\frac{\omega_{\text{з}}{}^2}{\left(j\omega\right)^2+\sqrt{2}\omega_{\text{з}}\left(j\omega\right)+\omega_{\text{з}}^2}=\frac{\omega_{\text{з}}{}^2}{j\omega\sqrt{2}\omega_{\text{з}}-\omega^2+\omega_{\text{з}}^2}=$$

$$=\frac{\omega_{\text{з}}^2}{\left(\omega_{\text{з}}^2-\omega^2\right)+j\left(\sqrt{2}\omega\omega_{\text{з}}\right)}=\underbrace{\frac{\omega_{\text{з}}^2}{\sqrt{\text{Re}^2\left(K\left(j\omega\right)\right)+\text{Im}^2\left(K\left(j\omega\right)\right)}}}_{\left|K\left(j\omega\right)\right|-\text{АЧХ}}e^{\overbrace{-j\text{arctg}\frac{\text{Im}}{\text{Re}}}^{\text{ФЧХ}}}. \quad (4.16)$$

Действительная часть коэффициента передачи: $\text{Re}\left[K\left(j\omega\right)\right]=\omega_{\text{з}}^2-\omega^2$. Мнимая часть коэффициента передачи: $\text{Im}\left[K\left(j\omega\right)\right]=\sqrt{2}\omega\omega_{\text{з}}$.

АЧХ аналогового ФНЧ Баттерворта представлена на рис. 4.16.

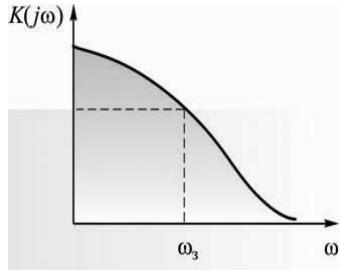


Рис. 4.16. АЧХ аналогового ФНЧ Баттерворта

Рассчитав коэффициенты передачи $K(p)$ и $K(j\omega)$ аналогового ФНЧ-прототипа Баттерворта, необходимо найти принципиальную аналоговую схему, которая могла бы реализовать полученный коэффициент передачи согласно заданной АЧХ.

Конечный этап синтеза фильтра состоит в нахождении принципиальной схемы устройства с определенным порядком.

Коэффициенты передачи K_1, K_2, ..., K_N должны быть такими, чтобы они могли реализовывать те полюса функции $K(p)$, которые были определены ранее на этапе аппроксимации частотной характеристики. Для создания фильтров необходимы звенья двух видов - звено 1-го порядка с единым действительным полюсом и звено 2-го порядка, имеющее пары комплексно-сопряженных полюсов.

Повышение порядка фильтра можно достичь путем структурного синтеза, при котором цепь образовывается каскадным включением некоторого количества звеньев, отделенных одно от другого идеальными элементами согласования (рис. 4.17). Частотный коэффициент передачи такой цепи: $K(j\omega) = K_1(j\omega)K_2(j\omega)...K_N(j\omega)$.

Этот метод дает возможность повысить порядок фильтра, т.е. качественно улучшить АЧХ системы (*приблизить характеристики к идеальным*), *используя в системе фильтры не более чем 2-го порядка.*

Давид Гилберт (David Hilbert, 1862 - 1943), немецкий математик-универсал, иностранный член-корреспондент РАН (1922) и иностранный почетный член АН СССР (1934). Закончил Кенигсбергский университет, в течение 1893 - 1895гг. - профессор этого университета, а со временем (1895 - 1930) - Геттингенского университета. Его работы существенно повлияли на развитие многих разделов математики, в которых он работал (теория инвариантов, теория алгебраических чисел, основы математики, математическая логика, вариационное исчисление, дифференциальные и интегральные уравнения, теория чисел, математическая физика).

Поэтому в дальнейшем будут рассматриваться примеры применения методики синтеза аналоговых фильтров 2-го порядка.

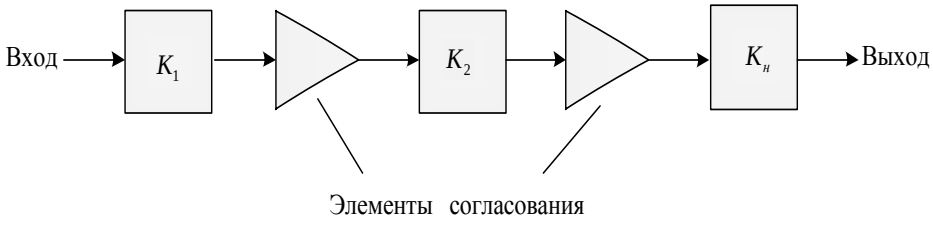


Рис. 4.17. Каскадная схема соединения фильтров

Найдем коэффициент передачи цепи 2-го порядка (рис. 4.18) в операторном виде

$$K(p)=\frac{Z_{\text{вых}}(p)}{Z_{\text{вх}}(p)}=\frac{\dfrac{R/pC}{R+1/pC}}{\left(pL+\dfrac{R/pC}{1/pC}\right)}=\ldots\frac{\dfrac{1}{LC}}{p^2+2\dfrac{1}{2RC}p+\dfrac{1}{LC}}=\frac{\omega_0^2}{p^2+2\delta p+\omega_0^2}.$$

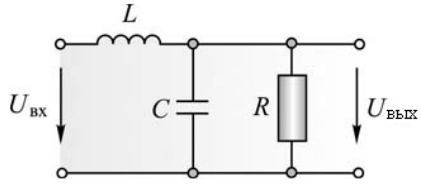


Рис. 4.18. ФНЧ второго порядка

Затем необходимо поставить этому коэффициенту передачи в соответствие операторный коэффициент передачи Баттерворта 2-го порядка:

$$K(p)=\frac{\omega_0^2}{p^2+2\delta p+\omega_0^2}\Leftrightarrow\frac{\omega_з^2}{p^2+\sqrt{2}\omega_з p+\omega_з^2}=K_{\text{Бат}}(p).$$

Чтобы заставить работать *RLC*-цепь как ФНЧ с характеристикой фильтра Баттерворта 2-го порядка необходимо выполнение таких соотношений (сопротивление R и частота среза $\omega_з$ задаются из начальных условий расчета фильтра):

$$\omega_0^2=\frac{1}{LC}=\omega_з^2\Rightarrow L=\frac{1}{C\omega_з^2},\quad\frac{2}{2RC}=\sqrt{2}\omega_з\Rightarrow C=\frac{1}{\sqrt{2}R\omega_з}.$$

Аналоговый ФВЧ Баттерворта. ФВЧ предназначен для пропускания с минимальным ослаблением колебаний, частоты которых превышают частоту среза. Колебания с частотами, содержащимися в пределах от 0 до $\omega_з$, должны заглушаться максимально.

Синтез ФВЧ получают путем предыдущего расчета ФНЧ-прототипа с той же частотой среза. При этом используют метод преобразования частоты $p = \omega_з^2 p'$. Место, соответствующее $p = 0$, будет в точке, отдаленной в бесконечность по оси частот; p' - новая операторная частота, предназначенная для синтеза ФВЧ:

$$K_{ВЧ}(p') = \frac{\omega_з^2}{\left(\frac{\omega_з^2}{p'}\right)^2 + \sqrt{2}\omega_з \frac{\omega_з^2}{p'} + \omega_з^2} \qquad (4.17)$$

Операторный коэффициент передачи получают также прямым путем - введением в операторный коэффициент передачи полюсов ФВЧ, содержащихся на правой плоскости комплексного круга, т.е.: $p_{н1}, p_{н4}$. Согласно теореме Пели - Хинчина для реализации ФВЧ используются полюса правой полуплоскости, и ФНЧ становится неаналитической функцией.

Полюса ФВЧ (нормированные):

$$p_{н1} = 1/\sqrt{2} + i/\sqrt{2}, \quad p_{н4} = 1/\sqrt{2} - i/\sqrt{2}.$$

Полюса ФВЧ (ненормированные):

$$p_1 = p_{н1}\omega_з, \; p_4 = p_{н4}\omega_з.$$

Тогда коэффициент передачи ФВЧ Баттерворта в операторном виде:

$$K_{ВЧ}(p) = \frac{1}{(p - p_1)(p - p_4)} =$$

$$= \frac{1}{\left(p - \left(1/\sqrt{2} + i1/\sqrt{2}\right)\omega_з\right)\left(p - \left(1/\sqrt{2} - i1/\sqrt{2}\right)\omega_з\right)} = \ldots \frac{p^2}{p^2 + 2\delta\omega_з p + \omega_з{}^2}.$$

На рис. 4.19 изображена АЧХ ФВЧ Баттерворта.

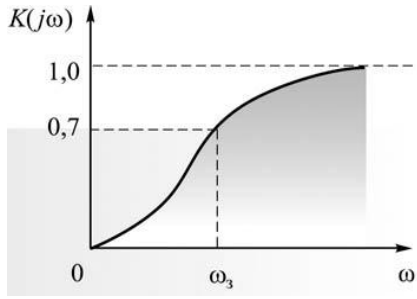


Рис. 4.19. АЧХ ФВЧ Баттерворта

Схема ФВЧ рассчитывается на основе схемы ФНЧ-прототипа путем замены емкости на индуктивность, а индуктивности - на емкость (рис. 4.20). При этом замена элементов эквивалентна выражениям

$$C_{\text{в}} = \frac{1}{\omega_{\text{з}}^2 L_H}; L_{\text{в}} = \frac{1}{\omega_{\text{з}}^2 C_H}.$$

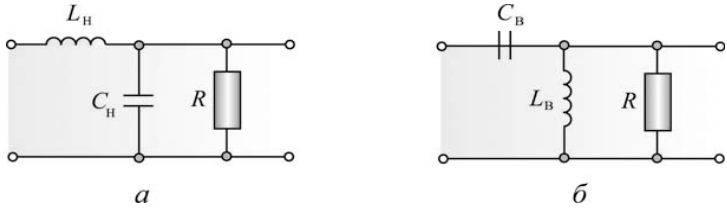


Рис. 4.20. Принципиальные схемы реализации фильтров Баттерворта (*а* - ФНЧ, *б* - ФВЧ)

Синтез аналоговых фильтров Чебышева. Широко применяется аппроксимация идеальной АЧХ фильтров полиномом Чебышева. Коэффициент передачи такого фильтра

$$K\left(p_{\text{н}}\right) = \frac{1}{1 + E^2 T_n^2\left(\omega_{\text{н}}\right)}, \qquad (4.18)$$

где $E \le 1$, называемый коэффициентом нелинейности характеристики Чебышева в полосе пропускания; $T_n\left(\omega_{\text{н}}\right)$ - многочлен Чебышева: $T_n\left(x\right) = \cos\left(n \operatorname{arcch} x\right)$. Функцию находят из соотношения $T_n\left(x\right) = 2xT_{n-1}\left(x\right) - T_{n-2}\left(x\right)$, причем $T_0\left(x\right) = 1$, а $T_1\left(x\right) = x$.

Основные преимущества многочлена Чебышева:

1. Все многочлены *n*-й степени с одинаковым коэффициентом при старшей степени аргумента почти не отклоняются от нуля на интервале $x \notin [-1,1]$.

2. При $|x| >> 1$ абсолютные значения многочлена Чебышева довольно велики. При этом $T_n(x) \approx 2^{n-1} x^n$. В пределах полосы пропускания величина $K(p)$ колеблется в границах $\left(1, \frac{1}{1+E^2}\right)$.

Если $\omega_н >> 1$, то фильтр будет обеспечивать довольно значительное ослабление сигнала.

Из графиков (рис. 4.21) видим, что в полосе пропускания характеристика фильтра немонотонная. Амплитуда пульсации прямо пропорциональна коэффициенту нелинейности E. Увеличение E ведет к ощутимому ослаблению сигнала вне полосы пропускания, и тем самым формируется вертикальный фронт характеристики и заданная частота среза $\omega_з$. Качество фильтра устанавливается оптимальным отбором двух параметров E и *n* (где *n* - порядок фильтра).

Рассмотрим передаточную функцию $K(p)$ ФНЧ Чебышева. Как видим из аппроксимирующего коэффициента передачи, полюса функций можно найти из уравнения

$$1 + E^2 T_n^2(\omega_н) = 0. \qquad (4.19)$$

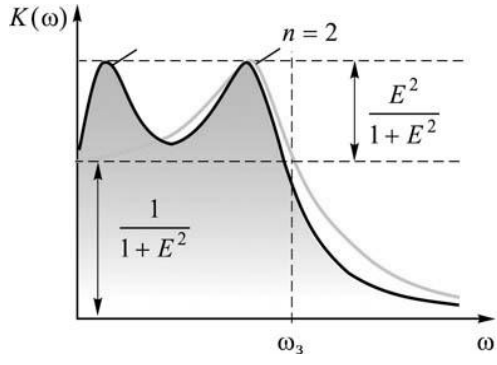


Рис. 4.21. АЧХ фильтров Чебышева

Метод решения этого уравнения довольно громоздкий, поэтому для расчета полюсов ФНЧ Чебышева используют полюсы ФНЧ Баттерворта.

Этапы расчета полюсов фильтра Чебышева.

1. *Рассчитывается параметр* a:

$$a = \frac{1}{n}\operatorname{arcsh}\left(\frac{1}{E}\right) = \frac{1}{n}\ln\left(\frac{1}{E} + \sqrt{\frac{1}{E^2}+1}\right), \qquad (4.20)$$

где n - порядок фильтра; E - коэффициент нелинейности.

2. *Находим полюсы передаточной характеристики фильтра Баттерворта того же порядка.* Для перехода от фильтра Баттерворта к фильтру Чебышева необходимо абсциссу каждого полюса функции Баттерворта умножить на $\operatorname{sh}(a)$, а ординату - на $\operatorname{ch}(a)$. Полюсы фильтра Чебышева лежат не на комплексном круге, а на комплексном эллипсе.

Для ФНЧ 2-го порядка при $n = 2,\ E = 1$ приведем графические изображения полюсов на комплексной плоскости (рис. 4.22):

$$a = \frac{1}{2}\ln\left(\frac{1}{1} + \sqrt{\frac{1}{1^2}+1}\right) = 0,407\ .$$

Полюсы ФНЧ Баттерворта	$\begin{cases} p_{н2} = -\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}} \\ p_{н3} = -\frac{1}{\sqrt{2}} - j\frac{1}{\sqrt{2}} \end{cases}$	Полюсы ФНЧ Чебышева	$\begin{cases} p_{н2} = -\frac{1}{\sqrt{2}}\operatorname{sh}a + j\frac{1}{\sqrt{2}}\operatorname{ch}a \\ p_{н3} = -\frac{1}{\sqrt{2}}\operatorname{sh}a - j\frac{1}{\sqrt{2}}\operatorname{ch}a \end{cases}$

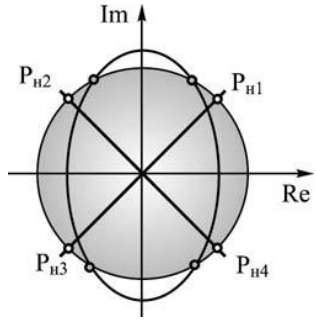


Рис. 4.22. Полюсы операторного коэффициента передачи фильтра Чебышева

Перейдем к ненормированным корням уравнения:

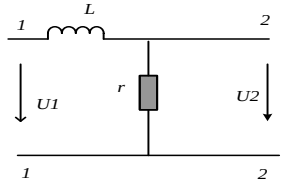
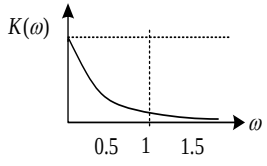
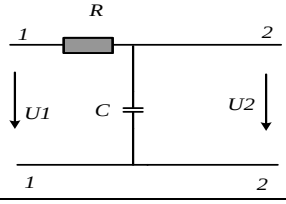
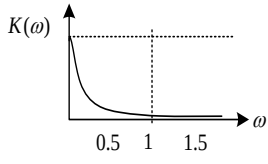
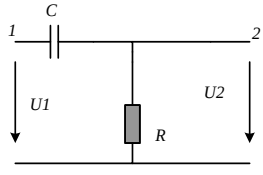
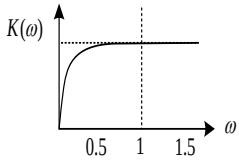
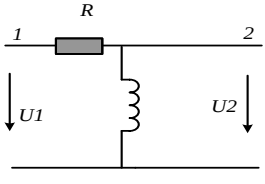
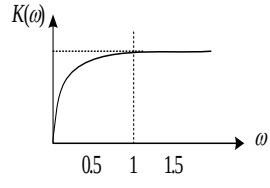
$$p_2 = \left(-\frac{1}{\sqrt{2}}\operatorname{ch}a + j\frac{1}{\sqrt{2}}\operatorname{ch}a\right)\omega_з, \quad p_3 = \left(-\frac{1}{\sqrt{2}}\operatorname{sh}a - j\frac{1}{\sqrt{2}}\operatorname{ch}a\right)\omega_з.$$

Тогда коэффициент передачи ФНЧ Чебышева

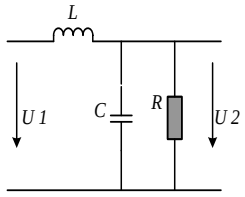
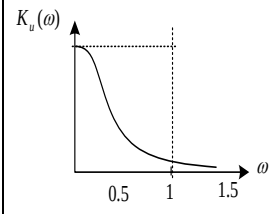
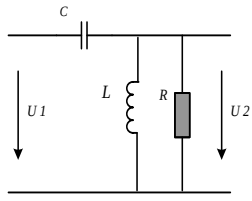
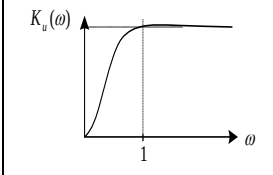
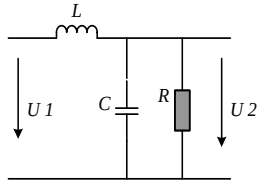
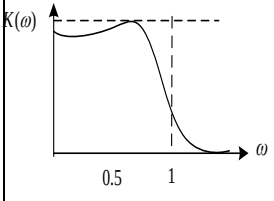
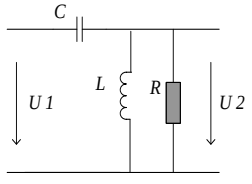
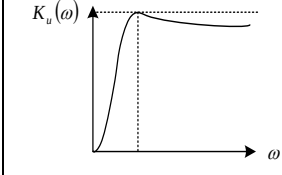
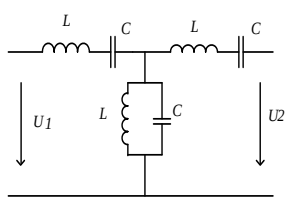
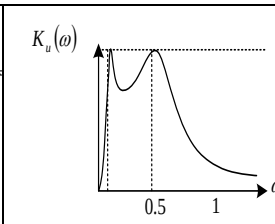
$$K(p) = \frac{1}{(p-p_2)(p-p_3)} = \ldots = \frac{\omega_з}{p^2 + \delta\operatorname{sh}ap + \omega_з^2}.$$

Далее расчет ФНЧ и ФВЧ Чебышева проводится согласно алгоритму синтеза аналоговых фильтров Батерворта. Принципиальные схемы ФНЧ и ФВЧ Баттерворта полностью эквивалентны фильтрам Чебышева. Различие заключается в значениях коэффициентов комплексной функции фильтров Баттерворта или Чебышева. Сформировав комплексную функцию согласно определенной аппроксимационной характеристике, получим принципиальную схему с АЧХ Батерворта или Чебышева определенного порядка. В табл. 4.1 приведены принципиальные схемы фильтров, частотные коэффициенты передачи, графические изображения АЧХ и ФЧХ.

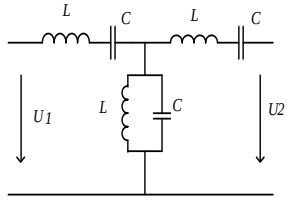
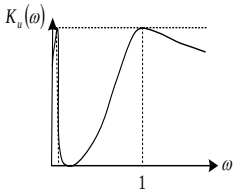
Таблица 4.1.

ФНЧ 1-го порядка		
	$\lvert K(\omega)\rvert = \frac{1}{\sqrt{1+(\omega\cdot\tau)^2}}$ $\varphi(\omega) = -arctg\,\omega\tau$ $\tau = R/L$	
ФНЧ 1-го порядка		
	$\lvert K(\omega)\rvert = \frac{1}{\sqrt{1+(\omega\cdot\tau)^2}}$ $\varphi(\omega) = -arctg\,\omega\tau$ $\tau = R\cdot C$	
ФВЧ 1-го порядка		
	$\lvert K(\omega)\rvert = \frac{\omega\cdot\tau}{\sqrt{1+(j\omega\cdot\tau)^2}}$ $\varphi(\omega) = \frac{\pi}{2} - arctg\,\omega\tau$ $\tau = R\cdot C$	
ФВЧ 1-го порядка		
	$\varphi(\omega) = \frac{\pi}{2} - arctg\,\omega\tau$ $\lvert K(\omega)\rvert = \frac{\omega\cdot\tau}{\sqrt{1+(\omega\cdot\tau)^2}}$ $\tau = L/R$	

Продолжение табл. 4.1

ФНЧ 2-го порядка Баттерворта		
	$\lvert K(\omega)\rvert = \frac{\omega_{ñð}^2}{\sqrt{(\omega_{ñð}^2-\omega^2)^2+(\sqrt{2}\cdot\omega\cdot\omega_{ñð})^2}}$ $\varphi(\omega) = -arctg\frac{\sqrt{2}\cdot\omega\cdot\omega_{cp}}{\omega_{cp}^2-\omega^2}$	
ФВЧ 2-го порядка Баттерворта		
	$\lvert K(\omega)\rvert = \frac{1}{\sqrt{(1-\frac{\omega_{cp}^2}{\omega^2})^2+(\sqrt{2}\frac{\omega_{cp}}{\omega})^2}}$ $\varphi(\omega) = arctg\frac{\sqrt{2}\frac{\omega_{cp}}{\omega}}{(1-\frac{\omega_{cp}^2}{\omega^2})}$	
ФНЧ 2-го порядка Чебышева		
	$\lvert K(\omega)\rvert = \frac{\omega_{ñð}^2}{\sqrt{(\omega_{ñð}^2-\omega^2)^2+(\sqrt{2}sha\cdot\omega_{ñð}\cdot\omega)^2}}$ $\varphi(\omega) = -arctg\frac{\sqrt{2}sha\cdot\omega_{cp}\cdot\omega}{\omega_{cp}^2-\omega^2}$	
ФВЧ 2-го порядка Чебышева		
	$\lvert K(\omega)\rvert = \frac{1}{\sqrt{(1-\frac{\omega_{cp}^2}{\omega^2})^2+(\sqrt{2}\frac{sha\cdot\omega_{cp}}{\omega})^2}}$ $\varphi(\omega) = arctg\frac{\sqrt{2}sha\frac{\omega_{cp}}{\omega}}{(1-\frac{\omega_{cp}^2}{\omega^2})}$	
Полосовой фильтр 2-го порядка		
	$\lvert K(\omega)\rvert = \frac{\frac{\omega^2}{\omega_{cp}^2}}{\sqrt{(\frac{\omega^2}{\omega_{cp}^2}-(1-Q\frac{\omega^2}{\omega_{cp}^2})^2-\sqrt{2}(1-Q\frac{\omega^2}{\omega_{cp}^2})\frac{\omega}{\omega_{cp}}}}$ $\varphi(\omega) = \frac{\pi}{2} - arctg\frac{-\sqrt{2}(1-Q\frac{\omega^2}{\omega_{cp}^2})\frac{\omega}{\omega_{cp}}}{(\frac{\omega^2}{\omega_{cp}^2}-(1-Q\frac{\omega^2}{\omega_{cp}^2})^2}$	

Окончание табл. 4.1

Режекторный фильтр 2-го порядка		
	$\lvert K(\omega)\rvert = \frac{1}{\sqrt{\left[(1-Q\frac{\omega^2}{\omega_{cp}^2})^2\cdot\omega_{cp}^2\cdot+1\right]^2+\left(\frac{\sqrt{2}\cdot\omega}{\left(1-Q\frac{\omega^2}{\omega_{cp}^2}\right)\cdot\omega_{cp}}\right)^2}}$ $\varphi(\omega) = arctg\frac{\frac{\sqrt{2}\cdot\omega}{\left(1-Q\frac{\omega^2}{\omega_{cp}^2}\right)\cdot\omega_{cp}}}{\left[(1-Q\frac{\omega^2}{\omega_{cp}^2})^2\cdot\omega_{cp}^2\cdot+1\right]^2}$	

Анализ прохождения сигналов через линейные и нелинейные цепи. Неискаженная передача сигналов. При прохождении через линейные и нелинейные цепи и системы информационные сигналы претерпевают различные изменения: сигнал на выходе устройства отличается от сигнала на входе. Выходным сигналом является реакция цепи на действие входного влияния. Под действием различных сигналов в системе возникают переходные процессы, существенным образом влияющие на характер реакции системы. В отличие от задач электротехники, в которых интересуются установлением режима при различных коммутациях, в радиотехнике, радиолокации, импульсной технике основное значение имеет влияние переходных процессов на форму сигнала. В радиотехнических системах это определяет влияние системы на информацию, содержащуюся в сигнале. *Искажениями называют такие изменения сигнала, которые приводят к искажению информации*, содержащейся в нем. В радиотехнике задача исследования процессов в системе чаще формулируется как задача исследования особенностей прохождения сигналов по цепям.

Такие задачи можно решать разными путями (рис. 4.23).

Рис. 4.23. Методы анализа сигналов

Рассмотрим суть *спектрального метода анализа*. Входной сигнал представляется в виде совокупности гармонических колебаний, возникающих в цепи задолго до момента наблюдения (на интервале времени $-\infty \le t \le +\infty$). При этом сигнал, появляющийся на выходе цепи, в соответствии с принципом суперпозиции, является суммой гармонических реакций, определяемых отдельно для каждой из составляющих входного сигнала. Таким образом, задача, в сущности, сводится к анализу постоянных режимов в цепи при синусоидальном действии.

Значения гармонических составляющих выходного сигнала легко отыскиваются, если известны спектр входного сигнала и частотные характеристики цепи.

Поскольку в основу спектрального метода положен принцип суперпозиции, то этот метод является линейным; он пригоден для решения линейных задач, т.е. для анализа прохождения сигналов через линейные цепи. Этот метод применим как в области действительных величин, так и при использовании плоскости комплексных величин. Более удобным является использование комплексной плоскости.

Анализ прохождения периодических сигналов через типовые фильтры. Предположим, на цепь влияет сложный информационный сигнал в виде периодической функции, которую можно представить рядом Фурье как сумму бесконечного количества гармоник. Нужно определить выходной сигнал, который является реакцией системы на определенное влияние.

Известно, что комплексные амплитуды синусоидального тока частоты ω, протекающего в цепи, и напряжения, приложенного к ней, связаны простым соотношением

$$\dot{I}_m e^{j\omega t} = Y\left(j\omega\right)\dot{U}_m e^{j\omega t},$$

где $X(j\omega)$ и $Z(j\omega)$ - комплексные проводимость и сопротивление системы, рассчитанные для частоты ω. Комплексные сопротивление $Z(j\omega)$ и проводимость $X(j\omega)$ являются частными случаями более общего понятия - *комплексной функции цепи*.

Представим сигнал, действующий на входе цепи, в виде ряда Фурье в комплексной форме (см. главу 3):

$$U_{\text{вх}}\left(t\right) = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty}\dot{U}_{\text{вх}_{mn}} e^{j\omega_n t} = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty} U_{\text{вх}_{mn}} e^{-j\psi_{\text{вх}_n}} e^{j\omega_n t}. \qquad (4.21)$$

Комплексная амплитуда каждой из гармоник выходного сигнала $\dot{U}_{\text{вых}_{mn}}$ определяется как произведение комплексной амплитуды соответствующей гармоники входного сигнала $\dot{U}_{\text{вх}_{mn}}$ на соответствующую комплексную функцию цепи $K\left(j\omega_n\right) = K\left(\omega_n\right)e^{j\varphi(\omega_n)}$:

$$\dot{U}_{\text{вых}_{mn}} = \dot{U}_{\text{вх}_{mn}} K\left(j\omega_n\right) = \dot{U}_{\text{вых}_{mn}} e^{-j\psi_{\text{вых}_n}} \tag{4.22}$$

где

$$\dot{U}_{\text{вых}\ mn} = \dot{U}_{\text{вх}\ mn} K\left(\omega_n\right) = \dot{U}_{\text{вых}\ mn} e^{j\varphi_{\text{вых}\ n}}\ .$$

Отсюда, на основании принципа суперпозиции, находим выражение для выходного сигнала

$$U_{\text{вых}}\left(t\right) = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty} \dot{U}_{\text{вых}_{mn}} e^{j\omega_n t} = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty} \dot{U}_{\text{вх}_{mn}} K\left(j\omega_n\right) e^{j\omega_n t}. \tag{4.23}$$

Если входной сигнал (4.17) представить рядом Фурье в действительной форме

$$U_{\text{вх}}\left(t\right) = U_{\text{вх}_0} + \sum_{n=1}^{\infty} U_{\text{вх}_{mn}} \cos\left(\omega_n t - \psi_{\text{вх}_n}\right), \tag{4.24}$$

то сигнал на выходе определится таким образом:

$$\begin{aligned} U_{\text{вх}}\left(t\right) &= U_{\text{вх}_0} + \sum_{n=1}^{\infty} U_{\text{вх}_{mn}} \cos\left(\omega_n t - \psi_{\text{вх}_n}\right) = \\ &= U_{\text{вх}_0}\left(K_0\right) + \sum_{n=1}^{\infty} U_{\text{вх}_{mn}} K\left(\omega_n\right) \cos[\omega_n t - \psi_{\text{вх}_n} + \varphi\left(\omega_n\right)]. \end{aligned} \tag{4.25}$$

АЧХ *сигнала на выходе можно получить путем умножения амплитудно-частотного спектра входного сигнала на модуль комплексной функции цепи*; *его ФЧХ - суммированием фазочастотных составляющих спектра входного сигнала и значений аргумента комплексной функции цепи на соответствующих частотах.*

Метод расчета базируется на использовании разложения сигналов в ряд Фурье и может быть разделен на следующие этапы:

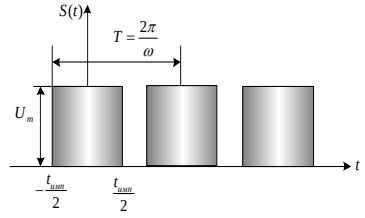
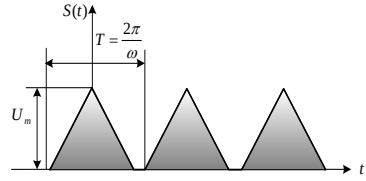
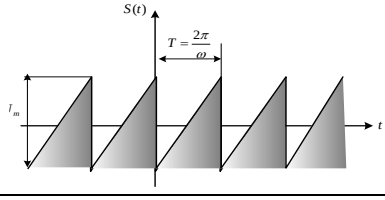
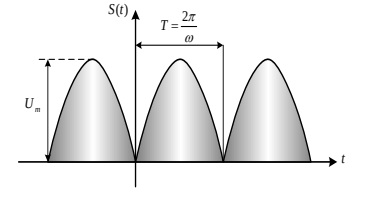
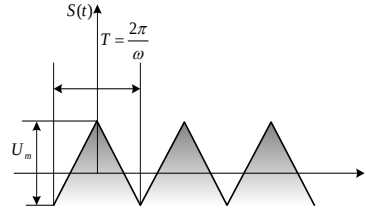
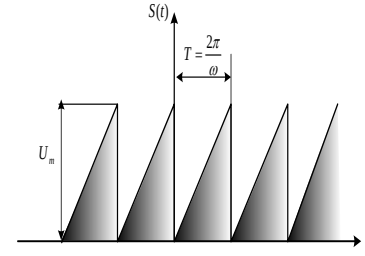
входной сигнал представляют в виде ряда Фурье;

определяют необходимую входную или передающую комплексную функцию цепи;

комплексные амплитуды гармонических составляющих выходного сигнала рассчитывают согласно формуле (4.22) как произведение комплексных амплитуд входного сигнала и комплексной функции цепи.

В табл. 4.1 определены коэффициенты передачи типичных фильтров, а для объяснения спектрального метода анализа целесообразно привести типичные радиотехнические сигналы и соответствующие им тригонометрические формы рядов Фурье (табл. 4.2).

Таблица 4.2

	Форма сигналов	Тригонометрическая форма ряда Фурье
1		$U(t)=\frac{2U_0 t_{имп}}{T}+\frac{2U_0}{\pi}\sum_{k=1}^{\infty}\frac{\sin k\omega_1 t_{имп}}{k}\cos k\omega_1 t$
2		$U(t)=\frac{U_0 t_{имп}}{T}+\frac{2U_0 T}{\pi^2 t_{имп}}\sum_{k=1}^{\infty}\frac{\sin^2 k\frac{\pi t_{имп}}{T}}{k^2}\cos k\omega_1 t$
3		$U(t)=\frac{2U_0}{\pi}\sum_{k=1}^{\infty}\frac{\cos k\pi}{k}\cos\left(k\omega_1 t+\frac{\pi}{2}\right)$
4		$U(t)=\frac{2U_0}{\pi}+\frac{4U_0}{\pi}\sum_{k=1}^{\infty}\frac{\cos(2k\omega_1 t+\pi)}{4k^2-1}$
5		$U(t)=\frac{8U_0}{\pi^2}\sum_{k=1}^{\infty}\frac{\sin^2 k\frac{\pi}{2}}{k^2}\cos k\omega_1 t$
6		$U(t)=\frac{U_0}{2}+\frac{U_0}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\cos\left(k\omega_1 t+\frac{\pi}{2}\right)$

Пример. На вход интегрирующей цепи (см. рис. 4.7) поступает последовательность прямоугольных видеоимпульсов напряжения (рис. 4.24). Найти напряжение, которое появляется на выходе фильтра, если $\omega RC >> 1$.

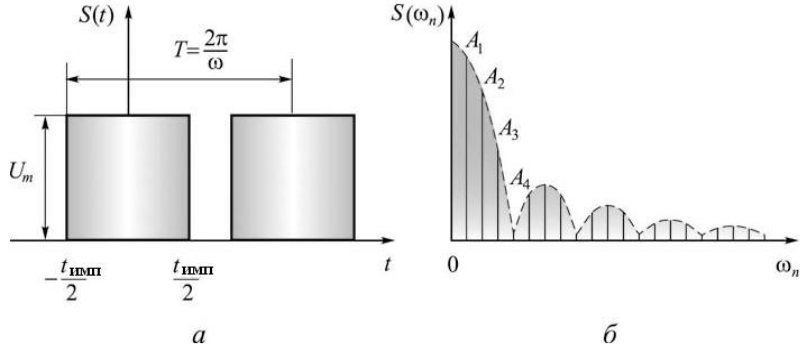


Рис.4.24. Периодическая последовательность прямоугольных видео-импульсов (*а*) и их спектр (*б*)

Р е ш е н и е. 1**.** Представление входного сигнала в виде ряда Фурье определяется выражением

$$u_{\text{вх}}(t) = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty} U_{\text{вх}_{nm}} e^{jn\Omega t} = \frac{U_m}{q}\sum_{n=-\infty}^{n=+\infty} \frac{\sin n\pi/q}{n\pi/q} e^{jn\Omega t},$$

а комплексный коэффициент передачи цепи по напряжению - выражением

$$K_U(j\omega) = \frac{1}{j\omega C}\frac{1}{R+\dfrac{1}{j\omega C}} = \frac{1}{1+j\omega\tau} = \frac{1}{\sqrt{1+(\omega\tau)^2}} e^{-j\ \text{arctg}\ \omega\tau},$$

где $\tau = RC$ - постоянная составляющая.

2. Напряжение на выходе цепи находим согласно (4.22):

$$u_{\text{вых}}(t) = \frac{1}{2}\sum_{n=-\infty}^{n=+\infty} \dot{U}_{\text{вх}_{nm}} K(jn\Omega) e^{jn\Omega t} = \frac{U_m}{\pi}\sum_{n=-\infty}^{n=+\infty} \frac{\sin\dfrac{n\pi}{q}}{n\sqrt{1+(n\Omega\tau)^2}} e^{j[n\Omega(t-t_0)-\text{arctg}\ n\Omega\tau]}.$$

При $\omega RC >> 1$

$$u_{\text{вых}}(t) \approx \frac{U_m}{q} + \frac{U_m}{\pi\Omega\tau}\sum_{n=1}^{\infty}\frac{\sin\dfrac{n\pi}{q}}{n^2}\cos[n\Omega(t-t_0)-\frac{\pi}{2}] = \frac{U_m}{q} + \frac{U_m}{\pi\Omega\tau}\sum_{n=1}^{\infty}\frac{\sin\dfrac{n\pi}{q}}{n^2}\sin\ n\Omega(t-t_0).$$

На рис. 4.25 приведен АЧС входного сигнала и зависимость модуля коэффициента передачи (АЧ) цепи от частоты, а также АЧС выходного сигнала, полученный в результате их перемножения.

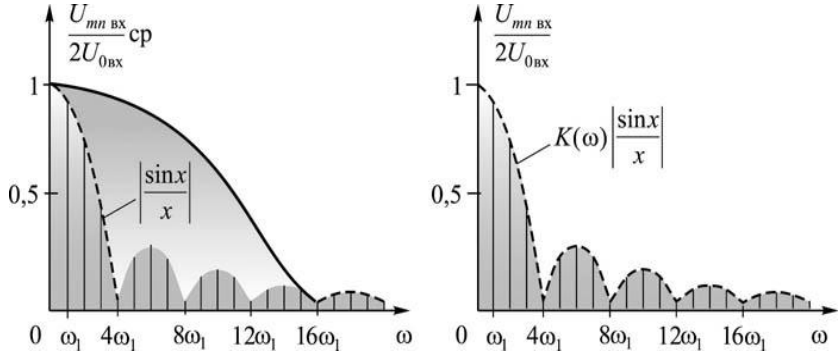


Рис. 4.25. Спектральное представление периодического сигнала

Расчет прохождения униполярных (одиночных) сигналов. *При анализе прохождения непериодических сигналов спектральным методом используются прямое и обратное преобразования Фурье*.

Представление непериодической функции в виде интеграла Фурье является результатом суммирования бесконечного количества незатухающих и бесконечно близких по частоте синусоидальных колебаний с бесконечно малыми амплитудами. Это дает возможность применять к бесконечно малым гармоническим составляющим тока и напряжения обычные методы расчета постоянных режимов, а далее, пользуясь методом наложения, определять результирующее напряжение и ток. *Ценность такого подхода заключается в том, что анализ переходного режима сводится к анализу постоянных режимов.*

Предположим, что на линейную цепь, комплексная функция $K(j\omega)$ которой известна, подается сигнал $U_{вх}(t)$. Интеграл Фурье дает возможность описать этот сигнал наложением бесконечного количества гармоник с комплексными амплитудами $\frac{1}{\pi}S_{вх}(j\omega)d\omega$.

Комплексные амплитуды соответствующих гармоник выходного сигнала определяются произведением

$$\frac{1}{\pi}S_{вх}(j\omega)K(j\omega)d\omega=\frac{1}{\pi}S_{вых}(j\omega)d\omega, \qquad (4.26)$$

а спектральная плотность исходного сигнала

$$S_{вых}(j\omega)=S_{вх}(j\omega)K(j\omega) \qquad (4.27)$$

равна произведению спектральной плотности входного сигнала на соответствующую комплексную функцию цепи. Значение выходного сигнала как функции времени находим с помощью обратного преобразования Фурье наложением бесконечно большого количества гармонических составляющих сигнала:

$$U_{\text{вых}}(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}S_{\text{вых}}(j\omega)e^{j\omega t}d\omega=\frac{1}{2\pi}\int_{-\infty}^{+\infty}S_{\text{вх}}(j\omega)K(j\omega)e^{j\omega t}d\omega. \qquad (4.28)$$

Таким образом, по известной комплексной функции цепи и спектральной плотности входного сигнала с помощью обратного преобразования Фурье можно рассчитать реакцию цепи на любое заданное действие (*влияние*).

Расчет имеет такую последовательность:

находят спектральную плотность $S_{\text{вх}}(j\omega)$ входного сигнала;

определяют необходимую входную или передающую функцию цепи;

рассчитывают спектральную плотность выходного сигнала по формуле (4.27) как произведение $S_{\text{вх}}(j\omega)$ на комплексную функцию цепи;

сигнал на выходе находят с помощью принципа суперпозиции как сумму реакций цепи на каждое из действий отдельно.

Пример. На простую *RC-цепь* (см. рис. 4.7) действует одиночный прямоугольный импульс конечной продолжительности $t_{\text{имп}}$ с амплитудой U_m (рис. 4.26). Найти напряжение на выходе.

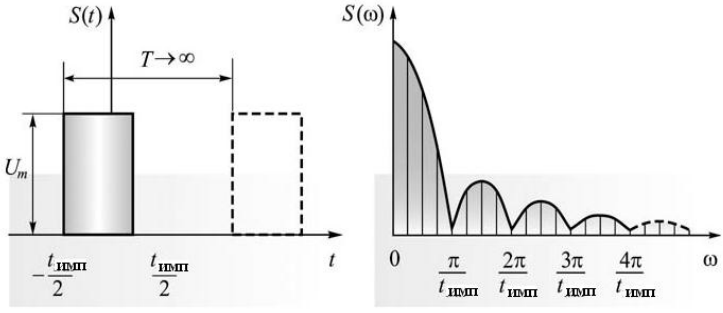


Рис. 4.26. Униполярный видео-импульс и его спектральная плотность

Р е ш е н и е. 1. Спектральная плотность исходного сигнала (где $(\tau = RC)$ - постоянная составляющая):

$$S_{\text{вых}}(j\omega)=S_{\text{вх}}(j\omega)K(j\omega)=U_m t_{\text{имп}}\,\frac{\sin\frac{\omega t_{\text{имп}}}{2}}{\frac{\omega t_{\text{имп}}}{2}}\,\frac{\omega}{\omega-j\frac{1}{\tau}}=2U_m\frac{\sin\frac{\omega t_{\text{имп}}}{2}}{\omega-j\frac{1}{\tau}}.$$

2. Сигнал на выходе цепи:

$$u_{\text{вых}}(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}S_{\text{вых}}(j\omega)e^{j\omega t}d\omega=\frac{U_m}{\pi}\int_{-\infty}^{+\infty}\frac{\sin\frac{\omega t_{\text{имп}}}{2}}{\omega-j\frac{1}{\tau}}e^{j\omega t}d\omega=$$

$$\frac{U_m}{\pi}2\pi j\sin\left(j\frac{1}{\tau}\frac{t_{\text{имп}}}{2}\right)e^{-\frac{1}{\tau}}=U_m\left[e^{-\frac{1}{\tau}\left(t+\frac{t_{\text{имп}}}{2}\right)}-e^{-\frac{1}{\tau}\left(t-\frac{t_{\text{имп}}}{2}\right)}\right],$$

поскольку $2j\sin jx=e^{-x}-e^{x}$.

Формы импульса на входе и выходе цепи при различных постоянных времени $\tau = RC$ приведены на рис. 4.27.

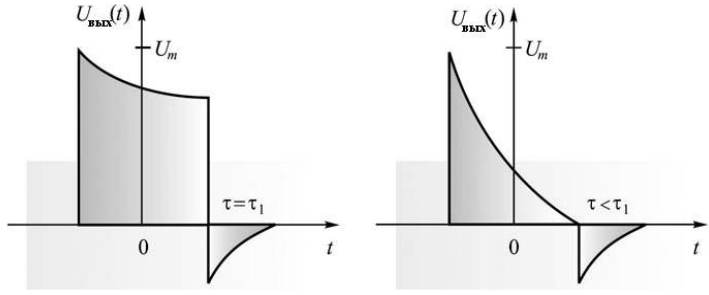


Рис. 4.27. Униполярные видео-импульсы на выходе фильтра

Неискаженная передача сигналов. Любые изменения сигнала сопровождаются изменением его спектра. При этом возможны два принципиально различных случая. В одном из них в спектре сигнала появляются составляющие с новыми частотами, которые отсутствовали на входе устройства. Такие изменения называются *нелинейными*, поскольку появление новых частот возможно только в нелинейных цепях. Во втором случае в спектре сигнала новые частоты не возникают, а его изменения определяются лишь изменениями амплитуд и начальных фаз гармоник. В этом случае говорят о линейных изменениях сигнала. В реальных устройствах всегда происходит искажение сигналов. Тем не менее, искажения стремятся сделать настолько малыми, чтобы они не превышали допустимого уровня, т.е. практически отсутствовали.

Будем считать, что информация, содержащаяся в сигнале, отображается его формой. Тогда неискаженная передача сигнала будет обозначать неизменность его формы (рис. 4.28).

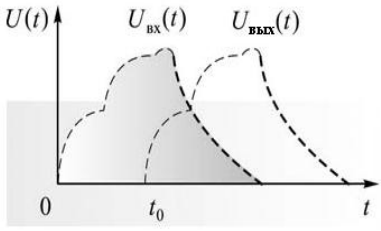


Рис. 4.28. Искажение сигнала при передаче

Выясним, какими характеристиками должны обладать линейные цепи или система для обеспечения неискаженной передачи сигнала. Очевидно, условием такой передачи при временном описании являются соотношения

$$U_{\text{вых}}(t) = kU_{\text{вх}}(t - t_0), \qquad (4.29)$$

где k - постоянный множитель, учитывающий изменения амплитуды; t_0 - время задержки, определяющее смещение сигнала во времени.

При спектральном описании, переходя от выражения (4.29) к уравнению для спектров входного и выходного сигналов, с учетом теоремы линейности и теоремы о сдвиге находим

$$S_{\text{вых}}(j\omega) = kS_{\text{вх}}(j\omega)e^{-j\omega t_i} = K(j\omega)S_{\text{вх}}(j\omega), \qquad (4.30)$$

где $K(j\omega) = ke^{-j\omega t_0} = K(\omega)e^{j\varphi(\omega)}$ - комплексная функция цепи.

Таким образом, для неискаженной передачи сигнала, сопровождающейся лишь изменением его амплитуды и задержкой на некоторое время t_0, АЧХ и ФЧХ цепи должны быть линейными во всем диапазоне частот $-\infty \le \omega \le +\infty$ (рис. 4.29):

$$\begin{cases} K(\omega) = k = \text{const}; \\ \text{t}\varphi(\omega) = -\omega t_0. \end{cases} \qquad (4.31)$$

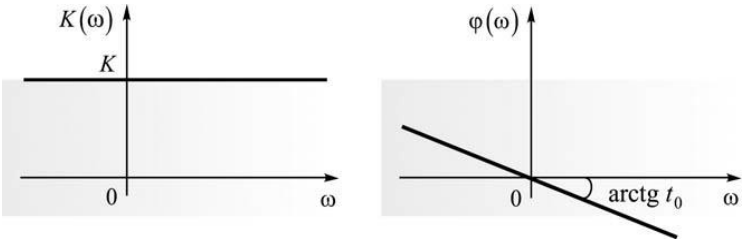


Рис. 4.29. Амплитудно-частотная и фазо-частотная характеристики цепи

Эбергард Фредерик Фердинанд Хопф (Eberhard Frederich Ferdinand Hopf, 1902 - 1983), австрийский математик, который сделал значительный вклад в развитие топологии и эргодичной теории. Основные работы касаются теории динамических систем и дифференциальных уравнений с частными производными. Ему принадлежит монография "Эргодичная теория", посвященная спектральной теории динамических систем. Есть работы и в области астрофизики. Получил степень доктора философии в области математики (1926). В 1929 г. получил квалификацию по математической астрономии от Берлинского универси-тета.

Это означает, что во всем диапазоне частот колебания на выходе прямо пропорциональны колебаниям на входе, причем все гармонические составляющие задерживаются на одно и то же время.

Нелинейность частотных характеристик цепи вызывает появление искажения сигналов. Реализовать цепи с линейными характеристиками во всем диапазоне частот невозможно. На практике это не нужно, поскольку реальные сигналы имеют ограниченный спектр. Поэтому вполне достаточно, чтобы частотные характеристики цепи были линейны только в ограниченной полосе частот (полосе пропускания), соответствующей ширине спектра сигнала. Тем не менее, при конечной полосе пропускания, если спектр сигнала шире, чем полоса пропускания системы, невозможно избежать искажений даже при идеальных ее характеристиках.

Частотные искажения *- изменения формы сигналов, обусловленные отклонением АЧХ цепи от равномерной. Аналогично искажения, вызванные нелинейностью ее ФЧХ, называются фазовыми.*

Причиной нелинейности частотных характеристик линейной цепи является наличие в ее составе реактивных элементов, предопределяющих частотную зависимость параметров этой цепи. Задержка во времени при прохождении сигналов через такие цепи объясняется возникновением переходных процессов, вызванных накоплением энергии в реактивных элементах.

Прохождение сигналов через идеальную линейную цепь с ограниченной полосой пропускания. Рассмотрим влияние ограничения полосы пропускания цепи на прохождение сигналов на примере идеального фильтра низких частот. Такая цепь имеет идеальные характеристики (5.31), но в ограниченном диапазоне частот (рис. 4.30):

$$K(j\omega)=\begin{cases} e^{j\omega t_0} & \text{при } -\omega_з < \omega < \omega_з \\ 0 & \text{при } \omega_з < \omega < -\omega_з \end{cases} \qquad (4.32)$$

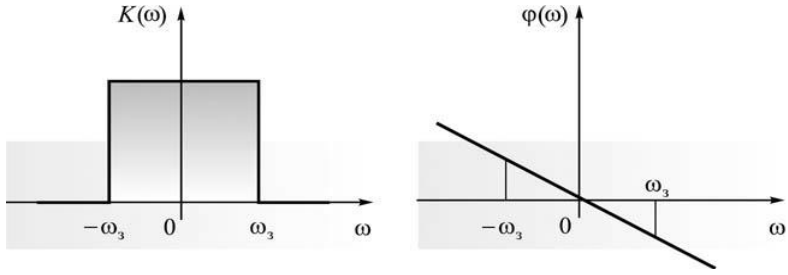


Рис. 4.30. Идеальные частотные характеристики фильтров

Через такую систему все гармоники спектра с частотами, более низкими, чем частота среза $\omega_з$, будут проходить без изменения амплитуд и со сдвигом фаз, пропорциональным частоте.

Гармоники с частотами, превышающими частоту среза $\omega_з$, не будут пропускаться, что приведет к изменению спектра и формы сигнала на выходе. Степень искажения сигнала будет определяться отсеченной частью спектра, зависящей от ширины полосы пропускания системы.

Заметим, что фильтр с указанными идеальными характеристиками не может быть реализован практически, поскольку его комплексная функция (4.32) не удовлетворяет условиям физической реализации. Тем не менее, такая идеализация удобна, поскольку дает возможность упростить анализ и выделить особенности явлений, происходящих в системе.

Пример. Рассчитать в общем виде прохождение прямоугольного радиоимпульса через идеальный полосовой фильтр, полоса пропускания которого, ограничена частотами $\omega_н \mp \omega_з$ (рис. 4.31).

Р е ш е н и е. Известно, что при действии прямоугольного видеоимпульса $S_{вх1} < 0$ на идеальный фильтр низких частот с частотой среза $\omega_з$ (см. рис. 4.30) будем иметь сигнал на выходе $S_{вых1}(t)$. Его спектральная плотность

$$S_{вых1}(j\omega) = K_1(j\omega) S_{вх1}(j\omega),$$

где $K_1(j\omega)$ - комплексный коэффициент передачи идеального ФНЧ; $S_{вх1}(t)$ - спектральная плотность прямоугольного видеоимпульса.

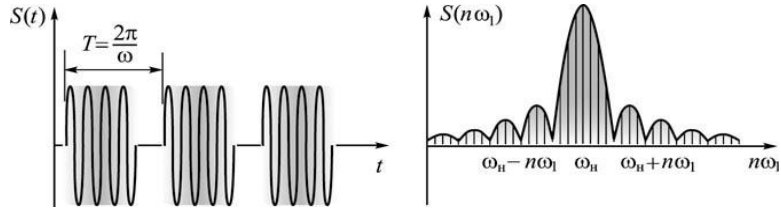


Рис. 4.31. Периодическая последовательность прямоугольных радиоимпульсов и их спектр

В соответствии с теоремой о модуляции, если

$$S_{\text{вых}}(t) = S_{\text{вых1}}(t)\cos\omega_{\text{н}}t, \tag{4.33}$$

то

$$S_{\text{вых}}(j\omega) = \frac{1}{2}\left\{S_{\text{вых1}}\left[j(\omega-\omega_{\text{н}})\right] + S_{\text{вых1}}\left[j(\omega+\omega_{\text{н}})\right]\right\} =$$

$$= \frac{1}{2}\left\{\begin{matrix} K_1\left[j(\omega-\omega_{\text{н}})\right]S_{\text{вх1}}\left[j(\omega-\omega_{\text{н}})\right] + \\ + K_1\left[j(\omega+\omega_{\text{н}})\right]S_{\text{вх1}}\left[j(\omega+\omega_{\text{н}})\right] \end{matrix}\right\} = K(j\omega)S_{\text{вх}}(j\omega), \tag{4.34}$$

что является произведением комплексного коэффициента передачи $K(j\omega)$ идеального полосового фильтра (он образован сдвигом по оси частот коэффициента передачи $K_1(j\omega)$ идеального ФНЧ в обе стороны от начала координат на величину $\pm\omega_{\text{н}}$) и спектральной плотности прямоугольного радиоимпульса (она образована аналогичным смещением по частоте спектральной плотности прямоугольного видеоимпульса, умноженной на ½). Учитываем, что

$$K_1[j(\omega-\omega_n)]S_{\text{вх1}}[j(\omega+\omega_n)] \approx 0,$$

$$K_1[j(\omega+\omega_n)]S_{\text{вх1}}[j(\omega-\omega_n)] \approx 0.$$

Отсюда вытекает, что выражения (4.33) и (4.34) описывают реакцию идеального полосового фильтра на действие прямоугольного радиоимпульса

$$S_{\text{вх}}(t) = S_{\text{вх1}}(t)\cos\omega_{\text{н}}t. \tag{4.35}$$

Таким образом, огибающая сигнала на выходе полосового фильтра при действии прямоугольного радиоимпульса (рис. 4.32) *определяется огибающей* сигнала на выходе ФНЧ при действии на его вход прямоугольного видеоимпульса. Этот вывод *можно распространить* и на случай других радиоимпульсов, сформулировав теорему об огибающей.

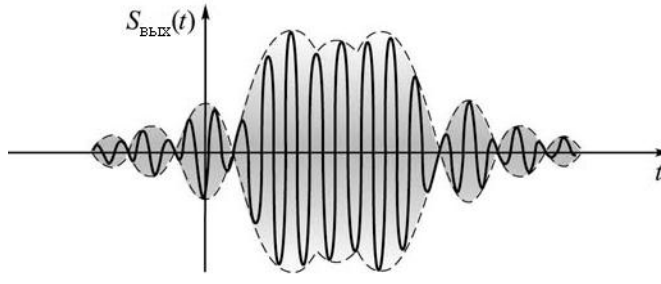


Рис. 4.32. Сигнал на выходе фильтра

Влияние частотных характеристик круга на спектр и форму исходного сигнала. При комплексной форме записи, соответственно выражениям

(4.22) и (4.27), спектр, а значит, и форма исходного сигнала определяются произведением комплексных функций входного сигнала и цепи. Частотные характеристики цепи полностью определяют изменения спектра и формы сигнала, возникающие при прохождении сигнала через цепь. *Спектр выходного сигнала зависит от частотной характеристики цепи так же, как и от спектра входного сигнала.*

Сравнение спектра входного сигнала с графиком частотных характеристик цепи дает возможность оценить систему передачи сигнала с точки зрения вносимых искажений. В зависимости от того, какая часть спектра сигнала задерживается, можно судить о характере и степени искажения. Это важно при проектировании систем. *Знание спектра сигнала дает возможность сделать выбор необходимой полосы пропускания системы и ее предельных частот.* По известным частотным характеристикам цепи и сигнала можно выбрать схемы и параметры устройств коррекции, необходимые для исправления частотных характеристик системы и формы выходного сигнала. Возможность оценивания искажений сигнала при прохождении по цепи является основной ценностью спектрального метода. Это важно, если известны не схемы цепей, а их частотные характеристики, полученные экспериментально. Чтобы отметить влияние различных участков частотной характеристики цепи на спектр и форму сигнала, обратимся к двум группам простых схем: дифференцируемых и интегрируемых. Их частотные характеристики имеют спад соответственно в области низких или высоких частот, что дает основания относить эти цепи к ФВЧ или ФНЧ. Комплексные передающие на входе функции таких схем приведены в табл. 4.3

Для случая ступенчатого воздействия:

$$\begin{cases} U_{\text{вх}}(t) = U_m 1(t), \\ U_{\text{вх}}(j\omega) = U_m \left[\dfrac{1}{j\omega} + \pi\delta(\omega) \right]. \end{cases} \tag{4.36}$$

Дифференцирующие цепи. Спектральная плотность сигнала на выходе цепей первой группы при ступенчатом действии на входе:

$$U_{\text{вых}1}(j\omega) = U_{\text{вх}}(j\omega) K_1(j\omega) = \frac{U_m}{a + j\omega} + \frac{U_m \pi\delta(\omega)\, j\omega}{a + j\omega} = \frac{U_m}{a + j\omega}, \tag{4.37}$$

поскольку $\delta(\omega) j\omega = 0$ для всех значений ω.

Такому спектру соответствует экспоненциальная функция

$$U_{\text{вых}\,1}(t) = U_m e^{-at} = U_m e^{1\frac{t}{\tau}}. \tag{4.38}$$

При больших значениях постоянной времени спад частотной характеристики сравнительно небольшой, и цепи этой группы слабо искажают сигнал. С уменьшением постоянной τ искажения возрастают, приобретая характер дифференцирования (рис. 4.33).

Дифференцирование происходит при

$$\omega_{max} << a = \frac{t}{\tau}, \text{ т.е. } \omega_{max}\tau << 1, \quad (4.39)$$

при

$$K_1(j\omega) = j\omega\frac{1}{a}, \quad (4.40)$$

где ω_{max} - максимальная частота, учитывающаяся в спектре входного сигнала.

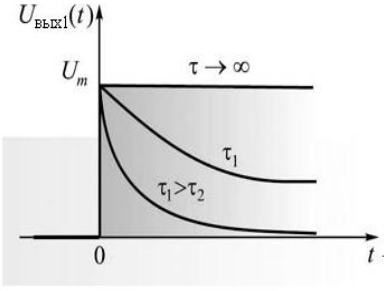


Рис. 4.33. Сигнал на выходе системы при разных значениях τ

При искажениях такого вида изменения фронта сигнала не происходят, изменяется, в основном, его вершина. В спектре сигнала этому соответствует уменьшение спектральной плотности низкочастотных составляющих.

Интегрирующие цепи. Спектральная плотность сигнала на выходе цепей второй группы при ступенчатом действии с учетом соотношений, приведенных в табл. 4.3, приобретает вид

$$U_{вых2}(j\omega) = U_{вх}(j\omega)K_2(j\omega) = U_{вх}(j\omega)\frac{a}{j\omega}K_1(j\omega) = \frac{a}{j\omega}U_{вых1}(j\omega). \quad (4.41)$$

Отсюда на основании теоремы интегрирования (см. табл. 4.3) находим

$$U_{вых2}(t) = a\int_0^1 U_{вых1}(t)dt = aU_m\int_0^1 e^{-at}dt = U_m\left(1-e^{-at}\right) = U_m\left(1-e^{-\frac{1}{t}}\right), \quad (4.42)$$

т.е. сигнал на выходе изменяется по закону обратной экспоненты.

Цепи этой группы слабо искажают сигнал при небольших значениях постоянной времени τ. Искажения возрастают с увеличением постоянной τ и имеют характер интегрирования (рис. 4.34).

Интегрирование происходит при

$$\omega_{max} >> a = \frac{t}{\tau}, \text{т.е. } \omega_{max}\tau >> 1 \quad (4.43)$$

где

$$K_2(j\omega) \approx a/j\omega. \quad (4.44)$$

При таком искажении происходит, в основном, изменение фронта сигнала, а его вершина изменяется слабо. В спектре сигнала этому соответствует уменьшение спектральной плотности высокочастотных составляющих.

Отмечая влияние различных участков частотной характеристики цепи на форму сигнала, можно сказать, что для неискаженной передачи фронта импульса необходимо обеспечить условия неискаженной передачи на высоких частотах, а для сохранности неизменной формы его вершины нужно обеспечить условия неискаженной передачи на низких частотах.

Для более сложных цепей частотные характеристики оказываются также составными функциями частоты. В этих случаях удобно применять приближенные методы, когда кривая частотных характеристик аппроксимируется прямолинейными отрезками.

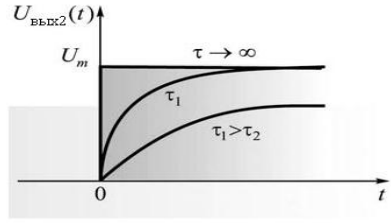


Рис. 4.34. Сигнал на выходе системы при разных значениях τ

Модуляция и детектирование сигналов. Передача информации от источника информации к ее потребителю - базовая задача информационно-коммуника-ционных систем. Для передачи информационных сигналов применяются высокочастотные электромагнитные колебания (ВЧ), которые эффективно используются при малых мощностях передатчика. Важным условием для выбора ВЧ передачи является способность электромагнитной волны данной длины распространяться в пространстве. Передаваемая информация (информационный сигнал), должна быть внесена во ВЧ колебание, которое в этом случае будет носителем информации. Такой процесс внесения информации во ВЧ сигнал осуществляется с помощью процесса модуляции.

Модуляцией *называют процесс изменения параметров несущего колебания по закону информационного сигнала.*

В общем случае модулированный сигнал можно представить в виде:

$$U(t) = A_m(t)\cos\left(\omega_н t + \psi_н(t)\right). \qquad (4.45)$$

В формуле (4.45) $A_m(t)$ и $\psi_н(t)$ изменяются по закону информационного сигнала. Если $A_m(t) = \psi_н(t) = \text{const}$, то выражение (4.45) преобразуется в простой ВЧ гармонический сигнал. В этом случае ВЧ сигнал никакой информации не несет. В зависимости от того, какой из параметров ВЧ колебания - $A(t)$ или $\psi(t)$ - изменяется, различают ***амплитудную модуляцию***

$$U(t) = A_m(t)\cos\left(\omega_н t + \psi_н\right) \qquad (4.46)$$

и ***угловую модуляцию***

$$U(t) = A_m\cos\left(\omega_н t + \psi_н(t)\right) \qquad (4.47)$$

Рудольф Эмиль Калман (Rudolf Emil Kalman, 1930) американский математик и специалист в области электронной инженерии. Закончил Массачусетский технологический институт (1954). Исследования касаются математического обеспечения теории автоматического регулирования. Создал фильтры (фильтры Калмана) для систем управления летательными аппаратами и разработал соответствующую математическую теорию. Сформулировал принципы статистической фильтрации. Указал на возможность применения методов современной алгебры к теории динамических систем.

Угловая модуляция, в свою очередь, в соответствии с изменением угла косинуса несущего колебания делится на *фазовую и частотную*. Функция параметров сигнала, который воссоздает информационное сообщение, настолько медленно изменяется, что ее можно считать *низкочастотной, или медленно изменяющейся*.

Модулированные колебания в общем случае не являются периодическими и принадлежат к квазигармоническим, почти периодическим функциям. Такие функции могут быть разложенные в тригонометрический ряд и представлены суммой гармонических составляющих, частоты которых в общем случае не являются кратными, а представляют собой комбинации частот и *называются комбинационными*. В отличие от такого ряда, ряд Фурье содержит гармонические составляющие с кратными частотами.

Амплитудная модуляция (АМ). АМ, принадлежащая к простейшим типам модуляции, получила широкое применение благодаря своей простоте технической реализации и использования.

При АМ амплитуда несущего колебания является функцией времени и имеет вид

$$A_m(t) = A_0[1 + F(t)],$$

где A_0 - постоянная составляющая (среднее значение амплитуды модулированного колебания); $F(t)$ - функция времени, изменяющаяся согласно закону информационного сообщения (информационного сигнала), является *модулирующей функцией несущего колебания*.

Для *однотональной модуляции* формула АМ колебания записывается как:

$$U(t) = A_0\left[1 + F(t)\right]\cos\left(\omega_н t + \psi_н\right).$$

Если $F(t) = U_m m\cos\left(\Omega t - \varphi_0\right)$, то формула для однотональной АМ при гармоническом законе изменения информационного сигнала приобретает вид

$$U_{АМ}(t) = A_0\left[1 + m\cos\left(\Omega t - \varphi_0\right)\right]\cos\left(\omega_н t + \psi_н\right), \quad (4.48)$$

где m - коэффициент модуляции; Ω - частота моду-

лированного колебания (низкочастотная составляющая); φ_0 - начальная фаза модулированного колебания. При условии $\omega_i \square \Omega$ *коэффициент модуляции* изменяется в пределах от нуля до единицы, или в относительных единицах от нуля до 100 %. Коэффициент модуляции пропорционален интенсивности переданного сигнала; его называют также *глубиной модуляции*.

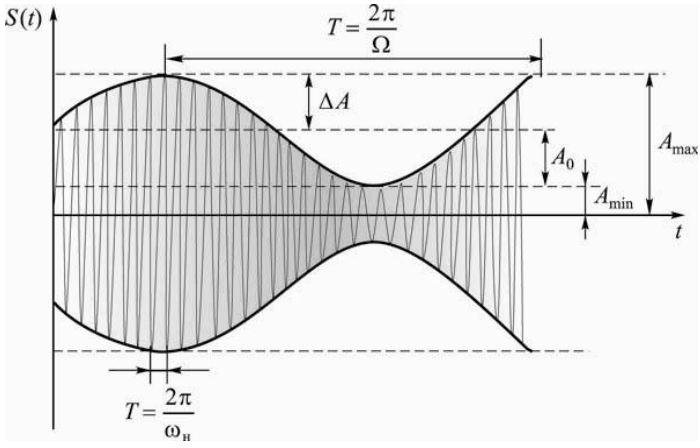


Рис. 4.35. Амплитудно-модулированное колебание (однотональная АМ)

При $0 < m < 1$ амплитуда АМ колебания не приобретает отрицательных значений. Такая модуляция называется *неискаженной* (рис. 4.35). При неискаженной модуляции амплитуда АМ колебания изменяется в пределах от $A_{m\min} = A_{m0}\left[1-m\right]$ до $A_{m\max} = A_{m0}\left[1+m\right]$. При этом коэффициент модуляции является отношением максимального увеличения ΔA амплитуды колебаний к среднему ее значению A_0 :

$$m = \frac{A_{\max} - A_0}{A_0} = \frac{A_0 - A_{\min}}{A_0} = \frac{\Delta A}{A_0},$$

где ΔA - глубина модуляции.

При m >1 значение $A_m(t)$ на некоторых интервалах времени становится отрицательным (см. рис. 4.35). В результате будет происходить так называемая *перемодуляция*, которая приводит к искажению информационной функции АМ колебания (рис. 4.36). Во избежание этого, на практике коэффициент модуляции выбирают не больше единицы.

В режиме перемодуляции происходит искажение информационного сигнала, что недопустимо при передаче информационного сообщения.

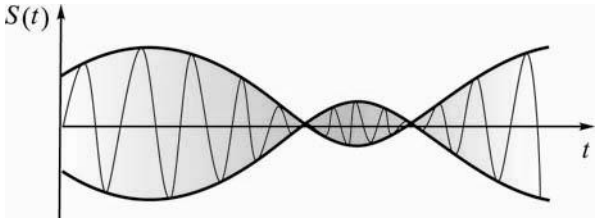


Рис. 4.36. Режим перемодуляции АМ колебания

Спектры амплитудно-модулированных колебаний. При модуляции даже простейшим гармоническим сигналом АМ колебание является сложным сигналом, содержащим ряд гармонических составляющих. На основе тригонометрических преобразований разложим АМ колебание на сумму элементарных функций:

$$U_{\text{АМ}}(t) = A_0\left[1 + m\cos(\Omega t - \varphi_0)\right]\cos(\omega_{\text{н}} t + \psi_{\text{н}}) =$$

$$= A_0\cos(\omega_{\text{н}} t + \psi_{\text{н}}) + A_0 m\cos(\Omega t - \varphi_0)\cos(\omega_{\text{н}} t + \varphi_{\text{н}}) =$$

$$= A_0\cos(\omega_{\text{н}} t + \psi_{\text{н}}) + \frac{A_0 m}{2}\cos[(\omega_{\text{н}} - \Omega)t + \psi_{\text{н}} - \varphi_0] +$$

$$+ \frac{A_0 m}{2}\cos[(\omega_{\text{н}} + \Omega)t + \psi_{\text{н}} + \varphi_0],$$

где $\omega_{\text{н.б}} = \omega_{\text{н}} - \Omega$ - частота нижней боковой составляющей; $\omega_{\text{в.б}} = \omega_{\text{н}} + \Omega$ - частота верхней боковой составляющей; $\varphi_{\text{н.б}} = \psi_{\text{н}} - \varphi_0$ - фаза нижней боковой составляющей; $\varphi_{\text{в.б}} = \psi_{\text{н}} + \varphi_0$ - фаза верхней боковой составляющей.

Первое слагаемое - несущее колебание с частотой $\omega_{\text{н}}$. Второе и третье слагаемые соответствуют новым гармоническим составляющим, которые появляются в процессе модуляции амплитуды. Они являются продуктом модуляции и называются *боковыми гармоническими составляющими*. Частоты этих колебаний $\omega_{\text{н}} + \Omega$ и $\omega_{\text{н}} - \Omega$ называют боковыми (верхней и нижней боковой частотой соответственно). Амплитуды этих составляющих - одинаковы (рис. 4.37), а фазы - симметричны фазе несущего колебания.

Амплитуды боковых гармонических составляющих зависят от глубины модуляции. Чем меньше коэффициент *m*, тем меньше амплитуда колебаний боковых частот, причем у границы при отсутствии модуляции ($m = 0$) они отсутствуют.

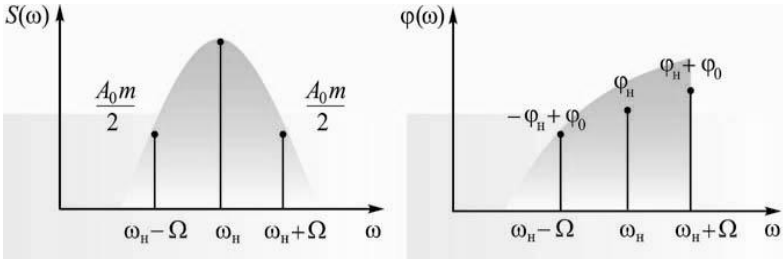


Рис. 4.37. АЧС та ФЧС однотональных амплитудно-модулированных колебаний

Введем понятие *полной фазы* и *мгновенной частоты* для модулированных колебаний (рис. 4.38).

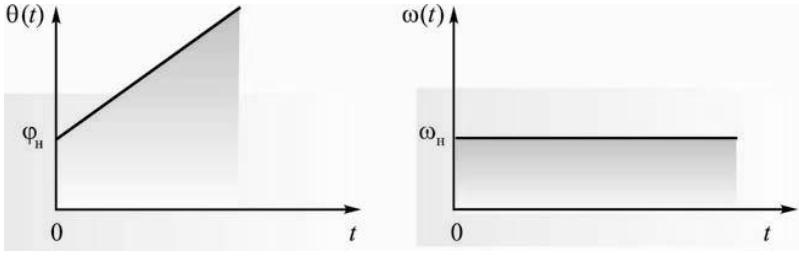


Рис. 4.38. Полная фаза $\theta(t)$ и мгновенная частота $\omega(t)$ АМ колебания

Полная фаза $\theta(t)$ для АМ колебания:

$$\theta(t) = \omega_н t + \varphi_н . \tag{4.49}$$

Мгновенная частота $\omega(t)$ - скорость изменения полной фазы косинуса несущего колебания во времени (первая производная от полной фазы):

$$\omega(t) = \frac{d\theta(t)}{dt} = \omega_н . \tag{4.50}$$

Определим мощность составляющих для АМ колебания.

Мгновенная мощность АМ сигнала равна квадрату суммарного напряжения: $P_{АМ}(t) = U^2_{АМ}(t)/R$, при $R = 1$ Ом;

$$P_{АМ}(t) = \frac{\left[U_{нес}(t) + U_{н.б}(t) + U_{в.б}(t)\right]^2}{1} = U^2_{нес}(t) + U^2_{н.б}(t) + U^2_{в.б}(t) + \tag{4.51}$$

$$+2U^2_{нес}(t)U^2_{н.б}(t) + 2U^2_{н.б}(t)U^2_{в.б}(t) + U^2_{нес}(t)U^2_{в.б}(t).$$

Суммарная мощность сигнала определяется путем усреднения мощностей на довольно большом отрезке времени:

$$\langle P_{\text{АМ}} \rangle = \lim_{T\to\infty} \int_0^T P_{\text{АМ}}(t)\, dt\,. \tag{4.52}$$

Подставив в выражение для определения средней мощности выражение для мгновенной мощности, получим

$$\begin{aligned}\frac{1}{T}\int_0^T P_{\text{АМ}}(t)dt = \frac{1}{T}\Bigg[&\int_0^T U_{\text{нес}}^2 dt + \int_0^T U_{\text{н.б}}^2 dt + \int_0^T U_{\text{в.б}}^2 dt + \\ &+2\int_0^T U_{\text{нес}} U_{\text{в.б}} dt + 2\int_0^T U_{\text{нес}} U_{\text{н.б}} dt + 2\int_0^T U_{\text{н.б}} U_{\text{в.б}} dt\Bigg].\end{aligned} \tag{4.53}$$

Рассмотрим первый интеграл:

$$P_{\text{нес}} = \frac{U_m^2}{T}\int_0^T \cos^2\left(\omega_{\text{н}} t + \varphi_{\text{н}}\right) dt = \ldots = \frac{U_m^2}{T}, \qquad T = \frac{2k\pi}{\omega_{\text{i}}};$$

$$P_{\text{в.б}} = \frac{A_m^2 m^2}{T\cdot 4}\int_0^T \cos^2\left(\left(\omega_{\text{н}} + \Omega\right)t + \varphi_{\text{н}} + \varphi_{\text{в.б}}\right) dt = \ldots = \frac{A_m^2 m^2}{8};$$

$$P_{\text{н.б}} = \frac{A_m^2 m^2}{T\cdot 4}\int_0^T \cos^2\left(\left(\omega_{\text{н}} + \Omega\right)t + \varphi_{\text{н}} + \varphi_{\text{н.б}}\right) dt = \ldots = \frac{A_m^2 m^2}{8};$$

$$\frac{2}{T}\int_0^T U_{\text{н.б}} U_{\text{в.б}} dt = \frac{2}{T}\int_0^T U_{\text{нес}} U_{\text{н.б}} dt = \frac{2}{T}\int_0^T U_{\text{нес}} U_{\text{в.б}} dt = 0\,.$$

Средняя мощность АМ колебания $(A_m = U_m)$:

$$\left\langle P_{\text{АМ}} \right\rangle = \left\langle P_{\text{нес}} \right\rangle + \left\langle P_{\text{н.б}} \right\rangle + \left\langle P_{\text{в.б}} \right\rangle = \frac{U_m^2}{2} + 2\frac{A_m^2 m^2}{8} = \frac{U_m^2}{2} + \frac{A_m^2 m^2}{4}\,. \tag{4.54}$$

Из формулы (4.54) следует, что даже при 100 % АМ модуляции слагаемое, соответствующее мощности боковых составляющих, занимает лишь 50 % мощности всего сигнала. Поскольку информация заложена только в боковых составляющих, то такая форма передачи информации невыгодна с точки зрения мощности характеристик.

Многотональная АМ. В том случае, когда *модулирующий информационный сигнал* (сигнал управления) является сложной функцией и представляется базисом Фурье, уравнение для АМ колебания имеет вид

$$U_{\text{АМ}}(t) = U_m\left[1 + \sum_{i=1}^{\infty} m_i \cos\left(i\Omega t - \varphi_i\right)\right]\cos\left(\omega_{\text{н}} t + \varphi_{\text{н}}\right) \tag{4.55}$$

и является выражением для многотональной АМ (рис. 4.39).

Коэффициенты $m_1, m_2, \ldots m_k$ называют *парциальными, или частичными, коэффициентами модуляции*. Они характеризуют влияние составляющих мо-

дулирующего колебания с частотами Ω_1, Ω_2, Ω_3, … Ω_k на общее изменение амплитуды модулированного (несущего) колебания.

Во избежание перемодуляции сумма этих коэффициентов $m_1, m_2, ... m_k$ (допускаем, что в какой-либо момент отрицательный максимум всех гармоник совпадает) не должна превышать единицу, т.е. $\sum_{k=1}^{n} m_k \le 1$.

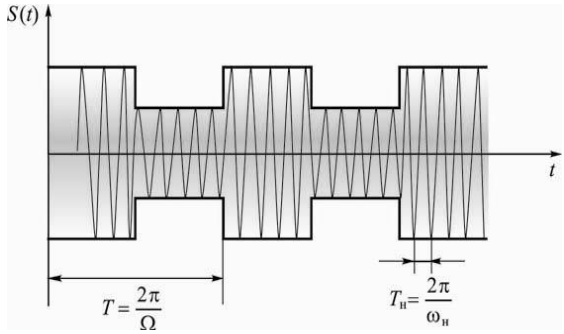


Рис. 4.39. Многотональная АМ

Амплитудно-частотный спектр (АЧС) многотональной АМ. Если *модулирующий информационный сигнал* (сигнал управления) является сложной функцией (4.55), то физический процесс образования спектральных составляющих АЧС не изменяется относительно однотональной АМ, причем каждая его гармоническая составляющая приводит к появлению пары соответственно верхних и нижних боковых частот. Результатом этого процесса является спектр, который состоит из двух полос частот, размещенных симметрично относительно несущей частоты $\omega_н$. Эти полосы частот, размещенные с обеих сторон от несущей, называются *верхней и нижней боковыми полосами* (рис. 4.40).

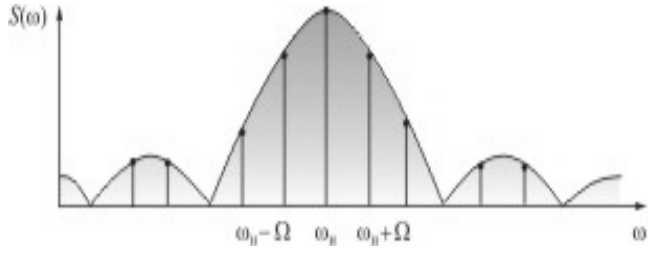


Рис. 4.40. АЧС многотональной АМ

Сравнивая спектры функций модулирующих сигналов, приходим к выводу, что спектр верхней боковой полосы АМ колебания подобен спектру информационного сигнала. При АМ происходит лишь трансформация спектра

сигнала по оси частот на величину $\omega_н$. Если полоса частот модулирующего сигнала ограничена частотой Ω_{max}, то соответствующий ему АМ сигнал будет иметь спектр (см. рис. 4.40), ширина которого вдвое больше частоты Ω_{max}:

$$\Delta\omega_с = 2\Omega_{max}.$$

Аналитическое выражение для АЧХ многотональной АМ можно получить на основе тригонометрических преобразований. Это выражение является показательным для представления АЧХ многотональной модуляции, так как разделяет указанную функцию АМ $U_{AM}(t)$ на сумму трех составляющих: несущей частоты, верхней и нижней боковых частотных полос:

$$U_{АМ}(t) = U_m\left[1+\sum_{i=1}^{\infty} m_i\cos(i\Omega t-\varphi_i)\right]\cos(\omega_н t+\varphi_н) = U_m\cos(\omega_н t+\varphi_н)+$$

$$+U_m\sum_{i=1}^{\infty} m_i\cos(i\Omega t-\varphi_i)\cos(\omega_н t+\varphi_н) = U_m\cos(\omega_н t+\varphi_н)+$$

$$+\underbrace{\frac{U_m}{2}\sum_{i=1}^{\infty} m_i\cos[(\omega_н - i\Omega)t+\varphi_н-\varphi_н]}_{\text{СПЕКТРАЛЬНЫЕ СОСТАВЛЯЮЩИЕ НИЖНЕЙ ПОЛОСЫ СПЕКТРА}} + \underbrace{\frac{U_m}{2}\sum_{i=1}^{\infty} m_i\cos[(\omega_н + i\Omega)t+\varphi_н+\varphi_i]}_{\text{СПЕКТРАЛЬНЫЕ СОСТАВЛЯЮЩИЕ ВЕРХНЕЙ ПОЛОСЫ СПЕКТРА}}.$$

Принцип работы амплитудного модулятора. Амплитудный модулятор создает на выходе АМ сигнал типа $U_{АМ}(t)=U_m(1+m\cos\Omega t)\cos\omega_н t$ при подаче на входы цепи гармонического несущего колебания $U_н(t)= =U_н\cos\omega_н t$ и низкочастотного моделирующего сигнала $U_{мод}(t)= =U_m\cos\Omega t$. Чаще всего амплитудные модуляторы строят, используя эффект преобразования спектра произведения двух сигналов в безинерционном линейном элементе. Простейшим амплитудным модулятором является умножитель, у которого резонансный контур в исходной цепи настроен на частоту исходного колебания $\omega_н$. К входу модулятора подают напряжение $U_{вх}(t)= U_0\cos\Omega t\ U_м\cos\omega_н t$. Принцип работы делает наглядным рис. 4.41.

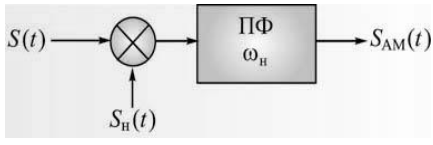


Рис. 4.41. Схема амплитудного модулятора

Балансная модуляция (БМ) формируется как разновидность АМ колебаний с подавленной составляющей спектра на частоте $\omega_н$.

В схеме фазовычитающего (фазоразностного) модулятора происходит затухание одной из боковых полос, а мощность другой боковой полосы удваивается. Недостатком является сложность фазовращателя (ФВ) для всей полосы частот модулирующего сигнала (рис. 4.42).

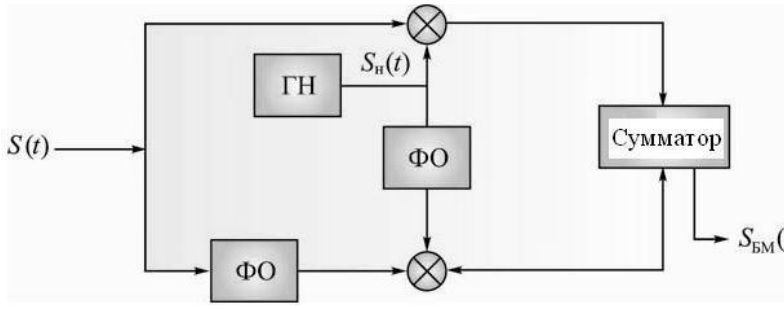


Рис. 4.42. Схемы некогерентного и квазикогерентного детектирования:
$S_{AM}(t)$ - амплитудно-модулированный сигнал; $S(t)$ - информационный сигнал;
$U_m \cos(\omega t)$ - опорный сигнал; Д - детектор информационного сообщения,
ФНЧ - фильтр низкой частоты

Несущая частота не переносит информационный сигнал, но на нее приходится значительная часть мощности сигнала АМ. Поэтому в ряде случаев несущую подавляют. Подавление несущей может быть полным или частичным. Сигнал БМ формируется перемножением несущей и модулирующего сигнала:

$$S_н(t) = U_н \cos(\omega_н t) \quad \text{и} \quad S(t) = U_0 \cos(\Omega t),$$

$$S_н(t) S(t) = 0{,}5 U_н U_0 \left\{ \cos\left[(\omega_н - \Omega) t \right] + \cos(\omega_н + \Omega) t \right\}.$$

Однополосная модуляция (ОМ) является разновидностью АМ без несущей или АМ с одной боковой полосой (АМ-ОБП). Такой вид модуляции можно получить с помощью линейного модулятора (рис. 4.43). Обогащенный сигнал после перемножения поступает на полосовой фильтр (ПФ). Задача ПФ - выделить верхнюю боковую полосу частот в диапазоне $\omega_н < \omega \le \omega_в$.

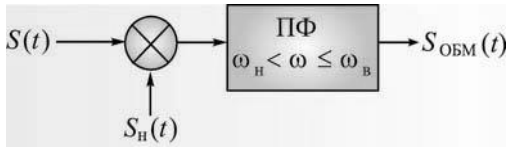


Рис. 4.43. Схема линейного модулятора

Недостатками АМ и, в частности, линейного модулятора являются:

необходимость подавления несущей; в АМ сигнале информация дублируется двумя боковыми полосами; сложность выполнения полосового фильтра.

Указанные недостатки большей частью устраняются при использовании *фазовычитающей* (фазоразностной) схемы (см. рис. 4.42).

Принцип детектирования АМ-сигналов. *Демодуляцией* называют процесс детектирования, т.е. выявление информационного сообщения на фоне параметров ВЧ несущего колебания.

Когерентными называют методы приема (демодуляции), для реализации которых необходимо точное априорное знание начальных фаз сигналов, действующих на входе.

Квазикогерентным называют прием в случае, когда данные о начальных фазах ожидаемых сигналов берутся из принятого сигнала.

Некогерентным называют прием, если сведения о начальных фазах сигналов, поступивших к приемнику, отсутствуют или их не используют (рис. 4.44). Задача ФНЧ - изъять из обогащенного спектра диапазон информационного сигнала от $0 < \omega \le \omega_в$.

Операция амплитудного детектирования прямо пропорциональна АМ. Имея на входе идеального детектора АМ колебание $S_{вх}(t) = U_{max}(1 + m\cos\Omega t)\cos\omega_н t,$ нужно получить на выходе низкочастотный сигнал $S_{вых}(t) = U_{m_{вых}}\cos\Omega t,$ пропорциональный передаваемому сообщению. Эффективность работы детектора оценивают *коэффициентом детектирования*

$$k_{дет} = U_{m_{вых}} / mU_{m_{вх}},$$

который определяется отношением амплитуды низкочастотного сигнала на выходе к изменению амплитуды высокочастотного сигнала на входе.

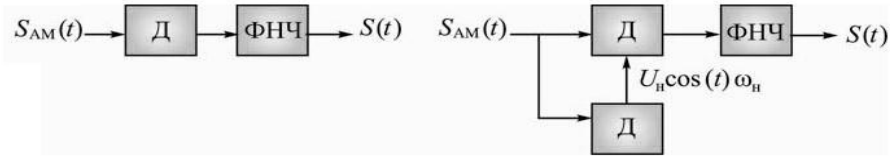


Рис. 4.44. Схемы некогерентного и квазикогерентного детектирования:
$S_{AM}(t)$ - амплитудно-модулированный сигнал; $S(t)$ - информационный сигнал;
$U_m\cos(\omega t)$ - опорный сигнал; Д - детектор информационного сообщения;
ФНЧ - фильтр низкой частоты

Можно осуществить детектирование, подав АМ сигнал на безынерционный нелинейный элемент и предусмотрев дальнейшую фильтрацию низкочастотных составляющих спектра. На выходе схемы осуществляется низкочастотная фильтрация. Для того чтобы цепь нагрузки выполняла роль частотно-

го фильтра, подавляющего высокочастотные спектральные составляющие, нужно выполнить неравенство

$$1/\omega_{н}C_{н} << R_{н},\ 1/\Omega C_{н} >> R_{н}.$$

Пусть входное напряжение на базе нелинейных систем подается в виде

$$S_{\text{вх.нел}}(t) = U_0 + U_{m_{\text{вх}}}(1 + m\cos\Omega t)\cos\omega_{н} t,$$

причем амплитуда $U_{m_{\text{вх}}}$ достаточно велика, для того чтобы воспользоваться частично-линейной аппроксимацией вольт-амперной характеристики нелинейной системы. Для упрощения выберем рабочую точку нелинейной характеристики системы $U_0 = U_{\text{í}}$ и угол отсечки тока $\vartheta = 90°$ независимо от изменения во времени амплитуды входного сигнала. Процессы в нелинейном детекторе изображены на рис. 4.45, а схема - на рис. 4.46.

Последовательность импульсов тока модулирована по амплитуде; нулевая составляющая тока медленно (с частотой Ω) изменяется во времени:

$$I_{\text{вых.нел}}(t) = S_x U_{m_{\text{вх}}}(1 + m\cos\Omega t)\gamma_0(90°) = 0{,}318 S U_{m_{\text{вх}}}(1 + m\cos\Omega t),$$

где S_x - крутизна нелинейной характеристики системы.

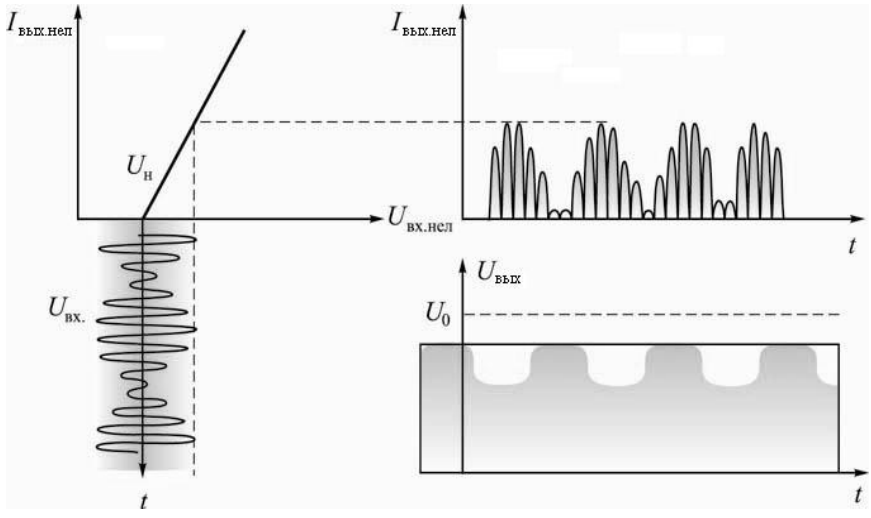


Рис. 4.45. Осциллограммы токов и напряжений на входе детектора

Исходное напряжение детектора

$$I_{\text{вых.нел}}(t) = U_0 - I_{\text{вых.нел}}(t)R_{н} = U_0 - 0{,}318 S U_{m_{\text{вх}}}(1 + m\cos\Omega t),$$

а коэффициент детектирования $k_{\text{дет}} = 0{,}318 S R_{\text{н}}$. Существенно, что амплитуды сигналов на входе и на выходе связаны прямо пропорционально. Поэтому

такой режим работы детектора называют *линейным* - из-за отсутствия искажений передаваемого сообщения.

Довольно широкий диапазон частот, занимаемый АМ сигналами, является недостатком этого вида модуляции. К другим недостаткам АМ следует отнести низкую помехозащищенность и низкую экономичность радиопередатчиков.

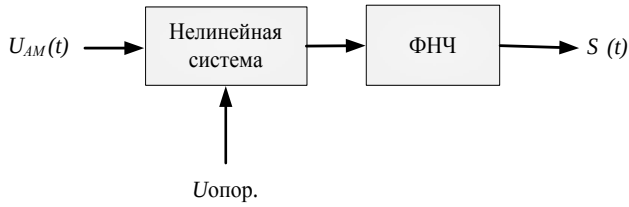


Рис. 4.46. Структурная схема нелинейного детектора

Указанные недостатки устраняются или в значительной мере снижаются при других видах модуляции, в частности при угловой модуляции.

Фазовая модуляция. *ФМ - это процесс влияния модулирующего сигнала на начальную фазу ВЧ несущего колебания с целью изменения ее по закону информационного сообщения.*

При ФМ начальная фаза несущего колебания становится переменной величиной по закону информационного сигнала $\psi(t) = F(t) + \psi_{н0}$, где $\psi_{н0}$ - постоянная величина начальной фазы; $F(t)$ - модулирующая функция, воссоздающая информационный сигнал.

Аналитически ФМ колебание в общем случае описываются как

$$U_{ФМ}(t) = A_{m0}\cos\left[\omega_н t + F(t) + \psi_{н0}\right] = A_{m0}\cos\left[\theta(t)\right],$$

где амплитуда неизменна, а фазовый угол изменяется во времени.

При гармонической (однотональной) модуляции при

$$\psi(t) = \Delta\psi\cos(\Omega t + \varphi_0) + \psi_{н0} = m\cos(\Omega t + \varphi_0) + \psi_{н0},$$

$$F(t) = \Delta\psi\cos(\Omega t + \varphi_0),$$

для ФМ колебания получим

$$U_{ФМ}(t) = U_m\cos\left[\omega_н t + m_\psi\cos(\Omega t - \varphi_0) + \varphi_н\right].$$

Величина m выражает максимальное отклонение фазы при модуляции и называется *фазовым отклонением* (*девиацией*, рис. 4.47), или *индексом фазовой модуляции*, который пропорционален амплитуде модулирующего сигнала и его интенсивности $m = \Delta\psi$, где $\Delta\psi$ - девиация фазы - отклонение фазы несущего колебания от фазы модулирующего.

Полная фаза *однотональной ФМ модуляции*:

$$\theta(t) = \omega_н t + m_\psi\cos(\Omega t - \varphi_0) + \varphi_н. \qquad (4.56)$$

Мгновенная частота *однотональной ФМ модуляции*:

$$\omega(t) = \frac{d\theta(t)}{dt} = \omega_{н} - m_{\psi}\Omega \sin(\Omega t - \varphi_0) \ . \qquad (4.57)$$

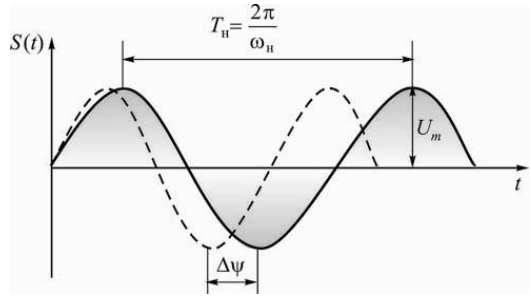


Рис. 4.47. Девиация фазы

С увеличением значений информационного сигнала полная фаза растает во времени быстрее, чем по линейному закону, а при уменьшении значения моделирующего сигнала происходит спад скорости роста полной фазы во времени (рис. 4.48).

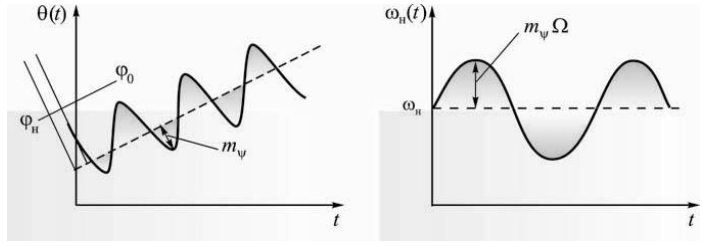


Рис. 4.48. Полная фаза и мгновенная частота для однотональной ФМ модуляции

При многотональной ФМ модуляции модулирующий сигнал имеет сложную форму функции информационного сообщения и может быть представлен тригонометрическим рядом в виде

$$\psi(t) = \psi_{н0} + \sum_{k=1}^{n} m_k \cos(\Omega_k t + \varphi_k) \, .$$

Для полной аналитической формы записи ФМ колебания получаем более сложное выражение

$$U_{фм}(t) = U_m \cos\left[\omega_{н} t + \sum_{k=1}^{n} m_k \cos(\Omega_k t + \varphi_k) + \psi_{н0}\right],$$

где m_k - парциальные индексы фазовой модуляции.

Карл Густав Яков Якоби (Carl Gustav Jacob Jacobi, 1804 - 1851), немецкий математик, член Берлинской АН (1836), член-корреспондент (1830) и почетный член (1833) Петербургской АН. Один из творцов теории эллиптических функций. Ввел и изучил тэта-функции и некоторые другие трансцендентные функции. Ввел функциональные определители и указал на их роль при замене переменных в кратных интегралах и при решении уравнений с частными производными. Исследовал класс ортогональных многочленов, что является обобщением многочленов Лежандра.

Фазовое детектирование. Известно много схем фазовых детекторов-устройств для демодуляции колебания с полной фазой

$$\psi(t) = \omega_0 t + \varphi(t),$$

промодулированных по фазовому углу. Работа таких детекторов базируется на нелинейном взаимодействии модулированного сигнала с немодулированным опорным колебанием, которое должно создаваться вспомогательным внешним источником.

Пример. К нелинейному безынерционному двухполюснику с вольт-амперной характеристикой

$$i(u) = a_0 + a_1 u + a_2 u^2$$

приложено сумму двух напряжений

$$u(t) = u_1 + u_2 = U_{m1}\sin[\omega_0 t + \varphi(t)] + U_{m2}\cos\omega_0 t\,.$$

Посредством квадратичного слагаемого характеристики будет присутствовать составляющая, описывающая нелинейное взаимодействие колебаний

$$i_{\text{вз}}(t) = 2a_0 U_{m1} U_{m2}\sin[\omega_0 t + \varphi(t)]\cos\omega_0 t =$$

$$= a_2 U_{m1} U_{m2}\sin\varphi(t) + a_2 U_{m1} U_{m2}\sin[2\omega_0 t + \varphi(t)].$$

Второму слагаемому в последней части приведенной формулы соответствует высококачественный сигнал со средней частотой $2\omega_0$, который подавляется линейным фильтром нижних частот (например, *RC-цепью*). Первое слагаемое описывает низкочастотный ток, пропорциональный передаваемому сообщению, если девиация фазы (индекс модуляции) детектированного сигнала достаточно невелика:

$$i_{\text{н.ч}}(t) = a_2 U_{m1} U_{m2}\sin\varphi(t) \approx a_2 U_{m1} U_{m2}\varphi(t).$$

При создании фазовых детекторов неминуемы трудности, связанные с требованиями жесткой стабилизации фазы колебаний сопротивляющегося генератора.

Частотная модуляция. *ЧМ – это процесс влияния модулирующего сигнала на мгновенное значение частоты ВЧ несущего колебания с целью изменения его по закону информационного сообщения*

$$\omega(t) = \omega_{\text{н}} + F(t).$$

При ЧМ между величинами информационного сигнала и полной фазы несущего колебания устанавливается зависимость вида

$$\omega(t) = \omega_н + \Delta\omega\cos(\Omega t + \varphi_0),$$
$$F(t) = \Delta\omega\cos(\Omega t + \varphi_0),$$

где $F(t)$ - функция информационного сигнала.

Общее аналитическое выражение для однотональной ЧМ модуляции:

$$U_{ЧМ}(t) = U_m\cos\left\{\int_0^t\left[\omega_н + \Delta\omega\cos(\Omega t + \varphi_0)\right]dt + \psi_{н0}\right\} =$$

$$= U_m\cos\left[\omega_н t + \frac{\Delta\omega}{\Omega}\sin(\Omega t + \varphi_0) + \psi_{н0}\right] = U_m\cos\left[\omega_н t + m_\omega\sin(\Omega t + \varphi_0) + \psi_{н0}\right].$$

Девиация частоты $\Delta\omega$ - отклонение частоты несущего колебания от номинальной частоты (рис. 4.49). Она пропорциональна интенсивности модулирующего сигнала и информативно эквивалентна коэффициенту модуляции m при АМ и фазовому отклонению (девиации) $m_ф = \Delta\psi$ при ФМ.

Отношение $m_\omega = \Delta\omega/\Omega$ называется *индексом частотной модуляции*. Индекс m_ω пропорционален девиации частоты и, в отличие от индекса ФМ, зависит от соотношения девиации $\Delta\omega$ и частоты модуляции и обратно пропорционален ей. Коэффициент модуляции всегда больше 1.

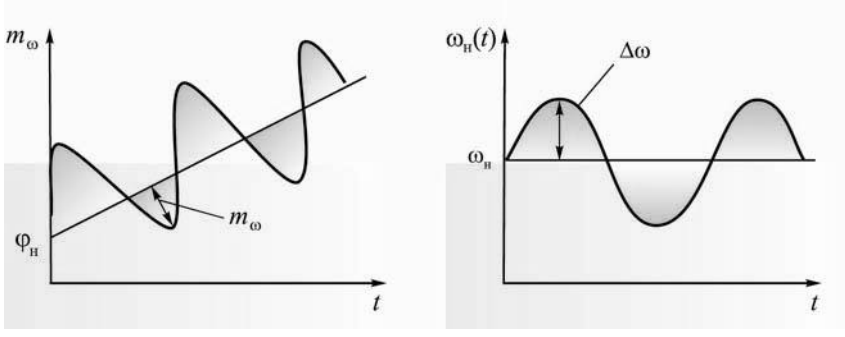


Рис. 4.49. Полная фаза и мгновенная частота для ЧМ модуляции

Полная фаза для ЧМ: $\theta(t) = \omega_н t + m_\omega\sin(\Omega t - \varphi_0) + \varphi_н$.

При многотональной ЧМ модуляции, когда модулирующий сигнал - сложное информационное сообщение, представленное рядом Фурье

$$\omega(t) = \omega_н + \sum_{k=1}^{n}\Delta m_k\cos(\Omega_k t + \varphi_k),$$

то для ЧМ колебания получим общую аналитическую форму записи в виде

$$U_{\text{ЧМ}}(t) = A_{m0}\cos\left[\omega_{\text{н}}t + \sum_{k=1}^{n}\frac{\Delta\omega_k}{\Omega_k}\sin(\Omega_k t + \varphi_k) + \psi_{\text{н}}\right] =$$

$$= A_{m0}\cos\left[\omega_{\text{н}}t + \sum_{k=1}^{n} m_{\omega k}\sin(\Omega_k t + \varphi_k) + \psi_{\text{н}}\right],$$

где $m_{\omega k}$ - парциальные индексы частотной модуляции.

Спектр сигналов с угловой модуляцией. Сравним полученные общие выражения для ФМ и для ЧМ колебаний:

$$U_{\text{ФМ}}(t) = A_{m0}\cos\left[\omega_{\text{н}}t + F(t) + \psi_{\text{н}0}\right],$$

$$U_{\text{ЧМ}}(t) = A_{m0}\cos\left[\omega_{\text{н}}t + \int_0^t F(t) + \psi_{\text{н}0}\right].$$

В обоих случаях физическая сущность явления аналогична: полная фаза угла косинуса ВЧ колебания изменяется по закону информационного сообщения, тем не менее, соотношения между фазовым углом и модулирующим сигналом разные**.**

В самом деле, при ФМ фазовый угол $\psi(t)$ пропорционален модулирующему сигналу

$$\psi(t) = F(t) + \psi_{\text{н}0},$$

а соответствующая мгновенная частота пропорциональна производной сигнала

$$\omega(t) = \frac{d\theta(t)}{dt} = \frac{d}{dt}\left[\omega_{\text{н}}t + F(t) + \psi_{\text{н}0}\right] = \left[\omega_{\text{н}}t + \frac{dF(t)}{dt}\right].$$

В случае же ЧМ, мгновенная частота пропорциональна модулирующему сигналу, а фазовый угол пропорционален интегралу сигнала:

$$\omega(t) = \frac{d\theta(t)}{dt} = \frac{d}{dt}\left[\omega_{\text{н}}t + \int_0^t F(t) + \psi_{\text{н}0}\right] = \left[\omega_{\text{н}} + F(t)\right],$$

$$\psi(t) = \int_0^t F(t)dt + \psi_{\text{н}0}.$$

Изложенные соображения важны при использовании ФМ и ЧМ, а также при построении их частотных спектров.

Пример. Если колебание модулируется по фазе сигналом, предварительно прошедшим через интегрирующую цепь, то образовывается колебание, модулированное по частоте исходным сигналом. Таким способом, в частности, с помощью ФМ формируют ЧМ колебание.

Как уже отмечалось, модулированное колебание, представленное в виде

$$U(t) = U_m\cos(\omega_{\text{н}}t + m\cos(\Omega t + \varphi_0) + \varphi_{\text{н}}), \qquad (4.58)$$

одинаково может касаться как ЧМ, так и ФМ, в зависимости от природы модуляции. Учитывая то, что

$$\cos\left(\omega_{н}t + m\sin\left(\Omega t + \varphi_0\right) + \varphi_{н}\right) = \mathrm{Re}\left[e^{j\left(\omega_{н}t + m\sin(\Omega t + \varphi_0) + \varphi_{н}\right)}\right], \tag{4.59}$$

выражение (4.59) представим как

$$U(t) = U_m\,\mathrm{Re}\left[e^{j\left(\omega_{н}t + m\sin(\Omega t + \varphi_0) + \varphi_{н}\right)}\right]. \tag{4.60}$$

Из теории функций Бесселя известно, что

$$e^{jm\sin x} = \sum_{-\infty}^{\infty} I_k\left(m\right)e^{jkx}, \tag{4.61}$$

$$\cos\left(y\sin x\right) = J_0\left(y\right) + 2J_2\left(y\right)\cos 2x + 2J_i\left(y\right)\cos 4x + \ldots + \sin\left(y\sin x\right) =$$
$$= 2J_1\left(y\right)\sin x + 2J_3\left(y\right)\sin 3x + 2J_5\left(y\right)\sin 5x + \ldots.$$

Посредством $J_n(y)$ обозначена функция Бесселя первого рода n-го порядка от аргумента y. Применяя эти соотношения с учетом того, что $x = \Omega t + \varphi_0$ $y = m$, после обычных алгебраических преобразований и раскрытия произведения тригонометрических функций получаем

$$\begin{aligned} U\left(t\right) &= U_m\,\mathrm{Re}\left[\sum_{-\infty}^{\infty} I_k\left(m\right)e^{j\omega_{i}t + \varphi_h}e^{jk(\Omega t + \varphi_0)}\right] = \\ &= U_m\,\mathrm{Re}\left[\sum_{-\infty}^{\infty} I_k\left(m\right)e^{j\left[(\omega_{i} + k\Omega)t + \varphi_{i} + \varphi_0 k\right]}\right]. \end{aligned} \tag{4.62}$$

Переходя от действительной части снова к косинусу, получим разложение сигнала с угловой модуляцией в ряд Фурье с учетом функций Бесселя:

$$U\left(t\right) = U_m\sum_{k=-\infty}^{\infty} I_k\left(m\right)\cos\left[\left(\omega_{i} + k\Omega\right)t + \varphi_{i} + \varphi_0 k\right].$$

На основании свойств функций Бесселя

$$I_k\left(m\right) = \left(-1\right)^k I_k\left(m\right) \tag{4.63}$$

получим окончательную форму записи ряда Фурье для сигнала с угловой модуляцией

$$U\left(t\right) = U_m\sum_{k=0}^{\infty}\left(-1\right)^k I_k\left(m\right)\cos\left[\left(\omega_{i} + k\Omega\right)t + \varphi_{i} + \varphi_0 k\right]. \tag{4.64}$$

Значения функций Бесселя разных порядков известны в виде таблиц и графиков (рис. 4.50). Они могут быть вычислены и с помощью разложения в статистический ряд. Спектры ЧМ или ФМ колебаний даже в случае простейшей - гармонической (однотональной) - модуляции имеют бесконечное множество гармоник, которые образовывают верхние и нижние боковые полосы

с частотами $\omega_н \pm \Omega$. В случае АМ колебаний при этом существуют лишь две боковые составляющие.

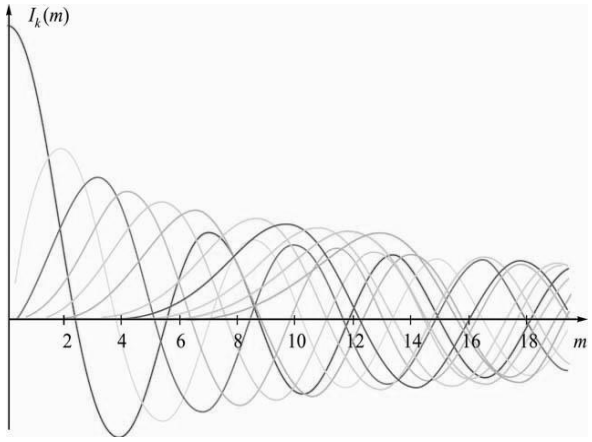


Рис. 4.50. Функции Бесселя

Амплитуда каждой гармонической составляющей частоты $(\omega_н \pm \Omega)$ *в спектре ЧМ и ФМ колебания определяется абсолютным значением функции Бесселя n-го порядка с аргументом* $m = m_\omega = m_\varphi$.

Амплитуды составляющих вычисляются при помощи кривых или таблиц функций Бесселя. Их значения изменяются в зависимости от индекса модуляции *m*. Порядок боковых составляющих спектра определяется номером гармоники *n*.

В качестве примера на рис. 4.51 приведены амплитудно-частотные спектры ЧМ колебания при различных значениях *m*. Анализируя спектры с разным значением индекса модуляции, видим, что с увеличением m порядок составляющих с максимальной амплитудой увеличивается, стремясь к m.

Для небольших индексов модуляции (до $m = 10$) значения максимальных функций Бесселя будут принадлежать гармонике с номером $n = m - 1$. Теоретически колебания с угловой модуляцией занимают бесконечную полосу частот. Тем не менее, для заданного индекса модуляции m абсолютное значение функции Бесселя быстро спадает с ростом k, и практически можно не учитывать боковые составляющие порядка $k = m + 2$. Отсюда вытекает оценка практической ширины спектра для сигналов с угловой модуляцией

$$\prod_{пр} = \Delta f_{пр} = 2(m+1)\Omega. \qquad (4.65)$$

Как правило, сигналы с ЧМ и ФМ передаются при условии $m >> 1$. В этом случае $\Delta f_{пр} = 2m\Omega$. Таким образом сигналы с угловой модуляцией занимают полосу частот, ширина которой равна удвоенной девиации частоты

$\Delta\omega = m\Omega$. При малых значениях коэффициента модуляции $m < 1$ можно взять функцию Бесселя нулевого порядка от $m \approx 1\ (I_0(m) \approx 1)$, а $I_1(m) \approx m/2$. Тогда ширина спектра будет равна 2Ω.

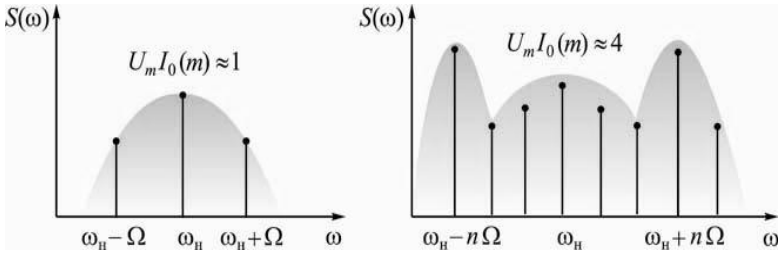


Рис. 4.51. АЧС сигналов с угловой модуляцией с различными значениями коэффициента модуляции *m*

Таким образом, *при угловой модуляции с малыми индексами модуляции* (*узкополосная ЧМ*) *ширина спектра полученного колебания совпадает с шириной спектра в случае АМ. При малых m АЧС спектры ФМ, ЧМ и АМ колебаний практически совпадают*.

Сравнение спектров разных ФМ и ЧМ колебаний показывает, что при одинаковых значениях $\omega_н$, Ω и m их спектры ничем не отличаются. Изменения $\omega_н$ и m вызывают одинаковые изменения в спектрах ФМ и ЧМ колебаний.

Отличительной чертой спектра ЧМ колебания сравнительно с ФМ является независимость его ширины от частоты модуляции. При ЧМ с уменьшением Ω индекс модуляции $m_\omega = \Delta\omega/\Omega$ увеличивается пропорционально Ω, а ширина спектра $\Delta\omega_c = 2m_c\Omega = 2\Omega$ остается неизменной. В случае ФМ индекс модуляции $m_ф = \Delta\psi$ не зависит от ψ, поэтому с изменением количества гармоник он остается неизменным, а ширина спектра изменяется: $\Delta\omega_c = 2m_ф\Omega = 2\Delta\psi\Omega$.

Таким образом, *ЧМ, в отличие от ФМ, характеризуется большим постоянством спектров сигналов. В этом заключается одна из причин более эффективного применения ЧМ на практике.*

Получение сигналов с угловой модуляцией. В 30-х годах Армстронг предложил эффективный способ получения радиосигналов с угловой модуляцией (ЧМ и ФМ сигналов). Структурная схема модулятора изображена на рис. 4.52. Здесь на одном из входов сумматора есть сигнал S_1, который поступает из балансного модулятора БМ. На втором входе сумматора есть немодулированный сигнал S_2 с выхода фазовращателя, который задерживает фазу гармонического сигнала несущей частоты на $90°$ в сторону запаздывания. Таким образом, сигнал на выходе данного модулятора

$$U_{\text{вых}}(t) = U_{m1} S(t) \cos \omega_{\text{н}} t + U_{m2} \sin \omega_{\text{н}} f,$$

где U_{m1} и U_{m2} - постоянные амплитуды.

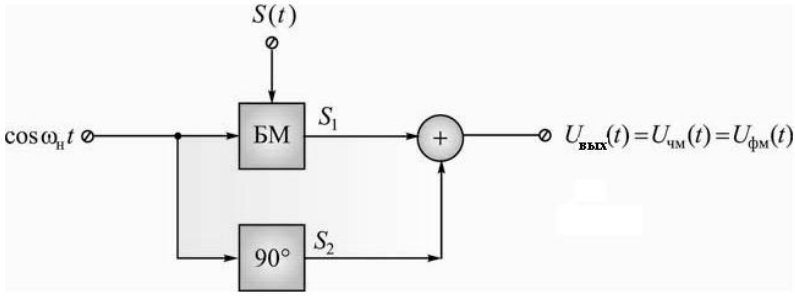


Рис. 4.52. Структурная схема модулятора Армстронга

Частотное детектирование. При частотной модуляции, как известно, полезное сообщение пропорционально отклонению мгновенной частоты сигнала от частоты несущего колебания. Рассмотрим некоторые способы демодуляции ЧМ сигналов.

ЧМ можно преобразовать в неглубокую АМ, подав демодулированный сигнал на линейный частотный фильтр, настроенный таким образом, чтобы в выражении для АЧХ

$$\left|K(j\omega)\right| = \left|K(j\omega_0)\right| + \left|K(j\omega_0)\right|' (\omega - \omega_0) + \ldots$$

коэффициент $\left|K(j\omega_0)\right|'$ был отличным от нуля. Тогда, учитывая, что частота детектированного сигнала $\omega(t) = \omega_{\text{i}} t + m\omega \cos \Omega t$, получим на выходе фильтра сигнал со сложной амплитудно-угловой модуляцией. Мгновенная амплитуда сменной составляющей этого сигнала изменяется во времени по закону

$$U_{m_{\text{вых}}}(t) = B_0 \left|K(j\omega_0)\right|' \Delta\omega \cos \Omega t,$$

где B_0 - постоянный коэффициент, который повторяет по форме передаваемое сообщение. Конечная обработка сигнала проводится обычным АМ детектором, включенным на выходе фильтра.

Такой метод частотного детектирования имеет недостатки: высокие требования к качеству ограничения возможной паразитной АМ на входе фильтра, а также недостаточная линейность характеристики детектирования.

Лучшие результаты обеспечивает способ, который базируется на преобразованиях ЧМ сигналов в ФМ сигнал с помощью линейного частотно-выборочного фильтра с последующим фазовым детектированием. При таком способе демодуляции ФЧХ выборочной узкополосной цепи в небольшом диапазоне частоты имеет вид

$$\varphi_к(\omega) = \varphi_к(\omega_0) - t_{гр}(\omega - \omega_0),$$

где $t_{гр}$ - групповое время задержки.

Если $\omega(t) = \omega_0 + \Delta\omega\cos\Omega t$, то узкополосный сигнал на выходе фильтра имеет полную фазу $\psi(t) = \omega_0 t + \varphi_к(\omega_0) - \Delta\omega t_{гр}\cos\Omega t$, т.е. в действительности является ФМ сигналом.

Сигналы с импульсной модуляцией. *Большей помехоустойчивостью характеризуется импульсная модуляция.* При использовании импульсной модуляции, широко применяемой в информационно-коммуникационных системах, информационным сообщением, которое несет ВЧ колебание, является последовательность радиоимпульсов. В зависимости от вида импульсной модуляции тот или иной параметр такой последовательности изменяется во времени по закону модулированного сигнала. Различают несколько видов импульсной модуляции: амплитудно-импульсную (АИМ); широтно-импульсную (ШИМ, ШИМ-А - ациклическая, ШИМ-С - симметричная); временно-импульсную (ВИМ); кодо-импульсную (КИМ); фазо-импульсную (ФИМ, ФИМ-А - ациклическая), отсчетно-импульсную (ОИМ) (рис. 4.53).

В системах связи с импульсной модуляцией носителем информации является периодическая последовательность импульсов единообразной формы

$$f(t) = \sum_{k=\infty}^{\infty} A_0 U(t - t_k),$$

где $U(t)$ - нормированная функция, характеризующая форму импульса; A_0 - амплитуда импульса; t_k - начало переднего фронта k-го импульса $t_k = kT_i + t_0$; T_i - период поступления импульсов; t_0 - начало отсчета последовательности; τ_k - продолжительность k-го импульса, который ведет свой отсчет на некотором заданном уровне.

При модуляции один из параметров последовательности изменяется соответственно передаваемому сигналу (см. рис. 4.53).

При АИМ изменяется амплитуда импульса

$$A(t) = A_0 + \Delta A u(t).$$

При ШИМ изменяется продолжительность импульса

$$\tau(t) = \tau_0 + 2\Delta\tau_m u(t), \qquad (4.66)$$

где $\Delta\tau_m$ - максимальное отклонение фронта импульсов в одну сторону.

При ФИМ изменяется сдвиг импульсов относительно тактовых точек kT_i:

$$t_k = \theta(t) = kT_i + \Delta\tau_m u(t). \qquad (4.67)$$

При ЧИМ соответственно передаваемому сообщению изменяется частота поступления импульсов. Так же, как при ФИМ, импульсы смещаются относительно тактовых точек, но с другой закономерностью.

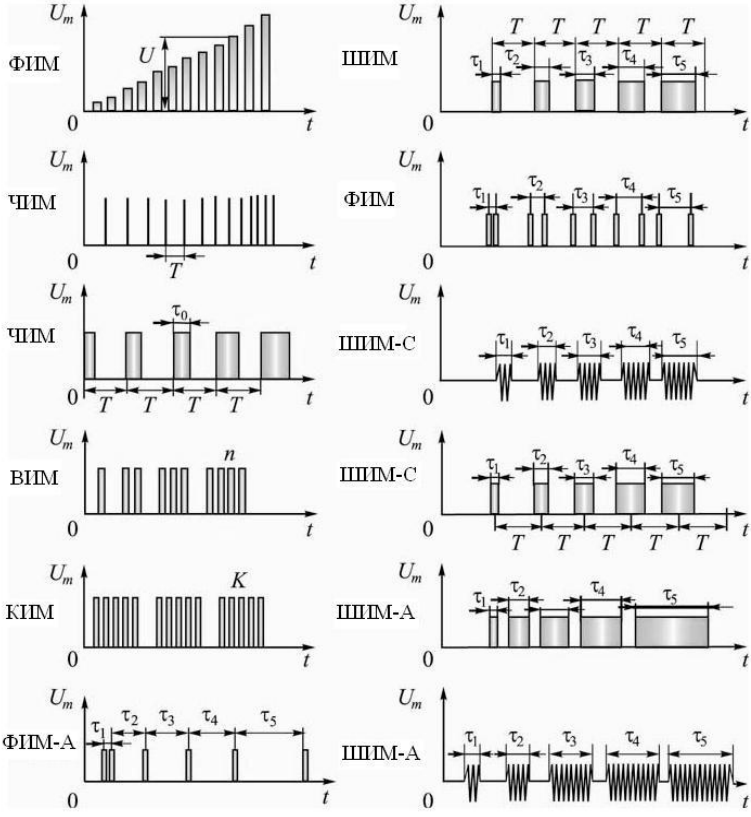


Рис. 4.53. Виды импульсной модуляции

Различие между ФИМ и ЧИМ аналогично различию между ФМ и ЧМ синусоидального носителя. Периодическую последовательность прямоугольных импульсов можно записать как

$$S(t)=\sum_{-\infty}^{\infty}A_0U(t-t_k)=\begin{cases}A_0 \text{ при } \theta_1+kT_l<t<\theta_2+kT_l,\\ 0 \text{ при } \theta_2+kT_l<t<\theta_1+(k+1)T_l\end{cases} \qquad (4.68)$$

Такую последовательность импульсов можно представить в виде ряда Фурье

$$S(t)=\frac{1}{2}\sum_{k=-\infty}^{\infty}A_{\text{к}}e^{ik\omega_1 t},\ A_k=\frac{2}{\Delta t}\int_{-\Delta t/2}^{\Delta t/2}S(t)e^{-i\omega_1 t}dt,\ \omega_1=\frac{2\pi}{\Delta t}.$$

В нашем случае

$$A_k = \frac{2A_0}{\Delta t}\int_{-\tau_0/2}^{\tau_0/2} e^{-ik\omega_1 t}dt = \frac{2A_0\tau_0}{ik\omega_1\Delta t}\left(e^{-i\frac{k\omega_1\tau_0}{2}} - e^{i\frac{k\omega_1\tau_0}{2}}\right) = \frac{2A_0\tau_0}{\Delta t}\frac{\sin\frac{k\omega_1\tau_0}{2}}{\frac{k\omega_1\tau_0}{2}}. \quad (4.69)$$

Тогда

$$S(t) = \frac{A_0\tau_0}{\Delta t}\sum_{k=-\infty}^{\infty}\frac{\sin\frac{k\omega_1\tau_0}{2}}{\frac{k\omega_1\tau_0}{2}}e^{ik\omega_1(t-t_0)} = \quad (4.70)$$

$$= \frac{A_0\tau_0}{\Delta t}\left[1 + 2\sum_{k=1}^{\infty}\frac{\sin\frac{k\omega_1\tau_0}{2}}{\frac{k\omega_1\tau_0}{2}}\cos k\omega_1(t-t_0)\right],$$

где $\tau_0 = \theta_2 - \theta_1$; $t_0 = \frac{\theta_1 + \theta_2}{2}$; $\omega_1 = \frac{2\pi}{T_i}$.

Спектр амплитуд периодической последовательности прямоугольных импульсов (рис. 4.54) изображен на рис. 4.55. Амплитуды спектральных компонентов A_k определяются значениями модуля спектральной плотности (4.69) на гармониках частоты повторения $\omega_1 = \frac{2\pi}{T}i$. Форма огибающей частотного спектра периодической последовательности определяется формой отдельного импульса.

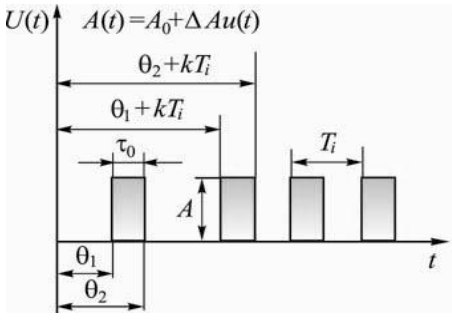


Рис. 4.54. График периодической последовательности прямоугольных импульсов

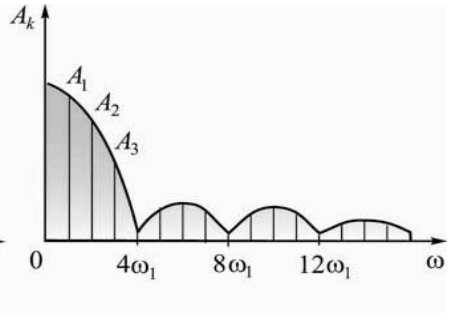


Рис. 4.55. Спектр периодической последовательности прямоугольных импульсов

С увеличением периода повторения интервал частот между соседними спектральными компонентами сокращается, их количество возрастает, а амплитуда каждого компонента уменьшается с сохранением постоянного отношения между ними. При неограниченном увеличении T периодическая последовательность превращается в единичный импульс, а дискретный спектр

становится сплошным. Спектр периодической последовательности радиоимпульсов получаем из спектра последовательности видеоимпульсов перенесением шкалы частот на несущую частоту $\omega_н$ и дополнением полученного спектра его зеркальным отображением.

При модуляции параметры, входящие в выражения (4.68) и (4.70), являются функциями времени:

$$A_0 = A(t);\ \theta_1 = \theta_1(t);\ \theta_2 = \theta_2(t).$$

Модулированная последовательность будет представлять теперь уже непериодическую функцию, деформированную относительно исходной

$$S(t) = \sum_{k=-\infty}^{\infty} A(t)U(t - t_k) = \begin{cases} U(t) \text{ при } \theta_1(t) + kT_i < t < \theta_2(t) + kT_i, \\ 0 \text{ при } \theta_2(t) + kT_i < t < \theta_1(t) + (k+1)T_i. \end{cases} \quad (4.71)$$

или согласно формуле (4.39)

$$S(t) = U(t)\frac{\theta_2(t) - \theta_1(t)}{\Delta t} \times$$

$$\times \left\{ 1 + 4\sum_{k=1}^{\infty} \frac{\sin\frac{k\omega_1}{2}\left[\theta_2(t) - \theta_1(t)\right]}{k\omega_1\left[\theta_2(t) - \theta_1(t)\right]} \times \cos k\omega_1 \left[t - \frac{\theta_1(t) + \theta_2(t)}{2} \right] \right\}. \quad (4.72)$$

Полученная формула определяет частотный спектр деформированной последовательности импульсов. Для получения спектров сигналов при различных видах модуляции в формулу (4.72) необходимо подставить соответствующие выражения модулированного параметра.

Амплитудно-манипулированные сигналы. Важным классом многотональных АМ сигналов являются так называемые манипулированные сигналы. В простом случае это - последовательности радиоимпульсов, отдаленные друг от друга паузами. Такие сигналы используются в радиотелеграфии и в системах передач дискретной информации по радиоканалам.

Если $s(t)$ - функция, в каждый момент времени принимающая значение или 0, или 1, то амплитудно-манипулированный сигнал представляется в виде

$$u_{\text{ман}}(t) = U_m s(t)\cos(\omega_0 t + \varphi_0). \quad (4.73)$$

Пусть, например**,** функция $s(t)$ отображает периодическую последовательность видеоимпульсов, рассмотренную в примере 2.1. Считая, что амплитуда этих импульсов $A = 1$, на основании выражения (4.14) при $\varphi_0 = 0$ имеем

$$u_{\text{ман}}(t) = \frac{U_m}{q}\cos\omega_0 t + \frac{U_m}{q}\sum_{n=1}^{\infty}\frac{\sin\frac{n\pi}{q}}{\frac{n\pi}{q}}\cos(\omega_0 + n\omega_1)t + \frac{U_m}{q}\sum_{n=1}^{\infty}\frac{\sin\frac{n\pi}{q}}{\frac{n\pi}{q}}\cos(\omega_0 + n\omega_1)$$

где q - скважность последовательности.

На рис. 4.56 изображен график спектра АМС.

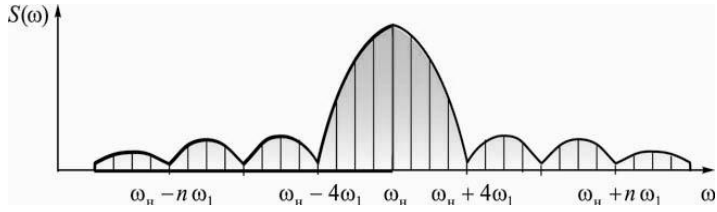


Рис. 4.56. Спектр сигнала с АИМ

При АИМ модулируется по амплитуде каждая составляющая спектра немодулированной последовательности импульсов как изолированная «несущая». В спектре получим низкочастотное модулирующее сообщение $u(t)$ с частотой Ω.

Итак, демодуляция при АИМ может быть осуществлена с помощью ФНЧ, который пропускает низкочастотное колебание $u(t)$.

Аналогично определяется спектр и для других видов импульсной модуляции. Для вычисления спектра при ФИМ в формуле (4.72) необходимо подставить выражение (4.67), определяющее изменение положения импульса относительно переданного сообщения, а при ШИМ - выражение (4.66), определяющее изменение продолжительности импульса.

При импульсно-кодовой модуляции (ИКМ) передача отдельных значений сигнала сводится к передаче определенных групп импульсов. Эти группы передаются друг за другом через большие, по сравнению с продолжительностью отдельных импульсов, промежутки. Каждая кодовая группа импульсов является регулярным непериодическим сигналом, спектр которого может быть вычислен обычным способом как преобразование Фурье.

Ширина спектра последовательности импульсов практически не зависит от частоты повторения ω_i и определяется большей частью шириной спектра одного импульса. При наличии модуляции любого вида спектр незначительно расширяется за счет боковых частот крайних составляющих спектра немодулированных импульсов. Поэтому рабочая полоса частот, занимаемая импульсными сигналами, практически не зависит от вида модуляции и определяется продолжительностью и формой импульса.

Сравнение помехоустойчивости разных методов модуляции. При различности двух сигналов за счет воздействия шума (помехи) наряду с правильными решениями возможны и ошибочные. Найдем вероятность искажения $P_{\text{пом}}$, что обеспечивается идеальным приемником Котельникова: если

двум сигналам на входе когерентного приемника соответствуют сигналы $u_1(t)$ и $u_2(t)$ одинаковой продолжительности t_c, а сумма сигнала и флуктуационной помехи - $N(t)$, то, согласно теории В. Котельникова, приемник обеспечит наименьшую вероятность искажения сигнала, если будет выдавать сигнал A при условии

$$I(U_1) >> I(U_2), \tag{4.74}$$

а в другом случае - сигнал B, где I - количество информации в сообщениях $u_1(t)$ и $u_2(t)$. Вероятность искажения (ошибочного решения приемника) равняется вероятности невыполнения условия (4.74).

Искажения возникают из-за того, что в принятый сигнал входит случайная составляющая - флуктуационная помеха. Большей частью помеху считают стационарным эргодическим процессом с нормальным законом распределения мгновенных значений напряжений

$$\omega(U_п) = \exp\left[-(U_п - \bar{U}_п)^2 / 2\sigma_п^2\right] / \left[\sigma_п^2 \sqrt{2\pi}\right],$$

где $U_п$ - мгновенное значение напряжения помехи; $\bar{U}_п$ - среднее значение напряжения помехи (постоянная составляющая); $\sigma_п^2$ - дисперсия (средний квадрат отклонения от постоянной составляющей) помехи. Поскольку постоянная составляющая помехи равна нулю, то

$$\omega(U_п) = \exp\left[-U_п^2 / (2\sigma_п^2)\right] / \left[\sigma_п^2 \sqrt{2\pi}\right].$$

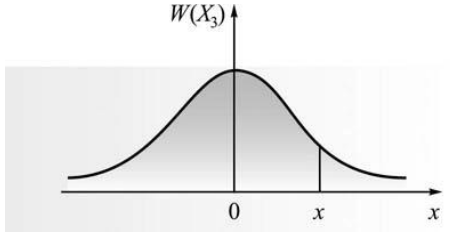


Рис. 4.57. Вероятность искажения при оптимальном приеме

Погрешность есть тогда, когда мгновенное значение напряжения помехи больше некоторого значения x. Вероятность этого события равна площади под кривой распределения в пределах от x до ∞ (рис. 4.57):

$$P_{пом} = P\left(|U_п| > x\right) = \int_x^{\infty} W(U_п)\, dU_п = -\Phi(x), \tag{4.75}$$

где $\Phi(\alpha) = \left(\int_{-\infty}^{\alpha} \exp\left[-z^2/2\right] z\right) / \left[\sigma_п^2 \sqrt{2\pi}\right]$ - функция Лапласа; α - обобщенная функция, обусловленная отношениям энергий сигнала и помехи, а также степенью расхождения сигналов U_1 и U_2:

$$\alpha = \sqrt{\left(\int_0^{\alpha} [S_1(t) - S_2(t)]^2 \, dt \right) / (2N_0)}, \tag{4.76}$$

где $N_0 = U^2_{\text{п.эфф}} / \Delta F$ - спектральная плотность мощности помехи; $U^2_{\text{п.эфф}}$ - интегральная мощность помехи; ΔF - полоса пропускания приемочного фильтра.

Преобразуем выражение (4.76):

$$\Delta E = \Delta P_{\text{с}} t_{\text{с}} = \int_0^{t_{\text{с}}} [\Delta S(t)]^2 dt, \tag{4.77}$$

где $\Delta P_{\text{с}}$ - удельная средняя мощность разностного сигнала $\Delta A(t)$. Тогда

$$\alpha = \sqrt{\Delta E / (2N_0)}. \tag{4.78}$$

Подставив это значение в выражение (4.75), получим

$$P_{\text{пом}} = 1 - \Phi\left(\sqrt{\Delta E / (2N_0)}\right). \tag{4.79}$$

Вероятность искажения (4.79) определяет потенциальную помехоустойчивость идеального и любого другого оптимального приемника. Уменьшить вероятность искажения можно за счет уменьшение спектральной плотности мощности помех N_0.

Помехоустойчивость ФМ. При ФМ сигналы противоположны по фазе, но равны по амплитуде и продолжительности

$$S(t) = -Y(t) \tag{4.80}$$

и энергии этих сигналов равны

$$E_{U_1} = E_{U_2} = E.$$

Разностный ФМ сигнал $\Delta S_{\text{ФМ}}(t)$ и его удельная энергия $\Delta E_{\text{ФМ}}$ соответственно равны между собой:

$$\Delta S_{\text{ФМ}}(t) = S(t) - Y(t) = 2S(t); \tag{4.81}$$

$$\Delta E_{\text{ФМ}} = \int_0^{t_{\text{с}}} [\Delta S_{\text{ФМ}}(t)]^2 dt = \int_0^{t_{\text{с}}} [2S(t)]^2 dt = 4E. \tag{4.82}$$

Значение $\Delta E_{\text{ФМ}}$ подставим в выражение (4.78) вместо ΔE. Тогда:

$$P_{\text{ФМ}} = 1 - \Phi\left(\sqrt{2E / (N_0)}\right) = 1 - \Phi(h\sqrt{2}), \tag{4.83}$$

где $h^2 = E / N_0$ - отношение сигнал/помеха.

Помехоустойчивость ЧМ. При ЧМ сигналы имеют одинаковую амплитуду и продолжительность, а потому их энергии также одинаковы $E_{U_1} = E_{U_2} = E$. При этом для ортогональных сигналов (сигналы с разрывом фазы)

$$\int_0^{t_{\text{с}}} S_1(t) \, S_2(t) \, dt = 0. \tag{4.84}$$

Для таких сигналов найдем удельную энергию их разности:

$$\Delta E_{\text{ЧМ}} = \int_0^{t_c} [S_1(t) - S_2(t)]^2 dt = \int_0^{t_c} [S_1(t)]^2 dt - 2\int_0^{t_c} S_1(t) S_2(t) dt +$$

$$+ \int_0^{t_c} [S_2(t)]^2 dt = 2E_A = 2E_B = 2E.$$

Тогда вероятность искажения

$$P_{\text{ЧМ}} = 1 - \Phi\left(\sqrt{E/(N_0)}\right) = 1 - \Phi(h). \qquad (4.85)$$

Помехоустойчивость АМ. При АМ (система с пассивной паузой) один из сигналов тождественно равен нулю (например, $в(t) = 0$). Для таких сигналов выполняется условие ортогональности. Нетрудно убедиться, что $\Delta E_{\text{АМ}} = E_{S_2} = E$,

$$PAM = 1 - \Phi\left(\sqrt{\Delta E/(2N_0)}\right) = 1 - \Phi(h/\sqrt{2}\ ). \qquad (4.86)$$

Для сравнения различных методов модуляции по помехоустойчивости проанализируем выражения для расчета вероятностей искажений (4.83), (4.85), (4.86). При условии довольно большого отношения сигнал/помеха на входе приемного устройства ($h^2 > 3$) и помехи типа *белый шум* все эти выражения могут быть сведены к виду

$$P_{\text{пом}} = 0{,}5 \exp(-\alpha^2 h^2/2),$$

где $\alpha^2 = 1/\sqrt{2}$ при АМ (сигналы с пассивной паузой), $\alpha^2 = 1$ при ЧМ и ФМ на угол $\pi/2$ (ортогональные сигналы), $\alpha^2 = \sqrt{2}$ при ФМ на угол π (противоположные сигналы).

При этом наибольшую помехоустойчивость имеют системы с ФМ, ЧМ, а наименьшую - системы с АМ. На рис. 4.58 приведены рассчитанные в соответствии с выражениями (4.83), (4.85), (4.86) графики зависимостей вероятностей искажений от величины h.

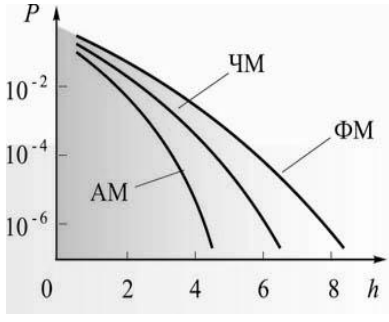


Рис. 4.58. Графики вероятностей искажений
при оптимальном приеме ФМ, ЧМ, АМ сигналов

ФМ и ЧМ колебания в сравнении с АМ занимают более широкую полосу частот, но имеют два важных преимущества:

высокая помехоустойчивость при передаче информационного сообщения;

возможность передачи более мощного сигнала при одной и той же мощности радиопередатчика.

Модуляция шумовой несущей. Как носитель можно использовать не только периодические колебания, но и узкополосный случайный процесс. Такие носители также находят практическое применение. Например, в оптических системах связи, использующих некогерентное излучение, сигналом является узкополосный гауссовский шум.

Узкополосный случайный процесс можно представить как квазигармоническое колебание вида

$$S(t)=U(t)\cos[\omega_0 t+\varphi(t)]=U(t)\cos\psi(t) \qquad (4.87)$$

с медленно меняющимися огибающей $U(t)$ и фазой $\psi(t)$. При АМ в соответствии с передаваемым сообщением, изменяется огибающая $U(t)$, при ФМ - фаза $\varphi(t)$ и при ЧМ - мгновенная частота $\omega(t)=\dfrac{d\psi}{dt}$.

Рассмотрим АМ шумовой несущей. Выражение для модулированной несущей в этом случае можно записать в виде

$$y(t)=[1+mu(t)]\,S(t), \qquad (4.88)$$

где $S(t)$ - носитель; $u(t)$ - модулирующая функция (видеосигнал); m - коэффициент модуляции.

Считается, что модулирующий процесс $u(t)$ также является стационарным нормальным процессом с нулевым средним значением $u(t) = 0$. Процессы $S(t)$ и $u(t)$ независимы. При этих ограничениях функция корреляции шумовой несущей, модулированной по амплитуде будет

$$B_y(\tau)=y(t)y(t+\tau)=[1+mu(t)]S(t)[1+mu(t+\tau)]S(t+\tau)= \qquad (4.89)$$

$$=[1+mu(t)+mu(t+\tau)+m^2u(t)u(t+\tau)]S(t)S(t+\tau)=[1+m^2B_u(\tau)]B_S(\tau).$$

Теперь определим энергетический спектр

$$G_y(\omega)=2\int_0^\infty B_y(\tau)\cos\omega\tau d\tau=2\int_0^\infty B_f(\tau)\cos\omega\tau d\tau+2m^2\int_0^\infty B_u(\tau)\cos\omega\tau d\tau.$$

Первый интеграл дает энергетический спектр шумовой несущей $G_S(\omega)$. Для второго интеграла на основании теоремы о спектре произведения имеем

$$\int_0^\infty B_u(\tau)B_S(\tau)\cos\omega\tau d\tau=\frac{1}{2}\int_{-\infty}^{\infty}G_S(v)G_u(\omega-v)dv.$$

Окончательно спектр модулированной несущей запишем как

$$G_y(\omega) = G_S(\omega) + m^2 \int_{-\infty}^{\infty} G_S(v-\omega) G_u(\omega_0 - v) dv. \qquad (4.90)$$

Таким образом, спектр шумовой несущей, модулированной по амплитуде, является суперпозицией спектра несущей и свертки этого спектра со спектром переданного сообщения, перенесенного в область высоких частот на величину $\omega_н$. Аналогично определяются функция корреляции и энергетический спектр при ФМ и ЧМ. Применение «шумовых» сигналов дает возможность ослабить влияние угасания в каналах с многолучевым распространением радиоволн.

Пример. Пусть на вход приемника поступают сигналы двух лучей $\xi(t)$ и $\xi(t - \tau)$ со сдвигом на время τ. Мощность результирующего сигнала, достигнутая за довольно большое время T:

$$P = [\xi(t) + \xi(t+\tau)]^2 = 2[P_0 + B_\xi(\tau)],$$

где $B_\xi(\tau)$ - функция корреляции сигнала; P_0 - его средняя мощность. Функция корреляции шума быстро спадает с увеличением τ и тем быстрее, чем более широк его спектр. Итак, при довольно большой ширине спектра можно считать $B_\xi(\tau) \approx 0$ и $P \approx 2P_0$, т.е. средняя мощность принятого сигнала, несмотря на угасание, остается приблизительно постоянной.

4.2. Квантование и дискретизация сигналов

Квантование. Для преобразования аналогового сигнала в цифровой после дискретизации по времени должна происходить *дискретизация по уровню* (*квантование*). Необходимость квантования вызвана тем, что любое вычислительное устройство может оперировать только числами, которые имеют конечное количество разрядов. Таким образом, *квантование есть округление передаваемых значений с заданной точностью.*

Характеристики квантования. Квантование сигналов можно описать графически с помощью *характеристики квантования* (рис. 4.59), где по оси абсцисс отложены значение непрерывного сигнала, а по оси ординат - значение квантованного сигнала. Величину *шага квантования* Δt выбирают, исходя из необходимой точности передачи сигнала. Квантование с постоянным шагом Δt называют *равномерным*. Равномерное квантование сигналов является наиболее простым и распространенным.

Тем не менее, равномерное квантование в отдельных случаях оказывается неудобным. Например, если переданный сигнал может приобретать очень большие и очень малые значения, то при постоянной величине интервала квантования относительная точность передачи малых значений сигнала оказывается значительно худшей, чем больших его значений. В этих случаях иногда применяют нелинейное, например *логарифмическое*, квантование

(рис. 4.60), когда шаг квантования пропорционален логарифму входного напряжения.

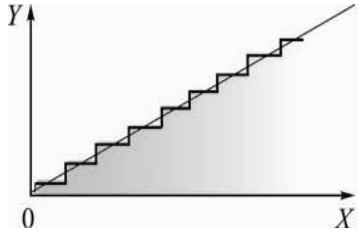


Рис. 4.59. Характеристика квантования

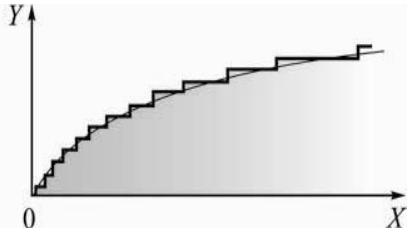


Рис. 4.60. Квантование с логарифмической характеристикой

При квантовании малых значений сигнала шаг квантования оказывается небольшим, а точность передачи сигнала - достаточно высокой. При больших значениях входного сигнала интервал квантования увеличивается. Таким образом, использование логарифмического квантования позволяет достичь высокую точность передачи сигнала при не очень большом количестве квантованных уровней сигнала.

Рассмотрим один из видов погрешности квантования, который называют *шумом квантования*.

Квантование - это округление значений сигнала к ближайшему дискретному значению. При этом каждое округленное значение отличается от начального (истинного) значения сигнала на величину ξ, которая является погрешностью округления и не превышает по величине половины шага квантования $\Delta t/2$. Если входной сигнал точно не известен, то погрешность округления ξ является случайной величиной. При малом шаге квантования распределение величины ξ близко к равномерному (см. рис. 4.61, *а*). Последовательность значений погрешности округления ξ, возникающей при квантовании дискретного сигнала $x(k\Delta t)$, образовывает дискретный случайный процесс $\xi(k\Delta t)$, который называют *шумом квантования* (рис. 4.61, *б)*. Квантованный сигнал можно представить в виде суммы не квантованного дискретного сигнала $x(k\Delta t)$ и шума квантования $\xi(k\Delta t)$.

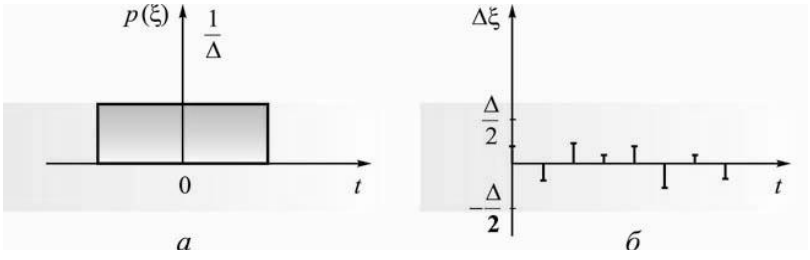


Рис. 4.61. К определению шума квантования:
а - плотность вероятности; *б* - реализация

Дисперсию шума квантования вычисляют как дисперсию закона равномерного распределения

$$\sigma_\xi^2 = \int_{-\Delta t/2} \xi^2 P(\xi) d\xi = \Delta t^2 / 12,$$

где Δt - шаг квантования.

При достаточно малом шаге квантования и правильно выбранном интервале дискретизации, за время между соседними отсчетами значение сигнала успевает измениться на много шагов квантования. При этом соседние значения погрешности $\xi(k\Delta t)$ оказываются некоррелированными.

Апертурные погрешности и шум квантования - это не единственные источники ошибок аналого-цифрового преобразования. Существуют и другие источники погрешностей, связанные с несовершенством работы схемы выборки и запоминания, с нелинейностью характеристик отдельных элементов и т.д.

Дискретизация сигналов. На рис. 4.62 приведена классификация признаков дискретизации.

Рис. 4.62. Классификация признаков дискретизации

Процесс дискретизации непрерывных сигналов состоит из двух этапов: *дискретизация по времени* и *дискретизация по уровню* (*квантование*). Сигнал, дискретизированный только по времени, называют *дискретным*; он еще не пригоден для обработки в цифровом устройстве.

Дискретный сигнал является последовательностью, элементы которой $S(k\Delta t)$ равны соответствующим значениям начального непрерывного сигнала $S(t)$ (рис. 4.63, *а*).

Примером дискретного сигнала может быть последовательность импульсов со сменной амплитудой - АИМ сигнал (рис. 4.63,*б*).

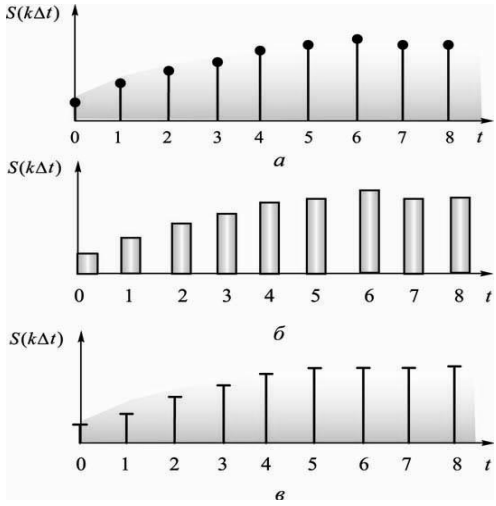


Рис. 4.63. Виды сигналов: *а* - дискретный сигнал; *б* – АИМ - сигнал; *в* - цифровой сигнал

Аналитически такой дискретный сигнал описывается выражением

$$\varphi(t) = \sum_{k=-\infty}^{+\infty} S(k\Delta t) F(t - k\Delta t), \qquad (4.91)$$

где $S(t)$ - начальный непрерывный сигнал; $F(t)$ - единичный импульс АИМ колебания.

Если уменьшать продолжительность импульса $F(t)$, сохраняя его площадь неизменной, то функция $F(t)$ приближается к δ-функции. Тогда выражение для дискретного сигнала можно представить в виде

$$\psi(t) = \sum_{h=-\infty}^{+\infty} S(k\Delta t)\, \delta(t - k\Delta t). \qquad (4.92)$$

Выбор частоты дискретизации. Непрерывный сигнал заменяется дискретными отсчетными значениями, взятыми через определенные интервалы времени. Важным является выбор интервала дискретизации. Эта задача решается для сигналов с ограниченным спектром на основании *теоремы Котельникова* (см. разд. 3). Вместо непрерывного сигнала $S(t)$ с ограниченным спектром можно передавать дискретную последовательность значений $S(k\Delta t)$, причем интервал дискретизации Δt должен быть не больше, чем $1/(2f_в)$. Если отсчеты взять реже, то это может привести к грубым ошибкам. Реальные сигналы имеют конечную длительность, и потому спектр их бесконечный (рис. 4.64).

Спектр дискретизированного сигнала является периодическим повторением спектров входного непрерывного сигнала $S(t)$.

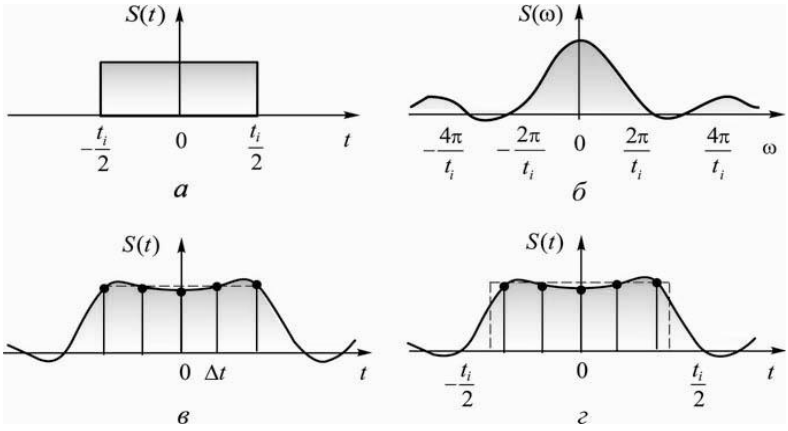


Рис. 4.64. Дискретизация прямоугольного ипульса:
а - начальный прямоугольный импульс; *б* - его спектр;
в - результат дискретизации при $\Delta t = t_{имп}/4$; *г* - то же, при $\Delta t = t_{имп}/5$

Если спектр непрерывного сигнала $S(\omega)$ ограничен по ширине (рис. 4.65, *а*), а интервал дискретизации Δt удовлетворяет условию $\Delta t \le \pi / \omega_в$, то периодом повторения спектра дискретизированного сигнала будет $2\pi/\Delta t \ge 2\omega_в$; поэтому соседние части спектра (соответствующие $S(\omega - k(2\pi/\Delta t)$ и $S[\omega - (k + 1)(2\pi/\Delta t)]$) не перекрываются (см. рис. 4.65, *б).*

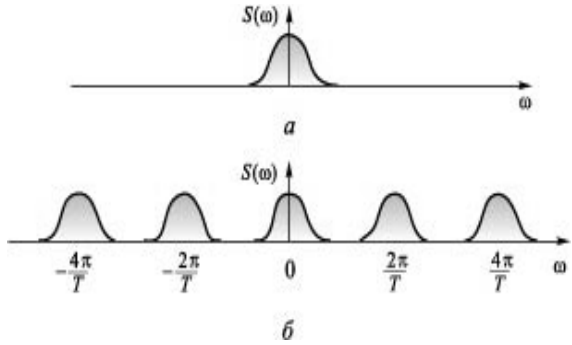


Рис. 4.65. Спектры дискретизированных сигналов: *а* - спектр начального сигнала; *б* - спектр модулированной последовательности *δ*-функций

Погрешности дискретизации и восстановление непрерывного сигнала. Если период дискретизации достаточно мал, так что выполняется условие $\Delta t < \pi /\omega_в$, то соседние составляющие спектра дискретизированного колебания не перекрываются (как показано на рис. 4.66, *а).*

В этом случае легко указать *способ восстановления* непрерывного колебания из дискретного, заключающийся в пропускании *дискретного сигнала через идеальный ФНЧ с полосой пропускания* $(-\omega_в, \omega_в)$ (см. рис. 4.66, *б).*

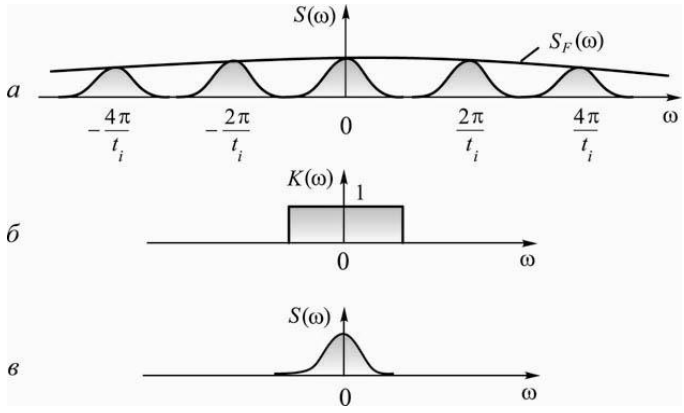


Рис. 4.66. Спектр дискретного колебания в виде последовательности модулированных импульсов (*а*), частотная характеристика ФНЧ (*б*) и спектр восстановленного сигнала (*в*)

При этом из спектра дискретизированного сигнала $S_\varphi(\omega)$ будет выделена средняя часть (см. рис. 4.66, *в)*, которая с точностью до постоянного множителя совпадает со спектром исходного непрерывного колебания $S(t)$.

Тем не менее, если начальное непрерывное колебание таково, что его спектр с ростом частоты не превращается строго в нуль, то при любом выборе интервала дискретизации соседние составляющие спектра дискретизированного колебания частично будут перекрываться (рис. 4.67, *а*). Если сигнал с таким спектром пропускать через идеальный фильтр нижних частот, то на выходе фильтра получим колебание, отличающееся от начального непрерывного сигнала $S(t)$. Это отличие заключается не только в том, что «отрезанная» часть спектра выше частоты $\omega_в = \pi / \Delta t$, но также и в том, что на спектр этого колебания накладываются «хвосты» от соседних спектральных составляющих (см. рис. 4.67, *б)*.

Наиболее простой и очевидный способ уменьшения погрешности дискретизации - повышение частоты дискретизации. Тем не менее, для получения достаточно малой погрешности частоту дискретизации приходится брать очень высокой, особенно если спектр сигнала спадает медленно, что нежелательно.

Для уменьшения погрешности дискретизации можно перед дискретизацией пропустить сигнал через ФНЧ с частотной характеристикой, близкой к прямоугольной. При этом спектр сигнала становится быстро спадающим, почти ограниченным, и дальнейшая дискретизация происходит практически без ошибок. Результирующая погрешность в этом случае определяется искажениями спектра при прохождении сигнала через ФНЧ. Поскольку на спектр сигнала в области частот $(-\omega_в, \omega_в)$ не накладываются «хвосты» от соседних составляющих, эта погрешность приблизительно вдвое меньше, чем при непосредственной дискретизации сигнала.

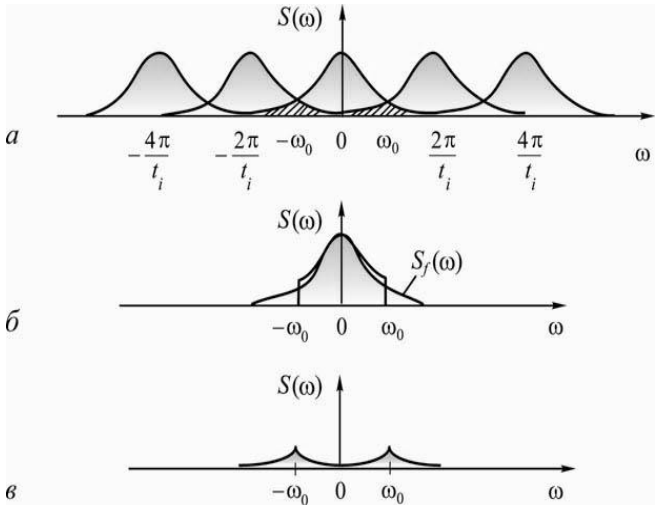


Рис. 4.67. Погрешности дискретизации сигнала с асимптотично спадающим спектром: *а* - спектр дискретизирванного сигнала; *б* - спектр сигнала после проходжения через идеальный ФНЧ; *в* - спектр сигнала погрешности

Пропускание сигнала через ФНЧ перед дискретизацией является полезной мерой для снижения погрешности, если дискретизация сигнала проводится при наличии широкополосного шума на входе. При прохождении через ФНЧ дисперсия шума уменьшается и, соответственно, уменьшается погрешность дискретизации.

Еще одним источником погрешности является неидеальная фильтрация в процессе восстановления непрерывного сигнала из дискретного. Идеальная прямоугольная форма частотной характеристики ФНЧ практически не может быть реализована; для сглаживания сигнала, по обыкновению, применяют фильтры, имеющие монотонно нисходящую характеристику (рис. 4.68, *б)*. Если на вход такого фильтра подать дискретизированный сигнал со спектром (см. рис. 4.68, *а*), то на выходе фильтра кроме основного сигнала, которому соответствует центральная часть спектра, появятся дополнительные составляющие, вызванные неполным подавлением боковых частей спектра (см. рис. 4.68, *в)*. Вследствие этого восстановленный сигнал будет отличаться формой от начального непрерывного сигнала.

Основной метод борьбы с этими погрешностями заключается в увеличении частоты дискретизации. Однако увеличение частоты дискретизации приводит к усложнению и удорожанию устройства обработки сигналов. Поэтому в каждом конкретном случае приходится искать компромиссное решение, исходя из характера сигнала, необходимой точности его восстановления, характеристик используемого сглаживающего фильтра и других факторов. Все

это приводит к тому, что в реальных устройствах частота дискретизации выбирается не $2f_{в}$,, а в 2 - 5 раз выше.

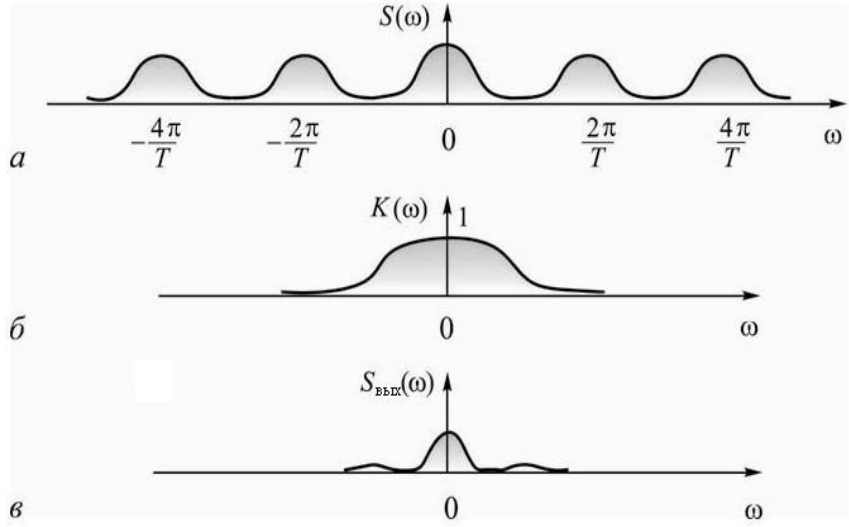


Рис. 4.68. Погрешности восстановления сигнала: *а* - спектр дискретизированного сигнала; *б* - характеристика ФНЧ; *в* - спектр сигнала на выходе ФНЧ

Спектр сигнала с конечной продолжительностью приведен на рис. 4.69.

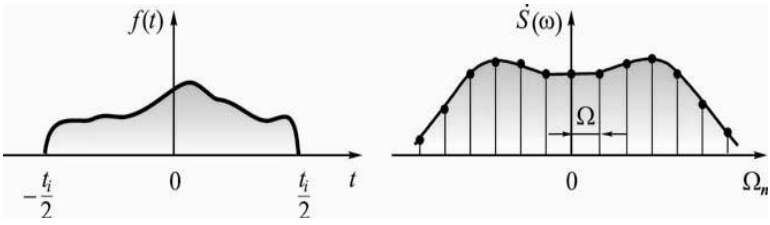


Рис. 4.69. Сигнал конечной продолжительности и его спектр

Обобщенная схема цифровой обработки сигналов. *Цифровая обработка сигналов* (ЦОС) - это область науки и техники, в которой изучаются общие для различных технических применений принципы, методы и алгоритмы обработки сигналов средствами цифровой вычислительной техники.

Обобщенная схема ЦОС (рис. 4.70) отображает последовательность процедур, необходимых для преобразования входного аналогового сигнала $x(t)$ в другой аналоговый сигнал $y(t)$ по заданному алгоритму средствами цифровой вычислительной техники.

В цифровой обработке сигнала можно выделить три основных этапа:

формирование цифрового сигнала $x_{ö}(k\Delta t)$ из входного аналогового сигнала $x(t)$;

преобразование цифрового сигнала $x_{\text{ö}}(k\Delta t)$ в сигнал $y_{\text{ö}}(k\Delta t)$ по заданному алгоритму;

формирование результирующего аналогового сигнала $y(t)$ из сигнала $y_{\text{ö}}(k\Delta t)$.

Рис. 4.70. Обобщенная схема ЦОС

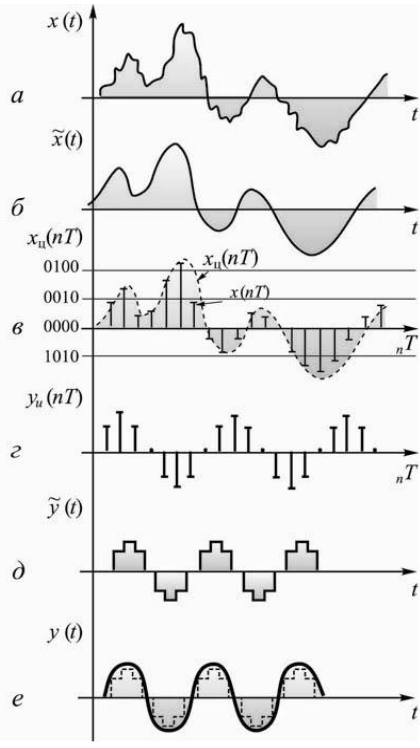


Рис. 4.71. Временные диаграммы поэтапного процесса ЦОС

В обобщенной схеме ЦОС этим этапам соответствуют три функциональные устройства: кодер; устройство ЦОС; декодер. Временные диаграммы поэтапного процесса ЦОС приведены на рис. 4.71. Рассмотрим каждый из этапов.

1. На первом этапе кодер из исходного аналогового сигнала $x(t)$ (см. рис. 4.71, *а*) формирует цифровой сигнал $x_{\text{ц}}(k\Delta t)$ (см. рис. 4.71, *б)*, без чего принципиально невозможна цифровая обработка. В состав кодера входят аналоговый ФНЧ и аналого-цифровой преобразователь (АЦП). *Аналоговый ФНЧ* предназначен для ограничения спектра $X(j\omega)$ исходного аналогового сигнала $x(t)$. Необходимость ограничения спектра вытекает из теоремы Ко-

тельникова, в соответствии с которой частота дискретизации $f_д$ выбирается из условия: $f_д \geq 2 f_в$, где $f_в$ - верхняя частота спектра сигнала.

Возможность ограничения спектра связана с особенностями частотного распределения энергии сигнала: основная часть его энергии сосредоточена в области $f \leq f_в$, т.е. амплитуды спектральных составляющих, начиная с некоторой частоты $f > f_в$, существенным образом снижаются (рис. 4.72, *а*). Выбор значения определяется конкретным типом сигнала и задачи. При обработке аудио- и видеосигналов выбор $f_в$ зависит от особенностей психофизического восприятия этих сигналов.

Пример. Для стандартного телефонного сигнала верхняя частота $f_в$ равна 3,4 кГц, а минимальная стандартная частота дискретизации $f_д$ - 8 кГц.

На выходе ФНЧ получают аналоговый сигнал $\tilde{x}(t)$ с финитным (ограниченным по частоте) спектром $\tilde{X}(j\omega)$ (см. рис. 4.72, *б)*.

АЦП формирует цифровой сигнал $x_ц(k\Delta t)$ с помощью дискретизации и квантования сигнала $\tilde{x}(t)$ (см. рис. 4.71, *в)*.

Дискретизация по времени (*дискретизация*) представляет собой процедуру взятия мгновенных значений - отсчетов - аналогового сигнала $\tilde{x}(t)$ с интервалом времени, равным периоду дискретизации Δt [по умолчанию будем понимать равномерную (эквидистантную) дискретизацию].

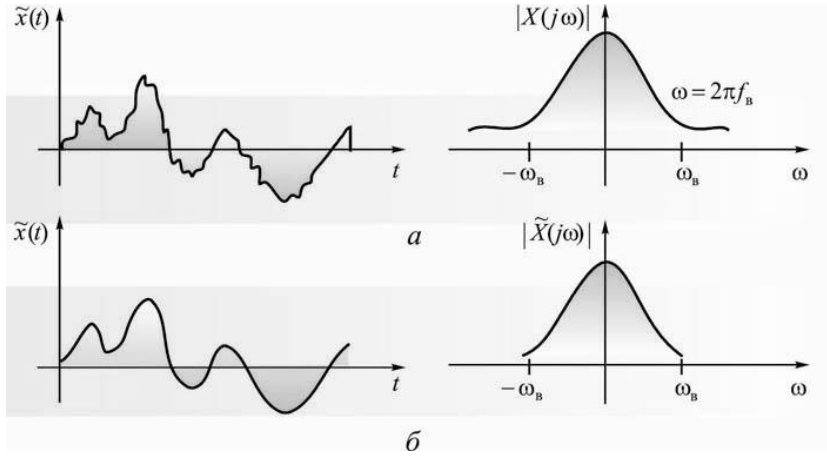


Рис. 4.72. Сигналы и их амплитудные спектры на входе (*а*) и на выходе (*б*) ФНЧ

Значения отсчетов $x(k\Delta t)$ совпадают со значениями сигнала $\tilde{x}(t)$ в моменты времени и $t = k\Delta t$:

$$x(k\Delta t) = \tilde{x}(t)_{t=k\Delta t}.$$

Совокупность отсчетов $x(k\Delta t)$, k = 0, 1, ... *называют* ***дискретным сигналом.***

Квантование сигнала по уровню проводится с целью представления точных значений отсчетов $x(k\Delta t)$ в виде двоичных чисел конечной разрядности - *квантованных отсчетов* $x_{ц}(k\Delta t)$. Для этого необходимо динамический диапазон дискретного сигнала $x(k\Delta t)$ разбить на конечное количество дискретных уровней - *уровней квантования;* каждому отсчету по определенному правилу поставить в соответствие значение одного из ближайших уровней, между которыми этот отсчет расположен. Равные квантования кодируются двоичными числами разрядности *b*, зависящими от числа уровней квантования *R:* $R \le 2^b$, откуда $b = \mathrm{int}(\log_2 R)$. На временной диаграмме (см. рис. 4.72, *б)* для примера взято 5 уровней квантования (без учета знака), поэтому $b = 3$, отсчеты $x_{ц}(k\Delta t)$ кодируются четырехразрядными двоичными числами: один разряд - знаковый, три - значащие.

Совокупность *квантованных отсчетов* $x_{ц}(k\Delta t)$ k = 0, 1,… *называют* ***цифровым сигналом***.

Процесс квантования сигналов обычно проводится одновременно с его кодированием, на выходе получаем сигнал, представленный в некотором цифровом коде.

АЦП при подаче управляющего сигнала формирует значение входного сигнала в цифровом коде. Темп подачи управляющих сигналов, т.е. темп взятия выборок, определяется шириной спектра сигнала и предназначением устройства, в котором используется АЦП. АЦП используют не только для цифровой обработки сигналов - он является основным узлом любого цифрового измерительного прибора.

АЦП может осуществлять одновременно и дискретизацию по времени, и квантование сигналов. Тем не менее, в этом случае при недостаточном быстродействии АЦП могут возникнуть специфические *погрешности*, называемые *апертурными*. Природа этих погрешностей состоит в следующем.

Преобразование аналогового сигнала в цифровой не происходит мгновенно. Процесс преобразования длится некоторое время Δt_a, который называют *апертурным временами*. Если превращаемый сигнал изменяется во времени, то за время Δt_a, его величина успевает как - либо измениться (рис. 4.73).

В результате сигнал на выходе АЦП не соответствует точно значению входного сигнала S_1 в отсчетный момент времени t_1. Возникает апертурная погрешность, которая может достигать величины $\Delta S = (d / dt)_t = t_1$. Величину апертурной погрешности можно определить, если задаться шириной спектра сигнала и временем преобразования (быстродействием) АЦП.

Апертурное время Δt_a, т.е. время обработки сигнала в АЦП, должно быть приблизительно в 2^n раз меньше интервала дискретизации. Это невыгодно, поскольку АЦП должны работать лишь малую часть периода, и потому быстродействие такой схемы дискретизации и кодирования оказывается невысокой.

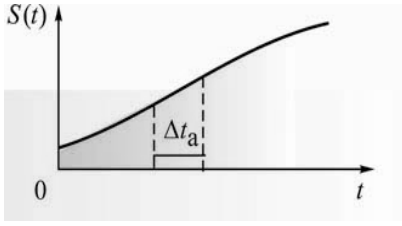


Рис. 4.73. К определению апертурной погрешности

Скорость дискретизации и квантования сигналов можно повысить на несколько порядков (в 2^n раз), если перед квантованием преобразовать сигнал так, чтобы он приобрел ступенчатую форму (рис. 4.74). Это осуществляется с помощью специального устройства выборки и запоминания, фиксирующего значение сигнала в отсчетные моменты времени t_1, t_2 ... и поддерживающего это значение постоянным к следующему отсчетному моменту. Теперь АЦП может обрабатывать сигнал на протяжении всего периода дискретизации Δt, вследствие чего быстродействие устройства в целом значительно повышается.

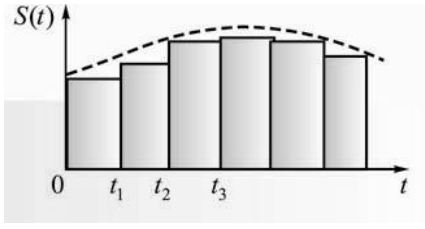


Рис. 4.74. Сигнал на выходе устройства выборки и хранения

Детерминированные и вероятностные оценки погрешностей квантования $e_{\text{кв}}(k)$ за счет АЦП: $e_{\text{кв}}(k) = x(k\Delta t) - x_{\text{ц}}(k\Delta t)$.

2. *Устройство ЦОС* превращает *цифровой сигнал* $x_{\text{ц}}(k\Delta t)$ (см. рис. 4.71, *в)* в *цифровой сигнал* $y_{\text{ц}}(k\Delta t)$ (см. рис. 4.71, *г)* по заданному алгоритму.

Устройство ЦОС может быть реализовано аппаратно или программно. В первом случае - в виде специализированного цифрового устройства, во втором - в виде программы на компьютере или цифровом процессоре обработки сигналов. Программная реализация преобладает. Устройства ЦОС могут рабо-

тать в реальном или нереальном времени. В реальном времени обработка сигналов должна выполняться в темпе поступления отсчетов входного сигнала $x_{\text{ц}}(k\Delta t)$, $k = 0, 1,\ldots$ и соответствовать таким требованиям:

время цикла $\Delta t_{\text{ö}}$ при вычислении отсчета $y_{\text{ц}}(k\Delta t)$ не должно превышать интервал между двумя соседними отсчетами $x_{\text{ц}}(k\Delta t)$, т.е. период дискретизации Δt:

$$\Delta t_{\text{ц}} \le T; \qquad (4.93)$$

тактовая частота f_{τ} процессора должна быть намного выше частоты дискретизации $f_{\text{д}}$ сигнала $x_{\text{ц}}(k\Delta t)$:

$$f_{\tau} \ge f_{\text{д}}.$$

Последнее вызвано тем, что в алгоритмах ЦОС количество операций в цикле, необходимое для вычисления одного отсчета $y_{\text{ц}}(k\Delta t)$, довольно велико.

Например, для стандартного телефонного сигнала с частотой дискретизации 8 кГц тактовая частота должна быть не меньше 6 мГц.

В реальном времени выполняется обработка сигналов, связанная с их передачей по каналам связи, в том числе по сети Интернет. К типичным задачам ЦОС в реальном времени принадлежат: выявление, фильтрация, сжатие, распознавание сигналов и др.

В нереальном времени выполняется обработка сигналов, связанная, прежде всего, с их исследованием. К типичным задачам ЦОС в нереальном времени принадлежат: студийная обработка аудио- и видеосигналов, обработка данных различной физической природы, полученных от датчиков, и т.п.

3. Декодер формирует результирующий аналоговый сигнал $y(t)$ из цифрового сигнала $y_{\text{ц}}(k\Delta t)$. В состав декодера входят цифро-аналоговый преобразователь (ЦАП) и сглаживающий фильтр.

ЦАП формирует из цифрового сигнала $y_{\text{ц}}(k\Delta t)$ (см. рис. 4.71, *г)* ступенчатый аналоговый сигнал $\tilde{y}(t)$ (см. рис. 4.71, *д)*. Сглаживая этот сигнал, ФНЧ устраняет ступенчатый эффект в исходном сигнале ЦАП $\tilde{y}(t)$. На выходе сглаживающего фильтра получаем аналоговый сигнал $y(t)$ (см. рис. 4.71, *е)* - результат преобразования исходного сигнала $x(t)$.

4.3. Цифровая обработка информации

Большинство операций в процессе восприятия, передачи и отображения информации требуют выполнения значительного объема вычислений, предварительного анализа и упорядочения. Все эти действия необходимо выполнять с высокой скоростью, определяемой ритмом функционирования информационной системы. К таким операциям принадлежат, например, статистическая обработка данных, вычисление корреляционных функций и спектров информационных сигналов с использованием быстрого преобразования Фурье, операции идентификации принятых кодов, технической диагностики и распознавания образов, прогнозирование, расчет управляющих влияний, информационный поиск, в частности в сети Интернет, и многое другое. Все эти операции связаны с обработкой информации, в результате которой содержание ее готовится к реализации следующего этапа, например представления информации человеку для принятия решения или наработка команд управления. Обработка информации связана с более высоким уровнем познания, чем простое восприятие и, тем более, ощущение. Познание ограничено, с одной стороны, недостаточностью объема полезных данных, которые есть в нашем распоряжении, а с другой - большим разнообразием данных об изучаемом объекте. В отличие от других информационных процессов при обработке информации воспринимается, толкуется и распознается лишь небольшая частица этого разнообразия, причем выделяются лишь наиважнейшие зависимости. Выбором соответствующих вариантов, упрощений и абстракций осуществляется ограничение этого разнообразия, выявляется существенное, что дает возможность человеку или устройству управления принимать решение в конкретной ситуации. Общее количество информации в процессе обработки, как правило, уменьшается за счет устранения избыточности, при этом ценность и содержательность ее увеличивается. Обработка информации осуществляется как в аналоговой (см. 4.1), так и в цифровой форме, но последняя на данный момент является основной. Рассмотрим один из самых используемых направлений цифровой обработки в информационных системах - цифровую фильтрацию.

Цифровая фильтрация выполняет функцию, эквивалентную аналоговым фильтрам, однако имеет ряд важных преимуществ:

высокую стабильность параметров;

возможность получения амплитудно-частотной характеристики и фазочастотной характеристики заведомо заданной формы;

улучшение формы амплитудно-частотной характеристики фильтра путем ее приближения к идеальной при одном порядке фильтра;

цифровые фильтры не нуждаются в настраивании, реализуются путем программирования процессора.

Цифровым фильтром (ЦФ) в широком смысле *называют любую цифровую систему, осуществляющую, согласно заданному оператору* $y(n) = F\{\tilde{x}(n)\}$, *преобразование аддитивной смеси* $\tilde{x}(n) = x(n) + \xi(n)$ *цифрового сигнала* $x(n)$ *или его параметров, который действует на входе системы, с помехой* $\xi(n)$.

Например, к цифровым фильтрам принадлежат: фильтры, согласованные с сигналами, адаптивные фильтры, амплитудные и фазовые корректоры, дифференциаторы, преобразователи Гильберта и др. Естественно, сигнал $y(n)$ на выходе реального ЦФ будет соответствовать переданному сигналу или его параметрам с некоторой точностью, определяемой свойствами алгоритма. Иначе говоря, на выходе реального ЦФ всегда имеет место различная степень приближения $y(n) \approx x(n)$.

Цифровой фильтр в узком смысле - *это частотно-избирательная цепь, обеспечивающая селекцию цифровых сигналов по частоте*

Рассмотрим структурно-логическую схему цифровой фильтрации (рис. 4.75). Линейные стационарные ЦФ принадлежат к классу систем с дискретной обработкой сигналов.

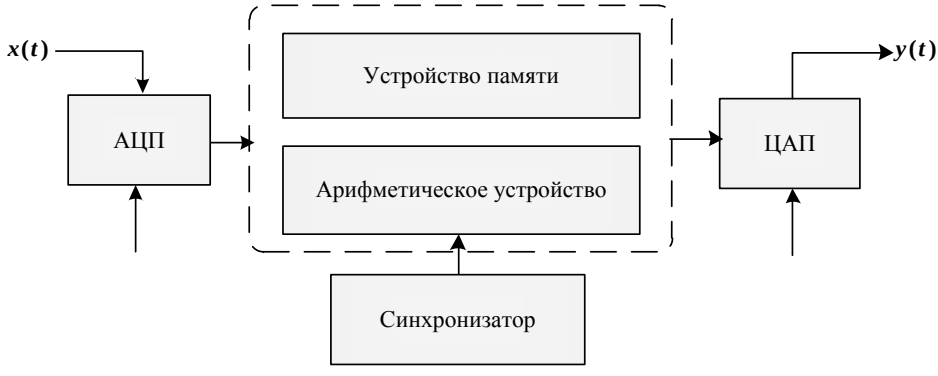


Рис. 4.75. Структурная схема ЦФ

Непрерывный сигнал *x*(*t*) поступает на АЦП, управляемый импульсами от генератора, задающего частоту дискретизации. В момент подачи тактового импульса на выходе АЦП возникает сигнал, равный по величине мгновенному значению амплитуды входного сигнала. Данное значение амплитуды переводится в двоичный код с фиксированным значением разрядов. Преобразованный цифровой сигнал поступает на вход процессора. В памяти сохраняется определенный алгоритм работы и числа, необходимые для работы алгоритмического устройства. Дальше отфильтрованный цифровой сигнал поступает на ЦАП и преобразуется в непрерывный.

К ЦФ принадлежат: фильтры нижних частот (ФНЧ), фильтры верхних частот (ФВЧ), полосовые фильтры (ПФ или СФ), режекторные фильтры (РФ).

Как и все цифровые системы, ЦФ делятся на два больших класса: *нерекурсивные* (*КИХ*) *и рекурсивные* (*БИХ*). В каждом из этих классов выделяют линейные и нелинейные фильтры. Фильтры, рассматриваемые здесь, являются линейными, так что оператор $F\left\{\tilde{x}(n)\right\}$ соответствует свойству линейности системы.

ЦФ, как и любая цифровая система, могут быть реализованы аппаратно, программно или аппаратно-программно, что определяется целью, назначением и местом ЦФ в предполагаемой системе.

Аппаратная реализация предусматривает использование разнообразных функциональных блоков (регистров, сумматоров, множителей, запоминающих устройств, логических элементов и др.), объединяемых в единое устройство.

Программная реализация означает, что фильтр представлен в виде программы, записанной на языке программирования, соответственно конкретному операционному блоку. Так, для персонального компьютера это будет любой из языков программирования высокого уровня, а для микропроцессорного комплекта или цифрового процессора - язык соответствующего ассемблера.

Под проектированием ЦФ понимают процесс, результатом которого является программа или цифровое устройство, отвечающее заданным требованиям и ограничениям.

Процесс проектирования ЦФ состоит из таких этапов:

1. *Синтез*, результатом которого является функциональная схема фильтра с коэффициентами. Собственно процедуры синтеза *нерекурсивных* и *рекурсивных* фильтров существенным образом различаются, но, тем не менее, имеют одинаковую последовательность действий:

задание требований к фильтрам;

решение задачи аппроксимации характеристик фильтра, в результате которой рассчитываются коэффициенты передаточной функции (разностного уравнения);

конструирование функциональной схемы ЦФ.

2. *Выбор или разработка эффективных алгоритмов вычислений* с учетом арифметики, используемой при заданном методе реализации: плавающая или фиксированная точка. Алгоритм зависит от разрядности регистров процессора, количества аккумуляторов; возможности распараллеливания операций, наличия устройств, выполняющих операцию умножения с накоплением и других особенностей процессора. Конечной целью этого этапа является обеспечение функционирования фильтра в реальном времени при минимальных потерях качества обработки сигналов.

3. *Проверка моделированием* запроектированного фильтра в нереальном времени по стандартным сигналам с использованием программных средств:

симуляторов системы команд (симуляторов), имитирующих работу конкретного процессора на уровне его команд;

настройщиков - буферных программ, предоставляющих разработчику необходимый интерфейс и обеспечивающих функциональные возможности симуляторов.

Задача проверки моделированием состоит в выявлении и устранении возможных логических и других скрытых ошибок, испытании на соответствие сконструированного фильтра заданным характеристикам, включая частотные, временные и шумовые.

4. *Практическая реализация и отладка* в реальном времени с помощью аппаратных средств: эмуляторов и проверочных модулей.

Результаты проверки моделированием и отладкой могут повлиять на изменение ряда решений в зависимости от выбора структурной схемы ЦФ вплоть до задания новых требований.

Основными показателями работы ЦФ являются скорость обработки данных и качество АЧХ (*ее приближение к идеальной*).

Алгоритмы линейной цифровой фильтрации. Математическая теория ЦФ полностью эквивалентна теории фильтров, работающих с непрерывными сигналами. Линейная стационарная система преобразует непрерывный входной сигнал $x(t)$ так, что на выходе возникает колебание $y(t)$, равное свертке функции $x(t)$ и импульсной характеристики $h(t)$:

$$y(t) = \int_{-\infty}^{\infty} x(t)h(t-\tau)dt. \qquad (4.94)$$

Линейный ЦФ *- это дискретная система, которая превращает последовательность* x_k *числовых отсчетов входных сигналов в последовательность* y_k *выходных отсчетов сигнала:*

$$\{x_k\} \Rightarrow \{y_k\}; \quad \{x_0, x_1, ..., x_N\} \Rightarrow \{y_0, y_1, ..., y_N\}.$$

Свойство линейности ЦФ *- сумма любого количества входных сигналов, умноженная на произвольные коэффициенты, преобразуется в сумму его отсчетов:* если $\{x_0\} \Rightarrow \{y_0\}$ или $\{x_i\} \Rightarrow \{y_i\}$, то

$$\alpha_0\{x_0\} + \alpha_1\{x_1\} + ... + \alpha_N\{x_N\} \Rightarrow \alpha_0\{y_0\} + \alpha_1\{y_1\} + ... + \alpha_N\{x_N\}$$

при любых α.

Для того чтобы применить основную формулу фильтрации, базирующуюся на свертке для ЦФ, необходимо описать и ввести понятие *импульсной характеристики ЦФ,* которую можно применить для работы с дискретным сигналом.

Импульсная характеристика ЦФ *- цифровой сигнал* h_k, *являющийся реакцией цепи на униполярный дельта-импульс* $\{h_k\} = \{h_0, h_1, h_2, ..., h_N\}$.

Стационарным ЦФ *называется такой фильтр, для которого при сдвиге входного униполярного импульса на любое количество интервалов дискретизации импульсная характеристика системы также смещается на это количество интервалов, не изменяя своей формы:*

$$\{0,1,0,0,0,0\} \Rightarrow \{0, h_0, h_1, h_2 ...\},$$

$$\{0,0,0,1,0,0\} \Rightarrow \{0,0,0, h_0, h_1 ...\}.$$

Используя указанные свойства линейности и стационарности, обоснуем алгоритм цифровой фильтрации.

Пусть на входе есть цифровой сигнал вида $\{x_k\} = \{x_0, x_1, x_2, ..., x_N\}$. Данный сигнал поступает на вход ЦФ с известной импульсной характеристикой. Тогда отфильтрованный *m-отклик* сигнала на выходе ЦФ будет определяться как

$$y_m = x_0 h_m + x_1 h_{m-1} + x_2 h_{m-2} + ... + x_{m-1} h_1 + x_m h_0,$$

или

$$y_m = \sum_{k=0}^{N=m} x_k h_{m-k}. \qquad (4.95)$$

Алгоритм цифровой фильтрации *- это процесс получения выходной последовательности как результат дискретной свертки входного сигнала и дискретной импульсной характеристики фильтра.*

Передаточные функции и структурные схемы цифровых фильтров. *Разностное уравнение фильтра.* Основные характеристики ЦФ (передаточная функция, частотные характеристики, временные характеристики) и принципы построения структурной схемы ЦФ удобно рассматривать без учета цифрового кодирования дискретных отсчетов сигнала, т.е. для случая дискретных сигналов и фильтров, описываемых разностными уравнениями.

Последовательностью выборок - отсчетов $x(nT)$ дискретного сигнала является решетчатый сигнал, который описывается решетчатой функцией $\{x(nT)\}$, или $x_T(t)$.

Математический аппарат решетчатых функций является дискретным аналогом дифференциального и интегрального исчисления. Его называют аппаратом конечных разностей и сумм. Так, обычному дифференциальному уравнению соответствует дискретное уравнение в конечных разностях.

Интегралу от функции времени $x(t)$ соответствует сумма ряда дискретных значений x_n:

$$\int_{t=0}^{\infty} x(t)\,dt \to \sum_{n=0}^{\infty} x_n = \sum_{k=0}^{\infty} x_{n-k}\ ,\ n = 0,1,2,\ ...,$$

Поль Адриен Жан Мари Констан Дюамель (Jean-Marie Constant Duhamel, 1797 - 1872), французский математик, член Парижской академии наук (1840), иностранный член-корреспондент Петербургской академии наук (1859). Учился в Политехнической школе и коллеж Луиле-Гран в Париже. Был профессором анализа в Нормальной и Политехнической школах в Париже (с 1834 г. - профессор). Важнейшие работы касаются математической физики, в частности теории колебаний, теории рядов и теории упругости. Сформулировал принцип - аналог метода вариаций постоянных.

имеющая конечное значение для реальных конечных сигналов $t \le T_c < \infty$, или по количеству отсчетов $n \le N < \infty$. *Дискретное уравнение в конечных разностях является математической основой построения алгоритма дискретной, или цифровой, обработки сигналов.*

Дискретный* (*цифровой*) *фильтр описывается *разностным уравнением, связывающим последовательности дискретных отсчетов сигнала как на входе, так и выходе фильтра.*

Принцип и цель цифровой фильтрации заключается в том, чтобы имеющуюся последовательность из N отсчетов $x(nT)$ входного сигнала с помощью операций над ней преобразовать в желаемую последовательность отсчетов $y(nT)$ на выходе фильтра.

Для линейного ЦФ справедливо следующее: суммы взвешенных выборок сигналов на входе и выходе фильтра равны между собой:

$$\sum_{m=0}^{M} b_m x\left[(n-m)T\right] = \sum_{m=0}^{N} a_m y\left[(n-m)T\right], \qquad (4.96)$$

где $x(nT)$ и $y(nT)$ - *n*-выборки-отсчеты сигнала соответственно на входе и выходе фильтра; a_m, b_m - весовые коэффициенты выборок (коэффициенты фильтра).

Из уравнения (4.96) можно найти *n-выборку-отсчет* выходного сигнала $y(nT)$, который формируется при $m = 0$ и $a_0 = 0$:

$$y(nT) = \sum_{m=0}^{M} b_m x\left[(n-m)T\right] - \sum_{m=0}^{N} a_m y\left[(n-m)T\right], \quad (4.97)$$

где всегда $M \le N$ - соответственно входных и выходных отсчетов, формирующих выходной сигнал, причем большее из чисел M и N - порядок фильтра и при условии положительного времени $t = TN \ge 0$ (при $n \ge 0$) и при нулевых начальных условиях.

ЦФ с линейными свойствами *- фильтр, весовые коэффициенты которого a_m, b_m зависят только от номера индекса m, но не зависят от величины сигнала $x(mT)$, $y(mT)$.*

Реакция фильтра $y(nT)$ в некоторый момент $t_n = nT$ определяется значением входного сигнала $x(nT)$ в этот момент, а также линейной комбинацией всех предыдущих значений входного $x[(n-m)T]$ и выходного $y[(n-m)T]$ сигналов.

Рекурсивный ЦФ (РЦФ) *- фильтр с обратными связями* $y[(n-m)T]$, *т.е. зависимость выходного сигнала* $y(nT)$ *от своих же предыдущих значений* $y[(n-m)T]$. Обратная связь $y[(n-m)T]$ улучшает функциональные возможности фильтра.

Если же обратных связей нет, фильтр называют *нерекурсивным.*

Не рекурсивный ЦФ (НЦФ) *- фильтр, выходной сигнал которого* $y(nT)$ *в момент времени* $t_n = nT$ *определяется только отсчетами входного сигнала в этот и все предыдущие моменты.* За счет $a_m = 0$ разностное уравнение НЦФ имеет вид

$$y(nT) = \sum_{m=0}^{M} b_m x[(n-m)T]. \tag{4.98}$$

Передаточные функции. ***Передаточной (системной) функцией*** $H(Z)$ *фильтра называется отношение Z-изображений выходного* $y(nT)$ *и входного* $x(nT)$ *сигналов при нулевых начальных условиях:*

$$H(Z) = \frac{Y(Z)}{X(Z)}, \tag{4.99}$$

$$H(Z) = \frac{Y(Z)}{X(Z)} = \frac{\sum_{m=0}^{M} b_m Z^{-m}}{1 + \sum_{m=0}^{M} a_m Z^{-m}}. \tag{4.100}$$

Если в формуле (4.100) $a_1 = a_2 = ... = a_n = 0,$ то получаем выражение для передаточной функции НЦФ

$$H(Z) = \sum_{m=0}^{M} b_m Z^{-m}. \tag{4.101}$$

Выражение (4.101) можно представить также в развернутом алгебраическом дробно-рациональном виде

$$H(Z)=\frac{b_0+b_1Z^{-1}+b_2Z^{-2}+...+b_MZ^{-M}}{1+a_1Z^{-1}+a_2Z^{-2}+...+a_MZ^{-M}}. \qquad (4.102)$$

Из выражения (4.102) следует, что передаточная функция РЦФ как ФКП имеет на комплексной *Z-плоскости* нули - корни многочлена числителя, а также полюса - корни многочлена знаменателя.

Количество и размещение полюсов $H(Z)$ *на Z-плоскости определяют устойчивость и реализуемость ЦФ.*

ЦФ устойчив и физически *реализуем, если в нем нет реакции к началу входного сигнала. А это возможно, если* $H(Z)$ *соответственно в формуле* (*4.99*) *является полиномом с отрицательными степенями* $|Z|<1$, *причем порядок полинома числителя не должен превышать порядок полинома знаменателя.*

В самом деле, поскольку Z^{-m} означает задержку на *m* тактов, то выражение с положительными степенями Z означало бы, что сигнал на выходе фильтра, «взвешенный» функцией $H(Z)$, опережает входной сигнал, что является невозможным.

Следовательно, в физически реализуемом ЦФ, количество элементов задержки на T (на Z^{-1}) в канале обратных связей (количество коэффициентов a_m) должно быть не меньше, чем количество таких элементов в канале прямых связей (количество коэффициентов b_m). При отсутствии обратных связей $(a_m=0)$ НЦФ является абсолютно устойчивым.

Таким образом, для физической реализуемости РЦФ его передаточная функция $H(Z)$ *должна иметь в числителе* (*4.99*) *полином, порядок которого не выше, чем в знаменателе:* $M \le N$.

При $M>N$ в разностном уравнении фильтра появляются выборки вида $x\left[(n+m)T\right]$, опережающие входной сигнал $x(nT)$, что является невозможным для фильтра, работающего в реальном времени, поскольку для определения $y(nT)$ нужно будет знать будущий, еще несуществующий сигнал $x(nT+mT)$.

Условие устойчивости $|Z|<1$ означает, что РЦФ тем устойчивее, чем дальше вглубь от единичного круга содержатся полюса его $H(Z)$. При $|Z_n|=1$РЦФ на границе неустойчивости легко возбуждается, что является недопустимым.

Из-за округления кодов или отображения коэффициентов ЦФ выражение для $H(Z)$ выходит неточным, полюса определяются неточно, и это по-

вышает риск неустойчивости РЦФ. Этот риск возрастает с увеличением порядка РЦФ.

Обобщенные структурные схемы ЦФ. ***Структурная схема ЦФ*** *должна содержать устройства, выполняющие все операции желательной обработки сигнала* $x(nT)$, *необходимые для преобразования его в выходной сигнал* $y(nT)$.

Набор и порядок этих операций образовывают алгоритм цифровой фильтрации, которая описывается разностным уравнением ЦФ

$$y(nT)=\sum_{m=0}^{M} b_m x\left[(n-m)T\right]-\sum_{m=0}^{N} a_m y\left[(n-m)T\right],$$

и является линейной комбинацией двух сверток отсчетов сигнала $x(nT)$ и $y(nT)$ с коэффициентами фильтра a_m и b_m соответственно в прямых и обратных связях. При этом прибавляются взвешенные предыдущие отсчеты, задержанные на m периодов дискретизации T.

Передаточная функция ЦФ имеет вид

$$H(Z)=\frac{\sum_{m=0}^{M} b_m Z^{-m}}{1+\sum_{m=0}^{M} a_m Z^{-m}}$$

и содержит задержанные на m тактов коэффициенты a_m и b_m (множитель задержки Z^{-m}). Поэтому указания о структурном составе схемы ЦФ одинаково присутствуют в обоих выражениях.

Отсюда следует, что структурную схему ЦФ можно построить в соответствии с видом разностного уравнения или передаточной функции.

В частности, алгоритм разностного уравнения, как и дискретная свертка, содержит три вида операций, поэтому в схеме ЦФ должны быть три вида операционных устройств:

1) *запоминающие устройства (регистры) - для задержки предыдущих отсчетов на шаг дискретизации T* (см. рис. 4.76, *а*);

2) *множительные устройства - для взвешивания отсчетов сигнала весовыми коэффициентами* a_m, b_m, *записанными в памяти* (рис. 4.76, *б)*;

3) *сумматоры - для добавления* (*вычитания*) *взвешенных задержанных отсчетов* (рис. 4.76, *в).*

Заметим, что большей частью структурная схема фильтра лишь указывает, какие именно операции и в какой последовательности должны быть выполнены для получения выходного сигнала, но не определяет аппаратной реализации фильтра. На рис. 4.76, *г* изображена структурная схема РЦФ, реализованная непосредственно в соответствии с выражением (4.99) или (4.102).

Такую схему называют *прямой формой реализации ЦФ*. Она содержит один сумматор, $M+N+1$ множителей и $M+N$ элементов задержки. Каждый элемент задержки обеспечивает задержку сигнала, поступающего на его вход, на время, равное интервалу дискретизации $T = M+N+1$.

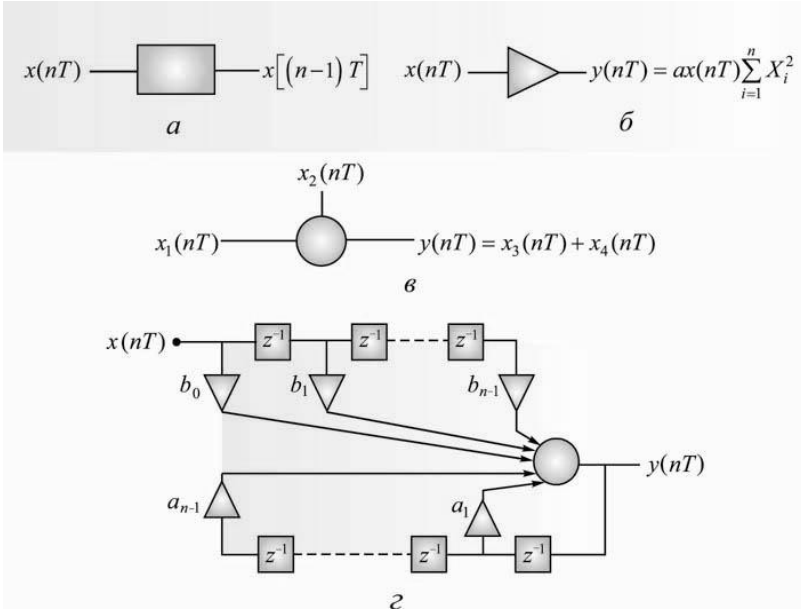


Рис. 4.76. Структурная схема РЦФ

Легко понять, что для реализации НЦФ в соответствии с алгоритмом его разностного уравнения

$$y(nT) = \sum_{m=0}^{M} b_m x\left[(n-m)T\right]$$

нет необходимости в канале обратных связей, поскольку все $a_m = 0$, благодаря чему прямая схема НЦФ существенным образом упрощается (рис. 4.77). Прямая форма структурной схемы НЦФ содержит M элементов задержки, $M+1$ множителей и сумматор на $M+1$ входов. *Эту форму называют также трансверсальным фильтром, или фильтром с многоотводной линией задержки.*

Рассмотрим возможные виды соединения схем ЦФ между собой.

1. Последовательное (каскадное) соединение элементов ЦФ: выходная последовательность отсчетов предыдущего фильтра является входной для следующего. При этом эквивалентная передаточная функция $H_e(Z)$ системы равна произведению передаточных функций $H_1(Z)$ и $H_2(Z)$ отдельных фильтров

$$H_{\text{å}}(Z) = H_1(Z)H_2(Z).$$

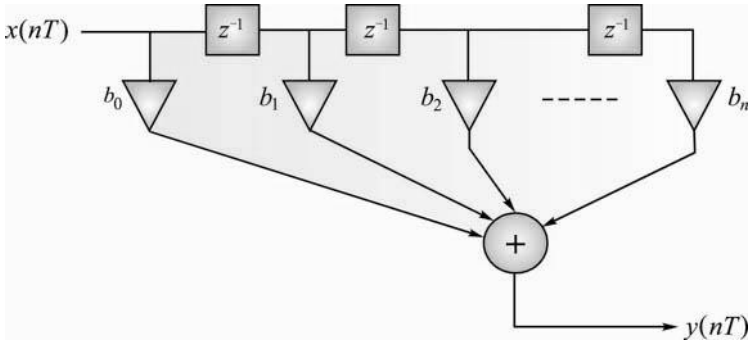


Рис. 4.77. Структурная схема НЦФ

2. *Параллельное соединение элементов ЦФ*: входная последовательность отсчетов во всех фильтрах одна и та же, а выходная последовательность отсчетов системы равна сумме выходных последовательностей отсчетов отдельных фильтров; при этом эквивалентная передаточная функция системы равна сумме передаточных функций отдельных фильтров:

$$H_{\text{Э}}(Z) = H_1(Z) + H_2(Z).$$

3. *Соединение с обратной связью элементов ЦФ*: выходная последовательность отсчетов одного фильтра подается на вход другого (рис. 4.78), причем возможна отрицательная и положительная обратная связь. Здесь эквивалентная передаточная функция системы:

$$H_{\text{Э}}(Z) = H_1(Z) / \left[1 \pm H_1(Z)H_2(Z)\right],$$

где знак «+» соответствует отрицательной обратной связи, а знак «-» - положительной.

А теперь рассмотрим *каноническую форму реализации РЦФ*, дающую возможность сократить количество элементов задержки до значения, равного порядку ЦФ, поскольку в ней те же регистры задержки используются в каналах прямой и обратной связи одновременно.

4. *Каноническая форма реализации ЦФ*: необходимо представить передаточную функцию РЦФ в виде произведения двух передаточных функций:

$$H(Z) = H_1(Z)H_2(Z),$$

где $H_1(Z) = 1 \Big/ \left(1 + \sum_{i=1}^{N} a_i Z^{-i}\right)$ - передаточная функция сугубо РЦФ без прямых задержанных связей; $H_2(Z) = \sum_{i=1}^{N} b_i Z^{-i}$ - передаточная функция РЦФ с каналом только прямых связей.

Тогда рекурсивный фильтр можно представить в виде каскадного соединения двух указанных фильтров (см. рис. 4.78), что и является канонической формой РЦФ. Поскольку в фильтрах, реализующих $H_1(Z)$ и $H_2(Z)$, происходит только задержка сигнала $y'(nT)$, то можно использовать только один набор элементов задержки. Хотя в канонической схеме добавляется еще один сумматор, эта форма реализации РЦФ аппаратурно более экономична за счет сокращения количества элементов задержки.

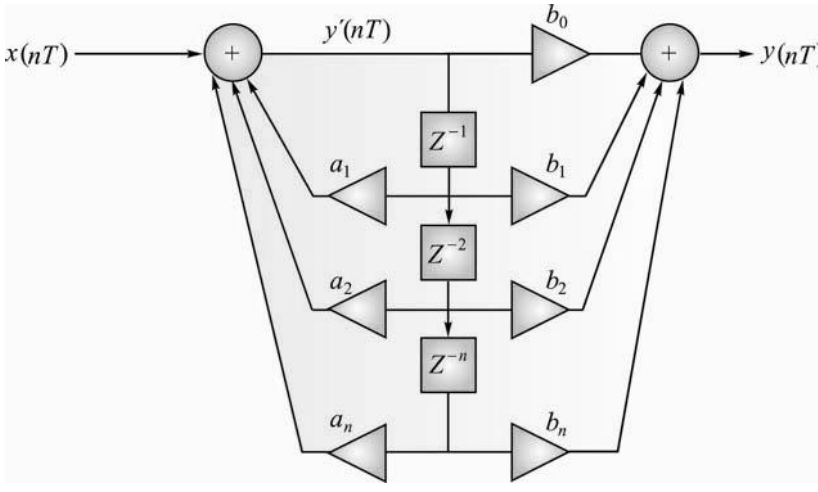


Рис. 4.78. Схема РЦФ

Универсальные базовые блоки ЦФ. Прямая и каноническая формы реализации ЦФ, рассмотренные ранее, не учитывают неточностей задания коэффициентов фильтра a_m, b_m двоичными кодами ограниченной разрядности. Чем выше порядок ЦФ, тем он более зависим от неточности коэффициентов a_m, b_m и тем больший риск неустойчивости фильтра из-за резкого возрастания зависимости положения нулей и полюсов передаточной функции от точности вычисления и задания коэффициентов a_m, b_m двоичным кодом (полюса ЦФ могут выйти за пределы единичной цепи *Z*-плоскости). Для уменьшения количества ошибок, связанных с округлением кодов, нужна большая разрядность кодов, усложняющих фильтр. Иначе возрастает риск неустойчивости (нарушение) фильтра и возрастают шумы округления на его выходе.

ЦФ высокого порядка никогда не реализуют в прямой или канонических формах. В этих формах реализуют ЦФ не выше второго порядка, т.е. с количеством элементов задержки, не превышающего двух.

Фильтры высокого порядка строят на простых базовых блоках 1-го или 2-го порядка, соединяемых последовательно или параллельно. Необходимо помнить, что параллельные соединения базовых блоков означают разбивку передаточной функции фильтра на части $H(Z)=H_1(Z)+H_2(Z)+...$ и одновременную ускоренную обработку сигнала по отдельным составляющим спектра.

Так же и последовательное (каскадное) соединение базовых блоков делает возможным ускоренную текущую обработку сигнала. При последовательном (каскадном) соединении блоков общий вид передаточной функции представляется произведением

$$H_{\text{пос}}=\prod_{i=0}^{S}\mathrm{H}_i^{(1)}(Z)\prod_{i=0}^{R}\mathrm{H}_i^{(2)}(Z). \qquad (4.103)$$

Рассмотрим каноническую форму реализации универсальных блоков 1-го и 2-го порядков.

Блок РЦФ 1-го порядка согласно формуле (4.99) описывается разностным уравнением 1-го порядка

$$y(nT)=b_0x(nT)+b_1x(nT-T)-a_1y(nT-T).$$

Он имеет передаточную функцию 1-го порядка со степенью Z и каноническую схему (рис. 4.79), содержащую один элемент задержки, общий для каналов прямой и обратной связи.

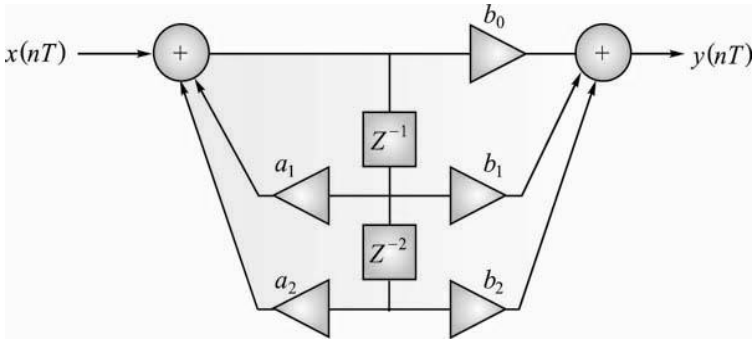


Рис. 4.79. Каноническая схема РЦФ 2-го порядка

Блок РЦФ 2-го порядка описывается разностным уравнением 2-го порядка

$$y(nT)=b_0x(nT)+b_1x(nT-T)-a_1y(nT-T)-a_2y(nT-2T).$$

Он имеет передаточную функцию 2-го порядка со степенью Z и каноническую схему (см. рис. 4.79), содержащую два элемента задержки, общих для каналов прямой и обратной связи.

Эти же стандартные блоки РЦФ 1-го и 2-го порядка в прямой форме реализации с одним сумматором, но удвоенным числом регистров задержки изображены соответственно на рис. 4.80,*а* и 4.80,*б*.

Применяя рассмотренные универсальные базовые блоки 1-го и 2-го порядков в каскадном и параллельном соединениях, получаем различные формы реализации ЦФ, которые при одной и той же передаточной функции могут иметь различные положительные свойства: невысокую чувствительность к неточному вычислению и заданию коэффициентов a_m, b_m и высокую устойчивость, малые шумы квантования и округление на выходе и др.

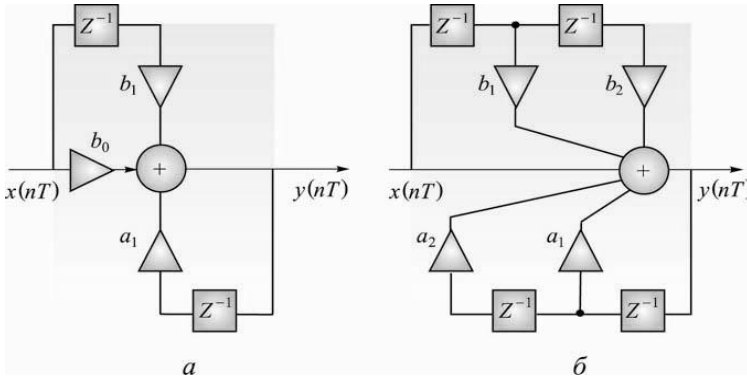


Рис. 4.80. Схемы РЦФ 1-го и 2-го порядков

Временные характеристики. *Временной характеристикой ЦФ* является импульсная характеристика $h(nT)$, т.е. реакция фильтра на входное влияние вида $\delta(nT)=1$ при нулевых начальных условиях.

Импульсную характеристику $h(nT)$ можно рассчитать, решив соответствующие разностные уравнения. Если известна импульсная характеристика ЦФ, то с помощью дискретной свертки можно рассчитать реакцию $y(nT)$ фильтра на любое влияние $x(nT)$:

$$y(nT)=\sum_{m=0}^{M} h(mT)\, x\left[(n-m)T\right]=\sum_{m=0}^{M} h\left[(n-m)T\right]x(mT).$$

Приведенное выражение дает возможность определить другую временную характеристику ЦФ - *переходную характеристику* $y(nT)$, являющуюся

реакцией фильтра на входное влияние вида $\delta(nT)=1$ при нулевых начальных условиях.

Связь между передаточной функцией и временными характеристиками фильтра можно определить с помощью пары *Z-преобразований*:

$$H(Z)=\sum_{n=0}^{\infty} h(nT)\,Z^{-n}, \tag{4.104}$$

$$h(nT)=\frac{1}{2\pi j}\oint_{|Z|=1} H(Z)Z^{n-1}dZ, \tag{4.105}$$

т.е. *Z-изображение* импульсной характеристики совпадает с передаточной функцией ЦФ. *Таким образом, импульсную характеристику можно найти как обратное Z-преобразование передаточной функции.*

Связь передаточной функции с переходной характеристикой фильтра имеет вид

$$g(nT)=\frac{1}{2\pi j}\oint_{|Z|=1}\frac{H(Z)}{1-Z^{n-1}}Z^{n-1}dZ. \tag{4.106}$$

В случае конечной импульсной характеристики $h(nT)$ при $n\geq M+1$ имеем

$$H(Z)=\sum_{n=0}^{\infty} h(nT)\,Z^{-n}. \tag{4.107}$$

Сравнивая полученное выражение (4.106) с выражением (4.99) для передаточной функции НЦФ, приходим к выводу, что коэффициенты b_m НЦФ являются отсчетами конечной импульсной характеристики $h(nT)$. Поэтому часто нерекурсивный ЦФ называют фильтром с конечной импульсной характеристикой - КИХ-фильтром. Импульсная характеристика дает также возможность судить о физической реализуемости и устойчивости нелинейного дискретного ЦФ.

Критерием физической реализуемости фильтра *является равенство нулю отсчетов импульсной характеристики при отрицательных значениях моментов отсчетов*

$$h(nT)=0 \text{ при } n<0.$$

Частотные характеристики фильтров. Комплексную частотную характеристику ЦФ можно вывести из выражения для его передаточной функции, перейдя из *Z-плоскости* в пространство частот, к гармоническим функциям, заменив $Z=e^{j\omega T}$:

$$Z^{-m}=e^{-j\omega T}=\cos(m\omega T)-j\sin(m\omega T);$$

$$H(j\omega)=\frac{\sum\limits_{m=0}^{M} b_m e^{-jm\omega T}}{1+\sum\limits_{n=0}^{N} a_m e^{-jm\omega T}}.$$

Модулем этой функции является ***амплитудно-частотная характеристика РЦФ***:

$$H(\omega)=\frac{\sqrt{\left[\sum\limits_{m=0}^{M} b_m \cos(m\omega T)\right]^2+\left[\sum\limits_{m=0}^{M} b_m \sin(m\omega T)\right]^2}}{\sqrt{\left[\sum\limits_{m=0}^{M} a_m \cos(m\omega T)\right]^2+\left[\sum\limits_{m=0}^{M} a_m \sin(m\omega T)\right]^2}} \qquad (4.108)$$

а аргументом - ***ФЧХ РЦФ***:

$$\theta(\omega)=\operatorname{arctg}\left[\frac{\sum\limits_{m=0}^{M} a_m \sin(m\omega T)}{1+\sum\limits_{m=0}^{M} a_m \cos(m\omega T)}\right]-\operatorname{arctg}\left[\frac{\sum\limits_{m=0}^{M} b_m \sin(m\omega T)}{\sum\limits_{m=0}^{M} b_m \cos(m\omega T)}\right]. \qquad (4.109)$$

Для НЦФ выражение АЧХ отличается от формулы (4.108) тем, что содержит только числитель (4.108), а выражение ФЧХ содержит только второй член (4.109). Выражения (4.108) и (4.109) дают возможность на любой частоте входного дискретного гармонического сигнала определять амплитуду и фазу этого сигнала на выходе ЦФ.

АЧХ и ФЧХ ЦФ - функции периодические, они повторяются через каждый интервал частот $\Delta\omega=2\pi/T$, который определяется шагом дискретизации T (шириной спектра обрабатываемого сигнала - рис. 4.81), поскольку в обоих случаях периодичность определяется дискретизацией процесса обработки.

Фильтрующие свойства ЦФ оценивают по форме АЧХ, задавая которую выполняют и расчет ЦФ. В этой связи полезно напомнить некоторые основные формы АЧХ выборочных (селективных) фильтров, широко применяемых на практике.

На рис. 4.81, *а* изображены АЧХ ФНЧ с полосой пропускания ППР до частоты среза $\omega_{ср}$, переходной полосой ПП от $\omega_{ср}$ к $\omega_з$ и полосой задержки на частотах, высших, чем $\omega_{ср}$. ФНЧ пропускает все низкие частоты $0\le\omega\le\omega_{ср}$, включая постоянную составляющую с частотой $\omega=0$, но подавляет высшие частоты спектра сигнала $\omega\ge\omega_{ср}$. Для повышения выборочных свойств ФНЧ нужна узкая переходная полоса ПП, т.е. крутой спад АЧХ.

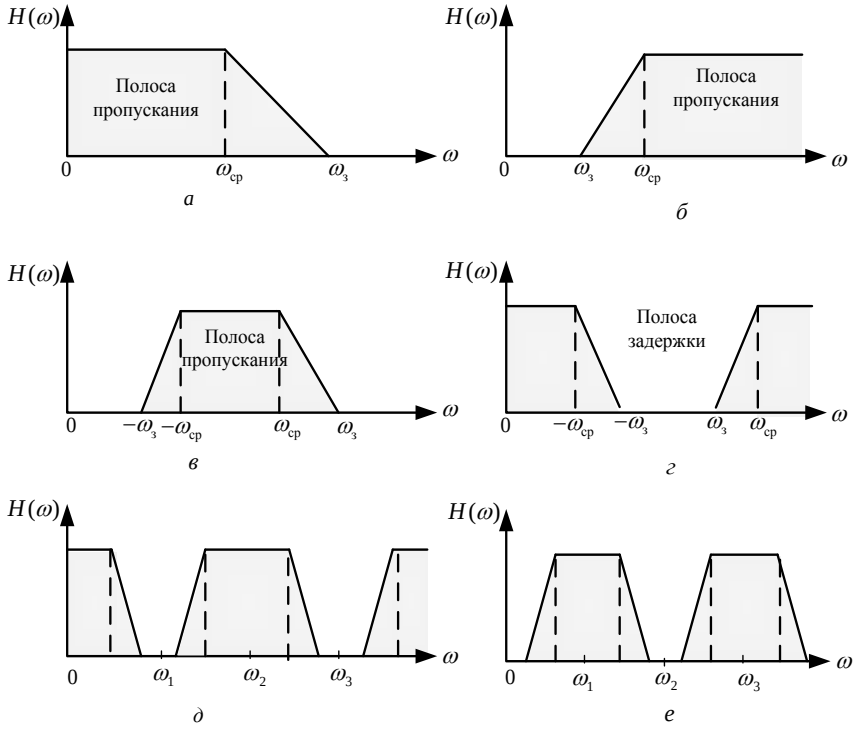


Рис. 4.81. Частотные характеристики фильтров

На рис. 4.81, *б* приведена АЧХ ФВЧ, которая не пропускает низких частот из постоянной составляющей, но есть прозрачной для высоких частот $\omega_з \geq \omega_{ср}$. Избирательность ФВЧ также прямо зависит от крутизны АЧХ в переходной полосе. Широко применяются *полосовые фильтры* (СФ) (см. рис. 4.81, *в)*, выделяющие заданную полосу частот шириной $\Delta\omega = \omega_з - \omega_{ср}$.

Фильтр, противоположный ПФ (рис. 4.81, *г)*, называют *режекторным* (РФ), он используется для подавления (режекции) определенной полосы частот шириной $\Delta\omega = \omega_{ср} - \omega_з$.

Указанные фильтры в дискретном варианте имеют периодически (с интервалом $\Omega_T = 2\pi / T$) повторяемую АЧХ вида «гребенка». Так образовываются *гребенчатые фильтры*. Если периодически повторяется АЧХ ФНЧ (см. рис. 4.81, *д)*, то образовывается *полосовой гребенчатый фильтр* (СГФ), прозрачный для частот $\omega = 0, \omega_2, \omega_4$, но этот же фильтр является режекторным гребенчатым (РГФ) для частот $\omega_1, \omega_3, \omega_5$. На рис. 4.81, *е* изображен РГФ для частот $\omega = 0, \omega_2, \omega_4$, образованный периодической АЧХ фильтра верхних частот, но он же является СГФ для частот $\omega_1, \omega_3, \omega_5$.

ПГФ широко применяют при согласованной фильтрации для накопления сигнала, а РГФ полезен, например, при «селекции движущихся целей» - для подавления мешающих отображений от неподвижных объектов радиолокационного наблюдения.

Приведенные на рис. 4.81 формы АЧХ идеализированы.

Анализ характеристик ЦФ 1-го и 2-го порядка. ***НЦФ 1-го порядка.*** НЦФ 1-го порядка имеет простейшую схему (рис. 4.82, *а*) с разностным уравнением

$$y(nT) = b_0 x(nT) + b_1 x(nT). \tag{4.110}$$

Из *Z-изображения* разностного уравнения $Y(Z) = b_0 X(Z) + b_1 X(Z) Z^{-1}$ легко получить передаточную функцию (см. рис. 4.82, *б)*

$$H(Z) = Y(Z) / X(Z) = b_0 + b_1 Z^{-1} = b_0 Z + b_1 / Z, \tag{4.111}$$

из которой следует, что единственный нуль фильтра находится в точке

$$Z_0 = -b_1 / b_0. \tag{4.112}$$

ФЧХ НЦФ 1-го порядка при $b_0 = 1$ описывается формулой

$$\theta(\omega) = -\operatorname{arctg} \frac{b_1 \sin \omega T}{1 + b_1 \cos \omega T}. \tag{4.113}$$

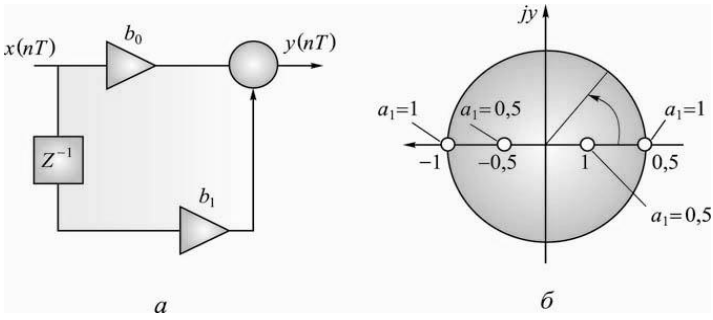


Рис. 4.82. Схема нерекурсивного НЦФ 1-го порядка

Импульсную и переходную характеристику НЦФ 1-го порядка при $b_0 = 1$ можно определить, воспользовавшись разностным уравнением. Соответственно имеем

$$h(nT) = \delta(nT) + b_1 \delta(nT - T), \tag{4.114}$$

$$y(nT) = 1(nT) + 1b_1 (nT - T). \tag{4.115}$$

НЦФ 2-го порядка. НЦФ 2-го порядка имеет схему (рис. 4.83), которая формируется согласно разностному уравнению

$$y(nT) = \sum_{m=0}^{N} b_m x\left[(n-m)T\right]$$

и алгоритму его работы

$$y(nT) = x(nT) + b_1 x(nT - T) + b_2 x(nT - 2T). \tag{4.116}$$

Передаточную функцию фильтра представляем в виде

$$H(Z) = \frac{Y(Z)}{X(Z)} = 1 + b_1 Z^{-1} + b_2 Z^{-2} = \frac{Z^2 + b_1 Z + b_1}{Z^2}. \tag{4.117}$$

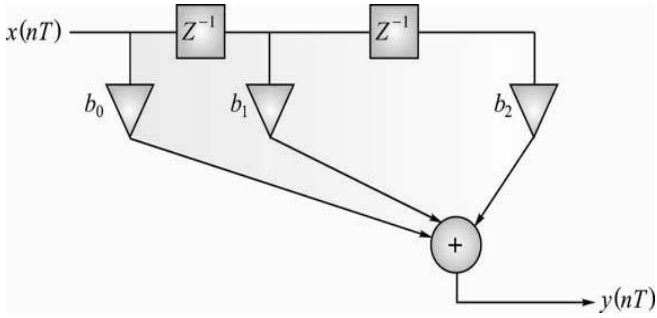


Рис. 4.83. Структурная схема НЦФ 2-го порядка

Результаты расчета характеристик НЦФ 2-го порядка для различных значений *b* изображены на рис. 4.84.

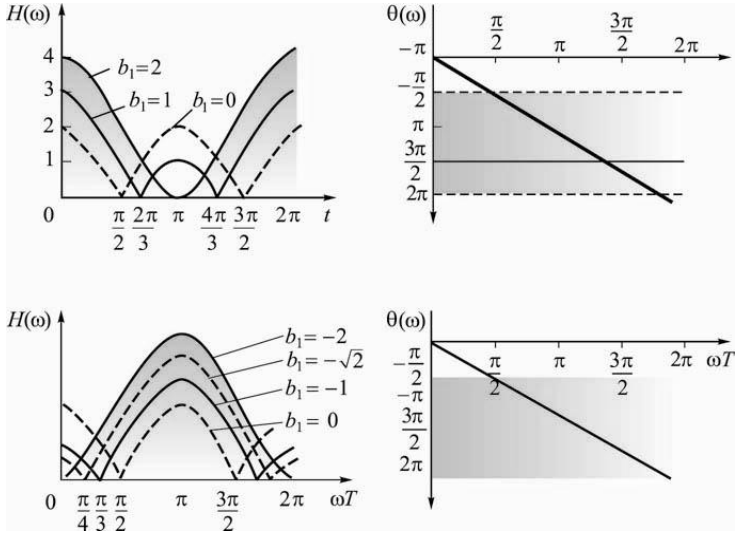


Рис. 4.84. Частотные характеристики НЦФ 2-го порядка

РЦФ 2-го порядка. РЦФ 2-го порядка - наиболее универсальный стандартный блок ЦФ, из которых обычно образовываются сложные ЦФ высокого порядка.

Для упрощения анализа РЦФ 2-го порядка рассмотрим его сугубо рекурсивный вариант без прямых связей, т.е. при $b_1 = b_2 = 1$. Схема такого фильтра приведена на рис. 4.85, *а*. Алгоритм его работы

$$y(nT) = x(nT) - a_1 y(nT - T) - a_2 y(nT - 2T),$$

или в виде *Z-образа* (см. рис. 4.85, *б)*

$$Y(Z) = X(Z) - a_1 Y(Z) Z^{-1} - a_2 Y(Z) Z^{-2}.$$

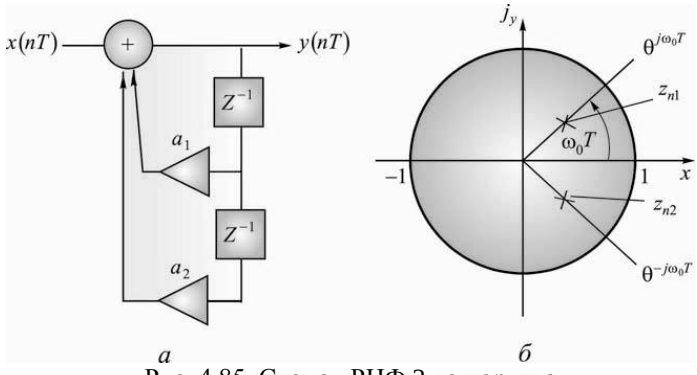


Рис. 4.85. Схема РЦФ 2-го порядка

Итак, передаточная функция

$$H(Z) = \frac{Y(Z)}{X(Z)} = \frac{1}{1 + a_1 Z^{-1} + a_2 Z^{-2}} = \frac{Z^2}{Z^2 + a_1 Z + a_2}.$$

Подставляя ее в последнее выражение для $H(j\omega)$ при $a_1 = 0$ и $a_2 = Z_{n2}$, получаем АЧХ блока РЦФ 2-го порядка в виде

$$H(\omega) = |H(j\omega)| = \frac{1}{\sqrt{(1 + a_2^2) + 2a_2 \cos 2\omega T}}. \qquad (4.118)$$

Графики АЧХ для различных значений *а* приведены на рис. 4.86.

С приближением a_2 к 1, а также с ростом $a_2 \neq 0$ добротность фильтра, острота максимумов его АЧХ и избирательные свойства возрастают. Но при этом полюса Z_n приближаются к единичному кругу, и возрастает риск потери устойчивости РЦФ.

Форму АЧХ можно рассматривать как полосовой фильтр с почти прямоугольной полосой пропускания или как режекторный фильтр высокоселективного глушения частот, близких к $f = 0$ и $f = F_T = 1/T$. Такие свойства

полезны, например, при селекции движущихся целей (для отсечения, подавления вблизи нулевых частот, присущих изображениям — помехам от недвижимых объектов и пассивных препятствий).

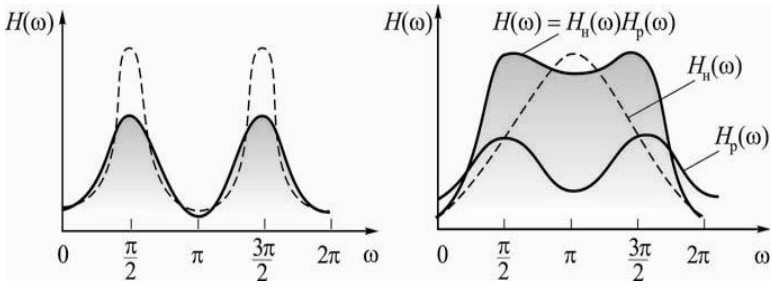


Рис. 4.86. Амплитудно-частотная характеристика РЦФ 2-го порядка

ФЧХ чистого РЦФ 2-го порядка (при $a_1 = 0$) описывается линейной функцией

$$Q(\omega) = \operatorname{arctg} \frac{a_2 \sin 2\omega T}{1 + a_2 \cos 2\omega T}. \qquad (4.119)$$

Заметим, что применение различных комбинаций фильтров 1-го и 2-го порядков при их каскадном или параллельном соединении дает возможность строить ЦФ с различными сложными формами частотных характеристик и разными фильтрующими свойствами, которые зависят от значений коэффициентов в прямых и обратных связях блоков.

Синтез цифровых фильтров. Для построения структурной схемы ЦФ с желательными частотно-селективными свойствами обычно задаются соответствующей этим свойствам идеальной частотной характеристикой (ЧХ), имеющей необходимые полосы пропускания и задержки, как показано на рис. 4.87. Однако идеальную гладкую ЧХ с резкими (крутыми) перепадами между полосами пропускания и задержки физически (технически) *реализовать невозможно*. Поэтому прибегают к ее приближенному образу - аппроксимации, график которой очень близок к идеальной ЧХ и вместе с тем физически реализуем.

На рис. 4.87, например, изображена идеальная прямоугольная АЧХ ФНЧ, заданная выражением

$$H(\omega) = \begin{cases} 1 \text{ при } 0 < \omega \le \omega_з, \\ 0 \text{ при } \omega > \omega_з, \end{cases}$$

и ее аппроксимация, реализованная физически с искажениями в виде пульсаций в полосах пропускания и задержки, а также растянутого перехода между полосами шириной $\omega_k - \omega_з$, где ω_k и $\omega_з$ - предельные частоты полос пропускания и задержки.

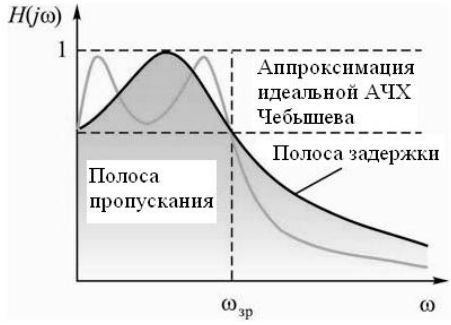


Рис. 4.87. Аппроксимация идеальной АЧХ

Чем меньше параметры искажения аппроксимации по сравнению с идеальной ЧХ, тем качественнее аппроксимация, а значит, и фильтр, построенный в соответствии с ней, в большей мере отвечает желательным селективным свойствам фильтра и его задачам.

Подавлением фильтра *называют величину* $a(\omega) = 20\lg\left[1/H(\omega)\right]$. *Идеальная АЧХ имеет нулевое подавление в полосе пропускания и бесконечное - в полосе задержки* (см. рис. 4.87). Любая аппроксимация АЧХ имеет конечное подавление (ослабление) в этих полосах $a_{\max}$ и $a_{\min}$.

Частотная характеристика ЦФ имеет три составляющих:

1) амплитудно-частотную (АЧХ) - модуль ЧХ $H(\omega) = \left|H(j\omega)\right|$;

2) фазочастотную (ФЧХ) - аргумент ЧХ $Q(\omega) = \arg H(j\omega)$;

3) групповое время задержки сигнала $t_г(\omega) = -\dfrac{dQ(\omega)}{d\omega}$.

В случае линейной ФЧХ $t_г(\omega) = r_0 = \text{const}$. Такой ЦФ задерживает все составляющие спектра входного сигнала на одинаковое время.

При нелинейной ФЧХ характеристика $t_г(\omega)$ неравномерна. Во многих практических применениях это не имеет значения, и тогда интересуются только амплитудной характеристикой ЦФ - аппроксимацией АЧХ.

Комплексная ЧХ ЦФ H(jω) *является результатом подстановки* $Z = e^{j\omega T}$ *в его передаточную функцию H(Z).*

На практике действуют наоборот: если выбрана надлежащая аппроксимация ЧХ $H(j\omega)$, ее дискретизируют (считая $t = nT$) и находят ее *Z-образ* $H(Z)$. Полученную передаточную функцию $H(Z)$ фильтра записывают

обычно в виде дробно-рационального выражения с коэффициентами a_m, b_m. Далее, в соответствии с выражением $H(Z)$, с найденными коэффициентами a_m, b_m осуществляют выбор конкретной структурной схемы ЦФ, которая может быть прямой, канонической, каскадной, параллельной и др.

Методы аппроксимации РЦФ. Универсального метода решения задачи аппроксимации для ЦФ пока не существует. Каждый из разработанных методов соответствует условиям лишь конкретных практических задач разработки ЦФ.

Выбор метода решения задачи аппроксимации - первая ответственная задача синтеза ЦФ.

Целью решения задачи аппроксимации является определение коэффициентов a_m, b_m передаточной функции $H(Z)$ фильтра.

К передаточной функции H(Z) РЦФ выдвигаются важные требования:

1. $H(Z)$ должна быть рациональной функцией (в виде алгебраического или тригонометрического полинома *N*-го порядка) с действительными коэффициентами a_m, b_m для того, чтобы входные и выходные сигналы регистров задержки, множителей, сумматоров и т.д. были также действительны.

2. Порядок числителя передаточной функции $H(Z)=Y(Z)/A(Z)$ должен быть не больше порядка знаменателя, с целью обеспечения физической реализации ЦФ. Это требование означает, что импульсная характеристика ЦФ должна быть определена только в области отрицательного времени: $h(nT)=0$ при $n<0$.

3. Полюса функции $H(Z)$ должны лежать внутри единичного круга *Z-плоскости* $(|Z_n<1|)$, чтобы ЦФ был устойчивым. Если передаточная функция $H(Z)$ имеет полюса, размещенные вне единичного круга или на нем, то рекурсивный ЦФ также может быть устойчив. Это возможно при условии, что числитель функции $H(Z)$ имеет корни в тех точках, в которых корни знаменателя размещены вне единичного круга или на нем.

Для решения задачи аппроксимации ЦФ применяют две группы методов: *прямые и прототипные методы.*

Прямые методы допускают расчет ЦФ непосредственно по заданной частотной или импульсной характеристике путем отбора их математических аппроксимаций (приближенных описаний, моделей).

В случае прямого расчета РЦФ (например, по заданным ЧХ) для описания квадрата АЧХ $H^2(\omega)$ подбирают тригонометрический полином; также в результате прямого расчета возможен подбор отсчетов импульсной характеристики. В обоих случаях прямой расчет РЦФ не обеспечивает высокой точности: он связан с громоздкими преобразованиями и применяется реже, чем

прототипные методы. Прямые методы расчета применяют, в основном, для решения задачи аппроксимаций НЦФ с конечной импульсной характеристикой, которая, не имея обратных связей, описывается более простыми выражениями. Такие выражения проще преобразовывать и вычислять по ним отсчеты. И, вдобавок, НЦФ не имеют аналогов среди пассивных *LC*-фильтров из-за своей конечной импульсной характеристики (ИХ).

Обратные (прототипные) методы, напротив, широко применяют для расчета РЦФ бесконечной ИХ. В этих методах используются приемы расчета аналоговых фильтров, соответствующих заданным выборочным свойствам искомого ЦФ, и которые являются прототипами рассчитываемых РЦФ.

При расчете РЦФ по аналоговому прототипу (АФ-прототипу) возможны три подхода в процедуре:

преобразование полосы - дискретизация;

дискретизация - преобразование полосы;

дискретизация с преобразованием полосы.

Дело в том, что при расчете аналоговых фильтров с различной полосой (ФНЧ, ФВЧ, СФ, РФ) за основу обычно берут нормированный (нормализованный, единичный) ФНЧ, имеющий единичную полосу пропускания, т.е. $H(Q)=1$ при $Q=1$.

Именно для него выведены формулы аппроксимации заданных АЧХ по Баттерворту, Чебышеву и другие, применяемые в справочной литературе.

Чтобы со временем перейти к заданному виду фильтра (ФНЧ с произвольной полосой, ФВЧ, СФ, РФ), необходимо выполнить «преобразования полосы», т.е. перейти от полосы нормализованного ФНЧ к полосе пропускания (задержки) заданного фильтра. Для этого в формулу передаточной функции $H_{\text{п}}(P)$ аналогового ФНЧ, найденную в справочнике для заданной аппроксимации ЧХ $Q = 1H(j\omega)$, нужно подставить вместо параметра P пересчитанную функцию частоты, вид которой приведен в табл. 4.4.

Таким образом, согласно первому подходу, сначала нормализованный ФНЧ с передаточной функцией $H_{\text{п}}(P)$ преобразуется в заданный фильтр путем замены в $H_{\text{п}}(P)$ оператора P на функцию от P . Далее, полученная $H(P)$ заданного фильтра дискретизируется заменой 1 на nT и преобразуется заменой e^{PT} на Z в дробно-рациональную функцию $H_{\text{п}}(P)$ ЦФ с заданными характеристиками.

Таблица 4.4

Преобразование	Формулы для расчета Ω_k	Замена P на функцию от P
ФНЧ — ФНЧ	$\omega_k/\omega_з$	$P \to P/\omega_з$
ФНЧ — ФВЧ	$\omega_з/\omega_k$	$P \to \omega_з/P$

Преобразование	Формулы для расчета Ω_k	Замена P на функцию от P
ФНЧ — СФ	$\frac{\omega_0}{\Delta\omega}\left(\frac{\omega_k}{\omega_0}-\frac{\omega_0}{\omega_k}\right)$	$P \to \frac{P^2+\omega_0{}^2}{\Delta\omega P}$
ФНЧ — РФ	$\frac{\Delta\omega}{\omega_0}\left(\frac{\omega_k}{\omega_0}-\frac{\omega_0}{\omega_k}\right)^{-1}$	$P \to \frac{\Delta\omega P}{P^2+\omega_0{}^2}$
Примечания	$\Delta\omega=\omega_з-\omega_{-з}$	$\omega_0=\sqrt{\omega_з-\omega_{-з}}$

Согласно второму подходу, наоборот, нормализованный ФНЧ сразу дискретизируется путем преобразования его аппроксимации $H_п(P)$ в функцию $H_п(Z)$, а дальше выполняется «преобразование полосы» уже в *Z-плоскости* путем замены оператора Z на соответствующую пересчитанную функцию согласно табл. 4.4. После такой замены оператора Z полученная функция $H(Z)$ ЦФ заданного вида преобразуется в дробно-рациональную функцию от Z, удобную для реализации схемы в ЦФ с заданными характеристиками.

Наиболее универсален, однако, третий подход, при котором для избранного АФ-прототипа (по заданным частотам и угасанию) преобразование полосы и дискретизация выполняются одновременно.

В этом случае оператор P передаточной функции $H(P)$ АФ-прототипа заменяют на пересчитанную функцию от Z.

При решении задачи аппроксимации с использованием АФ-прототипа наиболее применимы такие два метода:

1) метод инвариантности ИХ (стандартного Z-преобразования);

2) метод билинейного Z-преобразования.

Метод инвариантности ИХ. При таком методе предполагается известной импульсная характеристика АФ-прототипа $h(t)$, а ее дискретный вариант (инвариант) используется для получения передаточной функции $H(Z)$ ЦФ.

Метод дает хорошие результаты при использовании аналоговых аппроксимаций Баттерворта, Чебышева, Бесселя, если передаточная функция АФ-прототипа имеет только простые полюсы (без нулей) и описывается с помощью разложения на простые дроби вида

$$H(P)=\sum_{K=1}^{N}\frac{B_k}{P-P_k}, \qquad (4.120)$$

где Лаплас - образа $1/(P-P_k)$ соответствует табличная временная функция e^{P_kT}.

Отсюда ИХ такого АФ-прототипа, как обратное преобразование функции Лапласа (4.120), имеет вид

$$h(t)=\sum_{k=1}^{N}B_k e^{P_k t}\,.$$

Ее дискретный вариант (при $t=nT$): $h(nT)=\sum_{k=1}^{N}B_k e^{nP_k t}$, где взяты $h(0)=0$ и условие физической реализации фильтра $h(t)=h(nT)=0$ при $t<0$ или $n<0$.

Зная $h(nT)$, путем прямого *Z-преобразования* можно найти ее *Z-образ* $H(Z)$, т.е. искомую передаточную функцию ЦФ:

$$H(Z)=\sum_{n=0}^{\infty}h(nT)Z^{-n}=\sum_{n=0}^{\infty}\sum_{k}^{N}B_k e^{nP_k t}Z^{-n}\,.$$

Изменяя порядок суммирования и находя табличный *Z-образ* экспоненты $e^{nP_k t}$, окончательно имеем

$$H(Z)=\sum_{k=1}^{N}\frac{B_K}{1-e^{P_k t}Z^{-1}}\,. \qquad (4.121)$$

Полюсу $P_k=\sigma_k+j\omega_k$ АФ-прототипа в функции (4.120) соответствует полюс $Z_k=e^{Pkt}$ функции (4.121). Тогда устойчивому АФ-прототипу (при $\sigma_k<0$) соответствует устойчивый ЦФ (поскольку $|Z_k|=\left|e^{P_k T}\right|<1$). Сравнивая (4.120) и (4.121), видим, что при таком методе аппроксимации, имея выражение *H*(*P*) АФ-прототипа в виде (4.120), можно сразу получить передаточную функцию искомого ЦФ *H*(*Z*) путем замены

$$\frac{B_k}{P-P_k}\to\frac{B_k}{1-e^{P_k T}Z^{-1}}\,. \qquad (4.122)$$

Пример. Найти передаточную функцию $H(Z)$ РЦФ по устойчивому АФ-прототипу 2-го порядка с простыми полюсами P_1 и P_2 в левой *P*-полуплоскости, если $B_1=B_2=1$. К передаточной функции АФ-прототипа при $B_k=1$ и $N=2$, имеющей вид $H(P)=\sum_{k=1}^{2}\frac{1}{P-P_k}$, применить метод инвариантности ИХ.

Р е ш е н и е. Применяя замену, сразу получаем

$$H(Z)=\sum_{k=1}^{2}\frac{1}{1-e^{P_k T}Z^{-1}}=\frac{1}{1-e^{P_1 T}Z^{-1}}+\frac{1}{1-e^{P_2 T}Z^{-1}}=$$

$$=\frac{2-\left(e^{P_1 T}+e^{P_2 T}\right)Z^{-1}}{1-\left(e^{P_1 T}+e^{P_2 T}\right)Z^{-1}+e^{(P_1+P_2)T}Z^{-2}}=\frac{b_0+b_1Z^{-1}}{1-a_1Z^{-1}-a_2Z^{-2}}\,.$$

Получив $H(Z)$ в виде дробно-рационального полинома 2-го порядка с коэффициентами a_m, b_m, легко реализуем схему РНФ, например, в канонической форме (рис. 4.88).

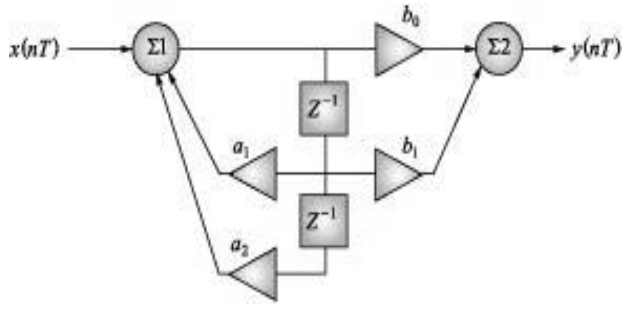


Рис. 4.88. Каноническая форма РЦФ

Преимуществом метода инвариантности ИХ является неизменность АЧХ и ФЧХ. Линейная ФЧХ АФ-прототипа остается также линейной и в рассчитанном РЦФ; причем форма АЧХ также не изменяется, но при следующем условии.

В результате дискретизация ИХ АФ-прототипа получается ЦФ, у которого ЧХ является периодическим повторением ЧХ прототипа с интервалом повторения $F_T = 1/T$. Чтобы периодическая ЧХ ЦФ не искажалась за счет положения соседних полос, полоса пропускания АФ-прототипа должна быть ограниченной шириной интервала повторения $2\Delta f_{\text{АФ}} \le F_T$. Для выполнения указанного неравенства либо повышают частоту дискретизации F_T, либо перед ЦФ вводят АФ-ограничитель спектра сигнала (фильтр-окно).

Метод билинейного Z-преобразования. Согласно методу инвариантности ИХ значения дискретной ИХ РЦФ совпадают с ИХ АФ-прототипа только в отсчетные моменты времени $t_{\text{п}} = nT$, а в другие моменты времени они могут быть различными, а значит, возможна и неточность представления $h(t)$ через $h(nT)$.

При билинейном *Z-преобразовании* реакции АФ и ЦФ практически совпадают для любых одинаковых влияний. Этим предопределяется ряд важных преимуществ билинейного *Z-преобразования*, главное из которых — возможность точно отображать максимально селективные АЧХ (с крутыми перепадами) физически реализованного устойчивого АФ-прототипа в АЧХ РЦФ с такими же свойствами, причем без искажений, присущих методу инвариантности ИХ.

Главные недостатки этого метода - искажение АЧХ и деформация частотной шкалы.

Рассмотрим суть метода билинейного *Z*-преобразования. Как уже отмечалось, задача аппроксимации РЦФ по АФ-прототипу состоит из двух частей:

дискретизации передаточной функции нормализованного ФНЧ выбранного АФ-прототипа;

преобразование его единичной полосы в полосу искомого вида ЦФ.

Обе эти части при билинейном *Z-преобразовании* объединяются, решаются одновременно применением надлежащей замены оператора P некоторой функцией от Z.

Под дискретизацией при расчете РЦФ понимают отображение передаточной функции $H(P)$ АФ-прототипа из комплексной *P*-плоскости на *Z-плоскость* с целью отыскания $H(Z)$ ЦФ. Осуществляется отображение надлежащей заменой оператора *P* на $Z = e^{PT} Z$.

При билинейном *Z-преобразовании* находим линейную замену *P* на *Z*, что сохраняет дробно-рациональный характер передаточной функции, удобный для построения его структурной схемы.

Известно, что *Z-преобразование* дискретной последовательности (*Z-образ*) дает возможность избежать трансцендентности ее Лаплас-образа относительно оператора *P* и придает ему удобную для анализа алгебраическую форму относительно оператора Z. Для этого в Лаплас-образе последовательности трансцендентную зависимость e^{PT} заменяют на оператор Z так что $Z = e^{PT}$, $P = \frac{1}{T} \ln Z$.

Если в дробно-рациональное выражение $H(P)$ физически реализованного АФ-прототипа $H(P) = \frac{b_0 + b_1 P + b_2 P^2 + ... + b_m P^m}{a_0 + a_1 P + a_2 P^2 + ... + a_m P^m}$ подставим значение Z, получим передаточную функцию $H(Z)$, но ее выражение не будет соответствовать ни одному реальному ЦФ, поскольку не будет дробно-рациональным. Устранить эту трансцендентность $H(Z)$ можно лишь приближенно.

Удобным и достаточно точным приближением оказалось представление функции логарифма степенным рядом

$$\ln Z = 2\left[\frac{Z-1}{Z+1} + \frac{1}{3}\left(\frac{Z-1}{Z+1}\right)^3 + ...\right],$$

из которого, ограничиваясь первым членом расписания, получаем билинейную замену выражения

$$P \approx \frac{2}{T} \frac{Z-1}{Z+1}. \qquad (4.123)$$

Теперь подстановка формулы (4.123) в функцию $H(P)$ АФ-прототипа сохраняет дробно-рациональный характер передаточной функции $H(Z)$ и обеспечивает техническую реализацию ЦФ.

В самом деле, легко показать, что при подстановке в формулу (4.123) $P = j\Omega$ и $Z = e^{j\omega T}$ можно, воспользовавшись формулами Эйлера и соотношениями для тригонометрических функций двойного угла, получить зависимость

$$\Omega = \frac{2}{T}\mathrm{tg}\left(\frac{\omega T}{2}\right) = \gamma \mathrm{tg}\left(\frac{\omega T}{2}\right), \qquad (4.124)$$

где γ - обобщенная константа (вместо $2/T$), значение которой не изменяет характера зависимости.

Итак, при изменении «аналоговой» частоты от $-\infty$ к ∞ «цифровая» частота изменяется лишь в интервале однозначности $\left[-\frac{\pi}{T}, \frac{\pi}{T}\right]$.

Нелинейная связь частот АФ и ЦФ деформирует шкалу частот так, что рассчитанная АЧХ (внизу на рис. 4.89) искажается сравнительно с АЧХ АФ-прототипа (вверху по левую сторону на рис. 4.89).

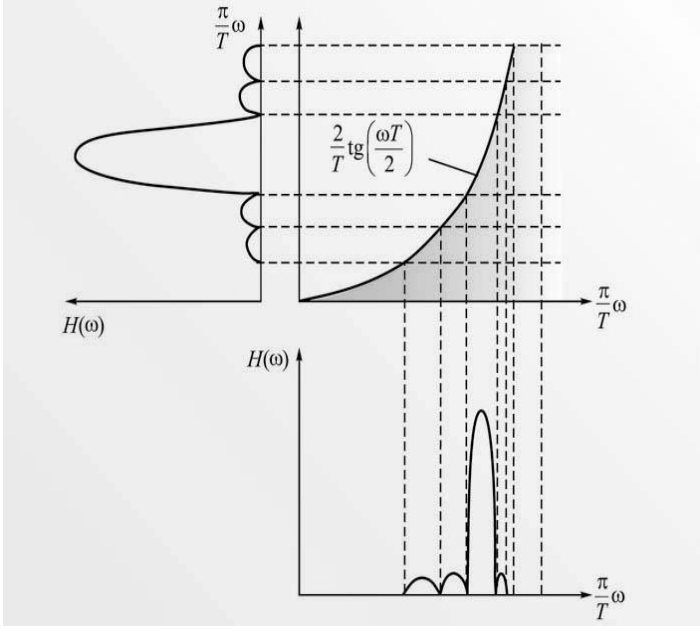


Рис. 4.89. Нелинейная связь частот АФ и ЦФ

Если интересоваться только АЧХ, то деформацию шкалы частот можно компенсировать введением предыдущих искажений в АФ-прототип. Для это-

го шкалу частот выбранного АФ-прототипа искусственно деформируют перед билинейным *Z-преобразованием* так, чтобы ось частот АФ-прототипа градуировалась в «аналоговых» частотах $\Omega_i = \gamma \mathrm{tg}(\omega_i T/2)$, где ω_i - заданные по условию задачи частоты рассчитанного РЦФ (частоты среза полос пропускания и задержки). При таком перерасчете искаженные заранее частоты АФ-прототипа в результате билинейного *Z-преобразования* деформируются в обратном порядке и превращаются в нормальные заданные частоты РЦФ: $\omega_i = \gamma \ \mathrm{arctg}(\Omega_i T/2)$.

Далее, чтобы при билинейном *Z*-преобразовании одновременно с дискретизацией решалась и задача преобразования полосы (задача перехода от нормализованного ФНЧ к другим фильтрам с иной полосой), каждый раз нужно конкретное билинейное преобразование, которое учитывает преобразование полосы. Иначе говоря, для каждого вида фильтра нужна конкретная константа γ в замене Ω_i.

В целом обобщенное билинейное *Z*-преобразование (табл. 4.5) обеспечивает такую замену оператора *P* передаточной функции *H*(*P*) АФ-прототипа, при которой само *Z*-преобразование (без множителя γ) осуществляет дискретизацию и преобразование полосы (вида) фильтра, а константа γ - преобразование шкалы частот. В результате получается полностью эквивалентная (адекватная) аппроксимация ЦФ, при которой передаточная функция $H_{\text{п}}(p)$ нормализованного ФНЧ-прототипа $(\Omega_\varepsilon = 1)$ превращается сразу в передаточную функцию $H(Z)$ искомого РЦФ определенного вида (ФНЧ, ФВЧ, СФ, РФ).

Таблица 4.5

Искомый ЦФ	Формулы замены оператора *P*	Значение константы γ	Связь аналоговых частот Ω_i с цифровыми ω_i
ФНЧ	$\gamma\dfrac{1-Z^{-1}}{1+Z^{-1}}$	$\gamma = \mathrm{ctg}\left(\dfrac{\omega_\varepsilon T}{2}\right)$	$\Omega_i = \gamma \ \mathrm{tg}\left(\dfrac{\omega_i T}{2}\right)$
ФВЧ	$\gamma\dfrac{1+Z^{-1}}{1-Z^{-1}}$	$\gamma = \mathrm{tg}\left(\dfrac{\omega_\varepsilon T}{2}\right)$	$\Omega_i = \gamma \ \mathrm{ctg}\left(\dfrac{\omega_i T}{2}\right)$
СФ	$\gamma\dfrac{1-2\alpha Z^{-1}+Z^{-2}}{1-Z^{-2}}$	$\gamma = \mathrm{ctg}\left(\dfrac{\omega_\varepsilon - \omega_{-\varepsilon}}{2}T\right)$ $\alpha = \dfrac{\cos[(\omega_\varepsilon + \omega_{-\varepsilon})T/2]}{\cos[(\omega_\varepsilon - \omega_{-\varepsilon})T/2]}$	$\Omega_i = \gamma\dfrac{\alpha - \cos\omega_i T}{\sin\omega_i T}$
РФ	$\gamma\dfrac{1-Z^{-2}}{1-2\alpha Z^{-1}+Z^{-2}}$	$\gamma = \mathrm{tg}\left(\dfrac{\omega_\varepsilon - \omega_{-\varepsilon}}{2}T\right)$ $\alpha = \dfrac{\cos[(\omega_\varepsilon + \omega_{-\varepsilon})T/2]}{\cos[(\omega_\varepsilon - \omega_{-\varepsilon})T/2]}$	$\Omega_i = \gamma\dfrac{\sin\omega_i T}{\alpha - \cos\omega_i T}$

Порядок решения задачи аппроксимации методом билинейного *Z*-преобразования следующий:

1. Анализируют задачу, вид ЧХ искомого ЦФ и ее параметры: предельные частоты, угасание на них, частоту дискретизации.

2. Определяют γ и предварительно искаженные «аналоговые» частоты по заданным «цифровым», т.е. требования к АЧХ ЦФ адресуют к АЧХ АФ-прототипу.

3. Выбирают тип АФ-прототипа (Баттерворта, Чебышева и др.) и определяют его порядок *n* по заданным предельным частотам и угасанию с помощью справочника, в котором изложена и методика действий.

4. По заданным требованиям к АЧХ для выбранного АФ-прототипа выписывают из справочника выражение его передаточной функции $H_N(p)$ *n*-го порядка и определяют значение ее коэффициентов.

5. Применяя к функции $H_N(P)$ билинейное преобразование для искомого вида РЦФ, соответственно табл. 4.5 определяют его передаточную функцию $H(Z)$.

В процессе ее преобразования в дробно-рациональный вид определяют коэффициенты фильтра. После этого переходят к построению структурной схемы РЦФ.

Пример. Передаточная функция нормализованного АФ Баттерворта 3-го порядка имеет такой вид: $H_N(P) = 1/\left(1+P\right)\left(1+P+P^2\right)$.

В соответствии с табл. 4.5, заменяя $P = \gamma\left(1-Z^{-1}\right)/\left(1+Z^{-1}\right)$ в выражении для $H_N(P)$ (где $\gamma = 3{,}08$), после преобразований получаем передаточную функцию заданного РЦФ

$$H(Z) = 0{,}018\frac{\left(1+Z^{-1}\right)\left(1+2\,Z^{-1}+Z^{-2}\right)}{\left(1-0{,}5098\,Z^{-1}\right)\left(1-1{,}2511\,Z^{-1}+0{,}5459\,Z^{-2}\right)},$$

которую можно реализовать в виде каскадной структуры при канонической форме реализации элементарных звеньев (рис. 4.90).

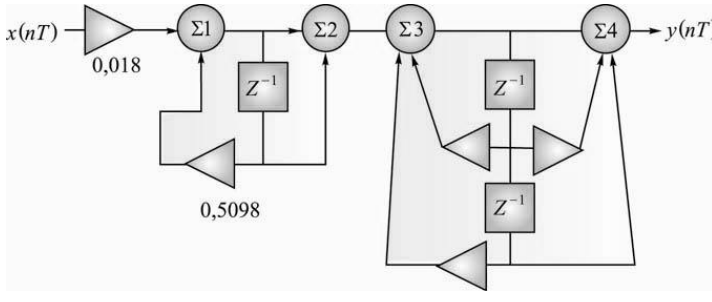


Рис. 4.90. Схема каскадной структуры ЦФ

Коэффициенты $H(Z)$ РЦФ обычно вычисляют вручную, на ЭВМ или с использованием таблиц справочников.

Аппроксимация РЦФ по цифровому прототипу. Для этого способа решения задачи аппроксимации используется преобразование ЦФ НЧ с предельной частотой полосы пропускания $\omega_з^0$ в ЦФ с другой полосой пропускания: ЦФ ВЧ, полосовой, режекторный или гребенчатый фильтр.

В табл. 4.6 приведены формулы преобразований, которые используются для решения задачи аппроксимации по цифровому прототипу.

Методика расчета по цифровому прототипу проще, чем методика расчета по аналоговому прототипу, поскольку в ней нет этапа перехода от АФ-прототипа к ЦФ. При использовании этой методики необходимо помнить, что частоты дискретизации ЦФ-прототипа и искомого ЦФ должны совпадать.

Пример. Рассчитать передаточную функцию ЦФ ВЧ с максимально гладкой АЧХ. Угасание в полосе 0...1,2 кГц не меньше, чем 20 дБ. Угасание на частоте 2,25 кГц и выше не больше 3 дБ. Частота дискретизации F_T = 10 кГц.

Р е ш е н и е. 1. Выбираем в качестве цифрового прототипа ЦФ НЧ 3-го порядка, рассчитанный методом билинейного Z-преобразования (см. предыдущий пример).

2.Для перехода от ЦФ НЧ с $\omega_з^0 T = 0,2\pi$ к ЦФ ВЧ с $\omega_з T = 0,45\pi$ используем замену вида 2 (см. табл. 4.6):

$$Z_{НЧ}^{-1} \to \frac{\alpha - Z^{-1}}{1 - \alpha Z^{-1}}, \text{ где } \alpha = \frac{\cos\left(\omega_з^0 + \omega_з\right)T/2}{\cos\left(\omega_з^0 - \omega_з\right)T/2} = \frac{\cos\left(0,325\pi\right)}{\cos\left(-0,125\pi\right)} \approx 0,5655.$$

Таблица 4.6

Тип преобразований	Выражение для замены	Примечания
1. ФНЧ → ФВЧ	$Z_{НЧ}^{-1} \to \frac{Z^{-1} - \alpha}{1 - \alpha Z^{-1}}$	$\alpha = \frac{\sin(\omega_з^0 - \omega_з)T/2}{\sin(\omega_з^0 + \omega_з)T/2}$
2. ФНЧ → ФВЧ	$Z_{НЧ}^{-1} \to \frac{\alpha - Z^{-1}}{1 - \alpha Z^{-1}}$	$\alpha = \frac{\cos(\omega_з^0 - \omega_з)T/2}{\cos(\omega_з^0 + \omega_з)T/2}$
3. ФНЧ → СФ с частотами $\omega_з$ и $\omega_{-з}$	$Z_{НЧ}^{-1} \to \frac{Z^{-2} - \frac{2\alpha\beta}{1+\beta}Z^{-1} + \frac{\beta-1}{\beta+1}}{\frac{\beta-1}{\beta+1}Z^{-2} - \frac{2\alpha\beta}{\beta+1}Z^{-1} + 1}$	$\alpha = \frac{\cos(\omega_з + \omega_{-з})T/2}{\cos(\omega_з - \omega_{-з})T/2}$ $\beta = \operatorname{ctg}\left(\frac{\omega_з - \omega_{-з}}{2}T\right) \cdot \operatorname{tg}\left(\frac{\omega_з}{2}T\right)$
4. ФНЧ → РФ с частотами $\omega_з$ и $\omega_{-з}$	$Z_{НЧ}^{-1} \to \frac{Z^{-2} - \frac{2\alpha}{1+\beta}Z^{-1} + \frac{1-\beta}{1+\beta}}{\frac{1-\beta}{1+\beta}Z^{-2} - \frac{2\alpha}{1+\beta}Z^{-1} + 1}$	$\alpha = \frac{\cos(\omega_з + \omega_{-з})T/2}{\cos(\omega_з - \omega_{-з})T/2}$ $\beta = \operatorname{ctg}\left(\frac{\omega_з - \omega_{-з}}{2}T\right) \cdot \operatorname{tg}\left(\frac{\omega_з}{2}T\right)$

Тогда передаточная функция ЦФ ВЧ после ряда преобразований приобретет вид

$$H(Z)=0,2077\frac{\left(1-Z^{-1}\right)\left(1-2Z^{-1}+Z^{-2}\right)}{\left(1-0,0783Z^{-1}\right)\left(1-0,1889Z^{-1}-0,3387Z^{-2}\right)}.$$

Такую функцию можно реализовать каскадно согласно схеме ЦФ.

Особенности выбора структурной схемы НЦФ. После определения коэффициентов передаточной функции ЦФ необходимо осуществить выбор структурной схемы ЦФ. Применять РЦФ в прямой или канонической форме нецелесообразно, если его порядок превышает 2, поскольку при этом оказывается значительным уровень шумов на выходе, обусловленных конечной разрядностью кодов, которые циркулируют в фильтре. Целесообразнее реализовать РЦФ в виде совокупности отдельных звеньев (биквадратных блоков), каждое из которых представляет собой фильтр с передаточной функцией

$$H_i(Z)=\frac{b_{0i}+b_{1i}Z^{-1}+b_{2i}Z^{-2}}{1+a_{1i}Z^{-1}+a_{2i}Z^{-2}}.$$

На практике используют две формы реализации РЦФ из отдельных звеньев типа биквадратного блока - каскадную (последовательную) и параллельную.

Каскадная реализация РЦФ - это последовательное соединение отдельных звеньев с передаточными функциями (рис. 4.91), причем передаточная функция фильтра

$$H(Z)=\prod_{i=1}^{L}H_i(Z)=\prod_{i=1}^{L}\frac{b_{0i}+b_{1i}Z^{-1}+b_{2i}Z^{-2}}{1+a_{1i}Z^{-1}+a_{2i}Z^{-2}}. \qquad (4.125)$$

В отдельных сомножителях выражения (4.125) некоторые коэффициенты могут быть равны нулю, например b_{2i}; a_{2i} и др.

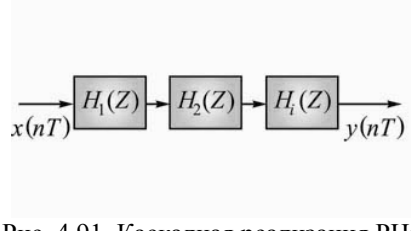


Рис. 4.91. Каскадная реализация РЦФ

Параллельная форма реализация РЦФ - это параллельное соединение звеньев (рис. 4.92), каждое из которых реализует одно из слагаемых выражения

$$H(Z)=\sum_{i=1}^{L}H_i(Z)=\sum_{i=1}^{L}\frac{b_{0i}+b_{1i}Z^{-1}}{1+a_{1i}Z^{-1}+a_{2i}Z^{-2}}.$$

В отдельных слагаемых некоторые коэффициенты могут быть равны нулю, например b_{1i}; a_{2i} и др.

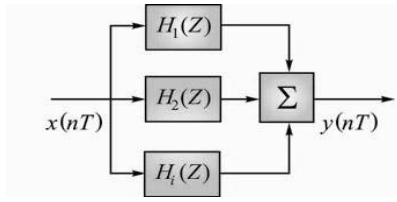


Рис. 4.92. Параллельная форма реализации РЦФ

Каскадную форму реализации РЦФ можно применять в том случае, когда нужно на выходе фильтра обеспечить низкий уровень выходных шумов квантования. Параллельная форма реализации РЦФ дает возможность получить ЦФ с высоким быстродействием, без введения в его схему любых дополнительных элементов.

НЦФ *могут быть реализованы как в прямой, так и в каскадной форме.* Прямая форма является непосредственной реализацией добытого выражения для передаточной функции $H(Z)$. Каскадная форма реализации НЦФ соответствует разложению его передаточной функции на множители

$$H(Z)=\sum_{k=0}^{N-1} b_k Z^{-k}=\prod_{i=1}^{L}\left(b_{0i}+b_{1i}Z^{-1}+b_{2i}Z^{-2}\right) \qquad (4.126)$$

и является последовательным соединением звеньев, каждое из которых реализует один из сомножителей формулы (4.126). Для прямой формы реализации каждого сомножителя в формулу (4.126) достаточно трех множительных устройств, двух элементов задержки и сумматора с тремя входами.

Выражение (4.126) можно представить в другой форме:

$$H(Z)=b_0\prod_{i=1}^{L}\left(1+b'_{1i}Z^{-1}+b'_{2i}Z^{-2}\right).$$

Для реализации $\mathrm{H}(Z)$ понадобится на $(\mathrm{a}-1)$ множительных устройств меньше.

Адаптивная цифровая фильтрация. ***Адаптация систем ЦФ*** *- изменение параметров или структуры системы ЦОС с целью достижения заданного эффекта в результате приспособления к неизвестным заранее внешним условиям.*

К таким условиям можно отнести случайные помехи, характеристики каналов распространения сигналов, принципиально непреодолимые шумы квантования. Главным свойством адаптивной системы можно считать изменяющееся во времени функционирование с саморегуляцией.

Адаптивные системы классифицируют по наличию *обратной связи.*

Адаптация без обратной связи *- это процесс измерения характеристик выходных влияний* (*сигналов и шумов*)*, введение этой информации в ал-*

горитм ЦОС и использование результатов для целенаправленного регулирования.

Адаптация с обратной связью *- это процесс автоматического оценивания влияния параметров регулирования на выходной сигнал.*

Одним из эффективных путей решения класса задач обработки сигналов в условиях априорной неопределенности может быть применение разнообразных методов адаптации. В этом случае задача решается так же, как при отсутствии неопределенности, а дальше в синтезированные алгоритмы обработки сигналов вместо неизвестных параметров подставляются их оценки (в статистическом смысле), полученные по входным выборкам. Естественно, что эффективность описанного алгоритма будет ниже, чем при наличии полной априорной информации, поскольку оценки неизвестных параметров исчисляются с определенной погрешностью.

Для оценки степени достижения необходимого качества адаптации вводят функционал качества (функция качества, рабочая функция, стоимостная функция), зависящая от входного сигнала и параметров системы ЦОС. Достижение экстремума этого функционала (локального или глобального) является целью функционирования адаптивной системы. Структурные схемы двух типов адаптивных систем приведены на рис. 4.93.

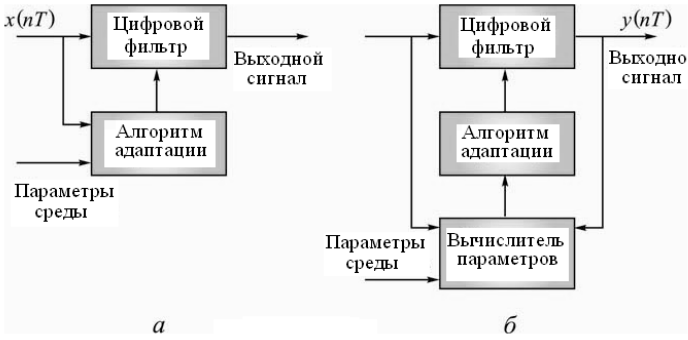


Рис. 4.93. Структурные схемы адаптивных систем:
а - без обратной связи; *б* - с обратной связью

В основном, особенно при обработке больших по объему выборок, оказывается, что определение оценок параметров системы и их функциональное преобразование связаны со значительными вычислительными трудностями. Поэтому используется другой метод адаптации, заключающийся в том, что оптимальное решение, соответствующее экстремуму функционала качества, достигается путем последовательных приближений. Сначала задача решается при полной определенности. Шаги последовательного приближения определяются по вычисленным значениям детерминированного функционала качества и его производных. Таким образом, строится детерминированный итерационный алгоритм. Поскольку в условиях априорной неопределенности функционал качества зависит от нескольких неизвестных параметров и не

может быть вычислен непосредственно, в синтезированных алгоритмах его значения (и значения его производных) заменяют их оценками. Алгоритм, образующийся при этом, называется итерационным стохастическим и соответствует теореме о разделении.

Оптимальное нерекурсивное оценивание. Принципы оптимального линейного оценивания являются фундаментальными при любом рассмотрении адаптивных систем обработки сигналов и, в частности, адаптивных фильтров.

Процесс адаптивной фильтрации включает два этапа проведения оценивания:

1) *оценивание выходного сигнала фильтра;*

2) *оценивание коэффициентов фильтра* (*отсчетов ИХ*), *необходимых для достижения поставленной цели.*

Второй этап необходим из-за априорной неопределенности входного сигнала, на которую влияет шумовая помеха.

Простейшей и наиболее распространенной адаптивной структурой является нерекурсивный фильтр с регулируемыми коэффициентами. Схема этого фильтра приведена на рис. 4.94. Иногда его называют *адаптивным линейным сумматором*.

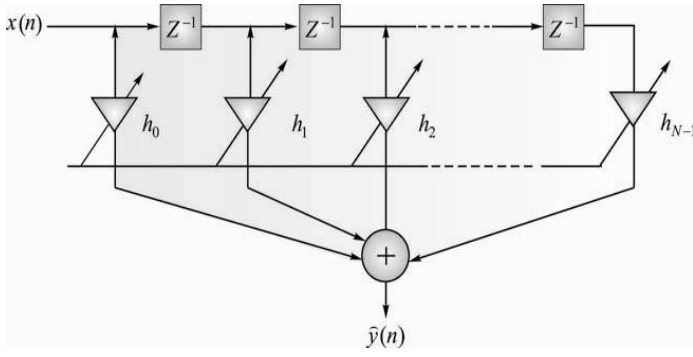


Рис. 4.94. Адаптивный нерекурсивный фильтр

На выходе фильтра необходимо получить оценку $\hat{y}(n)$, максимально соответствующую (в смысле выбранного критерия качества) незашумленному сигналу $y(n)$.

Синтез оцениваемого устройства на базе адаптивного КИХ-фильтра существенным образом зависит от определения стоимостной функции, соответственно которой качество оценивания характеризуется разностью между выходным сигналом оцениваемого устройства и значением, которое подлежит оцениванию:

$$e(n) = y(n) - \hat{y}(n),$$

где $e(n)$ - погрешность оценивания; $y(n)$ - оцениваемый случайный сигнал; $\hat{y}(n)$ - его статистическая оценка. В нашем случае оценка $\hat{y}(n)$ является линейной функцией последовательности входных отсчетов $x(n)$ и коэффициентов фильтра $h_n(n=0,1,...,N-1)$. Последовательность отсчетов $x(n)$ в общем виде можно представить как сигнал $y(n)$, искаженный аддитивным белым шумом $v(n)$ с дисперсией σ_v^2:

$$x(n)=y(n)+v(n). \tag{4.127}$$

Чаще всего при проведении оптимального оценивания $\tilde{o}(n)$ используется метод наименьших квадратов. При этом среднеквадратичная погрешность определяется как

$$E\left[e^2(n)\right]=E\left\{\left[y(n)-\hat{y}(n)\right]^2\right\}, \tag{4.128}$$

где $E[.]$ - оператор математического ожидания.

Среднеквадратичная погрешность минимизируется относительно весовых коэффициентов КИХ-фильтра для отыскания оптимальной оценки по критерию метода наименьших квадратов.

В нерекурсивном фильтре соответственно его разностному уравнению выходная оценка $\hat{y}(n)$ является конечным линейным полиномом

$$\hat{y}(n)=\sum_{k=0}^{N-1}h_k x(n-k) \tag{4.129}$$

Выражение можно переписать в векторно-матричной системе обозначений:

$$\hat{y}(n)=X^T(n)\cdot H=H^T X(n), \tag{4.130}$$

где $X(n)=\begin{bmatrix} x(n) \\ x(n-1) \\ \vdots \\ x(n-N+1) \end{bmatrix}$ и $H=\begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{N-1} \end{bmatrix}$ - вектор-столбец входного сигнала и коэффициентов фильтра.

Тогда функция среднеквадратичной погрешности приобретет вид

$$E\left[e^2(n)\right]=E\left[y(n)-H^T X(n)\right]^2. \tag{4.131}$$

Это выражение описывает стандартную поверхность гиперпараболлоида в $(N+1)$-мерном пространстве с единственным минимумом.

Если вектор коэффициентов H и вектор входного сигнала $X(n)$ не коррелированные, *функция среднеквадратичной погрешности* имеет вид

$$E\left[y(n)X^T(n)\right] = H_0 \cdot E\left[X(n)X^T(n)\right], \tag{4.132}$$

где H_0 - вектор оптимальных коэффициентов КИХ-фильтра, который обеспечивает минимум выражения (4.131).

Члены математических ожиданий в выражении (4.132) можно определить таким образом: $R = E\left[X(n)X^T(n)\right]$ - автокорреляционная квадратная матрица порядка N входных отсчетов сигнала; $P = E\left[y(n)X(n)\right]$ - вектор взаимной корреляции между оцениваемым сигналом и отсчетами входной последовательности размером $N \times 1$. С учетом обозначений формулу (4.132) можно переписать в виде

$$P^T = H_0^T R. \tag{4.133}$$

Уравнение (4.133) является известным матричным *уравнением Винера - Хопфа,* дающее оптимальное (по критерию минимума наименьших квадратов) решение для коэффициентов КИХ-фильтра

$$H_0 = R^{-1}P. \tag{4.134}$$

Выражение (4.134) получено с учетом симметричности корреляционной матрицы R, для которой $\left[R^{-1}\right]^T = R^{-1}$. Виннеровская оценка (4.134) есть одношаговым блочным процессом, который подходит для конечной выборки (блока) данных. В случае нестационарности входного сигнала восстановления матриц R и P должны происходить на каждом временном шаге.

Определим остаточную среднеквадратичную погрешность оценивания, используя оптимальный вектор коэффициентов H_0. Ее можно получить из соотношения

$$E\left[e(n)X(n)\right] = 0. \tag{4.135}$$

Преобразуем формулу для среднеквадратичной погрешности с учетом выражения (4.135) и вычисленного вектора H_0:

$$\begin{aligned} E\left[e^2(n)\right] &= E\left\{e(n)\left[e(n) - H_0^T X(n)\right]\right\} = \\ &= E\left[e(n)e(n)\right] - E\left\{\left[e(n) - H_0^T X(n)\right]y(n)\right\} = \\ &= E\left[y^2(n)\right] - H_0^T E\left[e(n)X(n)\right] = E\left[y^2(n)\right] - H_0^T P. \end{aligned} \tag{4.136}$$

Формула (4.136) дает возможность вычислять остаточную среднеквадратичную погрешность при известном полезном сигнале $y(n)$ и найденных векторах H_0 и P.

Пример. Найти оптимальный вектор коэффициентов адаптивного линейного фильтра 1-го порядка, имеющего в своей структуре два коэффициента h_0 и h_1.

Р е ш е н и е. Пусть входной сигнал является суммой дискретной синусоиды и стационарного белого шума с нулевым средним и дисперсией σ_v^2, т.е.

$$x(n) = \sin\left[\frac{\pi}{4}n\right] + v(n).$$

Таким образом, оцениваемый сигнал $y(n)$ имеет восемь отсчетов на один период синусоиды. Получим автокорреляционную матрицу

$$R = E\left\{\begin{bmatrix} x(n) \\ x(n-1) \end{bmatrix}\begin{bmatrix} x(n)x(n-1) \end{bmatrix}\right\} = E\begin{bmatrix} x^2(n) & x(n)x(n-1) \\ x(n)x(n-1) & x^2(n-1) \end{bmatrix}.$$

С учетом стационарности процессов и некоррелированности сигнала $y(n)$ и шума $v(n)$ найти значения элементов матрицы особенно просто, поскольку шумовая составляющая влияет лишь на ее диагональные элементы. Остальные элементы можно найти из детерминированного компонента сигнала $y(n) - \sin(\pi/4)n$ путем бесконечного усреднения в результате применения оператора математического ожидания. Далее несложно записать матрицу в явном виде

$$R = \begin{bmatrix} 1/2 + \sigma_v^2 & 1/2\sqrt{2} \\ 1/2\sqrt{2} & 1/2 + \sigma_v^2 \end{bmatrix}.$$

Свертывание матрицы соответственно процедуре дает

$$R^{-1} = \begin{bmatrix} 1/2 + \sigma_v^2 & -1/2\sqrt{2} \\ -1/2\sqrt{2} & 1/2 + \sigma_v^2 \end{bmatrix}.$$

Вектор взаимной корреляции P определяем из аналогичных соображений:

$$P = E\left\{y(n)\begin{bmatrix} x(n) \\ x(n-1) \end{bmatrix}\right\} = E\begin{bmatrix} \sin^2(n\pi/4) + \sin(n\pi/4)v(n) \\ \sin(\pi/4)\sin[(n-1)\pi/4] + \sin(\pi/4)v(n-1) \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2\sqrt{2} \end{bmatrix}.$$

Применение формулы дает вектор оптимальных коэффициентов КИХ-фильтра 1-го порядка:

$$H_0 = R^{-1}P = \frac{1}{\sigma_v^4 + \sigma_v^2 + 1/8}\begin{bmatrix} \sigma_v^2/2 + 1/8 \\ \sigma_v^2/2\sqrt{2} \end{bmatrix}.$$

Адаптация линейного фильтра заключается в том, что его коэффициенты (а значит, и частотные характеристики) зависят от мощности (дисперсии) аддитивного белого шума. Анализ выражения показывает, что при отсутствии шума $(\sigma_v^2 = 0)$ фильтр передает входную синусоиду прямо на выход $(h_0 = 1,\ h_1 = 0)$. Если мощность шума равна мощности сигнала $(\sigma_v^2 = 0{,}5)$, то вектор оптимальных коэффициентов имеет элементы

$$H_0^1 = \begin{bmatrix} 3/7 \\ \sqrt{2}/7 \end{bmatrix}.$$

Ненормированная АЧХ такого фильтра для частоты дискретизации 1000 Гц приведена на рис. 4.95.

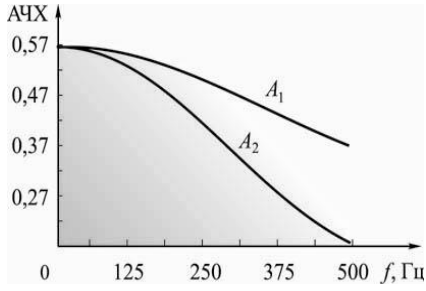


Рис. 4.95. A_1 - АЧХ аналогового фильтра; A_2 - АЧХ фильтра Калмана

Увеличив порядок КИХ-фильтра, можно найти лучшие оценки. В частности, снижается значение остаточной среднеквадратичной погрешности, однако для этого нужен намного больший объем вычислений.

Рекуррентные алгоритмы адаптации. Фильтры Калмана. Рассмотренная ранее винеровская оценка коэффициентов КИХ-фильтра нуждается в полном перечислении всех членов авто- и взаимнокорреляционных матриц для каждой новой выборки, что с точки зрения вычислений нерационально. Если иметь дело с продолжительным (теоретически бесконечным) рядом отсчетов входного сигнала, удобными являются рекуррентные алгоритмы получения оценок, которые вносят коррекцию на каждом шаге итерационного процесса.

Если входной сигнал $x(n)$ является случайным и марковским, то его можно представить в виде выхода линейно-дифференциальной системы 1-го порядка, создаваемой белым шумом $\omega(n)$ с нулевым средним и дисперсией σ_w^2. Модель генерирования сигнала описывается разностным уравнением 1-го порядка

$$x(n) = ax(n-1) + \omega(n-1). \qquad (4.137)$$

Структурная схема устройства, соответствующая уравнению (4.137), приведена на рис. 4.96.

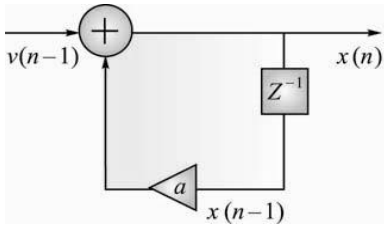


Рис. 4.96. Устройство генерации случайного сигнала

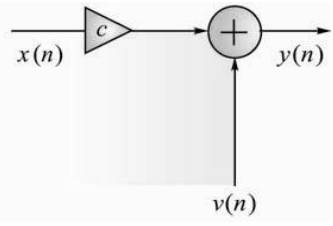


Рис. 4.97. Модель прохождения сигнала по каналу связи

Предполагается, что после прохождения канала связи сигнал $x(n)$ подвергся амплитудным изменениям, описываемым постоянным коэффициентом c, и на него оказывал влияние аддитивный белый шум $v(n)$ с нулевым средним и дисперсией σ_v^2. Модель влияния канала на сигнал описывается простым уравнением

$$y(n) = cx(n) + v(n). \tag{4.138}$$

Соответствующая модели структурная схема приведена на рис. 4.97.

Искаженный сигнал $y(n)$ поступает на вход синтезированного адаптивного калмановского фильтра. На его выходе необходимо получить рекуррентную оценку $\hat{x}(n)$, максимально приближенную к сигналу $x(n)$ по критерию метода наименьших квадратов. Рекурсивная формула оценки 1- го порядка имеет вид

$$\hat{x}(n) = b(n)\,\hat{x}\,(n-1) + k(n)\,y(n). \tag{4.139}$$

Заметим, что в общем случае коэффициенты $b(n)$ и $k(n)$ зависят от нормированного времени. Обобщенная структурная схема адаптивного оценивания, реализуемого алгоритмом (4.139), приведена на рис. 4.98.

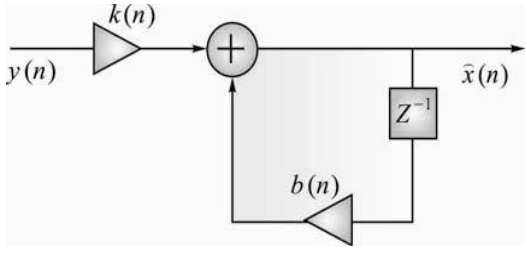


Рис. 4.98. Обобщенная структура рекурсивного оценивания 1-го порядка

Введем обозначения:

$$e(n) = \hat{x}(n) = x(n), \tag{4.140}$$

$$p(n) = E\left[\hat{x}(n) - x(n)\right]^2. \tag{4.141}$$

Выражение (4.140) называется *погрешностью оценки*, а выражение (4.141) - *среднеквадратичной погрешностью*. Подставляя выражение (4.139) в (4.141), получаем

$$p(n) = E\left[b(n)\,\hat{x}\,(n-1) + k(n)\,y(n) - x(n)\right]^2. \tag{4.142}$$

Чтобы получить оптимальную согласно методу наименьших квадратов оценку, выражение (4.142) дифференцируют по $b(n)$ и $k(n)$ с дальнейшим приравниванием результатов нулю:

$$\frac{\partial p(n)}{\partial b(n)} = 2E\left\{\left[b(n)\,\hat{x}\,(n-1) + k(n)\,y\,(n) - x(n)\right]\hat{x}\,(n-1)\right\} = 0, \tag{4.143}$$

$$\frac{\partial p(n)}{\partial k(n)} = 2E\left\{\left[b(n)\,\hat{x}\,(n-1) + k(n)\,y\,(n) - x(n)\right]y\,(n)\right\} = 0. \tag{4.144}$$

Преобразуем уравнение (4.143):

$$E\left\{\left[b(n)\,\hat{x}\,(n-1)\right]\hat{x}\,(n-1)\right\} = E\left\{-\left[k(n)\,y\,(n) - x(n)\right]\hat{x}\,(n-1)\right\}. \tag{4.145}$$

После несложных арифметических преобразований из формулы (4.145) имеем

$$\begin{aligned} b(n)E\left\{\left[\hat{x}(n-1) - x(n-1) + \hat{x}(n-1)\right]\hat{x}\,(n-1)\right\} = \\ = E\left\{\left[x(n) - k(n)\,y\,(n)\right]\hat{x}\,(n-1)\right\}. \end{aligned} \tag{4.146}$$

Подставив в формулу (4.146) значение $y(n)$ из формулы (4.138) с учетом выражения (4.140), получим

$$\begin{aligned} b(n)\,E\left[e(n-1)\,\hat{x}\,(n-1) + x\,(n-1)\,\hat{x}\,(n-1)\right] = \\ = E\left\{\left[x(n)\left[1 - ck(n)\right] - k(n)\,v\,(n)\right]\hat{x}\,(n-1)\right\}. \end{aligned} \tag{4.147}$$

Принцип ортогональности, минимизирующий погрешность, нуждается в некоррелированности погрешности $e(n)$ и оценки $\hat{x}(n-1)$, а также независимости шума $v(n)$ и $\hat{x}(n-1)$, что выполняется в рамках сделанных предположений. Это означает выполнение равенств

$$E\left[e(n)\,\hat{x}\,(n-1)\right] = 0, \tag{4.148}$$

$$E\left[v(n)\,\hat{x}\,(n-1)\right] = 0. \tag{4.149}$$

Тогда уравнение (4.148) с учетом уравнений (4.149) и (4.150) приобретает вид

$$b(n)\,E\left[x(n-1)\,\hat{x}\,(n-1)\right]=\left[1-ck(n)\right]E\left[x(n)\,\hat{x}\,(n-1)\right]. \quad (4.150)$$

Подставляя модель генерирования сигнала (4.51) в формулу (4.64), получаем

$$\begin{aligned} b(n)\,E\left[x(n-1)\,\hat{x}\,(n-1)\right]= \\ =\left[1-ck(n)\right]E\left[ax(n-1)\,\hat{x}\,(n-1)+\omega(n-1)\,\hat{x}\,(n-1)\right]. \end{aligned} \quad (4.151)$$

Последовательная подстановка формулы (4.137) в выражение (4.138), а далее в формулу (4.139) дает

$$\begin{aligned} \hat{x}(n-1)=b(n-1)\,\hat{x}\,(n-2)+ \\ +ack(n-1)\,x\,(n-2)+ck(n-1)\,\omega(n-2)+k(n-1)v(n-1). \end{aligned} \quad (4.152)$$

Помножим обе части равенства (4.152) на $\omega(n-1)$ и возьмем математическое ожидание

$$E\left[\hat{x}(n-1)\,\omega\,(n-1)\right]=0, \quad (4.153)$$

учитывая, что шум $\omega(n-1)$ не коррелирован со всеми членами в правой части формулы (4.152).

Воспользовавшись соотношением (4.153), преобразуем формулу (4.151):

$$b(n)E\left[x(n-1)\,\hat{x}\,(n-1)\right]=a\left[1-ck(n)\right]E\left[x(n-1)\,\hat{x}\,(n-1)\right],$$

что приводит к соотношению между коэффициентами $b(n)$ и $k(n)$:

$$b(n)=a\left[1-ck(n)\right]. \quad (4.154)$$

После несложных преобразований получаем

$$\hat{x}(n)=a\hat{x}(n-1)+k(n)\left[y(n)-ac\hat{x}(n-1)\right]. \quad (4.155)$$

Равенство (4.155) является решением для построения адаптивного рекурсивного оценщика 1-го порядка, названного *скалярным фильтром Калмана*. Его структурная схема изображена на рис. 4.99.

Адаптация в этом устройстве оценивания происходит следующим образом. Предыдущая оценка $\hat{x}(n-1)$ после умножения на коэффициенты a и c определяет очередной отсчет зашумленого сигнала $\hat{y}(n)$. Он сравнивается с текущим отсчетом $y(n)$.Разность между ними с коэффициентом «доверия» $k(n)$ подытоживается с оценкой $a\hat{x}(n-1)$, в результате чего получают текущую оценку $\hat{x}(n)$. Нетрудно предположить, что изменяемый во времени

коэффициент «доверия» $k(n)$ должен зависеть от шумовых параметров модели и текущего значения $p(n)$ среднеквадратичной погрешности.

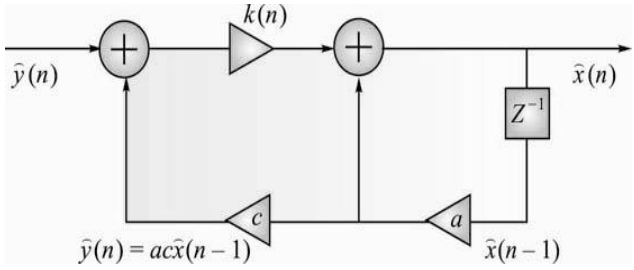


Рис. 4.99. Скалярный фильтр Калмана

Явное выражение для $k(n)$:

$$k(n)-\frac{c\left[a^2 p(n-1)+\sigma_w^2\right]}{\sigma_v^2+c^2\sigma_w^2+c^2a^2(n-1)}, \tag{4.156}$$

где

$$p(n)=\frac{1}{c}\sigma_v^2 k(n). \tag{4.157}$$

Анализ выражений (4.157) и (4.158) указывает на гибкое (адаптивное) изменение коэффициента $k(n)$ в зависимости от дисперсий действующих шумов σ_v^2 и σ_w^2 и значения текущей среднеквадратичной погрешности $p(n)$.

Для перехода к векторному фильтру Калмана необходимо перейти к авторегрессионной модели генерирования сигнала более высокого порядка с дальнейшей редукцией к многомерному пространству состояний.

Рассмотрим пример, который иллюстрирует адаптацию коэффициента $k(n)$ к помехоизменяемой обстановке.

Пример. В рамках описанной модели для определенности возьмем $a=0,5;\ c=1;\ \sigma_w^2=1$. Пусть дисперсия аддитивного шума $\sigma_v^2(n)$ в четные моменты времени равняется 1, а в нечетные - 2. Необходимо проследить изменения коэффициента $k(n)$, определяющего адаптивные свойства фильтра Калмана. После сделанных предположений формулы (4.157) и (4.158) приобретут упрощенный вид

$$k(n)=\frac{1+0,25p(n-1)}{1+0,25p(n)+\sigma_v^2(n)};\ \ p(n)=\sigma_v^2(n)k(n).$$

При $p(-1)=0$ получаем расчетную таблицу, анализ которой показывает наличие в фильтре Калмана вплоть до момента времени $n=2$ переходного

процесса адаптации, связанного с произвольным начальным заданием среднеквадратичной погрешности $p(-1)=0$. Начиная с $n=3$ наблюдается постоянный режим, причем степень «доверия» к искаженным отсчетам (нечетные моменты времени) значительно ниже, чем к «чистым».

n	1	0	1	2	3	4	5
$\sigma_v^2(n)$	—	1	2	1	2	1	2
$p(n)$	0	0,5	0,36	0,52	0,72	0,54	0,72
$k(n)$	—	0,5	0,72	0,52	0,36	0,54	0,36

В сущности, *калмановское оценивание* реализует рекурсивную процедуру адаптации, которая базируется на авторегрессионной модели процесса генерирования сигнала.

Основные выводы

Реакция линейной электрической цепи, не содержащей независимых источников энергии, на гармоническое влияние определяется комплексной функцией этой цепи.

Компоненты комплексной функции выражают определенные частотные характеристики цепей, т.е. характеризуют их частотные свойства.

Электрические цепи 1-го или 2-го порядка представляют собой в простейшем случае реализованные на практике ФНЧ или ФВЧ.

Комплексная частотная функция цепи равна спектральной плотности ее импульсной временной характеристики, тогда как импульсная характеристика является обратным преобразованием Фурье (оригиналом) ее комплексной функции.

Модуляция - процесс изменения параметров несущего колебания по закону информационного сигнала. Демодуляцией называют процесс детектирования, или выявление информационного сообщения на фоне параметров ВЧ несущего колебания.

Различают два вида модуляции: амплитудную и угловую. Угловая модуляция, в свою очередь, подразделяется на фазовую и частотную. Разновидностью АМ является балансовая модуляция и однополосная модуляция.

Одна из основных характеристик модулированных сигналов - это глубина модуляции, пропорциональная интенсивности переданного информационного сигнала.

Спектр верхней боковой полосы АМ колебания подобен спектру информационного сигнала (модулируемого). При АМ происходит лишь трансформация спектра сигнала по оси частот на величину $\omega_н$.

Наибольшую помехоустойчивость имеют системы с ФМ, ЧМ.

Как носитель можно использовать не только периодические колебания, а и узкополосный случайный процесс.

ЧМ, в отличие от ФМ, характеризуется большим постоянством спектров сигналов. В этом заключается одна из причин более частого применения ЧМ на практике.

При импульсно-кодовой модуляции (ИКМ) передача отдельных значений сигнала сводится к передаче определенных групп импульсов. Эти группы передаются одна за одной через промежутки времени, большие по сравнению с продолжительностью отдельных импульсов. Каждая кодовая группа импульсов представляет собой регулярный непериодический сигнал, спектр которого можно вычислить с помощью преобразований Фурье обычным способом.

При наличии модуляции любого вида спектр расширяется незначительно за счет боковых частот крайних составляющих спектра немодулированных импульсов. Поэтому рабочая полоса частот, занимаемая импульсными сигналами, практически не зависит от вида модуляции и определяется продолжительностью и формой импульса.

Фазо- и частотно-модулированные колебания, сравнительно с АМ, занимают более широкую полосу частот; они имеют преимущества: высокую помехоустойчивость при передаче информационного сообщения и возможность передачи более мощного сигнала при одинаковой мощности радиопередатчика.

Процесс преобразования непрерывных сигналов в цифровые состоит из двух этапов: дискретизации по времени и дискретизации по уровню (квантование).

В результате дискретизации непрерывный сигнал определяется дискретными отсчетными значениями, взятыми через определенные интервалы времени. Задача выбора интервала дискретизации решается на основании теоремы Котельникова (теоремы отсчетов)

Повышение частоты дискретизации - это наиболее простой и очевидный способ уменьшения погрешности дискретизации.

Основные направления развития методов обработки информации: аналоговая обработка; дискретно-аналоговая обработка; цифровая обработка.

Цифровой фильтр (ЦФ) - это частотно-избирательная цепь, обеспечивающая селекцию цифровых сигналов по частоте. К ЦФ принадлежат: фильтры нижних частот, фильтры верхних частот, полосовые фильтры, режекторные фильтры.

Как и все цифровые системы, цифровые фильтры подразделяются на два больших класса: нерекурсивные и рекурсивные. В каждом из этих классов выделяют линейные и нелинейные фильтры.

Процесс проектирования ЦФ включает такие этапы: синтез; выбор или разработка алгоритмов вычислений; проверка моделированием; практическая реализация и настройка.

Основными показателями работы ЦФ является скорость обработки данных и качество амплитудно-частотной характеристики (ее приближение к идеальной).

Алгоритм разностного уравнения включает три вида операций и предусматривает наличие в схеме ЦФ трех видов операционных устройств: запоминающих устройств (регистров); множителей; сумматоров.

Возможные виды соединения схем ЦФ между собой: последовательное (каскадное) соединение элементов ЦФ; параллельное соединение элементов ЦФ; соединение с обратной связью элементов ЦФ; каноническая форма реализации ЦФ.

Для решения задачи аппроксимации ЦФ применяют две группы методов: прямые и прототипные.

Адаптация систем ЦФ - изменение параметров или структуры системы ЦОС с целью достижения заданного эффекта в результате приспособления к неизвестным заранее внешним условиям.

Вопросы для самоконтроля

1. Дайте определение понятия и проклассифицируйте комплексные функции электрических цепей.

2. Приведите определение электрических фильтров и их основных параметров согласно классификации.

3. Укажите основные операторы перехода от известных характеристик фильтров нижних частот с целью синтеза фильтров верхних частот, полосовых и режекторных.

4. Приведите примеры прохождения непериодических сигналов на базе спектрального метода анализа для фильтров 1-го порядка.

5. Приведите определение взаимосвязи комплексной частотной характеристики фильтров с их импульсной временной характеристикой. Раскройте понятие амплитудной и угловой модуляции.

6. Укажите причины возникновения режима перемодуляции в АМ колебании.

7. Запишите аналитическое выражение для АЧС многотональной АМ.

8. Назовите недостатки амплитудной модуляции.

9. Укажите отличия когерентного, некогерентного и квазикогерентного методов демодуляции.

10. Какой вид модуляции является основным в радиовещании, телевидении, радиорелейной связи, радиотелеграфии, радиолокации, в измерительной и ядерной технике?

11. Как изменяется амплитудно-частотный спектр с угловой модуляцией при различных значениях коэффициента модуляции?

12. Сравните различные методы модуляции по помехоустойчивости.

13. В каких случаях целесообразнее использовать логарифмическое квантование?

14. Раскройте понятие "шум квантования".

15. Приведите аналитическое описание дискретного сигнала.

16. Как выбирается интервал дискретизации?

17. Какой вид имеет спектр дискретизированного сигнала?

18. В чем заключается способ восстановления непрерывного колебания из дискретного?

19. Укажите наилучший способ уменьшения ошибки дискретизации?

20. Укажите этапы цифровой обработки сигналов?

21. Укажите цель использования сглаживающего фильтра и этап цифровой обработки сигнала, на котором этот фильтр используется.

22. Назовите основные преимущества использования цифровой фильтрации.

23. Что понимают под проектированием ЦФ и какие этапы при этом выделяют?

24. Опишите передаточные функции ЦФ.

25. Назовите основные устройства, включаемые в структурную схему ЦФ?

26. Какие временные и частотные характеристики ЦФ вы знаете?

27. Запишите критерий физической реализуемости цифрового фильтра.

28. Какой метод аппроксимации рекурсивных ЦФ целесообразнее использовать?

29. Укажите порядок решения задачи аппроксимации методом билинейного Z-преобразования?

30. Опишите метод инвариантности импульсной характеристики.

The main conclusions

Response of a linear electrical circle that does not contain independent sources of energy on harmonic influence is determined by complex function of this circle.

Components of complex function express the certain frequency characteristics of circles, in other words they characterize its frequency qualities.

Electrical circles of the first or second order are realized in practice low-pass filters or high-pass filters in the simplest case.

Complex frequency function of a circle is equal to spectral density of its impulse time characteristic whereas the impulse characteristic is return transformation of Furie (original) of its complex function.

Modulation is the process of changing the parameters carrying oscillation under the law of an informational signal. Demodulation is process of detecting or revealing information message on a background of parameters of upper frequencies of carrying oscillation.

Two types of modulations can be distinguished: amplitude and angle modulation. Angle modulation is in turn divided into phase and frequency. A kind of AM is balanced modulation and single-sideband modulation.

One of the main characteristics of modulated signals is depth of modulation that is proportional to intensity of the transmitted informational signal.

The spectrum of upper side bar of AM oscillation is similar to a spectrum of an informational signal (modulating). Only transformation of a spectrum of a signal occurs at AM, along the axis of frequencies on the value ω_i .

The systems in PM, FM have the greatest noise immunity.

It is possible to use not only periodic oscillations as the carrier but also narrow-band stochastic process.

FM, unlike PM, is characterized by greater persistence of signals spectrum. It is one of the reasons of best application of FM in practice.

During pulse code modulation (PCM) the transmission of separate values of a signal is brought to transmission of the certain groups of impulses. These groups are transferred one by one through relatively greater intervals of time in comparing to duration of separate impulses. Each code group of impulses is regular acyclic signal, the spectrum of which can be calculated on the basis of transformations of Furie in ordinary way.

Involving the modulation of any type the spectrum extends slightly due to side frequencies of extreme components of a spectrum of not modulated impulses. Therefore the working band occupied by impulse signals, practically does not depend on type of modulation and it is determined by duration and the form of an impulse.

Phase- and frequency-modulated oscillations comparing with AM occupy wider band, but have advantages: high noise immunity during transmission of an information message and possibility of transmission of more powerful signal at equal power of a radio transmitter.

Process of transformation of continuous signals in digital consists of two stages: sampling and digitization on a level (quantization).

As a result of digitization the continuous signal is determined by discrete measuring values taken in certain time slices. The task about a choice of an interval of digitization is solved on the basis of theorem of Kotelnikov (sampling theorem).

An increase of sampling rate is the simplest and the most obvious way of decrease of an error of digitization.

The main directions of development of methods of information processing are the analogue processing, the digital-analog processing and the digital processing.

Digital filter is frequency-selective circle that provides the selection of digital signals on frequency.

Low-pass filters, high-pass filters, band-pass filters, notch filters belong to DF.

As well as all digital systems, digital filters are divided into two extensive classes: nonrecursive and recursive. In turn, linear and nonlinear filters are singled out in each of these classes.

Process of designing of DF includes following stages: synthesis; a choice or development of algorithms of calculations; simulation test; practical realization and debugging.

The main factors of operation of DF are the speed of data processing and the quality of gain-frequency characteristic (its approximation to ideal).

The algorithm of a difference equation includes three kinds of operations and demands to have three sorts of operational devices in DF scheme: storage devices (registers), multipliers devices, adders.

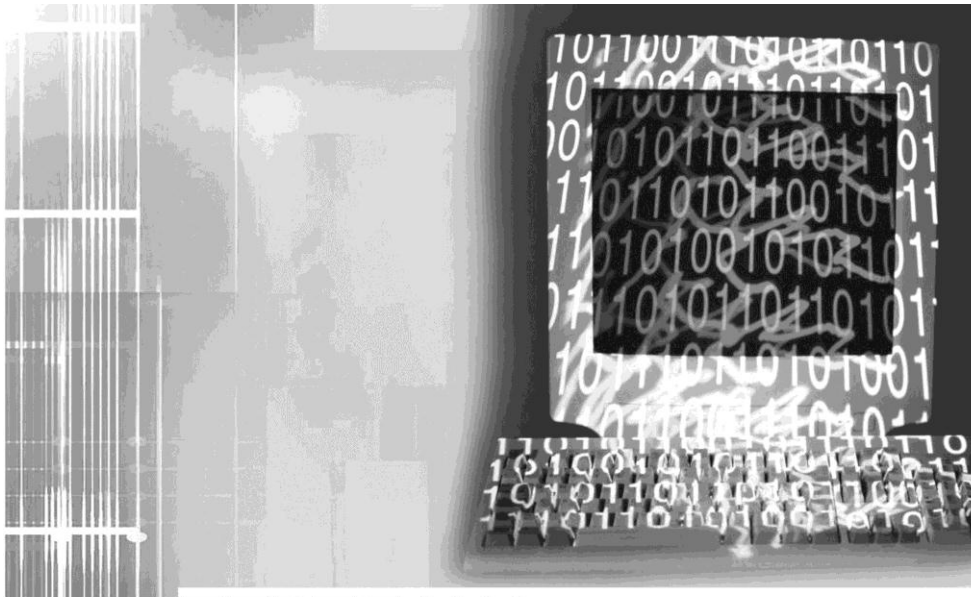
Such kinds of connection of DF schemes between themselves are possible: consistent (cascading) connection of DF units; equal connection of DF units; feedback connection of DF units; canonical form of realization of DF.

Two groups of methods such as direct and prototypes methods are used for solution of the task of approximation of DF.

Adaptation of DF systems is a change of parameters or structure of digital signal processing system with the purpose of achieving the set effect as a result of the adaptation to the unknown beforehand external conditions.

Ключевые слова

Русский	**Английский**
электрический фильтр	electric filter
амплитудная модуляция	amplitude modulation
фазовая модуляция	phase modulation
частотная модуляция	frequency modulation
импульсная модуляция	impulse modulation
детектирование	detection
дискретизация	dyskretyzation
аналого-цифровой преобразователь	analogy-digital transformer
цифровой фильтр	digital filter
адаптивная фильтрация	adaptive filtration



ПЕРЕДАЧА ИНФОРМАЦИИ

5

5.1. Информационные системы передачи данных
5.2. Виды информационных каналов, их математические модели и характеристики
5.3. Скорость передачи информации в каналах связи
5.4. Синтез элементов информационных систем. Оптимальный приёмник
5.5. Многоканальные сети передачи данных. Разделение информационных каналов
5.6. Помехоустойчивость систем передачи информации

5.1. Информационные системы передачи данных

Современный этап развития общества характеризуется возрастающей ролью информационной сферы - совокупности информации, информационной инфраструктуры, субъектов, осуществляющих сбор, формирование, распространение, обработку и использование информации, а также системы регулирования возникающего при этом общественного отношения. Информационная сфера является системообразующим фактором жизни общества, активно влияет на состояние политической, экономической, оборонительной и других составляющих безопасности государства.

Трудно представить современное общество вне широко разветвленных информационно-коммуникационных систем, без которых не функционируют важнейшие области производства, медицины, экономики, сельского хозяйства и т.п. Важно обеспечить верность, целостность, доступность информации на всех этапах обработки, хранения и передачи данных по информационным каналам связи.

Сигналы, передаваемые по каналам связи, под влиянием помех искажаются. Вследствие этого принятое сообщение лишь в некоторой степени соответствует переданному. Степень соответствия принятого сообщения переданному называется *верностью* информации. Верность информации является одним из основных показателей качества систем передачи данных. Верность информации тесно связана с понятием *помехоустойчивости*.

Канал связи *является совокупностью технических средств между источником сообщения и потребителем*. Технические устройства, входящие в состав канала связи, предназначены для того, чтобы сообщение дошли к потребителю наилучшим образом, и с этой целью сигналы обрабатывают и преобразуют согласно определенным алгоритмам. Такими преобразованиями сигнала является, например, модуляция и преобразование непрерывных сигналов в дискретные. Соответственно типу сигналов каналы разделяют на непрерывные и дискретные.

Преобразование дискретного сообщения в информационный сигнал канала связи осуществляется в виде операций - *форматирования, кодирования и модуляции.* Кодирование является преобразованием сообщений в последовательность кодовых символов, а модуляция - преобразованием этих символов в сигналы, пригодные для передачи по информационному каналу.

При кодировании происходит процесс преобразования элементов сообщений в соответствующие им числа (кодовые символы). Каждому элементу сообщения присваивается определенная совокупность кодовых символов, называемая *кодовой комбинацией*. Совокупность кодовых комбинаций образовывает код.

Показатели качества дискретных систем связи приведены на рис. 5.1.

В результате преобразований непрерывного сигнала, называемых *дискретизацией* и *квантованием*, получают отсчеты, рассматриваемые как числа

в той или иной системе исчисления, которые являются дискретными сигналами. Эти числа преобразовывают в кодовые комбинации электрических сигналов, которые и передают по линиям связи как непрерывные. При использовании в качестве носителя информации постоянный ток, получают последовательность видеоимпульсов. При необходимости эту последовательность модулируют гармоническим колебанием и получают последовательность радиоимпульсов.

Рис. 5.1. Основные показатели качества дискретных систем связи

В общем виде модель системы передачи данных изображена на рис. 5.2. Хотя эта модель и содержит основные элементы, присущие любой системе передачи информации, она может служить лишь простой иллюстрацией к описанию *информационной системы передачи данных*, поскольку практически не отображает тех действий, которые должны (или могут) выполняться над информацией в процессе ее передачи от источника к потребителю.

Рис. 5.2. Обобщенная схема передачи данных

Более полной в этом смысле является модель системы передачи (и хранения) информации, приведенная на рис. 5.3. Следует указать, что на самом деле проблемы, возникающие при передаче (причем не только с использованием радиоволн) и хранении информации (на оптических дисках, магнитных носителях и в памяти компьютеров), похожи, поэтому методы их решения и

структура технических устройств также во многом идентичны.

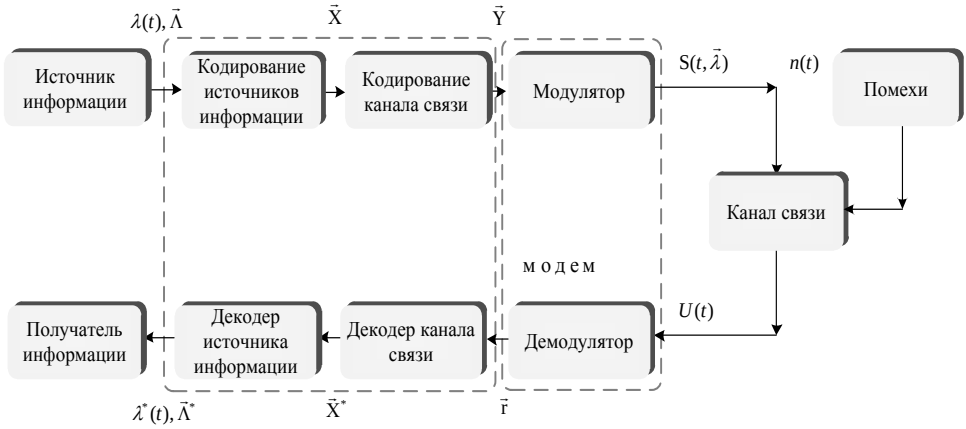


Рис. 5.3. Структурная схема системы передачи информации

Кратко охарактеризуем назначение и функции элементов такой модели.

1. ***Источник информации или сообщение*** - *это физический объект, система или явление, формирующие сообщение с целью его передачи от источника информации к ее потребителю.*

Сообщение - *это значение или изменение некоторой физической величины, отображающей состояние объекта* (*системы или явления*). Как правило, первичные сообщения - язык, музыка, изображение, измерение параметров окружающей среды и др. - являются функциями времени неэлектрической природы.

С целью передачи по каналу связи эти сообщения преобразуются в электрический сигнал, изменение которого во времени $\lambda(t)$ отображает передаваемое сообщение. Значительная часть таких сообщений по своей природе не являются сигналами - это массивы чисел, текст или другие файлы. Сообщение можно представить в виде некоторых векторов $\vec{\Lambda}$.

2. ***Кодирование источника информации***. Большей частью информационные сообщения содержат избыточную информативность и имеют медленно изменяемую частоту, которая в общем случае неудовлетворительно приспособлена для эффективной передачи по каналам связи. Поэтому сообщения ($\lambda(t)$ или $\vec{\Lambda}$), как правило, подвергаются кодированию. Процедура кодирования включает предварительную дискретизацию непрерывного сообщения $\lambda(t)$, т.е. его преобразование в последовательность элементарных дискретных сообщений $\{\lambda_i\}$.

Кодирование источника информации может иметь различные цели: сокращение объема переданных данных (сжатие данных), увеличение количества переданной за единицу времени информации, повышение достоверности

передачи, обеспечение секретности при передаче и др.

Таким образом, на выходе *кодера источника* по переданным сообщениям $\lambda(t)$ или $\vec{\Lambda}$ формируется последовательность кодовых символов $\vec{X}$ - так называемая *информационная последовательность*, допускающая точное или приближенное восстановления начального информационного сообщения.

3. ***Кодирование в каналах связи***. При передаче информации по каналу связи с помехами в принятых данных могут возникать ошибки. Если такие ошибки не велики или возникают достаточно редко, информация может быть использована потребителем. При значительном количестве ошибок полученной информацией пользоваться невозможно.

Кодирование в канале, или *помехоустойчивое кодирование*, - это способ обработки переданных данных, обеспечивающий *уменьшение количества ошибок*, возникающих в процессе передачи по каналу с помехами. Существует много различных методов помехоустойчивого кодирования информации, но все они базируются на следующем принципе: при помехоустойчивом кодировании в передаваемые сообщения вносится специальным образом организованная избыточность (в передаваемые кодовые последовательности вводятся избыточные символы), что дает возможность на приемной стороне обнаруживать и исправлять возникающие ошибки. Таким образом, если при кодировании источника проводится устранение естественной, существующей в сообщении избыточности, то при кодировании в канале избыточность в передаваемое сообщение вносится сознательно. В результате на выходе кодера канала формируется последовательность кодовых символов $Y(X)$, называемая *кодовой последовательностью канала связи.*

Заметим, что как помехоустойчивое кодирование, так и сжатие данных не являются обязательными операциями при передаче информации. Эти процедуры (и соответствующие блоки в структурной схеме системы передачи информации) могут отсутствовать. Тем не менее, их отсутствие может привести к серьезным потерям в помехоустойчивости системы, уменьшению скорости передачи и снижению качества передачи информации. Поэтому все современные системы включают у себя и эффективное, и помехоустойчивое кодирование данных.

4. ***Модулятор***. *Функции модулятора в информационных системах передачи данных — это согласование сообщения источника или кодовых последовательностей, вырабатываемых кодером, со свойствами канала связи и обеспечение возможности одновременной передачи большого количества сообщений по общему информационному каналу связи.*

В самом деле, большинство подлежащих передаче непрерывных $\lambda(t)$ и дискретных $\vec{\Lambda}$ сообщений, а также результаты их кодирования - последовательности кодовых символов $\vec{X}$ и $\vec{Y}$ - представляют собой сравнительно низкочастотные сигналы с широкой полосой ($\Delta F \le 1$МГц, $\Delta F \sim f_0$). Вместе с тем эффективная передача с использованием электромагнитных колебаний

(радиоволн) возможна лишь для достаточно высокочастотных сигналов $\left(f_0 \geq 1...1000 \text{ МГц}\right)$ со сравнительно узкополосными спектрами $\left(\Delta F << f_0\right)$. Поэтому модулятор должен преобразовать сообщения *источника* ($\lambda(t)$ *или* $\vec{\Lambda}$) *или соответствующие им кодовые последовательности* ($\vec{X}$ *и* $\vec{Y}$) в сигналы ($S\left(t,\lambda(t)\right)$, $S\left(t,Y\left(\lambda(t)\right)\right)$), т.е. наложить сообщения на сигналы, свойства которых обеспечивали бы возможность эффективной их передачи по радиоканалу (или по другим существующим каналам связи - телефонным, оптическим и т.д.).

Существует значительное количество методов модуляции сигналов, обеспечивающих передачу информации с различной эффективностью и качеством. Простейшими из них является амплитудная, частотная и фазовая модуляция непрерывных сигналов.

5**.** ***Канал связи*** *- это система передачи информации, использующая в качестве носителя от источника к потребителю электромагнитные волны или радиоволны, а в качестве среды распространения - окружающее пространство или радиоканал.*

Рассмотрим информационный канал связи в виде звена радиотехнической системы передачи информации, на вход которого поступает *сигнал передатчика* $S\left(t,\lambda(t)\right)$, *а на выходе образовывается сигнал,* называемый, по обыкновению, *принятым колебанием.*

Существует достаточное количество моделей информационных каналов большей или меньшей сложности, тем не менее, в общем случае сигнал $S\left(t,Y\left(\lambda(t)\right)\right)$, проходя по каналу связи, ослабевает, приобретает некоторую временную задержку (или фазовый сдвиг) и искривляется. Принимаемое колебание $U(t)$ в этом случае будет иметь вид

$$U(t)=\varepsilon\, S\left(t-\tau, Y\left(\lambda(t)\right)+n(t)\right), \qquad (5.1)$$

где ε - угасание; τ - задержка времени; $n(t)$ - помехи в канале связи.

6. ***Получатель информации***. *Функции получателя системы передачи информации состоят в том, чтобы с максимально возможной точностью по принятому колебанию* $U(t)$ *воссоздать на своем выходе переданное сообщение (* $\lambda(t)$ или $\vec{\Lambda}$ *).*

Принятое (воспроизведенное) сообщение из-за помех, в общем случае, отличается от переданного. Принятое сообщение будем называть *оценкой* (имеется в виду оценка сообщения) и обозначать тем же символом, что и отправленное сообщение, но со знаком *: $\lambda^*(t)$ или $\vec{\Lambda}^*$.

7. ***Демодулятор***. Для воспроизведения оценки сообщения $\lambda^*(t)$ или $\vec{\Lambda}^*$ приемник системы должен по принятому колебанию $U(t)$, учитывая сведения

об использованных при передаче сигналах и способах модуляции, получить оценку кодовой последовательности $Y\left(\lambda^*(t)\right),$; эта оценка называется принятой *последовательностью* $\vec{r}$. Такая процедура называется *демодуляцией*, *детектированием* или *приемом сигнала*. При этом демодуляция должна выполняться так, чтобы принятая последовательность $\vec{r}$ минимально отличалась от переданной кодовой последовательности $\vec{Y}$.

Задача демодуляции принятого колебания $U(t)$ совпадает с различными вариантами задачи оптимального приема сигнала на фоне помех (оптимальное выявление, оптимальное различение двух или нескольких сигналов и т.д.).

8. ***Декодер канала связи***. Принятые последовательности $\vec{r}$ в общем случае могут отличаться от переданных кодовых слов $\vec{Y}$, т.е. содержать ошибки. Количество таких ошибок зависит от уровня помех в канале связи, скорости передачи, выбранного для передачи сигнала и способа модуляции, а также от способа приема (демодуляции) колебания $U(t)$.

Функция декодера канала связи заключается в выявлении и, по возможности, исправлении ошибок, возникших под влиянием помех, и воссоздании с максимальным приближением переданного от источника информации к потребителю сигнала. Процедура воспроизведения сообщения, выявления и исправления ошибок в принятой последовательности $\vec{r}$ *называется декодированием канала.*

Результатом декодирования $\vec{r}$ является оценка информационной последовательности $\vec{X}^*$. Выбор помехоустойчивого кода, способа кодирования, а также метода декодирования может осуществляться так, чтобы на выходе декодера канала осталось по возможности меньше неисправленных ошибок.

Вопросом помехоустойчивого кодирования/декодирования в системах передачи (и хранения) информации отводится большое внимание, поскольку этот прием дает возможность существенным образом повысить качество передачи. Во многих случаях, при высоких требованиях к достоверности принимаемой информации (в компьютерных сетях передачи данных, в дистанционных системах управления и т.д.), передача без помехоустойчивого кодирования вообще невозможна.

9. ***Декодер источника информации***. Поскольку информация источника $(\lambda(t),\ \Lambda)$ в процессе передачи подвергалась кодированию с целью ее более компактного (или удобного) представления (сжатие данных, экономное кодирование, кодирование источника), необходимо восстановить ее к результирующему (начальному) виду по принятой последовательности $\vec{X}^*$.

Процедура восстановления $\vec{\Lambda}^*$ по $\vec{X}^*$ называется *декодированием источника* и может быть обратной к операции кодирования (неразрушительное кодирование/декодирование) или восстанавливать приближенное значение

$\vec{\Lambda}^*$, в большей или меньшей степени отличающееся от $\vec{\Lambda}$ (разрушительное кодирование/декодирование). К операции восстановления $\vec{\Lambda}^*$ по $\vec{X}^*$ будем относить также восстановление, если в этом есть необходимость, непрерывной функции $\lambda^*(t)$ в соответствии с набором дискретных значений оценок $\vec{\Lambda}^*$.

Таким образом, коротко описав общую структуру информационной системы передачи данных, перейдем к более подробному изучению ее основных элементов.

5.2. Виды информационных каналов, их математические модели и характеристики

Классификация каналов по частотным распределениям. В методах и средствах передачи данных находят применение механические, акустические, оптические, электрические и радиоканалы, которые различаются по техническим характеристикам и физической природе сигналов. Основным признаком каждого из перечисленных видов каналов обычно является диапазон рабочих частот. Классификация каналов в соответствии с природой сигналов приведена на рис. 5.4, а в табл. 5.1 представлены соответствующие данные по частотным диапазонам.

Рис. 5.4. Классификация информационных каналов

Механический канал передачи данных. Механические каналы применяются для передачи на короткие (до 500 м) расстояния сигналов в виде механических усилий или давлений. Применяются такие разновидности механических каналов.

Жесткие, или собственно механические, каналы. Простейшим примером является трос управления дроссельной заслонкой карбюратора. Длина таких каналов может достигать нескольких десятков метров.

Гидравлические каналы, в которых передающей средой является жидкость. Их недостатком являются неудовлетворительные динамические свойства, из-за чего длина этих каналов ограничивается несколькими метрами.

Пневматические каналы. По пневматическим каналам передается сигнал в виде давления. Средой для передачи обычно является воздух. Длина пневматических каналов достигает нескольких сотен метров.

Таблица 5.1

Вид канала	**Классификационный признак**	**Разновидности каналов**	**Частота сигналов f**	**Длина волны λ**
Механический	Среда передачи	Жесткие Гидравлические Пневматические		
Акустический	Диапазон частот	Звуковые Ультразвуковые	< 20 кГц > 20 кГц	> 15 км < 15 км
Оптический	>>	Видимого спектра Инфракрасное излучение	400...…1000 ГГц 0,3…400 ГГц 1000…3000 ГГц	0,3... 0,75 мкм 0,75…1000 мкм 0,1... 0,3 мкм
Электрический	>>	Ультрафиолетовое излучение Подтональных частот Тональных частот	200 ГГц 300…3400 Гц 4000…8500 Гц >10 кГц	> 1500 км 90…1000 км 32...75…75 км < 30 км
Радио	>>	Сверхтональных частот Высокочастотные Длинноволнового диапазона Средневолнового диапазона Промежуточный диапазон Коротковолнового диапазона Ультракоротковолнового диапазона	< 300 кГц 300…1500 кГц 1,5... 6 МГц 6... 30 МГц 30... $30 \cdot 10^5$ МГц	> 1000 м 200…1000 м 50…200 м 10... 50 м 0,0001…10 м

Среди механических каналов наибольшее распространение приобрели пневматические каналы в связи с широким применением унифицированных пневматических систем контроля и регулирования на предприятиях с пожароопасной средой. Данный канал состоит из следующих компонентов: пнев-

матический датчик, или преобразователь, вырабатывающий аналоговый пневматической сигнал (в виде давления сжатого воздуха в унифицированной шкале), пропорциональный измеренному технологическому параметру; пневматическая сеть связи; регистрирующий, регулировочный или преобразовательный прибор на выходе сети. Основным препятствием, ограничивающим применение пневматических систем, являются продолжительные переходные процессы в пневматических сетях связи, особенно в линиях большой длины.

Акустический канал передачи данных. Акустические каналы предназначаются для передачи колебаний. Средой для передачи могут служить любые звукопроводящие материалы и среды.

Акустические сигналы и каналы нашли разнообразное применение в технике автоматического контроля, выявления и связи: акустический контроль состояния работающих механических объектов, ультразвуковая дефектоскопия, акустическое выявление объектов (подводных лодок, самолетов), гидролокация, акустическая связь и т.д. Справочные данные относительно скорости распространения звуковых волн в различных средах приведены в табл. 5.2.

Таблица 5.2

Среда	Скорость распространения звуковых волн, м/с	Среда	Скорость распространения звуковых волн, м/с
Воздух	331,45	Дерево	3350
Вода пресная	1430	Стекло	5400
Вода морская	1500	Сталь	6100

При пассивной передаче сигналов в процессе контроля или выявления источниками звука являются контролируемые или выявленные объекты. При активной передаче (ультразвуковая дефектоскопия, локация, связь) акустические сигналы создаются специальными генераторами.

Оптический канал передачи данных. По диапазону используемых частот (или длиной волны) оптические каналы подразделяются (см. табл. 5.1) на такие группы:

каналы видимой части спектра оптических сигналов (с длиной волны $0,3 < \lambda < 75$ мкм);

каналы инфракрасной части спектра ($0,75 < \lambda < 1000$ мкм);

каналы ультрафиолетовой части спектра ($\lambda < 0,3$ мкм).

Устройства, работающие с инфракрасным излучением, в отличие от устройств с видимым и ультрафиолетовым излучением, применяются чаще благодаря ряду преимуществ: меньшее ослабление инфракрасного излучения атмосферой сравнительно с излучением видимой и ультрафиолетовой частей спектра; распространение инфракрасного излучения в темноте, тайность пе-

редачи.

Перспективным в технике передачи информации является применение квантовых генераторов света - лазеров. Разработаны многочисленные конструкции кристаллических и газовых лазеров, работающих в разных частях оптического диапазона. Как приемники используются фоторезисторы, фотодиоды и фотомножитель.

Электрический канал. Электрические каналы - это каналы с применением коммутативных (проводниковых) сетей связи. Для передачи информации используются как специально выделенные сети, так и сети, созданные для иных целей. Например, широко применяются сети энергоснабжения, высоковольтные сети электропередачи, телекоммуникационные сети.

Шкала частот, которую занимают сигналами в электрическом канале связи, условно делится на четыре диапазона:

подтональные частоты 300…3400 Гц;

тональные частоты 4000…8500 Гц;

сверхтональные частоты – меньше, чем 300 кГц;

высокие частоты 300…1500 кГц.

Радиоканалы передачи данных. По диапазону частот радиосигналов различают такие каналы:

длинноволнового диапазона (300…30кГц);

средневолнового диапазона (З…0…0,3МГц);

коротковолнового диапазона (З…30…30 МГц);

ультракоротковолнового диапазона (< 20 МГц).

На распространение радиоволн влияют отражательные и поглотительные свойства земной поверхности и атмосферы, особенно слоя, называющегося ионосферой и расположенного над стратосферой. Ионосфера состоит из заряженных частичек газов — электронов и ионов, которые образовываются в результате влияния солнечных лучей, космического излучения и метеоритных частичек.

Изменение степени концентрации ионов с высотой предопределяет непрерывное изменение угла преломления радиоволн, в результате чего волны распространяются криволинейно. Если направление распространения становится горизонтальным, не достигнув уровня максимальной ионизации, происходит отражение радиоволн в сторону Земли. Преломляющая способность ионосферы не одинакова для различных типов волн. Она уменьшается с уменьшением длины волны. Чем больше длина волны, тем меньше степень ионизации нужна для ее отражения. Волны, распространяемые вследствие отражения от ионосферы, *называются пространственными*. Кроме пространственных существуют так называемые *поверхностные волны*, распространяющиеся вдоль поверхности Земли благодаря дифракции. Чем меньшая длина волны, тем быстрее угасает поверхностная волна (вследствие потерь в земной поверхности) и тем медленнее угасает пространственная волна. Ультракоротковолновые сигналы не отражаются и не выходят за пределы земной

атмосферы, вследствие чего они эффективно используются для космической связи. Сигналы других диапазонов отражаются ионосферой.

Для передачи сигналов и данных широко применяются радиорелейные каналы, осуществляющие связь на ультракоротких волнах. Приемо-передающие станции располагаются в пределах прямой видимости (рис. 5.5).

Для передачи информации на большие расстояния используются промежуточные ретрансляционные станции, служащие одновременно и для восстановления сигнала, искаженного и ослабленного в процессе передачи. Если прямая видимость ограничена только кривизной поверхности Земли, то расстояние между станциями (в километрах) рассчитывается по формуле $l = 7,2\sqrt{h}$, где h - высота антенны системы. Таким образом, при $h = 60$ м имеем $l = 54$ км.

Рис. 5.5. Радиорелейный канал связи

Реальная структура атмосферы (рис. 5.6) более сложна, и приведенное деление на тропосферу, стратосферу и ионосферу довольно условно.

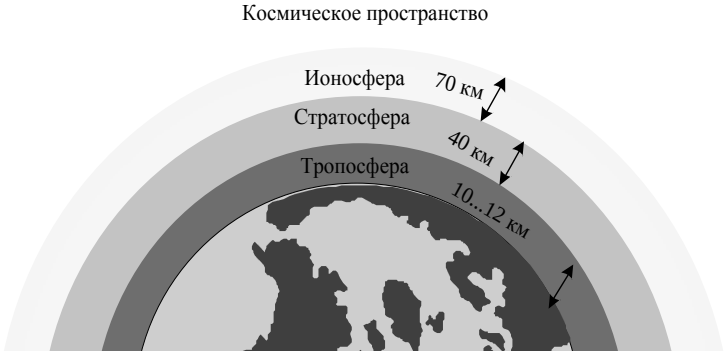


Рис. 5.6. Структура атмосферы Земли

Высота слоев атмосферы приведена приближенно, причем она различна для разных географических точек Земли. Около 80 % массы атмосферы со-

средоточенно в тропосфере и около 20 % — в стратосфере.

Плотность атмосферы в ионосфере крайне мала, граница между ионосферой и космическим пространством является условным понятием, поскольку следы атмосферы встречаются даже на высотах свыше 400 км. Считается, что плотные слои атмосферы заканчиваются на высоте около 120 км.

Типичная схема радиолинии изображена на рис. 5.7. Линия может состоять из двух конечных станций.

Рис. 5.7. Типичная схема радиолинии

Классификация и способы распространения радиоволны приведены в табл. 5.3 и 5.4. Деление радиоволн на диапазоны установлены Международным регламентом радиосвязи МСЕ-Р. Радиоволны, излучаемые передающей антенной, прежде чем попасть на приемную антенну, проходят в общем случае сложный путь.

Таблица 5.3

Вид радиоволн	Тип радио-волн	Диапазон радиоволн (длина волны)	Номер диапа-зона	Диапазон частот	Вид радиочастот
1	2	3	4	5	6
Мириаметровые	Сверх-длинные	10...100 км	4	3...30 кГц	Очень низкие (ОНЧ)
Километровые	Длинные	1...10 км	5	30...300 кГц	Низкие (НЧ)
Гектометровые	Средние	100...1000 м	6	300...3000 кГц	Средние (СЧ)
Декаметровые	Короткие	10...100 м	7	3...30 МГц	Высокие (ВЧ)
Метровые		1...10 м	8	30...300 МГц	Очень высокие (ОВЧ)
Дециметровые	Ультра-короткие	10…100 дм	9	300...3000МГц	Ультравысокие (УВЧ)
Сантиметровые		1...10 дм	10	3...30 ГГц	Сверхвысокие

Вид радиоволн	Тип радио-волн	Диапазон радиоволн (длина волны)	Номер диапа-зона	Диапазон частот	Вид радиочастот
1	2	3	4	5	6
					(СВЧ)
Миллиметровые		1…10 мм	11	30...300 Ггц	Крайне высокие (КВЧ)
Децимилли-метровые		0.1...1 мм	12	300... 3000 Ггц	Гипервысокие (ГВЧ)

На значение напряженности поля в точке приема влияют такие факторы: отражение электромагнитных волн от поверхности Земли; преломление в ионизированных слоях атмосферы (ионосфере); рассеяние на диэлектрических неоднородностях нижних слоев атмосферы (тропосфере); дифракция на сферической поверхности Земли.

Напряженность поля в точке приема зависит также от длины волны, освещенности земной атмосферы солнцем и ряда других факторов.

Таблица 5.4

Вид радиоволн	Основные способы распространения радиоволн	Дальность связи
Мириаметровые и километровые (сверхдлинные и длинные)	Дифракция Отражения от Земли и ионосферы	До тысячи км Тысячи км
Гектометровые (средние)	Дифракция Преломления в ионосфере	Сотни км Тысячи км
Декаметровые (короткие)	Преломление в ионосфере и отражение от Земли	Тысячи км
Метровые и более короткие	Свободное распространение и отражение от Земли Рассеяние в тропосфере	Десятки км Сотни км

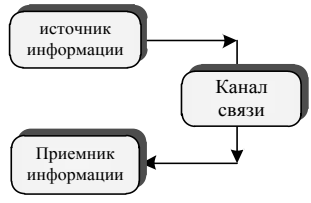
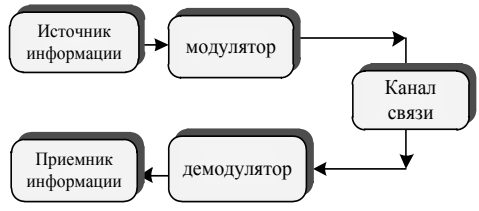
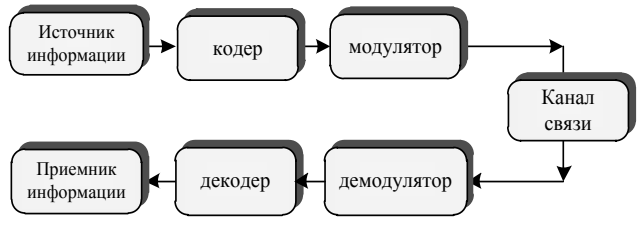
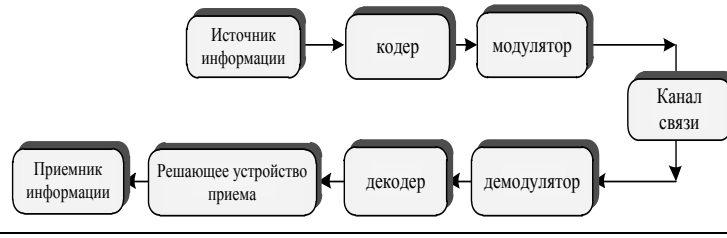
Классификация каналов по структуре. Канал передачи информации состоит из сети связи, модулятора и демодулятора (кроме случая передачи данных с использованием простой модуляции, при которой сигнал в сети совпадает с сигналом кодирующего и декодирующего датчика), а также устройств принятия решения, дающих возможность с высокой степенью достоверности принять и передать сообщение. Для увеличения надежности передачи применяются также каналы обратной связи. Варианты структур каналов приведены в табл. 5.5. Решающие устройства служат для классификации сомнительных сигналов, отождествляя их с достаточно высокой степенью вероятности с состоянием источника информации или с определенным кодом.

Количество информации, содержащейся в отдельном сигнале, имеющем

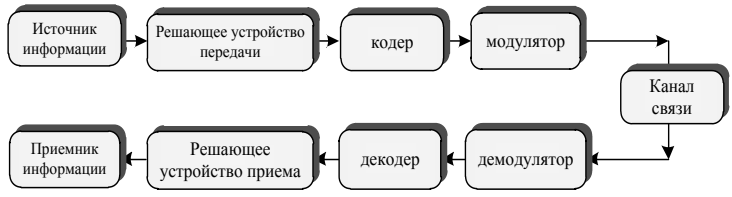
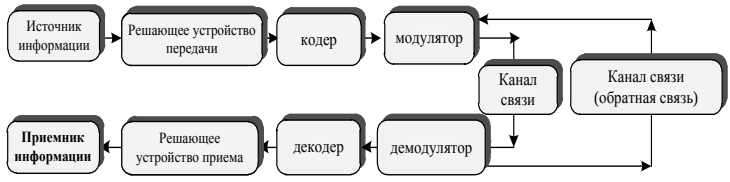
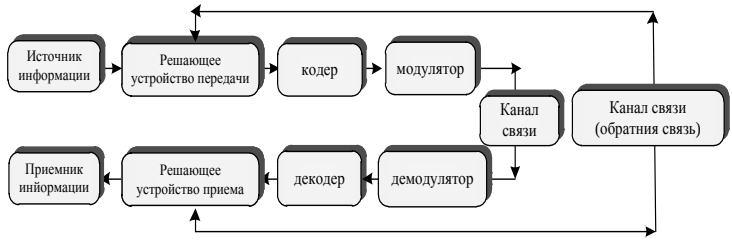
вероятность p, составляет $\log_2(1/p)$ бит и может быть значительно большей. Однако в среднем, в довольно длинном сообщении двоичный сигнал переносит не более, чем один бит информации. В общем случае в сообщении, состоящем из символов алфавита емкостью h, среднее количество информации не превышает одной единицы информации этого алфавита на символ. Тем не менее, не все сигналы несут полезную информацию.

Импульс помехи, случайно возникающей в сети, информации не несет и мешает передаче. Иногда дополнительные (избыточные) сигналы вводятся специально для повышения помехоустойчивости. В этих случаях среднее количество информации, переносимой одиночным сигналом, уменьшается.

Таблица 5.5

Вид структур	Схема
Элементарные	
С модуляцией	
С модуляцией и кодированием	
С решающим устройством на приеме	

Продолжение таблицы 5.5

Вид структур	Схема
С решающим устройством на приеме и передаче	
С информационной обратной связью	
С решающей обратной связью	

Случай, когда сообщение состоит из символов, несущих максимальное количество информации, встречается редко и возникает при отсутствии помех в сети, при равномерном распределении вероятностей между символами и при отсутствии статистической зависимости между последовательно переданными символами.

В случае невыполнения одного из этих условий количество информации оказывается меньшим. При заданном характере и уровне помех задача повышения эффективности передачи решается, прежде всего, путем определенного преобразования, приводящего к увеличению среднего количества информации.

Непосредственным носителем информации является случайно изменяющийся сигнал. На практике при рассмотрении информационных процессов удобно оперировать обобщенными показателями, характерными для сигналов данного вида и наиболее важными с точки зрения передачи заложенной в

них информации.

Каждый сигнал имеет определенную длительность. Длительность сигнала характеризует время передачи сообщений, продолжительность занятости информационного канала, т.е. *время передачи сигнала* T_c.

Каждый сигнал характеризуется определенным частотным спектром. Теоретически ширина спектра сигнала конечной продолжительности неограниченна. Тем не менее, изучение спектров реальных сигналов показывает, что их спектральная плотность спадает с ростом частоты. Это дает возможность при определенных условиях рассматривать сигналы как процессы с ограниченным спектром (см. гл. 3 и 4). Существуют различные критерии ограничения спектра сигнала. Одним из таких критериев являются допустимые искажения сигнала. Например, при передаче речевого сигнала разборчивость и качество речи практически полностью сохраняются при ширине спектра от 300 до 3400 Гц. Таким образом, второй обобщенной характеристикой сигнала должна быть *ширина частотного спектра* F_c.

Третьей важной характеристикой сигнала является его энергетическая характеристика - *средняя мощность* P_c.

Однако поскольку при передаче на сигналы всегда влияют помехи, то в качестве энергетической характеристики сигнала целесообразно брать отношения средней мощности сигнала $P_с$ к средней мощности помехи $P_п$.

Динамическим диапазоном информационного канала *называют логарифмическую меру, выраженную в отношении средней мощности сигнала* $P_с$ *к средней мощности помехи* $P_п$:

$$D_к = \log_2 P_с / P_п.$$

При оценивании информационной содержательности удобно выражать динамический диапазон посредством логарифма с основанием 2.

Объемом сигнала называется *произведение времени передачи сигнала* $T_с$, *ширины частотного спектра* $F_с$ *и средней мощности* $P_с$ *этого сигнала*:

$$V_с = T_с F_с P_с. \qquad (5.2)$$

В геометрическом изображении объем сигнала имеет вид параллелепипеда с ребрами $T_с, F_с$ и $D_с$ (рис. 5.8, *а*).

Информационный канал можно охарактеризовать также тремя соответствующими параметрами: временем использования канала $T_к$, шириной полосы частот пропускания канала $F_{к}$ и динамическим диапазоном канала $D_{к}$, характеризующим способность передавать различные уровни сигнала (см. рис. 5.8, *б)*.

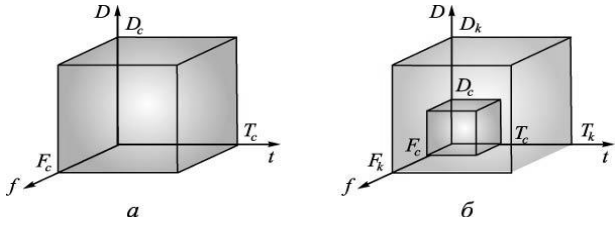


Рис. 5.8 . Геометрическая модель объема сигнала (*а*)
и емкости канала святи (*б*)

Емкостью канала *называется произведение времени занятости канала* $T_к$ *, ширины частотного спектра канала* $F_к$ *и динамического диапазона канала* $D_к$:

$$V_к = T_к F_к D_к. \tag{5.3}$$

Неискаженная передача сигналов возможна только при условии, что объем сигнала меньше или такой же, как и емкость канала связи («вмещается» в емкость канала - рис. 5.9).

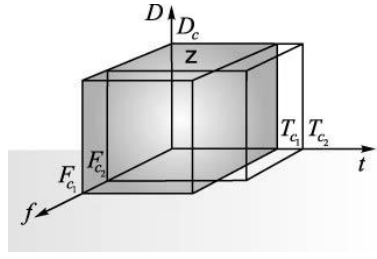


Рис. 5.9. Геометрическое представление трансформации сигнала
по параметрам F_c и T_c

Общее условие согласования *сигнала с каналом передачи информации определяется соотношением*

$$V_с \le V_к. \tag{5.4}$$

Однако соотношение (5.4) выражает необходимое, но не достаточное условие согласования сигнала с каналом.

Достаточное условие согласования *по всем параметрам*:

$$T_с \le T_к,\ F_с \le F_к,\ D_с \le D_к. \tag{5.5}$$

Если при выполнении условия (5.4) не обеспечивается часть условий

(5.5), то согласование можно достичь трансформацией сигнала с сохранением его объема.

Например, при *отсутствии согласования речевого сигнала с каналом по частоте* выполняются такие соотношения:

$$T_с < T_к,\ F_с > F_к,\ D_0 < D_к$$

Тогда согласование по частоте при передаче можно достичь, записывая сигнал на стационарный накопитель информации с одной скоростью и воссоздавая его с меньшей скоростью.

Предположим, что при выполнении условий $V_с \le V_к$ и $D_с \le D_к$ частотный спектр сигнала в n раз шире полосы пропускания канала

$$F_с = nF_к.$$

Для согласования сигнала с каналом можно записать сигнал на стационарный накопитель информации со скоростью u_1, а передавать со скоростью u_2, в n раз меньшей скорости u_1. При этом продолжительность сигнала увеличивается в n раз и во столько же раз уменьшается ширина его спектра. Объем сигнала при этом остается неизменным.

Скоростью ***передачи информации*** *называется среднее количество информации, которое передается по каналу связи за единицу времени.*

В общем случае эта скорость зависит от продолжительности T соответствующих сигналов. При достаточно длинных сообщениях скорость передачи остается постоянной. Учитывая это, скорость передачи информации аналитически представляется как

$$C = \bar{I}(Z,Y) = \lim_{T\to\infty} \frac{I(Z,Y)}{T}, \qquad (5.6)$$

где $I(Z, Y)$ - количество информации, переданное сигналом длительностью T.

Пропускной способностью канала *называется максимальная теоретически достижимая для этого канала скорость передачи информации.*

Пропускная способность информационного канала $C = \max\{\bar{I}(Z,Y)\}$.

Скорость передачи информации в общем случае зависит от статистических свойств сообщения, метода кодирования и свойств канала. Пропускная способность - это характеристика канала. Она не зависит от фактической скорости передачи информации.

С целью наиболее эффективного использования информационного канала необходимо принимать меры, чтобы скорость передачи информации была по возможности более близкой к пропускной способности канала. Тем не менее,

скорость введения информации в канал не должна превышать пропускную способность канала, иначе не вся информация будет передана по каналу.

Скорость введения информации *- аналитическое отношение среднего количества информации, заложенной в информационном потоке на входе канала, к продолжительности сообщения*

$$\bar{I}(X) = \lim_{T \to \infty} \frac{I(X)}{T}, \tag{5.7}$$

где $I(X)$ - среднее количество информации, заложенной в сообщении на входе канала; T - продолжительность сообщения. Таким образом, должно выполняться *основное условие динамического согласования информационного потока источника сообщений и информационного канала*

$$\bar{I}(X) \le C. \tag{5.8}$$

Одним из основных вопросов в теории передачи информации является определение зависимости скорости передачи информации и пропускной способности канала от его параметров и характеристик сигналов и помех. Эти вопросы впервые глубоко исследовал Клод Шеннон.

Рассмотрим три вида каналов: дискретный канал без помех, дискретный канал с помехами и непрерывный канал с помехами.

Помехи в непрерывном канале. ***Помехой*** *называется любой нежелательный процесс или действие, влияющее на сигнал и усложняющее его достоверный прием.*

Для рассмотрения помех сигнал на выходе канала подают в виде

$$y(t) = x(t) n_{\text{м}}(\text{t}) + n_{\text{а}}(t),$$

где $x(t)$ - чистый, или полезный, информационный сигнал на входе передатчика; $y(t)$ - сигнал на входе приемника; $n_{\text{м}}(t)$ - *мультипликативная* помеха, обусловленная случайными изменениями параметров канала; $n_{\text{а}}(t)$ - *аддитивная* помеха, которая имеет размерность сигнала и прибавляется к сигналу. Обычно считают, что аддитивная помеха возникает в сети связи.

Аддитивные помехи предопределяются многочисленными факторами, такими как флуктуационные шумы, вызванные тепловыми процессами в резисторах, лампах и других элементах схем, индустриальные помехи, в частности от наведений линий электропитания, контактной сети, радиостанций, других сетей связи.

Аддитивные помехи делятся на *сосредоточенные и флуктуационные*. Сосредоточенные помехи бывают узкополосные (сосредоточенные в узкой полосе частот) и импульсные (сосредоточенные во времени). Узкополосные помехи характерны для радиосвязи (помехи от соседних станций); борьба с

ними ведется методами повышения селективности систем.

Импульсные помехи - случайные последовательности импульсов, создаваемые промышленными установками (например, в цепи контактный провод-пантограф). Помеха считается импульсной, если ее длительность намного меньше длительности сигнала $\tau_{\text{п}} << T_{\text{с}}$. Борьба с импульсной помехой - применение систем с широкополосными усилителями, ограничителями, узкополосными усилителями, с использованием низкочастотных и сверхвысокочастотных диапазонов, где спектральная плотность мощности помех спадает.

Флуктуационная помеха распределена в широком спектре частот. Мощность теплового шума на 1 Ом нагрузки в полосе частот Δf определяется по формуле $P_{\text{п}} = \sigma^2 = \Delta f\, 4KT$, где $K = 1{,}37 \cdot 10^{-23}$ Дж/К - постоянная Больцмана; T - абсолютная температура.

Общая характеристика дискретного канала. Как уже отмечалось, дискретный канал имеет в своем составе непрерывный канал, через который дискретные сигналы (последовательность символов) проходят как непрерывные, отличающиеся друг от друга кодовыми признаками. На эти сигналы действуют помехи, вследствие чего на выходе дискретного канала появляется последовательность символов, среди которых есть искаженные. Это означает, что в канале возникли ошибки.

Дискретный канал в общем виде представляется совокупностью дискретного модулятора на входе, непрерывного канала и дискретного демодулятора на выходе (рис. 5.10).

Рис. 5.10. Дискретный канал в общем виде

Дискретный канал характеризуется:

Владимир Александрович Котельников (1908—2005), выдающийся советский и русский ученый в области радиотехники, радиосвязи и радиоастрономии. Основные работы посвящены проблемам усовершенствования методов радиоприема, изучению радиопомех и разработке методов борьбы с ними. К его наибольшим научным достижениям следует отнести открытие теоремы отсчетов, которая носит его имя, создание теории потенциальной помехоустойчивости, а также разработку планетарных радиолокаторов и проведение с их помощью фундаментальных астрономических исследований.

алфавитом входа $B_{\text{ё}} = (b_1; b_2; ...; b_i; ...; b_{m-1})$, например 0, 1, 2…;

алфавитом выхода $B'_{\text{ё}} = (b'_1; b'_2; ...; b'_i; ...; b'_{m-1})$, причем алфавиты входа и выхода не обязательно совпадают, например, выходной алфавит может иметь избыточные символы (*m* - основа кода);

скоростью передачи V_k символов/с. Эта скорость определяется в основном свойствами непрерывного канала - его памятью;

матрицей или графом переходов, т.е. совокупностью условных вероятностей $P(b'_i / b_j)$ того, что при входном символе b_j на выходе будет b'_i.

Дискретный канал, кроме перечисленных (см. рис. 5.11), имеет такие характеристики, как *пропускная способность и количество переданной по каналу информации*. Эти характеристики будет рассмотрено в следующем разделе.

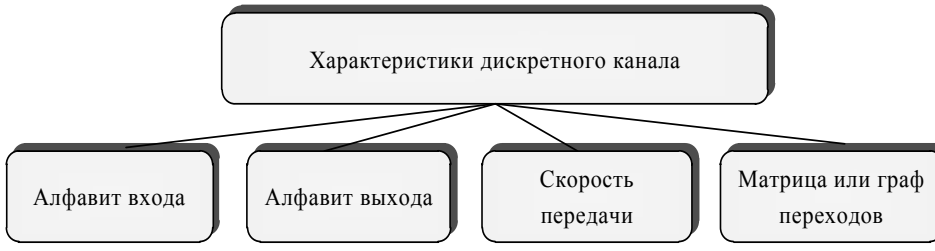


Рис. 5.11. Характеристики дискретного канала

Модели дискретных каналов передачи данных. Помехи и искажения в непрерывном канале вызывают появление в дискретном канале потока ошибок.

В зависимости от свойств потока ошибок дискретные каналы могут описываться такими моделями.

1. ***Двоичный симметричный канал без памяти*** *характеризуется тем, что каждый переданный символ может быть принят или ошибочно с вероятностью* P_0, *или правильно с вероятностью* $1 - P_0$, *причем в случае ошибки переходы* $1 \to 0$ *и* $0 \to 1$ *равновероятны.*

Вероятность того, что при передаче b_j будет получен b'_i:

$$P(b'_i / b_j) = P_0 \text{ при } i \neq j, \; P(b'_i / b_j) = 1 - P_0 \text{ при } i = j.$$

Отсутствие памяти проявляется в том, что условная вероятность $P(b'_i / b_j)$ не зависит от предыдущих событий, т.е. от того, какие символы передавались ранее и как они были приняты. Вероятность $P(b'_i / b_j)$ называется *априорной вероятностью и* для данного канала определяется на основе про-

должительных экспериментов.

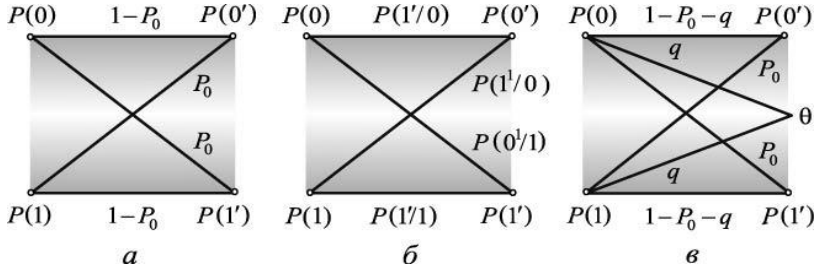


Рис. 5.12 Графы вероятностей переходов в двоичных симметричных каналах без памяти: *а* - с равными переходами; *б* - с неравными переходами; *в* - со стиранием памяти

Физический смысл симметрии канала заключается в том, что в дискретном демодуляторе, называемом также *первой расчетной схемой*, пороговый уровень выбран точно внутри между средними значениями сигналов, которые соответствуют нулю и единице. Если бы этот уровень был выбран возле значения 0, то за счет помехи вероятность перехода 0 > 1 была бы большей, чем 1 > 0.

Вероятность переходов в двоичном симметричном канале иллюстрирует рис. 5.12, *а*.

2. ***Двоичный несимметричный канал без памяти*** *отличается от предыдущей модели неодинаковой вероятностью переходов от* $0 \to 1$ и от $1 \to 0$. Графы переходов для данной модели изображены на рис. 5.12, *б*.

Полная вероятность приема *символа* b_i' *для двоичного канала на базе вероятности входных символов и условной вероятности переходов определяется как*

$$P(0') = P(0)P(0'/0) + P(1)P(0'/1);\ P(1') = P(1)P(1'/1) + P(0)P(1'/0).$$

Средняя вероятность ошибки *в канале*

$$P_0 = P(0)\ P(1'/0) + P(1)\ P(0'/1). \qquad (5.9)$$

3. ***Симметричный канал без памяти со стиранием*** *отличается от симметричного канала наличием в алфавите на выходе канала дополнительного символа, который появляется при ненадежном опознавании демодулятором переданного символа.*

Естественно, что в таком канале вероятность ошибочных переходов уменьшается за счет вероятности q появления символа стирания θ.

Граф переходов для рассматриваемого канала приведен на рис. 5.12, *в*.

Физический смысл канала со стиранием заключается в том, что в демодуляторе создаются два пороговых уровня; при этом фиксируется нуль, если уровень принятого непрерывного сигнала меньше нижнего уровня; фиксируется единица, если уровень сигнала больше верхнего уровня. Символ стирания θ фиксируется (его можно и не фиксировать) в случае попадания сигнала в промежуток между уровнями.

Основная задача приема дискретных сигналов. *Заданным* называется дискретный канал, если известны его алфавиты входа $(0,1)$ и выхода $(0',1')$, а априорная вероятность перехода символов входного алфавита в символы выходного определена условными вероятностями: $P(0'/0)$, $P(1'/0)$, $P(1'/1)$, $P(0'/1)$.

Обратная задача приема - *определение апостериорной вероятности того, что при приеме была подтверждена указанная гипотеза, например, при приеме нуля был передан нуль.* Поскольку нуль в приемнике можно было получить также и за счет ошибочного перехода $1 \to 0$, в случае двоичного канала необходимо рассматривать две гипотезы появления нуля на выходе приемника и определять вероятность этих гипотез по формуле Байеса (рис. 5.13):

$$P(0/0') = \frac{P(0)P(0'/0)}{P(0)P(0'/0) + P(1)P(0'/1)};$$
$$P(1/0') = \frac{P(1)P(0'/1)}{P(1)P(0'/1) + P(0)P(0'/0)}. \qquad (5.10)$$

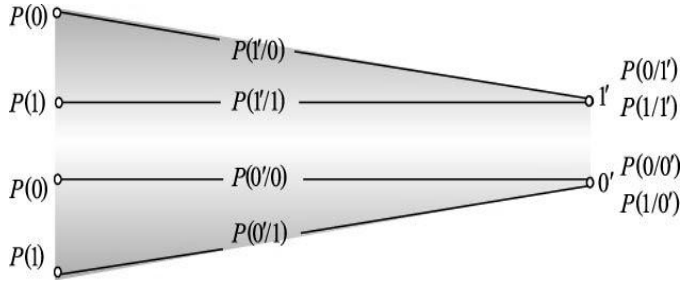


Рис. 5.13. Граф переходов в двоичном канале

Вероятность каждой гипотезы определяется отношением вероятности приема данного символа к полной вероятности его приема. Сравнивая вероятность гипотез, особенно при приеме маловероятного символа, когда вероятность гипотезы относительно правильного перехода приближенно равна вероятности ошибочного перехода, приходят к выводу о неудовлетворительной работе канала. Улучшения можно достичь изменением порогового уров-

ня так, чтобы априорная вероятность, например, приема нуля при передаче единицы $P(0'/1)$, была намного меньше вероятности $P(1)$, т.е. $P(0'/1) \ll P(1)$.

5.3. Скорость передачи информации в каналах связи

Скорость передачи информации и пропускная способность дискретного канала без помех. *Дискретный канал передачи информации - совокупность методов и средств, предназначенных для передачи дискретных сигналов от источника сообщения к потребителю.*

На вход такого канала подаются дискретные сообщения X создаваемые из первичного алфавита $x_1, x_2,..., x_n$. Эти сообщения кодируются с помощью кодеров (см. рис. 5.3) и превращаются в кодированные сообщения Y. Для кодирования используется некоторый алфавит символов $y_1, y_2,..., y_m$. Суть кодирования сводится к представлению отдельных сообщений или последовательностей сообщений определенными комбинациями символов используемого алфавита.

Скорость введения *информации дискретного канала*

$$\bar{I}(X) = \frac{H(X)}{\bar{\tau}_X} = \bar{V}_X H(X), \tag{5.11}$$

где $H(X)$ - средняя энтропия одного сообщения; $\bar{\tau}_X$ - средняя продолжительность сообщения.

Скорость выдачи символов *сообщения источником определяется как*

$$\bar{V}_X = 1/\bar{\tau}_X .$$

Под *продолжительностью сообщения* понимается интервал времени, в течение которого сообщение на выходе источника информации формируется или существует.

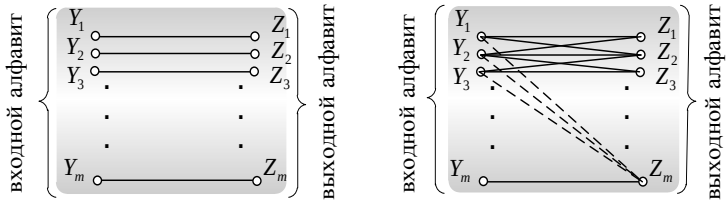
Средняя продолжительность $\bar{\tau}_x$ *информационного сообщения на фоне источника при отсутствии статистических зависимостей между сообщениями определяется выражением*

$$\bar{\tau}_X = \sum_{i=1}^{n} p(x_i)\tau_{X_i}, \tag{5.12}$$

где $p(x_i)$ и τ_{x_i} - априорная вероятность и продолжительность i-го сообщения соответственно; n - количество сообщений.

В канале без помех каждому определенному входному сигналу всегда будет соответствовать тот же сигнал на выходе канала, другими словами,

входные и выходные сигналы связаны между собой однозначной функциональной зависимостью (рис. 5.14, *а*).



а *б*

Рис. 5.14. Графы функциональных зависимостей между входом и выходом дискретного канала без помех (*а)* и с помехами (*б*)

Среднее количество информации *в этом случае равно энтропии символа на входе канала*, дв. ед./символ:

$$I(Y) = H(Y).$$

Скорость передачи информации *в дискретном канале без помех*, дв.ед./символ:

$$I(Y) = \overline{U}_Y \cdot H(Y), \tag{5.13}$$

где $\overline{U}_Y = 1 / \overline{\tau}_Y$ - скорость передачи элементарных символов сигнала; τ_Y - средняя продолжительность элементарных сигналов.

Пропускная способность *дискретного канала без помех*

$$C = \max\{\overline{U}_Y H(Y)\}.$$

Максимальная скорость передачи информации $\overline{U}_Y$ *будет обеспечена при максимальном значении энтропии кодированного сигнала*

$$C = \overline{U}_Y \max\{H(Y)\} = \overline{U}_Y \log_2 n, \tag{5.14}$$

т.е. в случае равномерного распределения вероятностей и статистической независимости символов алфавита сигналов.

Таким образом, скорость передачи информации может быть максимальной при условии согласования определенным образом статистических характеристик источника сообщений со свойствами информационного канала. Для каждого источника сообщений подобное согласование может быть достигнуто специальным выбором способа кодирования сигналов.

На вопрос о степени, в которой скорость передачи информации может быть приближена к пропускной способности информационного канала, отвечает теорема К. Шеннона для дискретного канала без помех.

Теорема К. Шеннона для дискретного канала связи без помех: *если поток информации от источника достаточно близок к пропускной способности канала, т.е. если выполняется равенство*

$$\overline{I}(X) = C - \delta, \tag{5.15}$$

где δ *- достаточно малая величина, то всегда можно найти такой способ кодирования, который обеспечит передачу всех сообщений от источника к потребителю, причем скорость передачи информации будет достаточно близкой к пропускной способности канала*

$$\overline{I}(Z, Y) = C - \delta.$$

Обратное утверждение теоремы *заключается в том, что невозможно обеспечить продолжительную передачу всех сообщений, если поток информации от источника превышает пропускную способность канала*

$$\overline{I}(X) > C.$$

Итак, теорема К. Шеннона утверждает, что при выполнении условия (5.15) скорость передачи информации может быть в принципе в достаточной степени приближена к пропускной способности канала. Это может быть обеспечено соответствующим кодированием сигналов. Тем не менее, рассмотренная теорема не отвечает на вопрос, каким образом нужно осуществлять кодирование.

Скорость передачи информации и пропускная способность дискретного канала с помехами. При наличии помех в канале передачи информации нарушается однозначное соответствие между входным и выходным алфавитами канала. Одному входному сигналу могут соответствовать различные выходные сигналы (см. рис. 5.14, *б)*. Из-за случайного характера помех невозможно заранее точно установить, какой сигнал может быть принят на выходе канала при посылке определенного входного сигнала. Речь может идти только о вероятностях получения на выходе канала элементарного сигнала z_Y при условии, что был отправлен соответствующий элементарный сигнал.

Вероятностный характер связи *между входным и выходным алфавитами канала передачи информации полностью определяется матрицей переходных вероятностей*

$$\begin{vmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{vmatrix},$$

где p_{ij} - условная вероятность перехода i-го символа входного алфавита в j-и символ выходного алфавита.

Очевидно, выполняется такое равенство:

$$\sum_{j=1}^{n} p_{ij} = 1.$$

Бинарным каналом называется *дискретный канал, по которому передаются только два типа элементарных сигналов.*

Матрица переходных вероятностей для такого канала:

$$\begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}.$$

Симметричным называется канал *при условии, что все вероятности правильной передачи сигналов одинаковы, а также одинаковы все вероятности искаженной передачи.*

Для симметричного бинарного канала матрица переходных вероятностей

$$\begin{vmatrix} p & q \\ q & p \end{vmatrix},$$

где $p = p_{11} = p_{22}$ - вероятность правильной передачи; $q = p_{12} = p_{21}$ - вероятность искаженной передачи.

Поскольку в симметричном канале вероятности искажения всех символов сигнала одинаковы, то можно утверждать, что в таком канале помехи не зависят от передаваемых сигналов.

На рис. 5.15 приведен график переходных вероятностей двоичного симметричного канала.

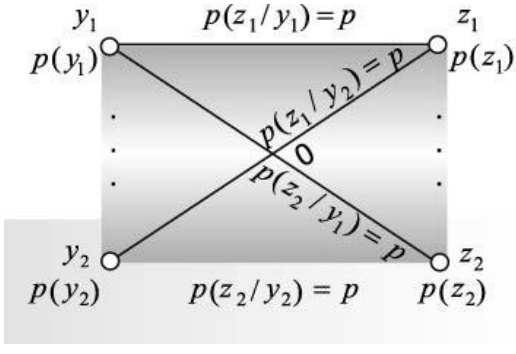


Рис. 5.15. Граф вероятностей переходов в двоичном симметричном канале без памяти с помехами

Скорость передачи информации *дискретного канала с помехами*

$$\bar{I}(Z,Y) = \bar{U}_Y[H(Y) - H(Y/Z)], \tag{5.16}$$

где $H(Y/Z)$ - остаточная энтропия сигнала, обусловленная действием помех.

Выражение для скорости передачи информации можно представить также в виде

$$\bar{I}(Z,Y) = \bar{U}_Y[H(Z) - H(Z/Y)], \tag{5.17}$$

где $H(Z)$ - энтропия выходного сигнала; $H(Z/Y)$ - условная энтропия выходного сигнала по известной энтропии входного сигнала.

Каналом без памяти *называется дискретный канал, в котором помехи влияют на каждый переданный символ информационного сигнала независимо от того, какие сигналы передавались раньше.*

В таких каналах помехи не вызывают дополнительных коррелятивных связей между символами.

В случае независимости отдельных символов сигнала выражения (5.16) и (5.17) для канала без памяти приобретают следующий вид:

$$\bar{I}(Z,Y) = \bar{U}_Y[-\sum_{i=1}^{n} p(y_i)\log_2 p(y_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} p(z_i)\,p(y_i/z_i)\log_2 p(y_i/z_i)]; \tag{5.18}$$

$$\bar{I}(Z,Y) = \bar{U}_Y[-\sum_{i=1}^{n} p(z_i)\log_2 p(z_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} p(y_i)\,p(z_i/y_i)\log_2 p(z_i/y_i)]. \tag{5.19}$$

В качестве примера рассмотрим бинарный канал. Для такого канала алфавиты входного *Y* и выходного *Z* сигналов состоят из двух символов

$$Y = \{y_1, y_2\};\ Z = \{z_1, z_2\}.$$

Для сокращения записи обозначим вероятности искажения сигналов

$$p(z_1/y_2) = q_1\ ^{з}\ p(z_2/y_2) = q_2.$$

Очевидно, что вероятности правильной передачи будут

$$p(z_1/y_1) = 1 - q_2;\ p(z_2/y_2) = 1 - q_1.$$

При этом выражение (5.19) приобретет вид

$$\bar{I}(Z,Y) = \bar{U}_Y\{-p(z_1)\log_2 p(z_1) - p(z_2)\log_2 p(z_2) +$$
$$+ p(y_1)[(1-q_2)\log_2(1-q_2) + q_2\log_2 q_2] + p(y_2)[q_1\log_2 q_1 +$$
$$+ (1-q_1)\log_2(1-q_1)]\}.$$

Максимизируя правую часть выражения (5.19), можем определить пропускную способность канала. Очевидно, что этого можно достичь за счет оп-

тимизации значений априорных вероятностей $p(y_1)$ и $p(y_2)$ передачи сигналов y_1 и y_2, поскольку никаких других параметров канала изменять мы не можем.

Рассмотрим частные случаи.

1. Вероятности искажения сигналов $q_1 = q_2 = q$. Этот случай соответствует симметричному каналу. Для симметричного канала условная энтропия

$$H(Z/Y) = -[p(y_1) + p(y_2)][q\log_2 q + (1-q)\log_2(1-q)] = \\ = q\log_2 q + (1-q)\log_2(1-q), \tag{5.20}$$

поскольку $p(y_1) + p(y_2) = 1$.

Из выражения (5.20) следует, что условная энтропия не зависит от априорных вероятностей $p(y_1)$ è $p(y_2)$. Итак, максимальная скорость передачи информации достигается в этом случае при таком распределения вероятностей $p(y_1)$ и $p(y_2)$, при котором энтропия $H(Z)$ оказывается максимальной. Это будет в случае равенства априорных вероятностей $p(z_1) = p(z_2)$. Тогда максимальное значение энтропии $H(Z_{\max}) = 1$ дв. ед. и пропускная способность канала определяется выражением

$$C = \bar{U}_Y[\log_2 2 + q\log_2 q + (1-q)\log_2(1-q)]. \tag{5.21}$$

Используя известное правило теории вероятностей

$$p(z_i) = \sum_{i=1}^{n} p(y_i)\,p(z_i/y_i), \tag{5.22}$$

можно показать, что равенство априорных вероятностей выходных сигналов z_1 и z_2 для симметричного канала достигается в случае равенства априорных вероятностей входных сигналов.

Таким образом, в симметричном бинарном канале с помехами максимальная скорость передачи информации достигается при таком же условии, как и в канале без помех. Однако из сравнения формул (5.21) и (5.14) следует, что наличие помех в канале приводит к уменьшению пропускной способности канала.

2. Вероятности искажения сигналов $q_1 = 0,\ q_2 \neq 0,$ т.е. искажения возникают лишь при передаче сигналов y_2.

Из формулы (5.19) с использованием формулы (5.22) для этого случая получаем

$$\bar{I}(Z,Y) = \bar{U}_Y\{-p(y_1)\log_2 p(y_1)q_2\log_2 q_2 - [p(y_2) + \\ + p(y_1)q_2]\log_2[p(y_2) + p(y_1)q_2]\}.$$

Максимальное значение скорости передачи информации достигается при условии

$$p(y_1)_{\text{опт}} = \frac{1}{1 + q_2^{\frac{q_2}{1-q_2}} - q_2}. \tag{5.23}$$

Пропускная способность такого канала определяется выражением:

$$\begin{aligned} C = \bar{U}_Y \{ &- p(y_1)_{\text{опт}} + \log_2 p(y_1)_{\text{опт}} + p(y_1)_{\text{опт}} q_2 \log_2 q_2 - \\ &- [1 - p(y_1)_{\text{опт}} + p(y_1)_{\text{опт}} q_2] \log_2 [1 - p(y_1)_{\text{опт}} + p(y_1)_{\text{опт}} q_2] \}. \end{aligned} \tag{5.24}$$

На рис. 5.16 приведены графики $\frac{C}{\bar{U}_Y} = C\bar{\tau}_Y$, $p(y_1)_{\text{опт}}$ и $p(y_2)_{\text{опт}} = 1 - p(y_1)_{\text{опт}}$ как функции от вероятности q_2. Как следует из графиков, с увеличением q_2 от 0 до 1 пропускная способность спадает от $U_Y = 1/\tau_Y$ до нуля.

Теорема ***К. Шеннона для дискретного канала с помехами***: *если поток информации от источника сообщений достаточно близок к пропускной способности канала (если выполняется равенство $\bar{I}(X) = C - \delta$, где δ - достаточно малая величина), то всегда можно найти такой способ кодирования, который обеспечит передачу всей информации, создаваемой источником сообщений, со сколь угодно малой вероятностью неправильного (ошибочного) выявления любого переданного сообщения $p_{\text{н.в}}$ (т.е. $p_{\text{н.в}} < \eta$, где η - бесконечно малая величина).*

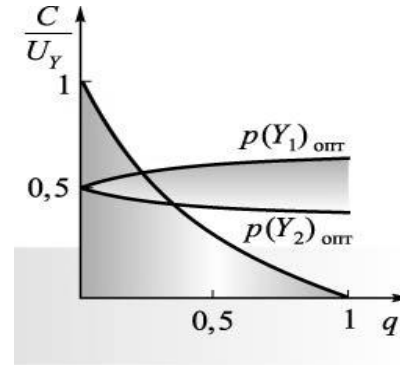


Рис. 5.16. Пропускная способность канала в зависимости от вероятностей искажений

Обратное утверждение теоремы: *если поток информации источника превышает пропускную способность канала, то не существует способа кодирования, который обеспечивает передачу любого сообщения с малой вероятностью ошибки.*

Таким образом, рассмотренная теорема определяет соотношение между скоростью создания сообщений источником, пропускной способностью канала при наличии помех и вероятностью передачи. Если для канала без помех характерной является эффективность передачи, то для канала с помехами — эффективность и вероятность передачи.

Эта теорема, как и теорема для канала без помех, не отвечает на вопрос, каким образом нужно осуществлять кодирование, чтобы приблизить скорость передачи информации к пропускной способности канала. Но для приближе-

ния скорости передачи к предельной общим методом как для канала с помехами, так и для канала без помех является кодирование длинных сообщений.

Скорость передачи информации и пропускная способность непрерывного канала с помехами. ***Непрерывный канал передачи информации*** - *совокупность методов и средств, предназначенных для передачи непрерывных сигналов от источника сообщения к потребителю.*

В отличие от дискретных каналов, в непрерывных каналах вместо кодирующих и декодирующих устройств может использоваться широкий класс различных преобразователей. Для передачи информации по каналу может применяться модуляция одного или нескольких параметров сигнала. Независимо от конкретного характера преобразования сигналов входные и выходные сигналы непрерывного канала задаются в виде ансамблей непрерывных функций с соответствующими функциями плотности распределения вероятностей.

Пусть на вход канала поступает непрерывный сигнал $Y(t)$ продолжительности T. Вследствие влияния помех $\xi(t)$ выходной сигнал $Z(t)$ будет отличаться от входного.

Количество информации в случайном сигнале $Z(t)$ относительно случайного сигнала $Y(t)$ определяется известной зависимостью

$$I_T(Z,Y) = H_T(Z) - H_T(Z/Y) \tag{5.25}$$

В соответствии с теоремой Котельникова, непрерывные сигналы $Y(t)$ и $Z(t)$ можно представить совокупностями отсчетов y_i и z_i в дискретные моменты времени (рис. 5.17), являющиеся случайными величинами.

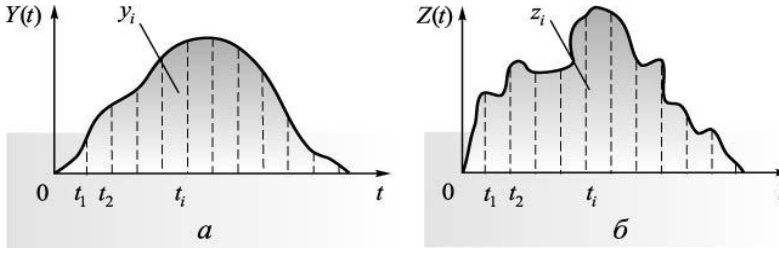


Рис. 5.17. Случайный информационный сигнал (*а*), смесь информационного сигнала и помех (*б*)

Распределение совокупности случайных величин описывается многомерными плотностями распределения вероятности $w(y_1, y_2, ..., y_m)$ и $w(z_1, z_2, ..., z_m)$.

Тогда дифференциальная энтропия сигнала на выходе канала

$$h_t(Z) = -\int_{-\infty}^{\infty} ...m ... \int_{-\infty}^{\infty} w(z_1, z_2, ..., z_m)\log_2 w(z_1, z_2, ..., z_m)dz_1,\ dz_2, ..., dz_m.$$

В соответствии с критерием М. А. Железнова при квантовании случайных сигналов по времени интервал квантования необходимо брать равным интервалу корреляции функции τ_0. Тогда случайные величины $z_1, z_2, ..., z_i, ... z_m$ можно считать независимыми и такими, как

$$w(z_1, z_2, ..., z_m) = w(z_1)w(z_2) ... w(z_m).$$

Исходя из равенства энтропии совокупности независимых случайных величин сумме энтропии случайных величин, получаем следующее выражение для ***дифференциальной энтропии сигнала***:

$$h_T(Z) = \sum_{i=1}^{m} h(z_i),$$

где $h(z_i) = -\int_{-\infty}^{\infty} w(z_i)\log_2 w(z_i)dz_i$ - дифференциальная энтропия *i*-го отсчета сигнала Z; $m = T/\Delta t$ - общее количество отсчетов сигнала Z продолжительности T; Δt - интервал временного квантования.

Ограничиваясь рассмотрением стационарных процессов, получаем

$$w(z_1) = w(z_2) = ... = w(z_m);$$
$$h(z_1) = h(z_2) = ... = h(z_m) = h(Z).$$

Тогда $h_T(Z) = m\ h(Z)$, где $h(Z)$ - дифференциальная энтропия одного отсчета.

Аналогично можно утверждать, что условная дифференциальная энтропия

$$h_T(Z/Y) = m\ h(Z/Y),$$

где $h(Z/Y)$ - условная дифференциальная энтропия одного отсчета.

Таким образом, выражение для количества информации приобретает вид

$$I_T(Z,Y) = m[h(z) - h(Z/Y)].$$

Скорость передачи информации *в непрерывном канале с помехами*

$$\bar{I}_T(Z,Y) = \frac{m}{T}[h(z) - n(Z/Y)] = F_0[h(Z) - h(Z/Y)], \qquad (5.26)$$

где $F_0 = \frac{m}{T} = \frac{1}{\Delta t}$ - *частота временного квантования* (*отсчета*).

Пропускная способность канала *в непрерывном канале с помехами*

$$C = \max[\bar{I}_T(Z,Y)] = F_0 \max[h(Z) - h(Z/Y)]. \qquad (5.27)$$

Рассмотрим некоторые частные случаи.

1. Сигнал ограниченной мощности передается по каналу, в котором действует аддитивная помеха ограниченной мощности типа белого гауссового шума.

При аддитивной помехе сигнал $Z(t)$ на выходе канала будет равен

$$Z(t) = Y(t) + \xi(t),$$

где $\xi(t)$ - помеха, действующая в канале передачи информации.

Средние мощности сигнала и помехи соответственно равны P_Y и $P_\xi = \sigma_\xi^2$. Полоса пропускания канала ограничена значениями 0 и $F_к$. Ширина спектра сигнала и помехи ограничивается полосой пропускания канала. Частота квантования ограниченного по спектру сигнала в соответствии с теоремой Котельникова $F_0 = 2F_{\text{ê}}$. Тогда выражение (5.27) для пропускной способности канала приобретет вид

$$C = 2F_к \max[h(Z) - h(Z/Y)]. \tag{5.28}$$

При взаимно независимым сигнале Y и помехе ξ вероятность того, что при передаче сигнала Y выходной сигнал будет равен $Z = Y + \xi$, должна определяться вероятностью того, что помеха приобретет данное значение $\xi = Z - Y$, т.е.

$$w(Z/Y)dY = w(\xi)d\xi.$$

При этом

$$\begin{aligned} w(Z/Y) &= w(Y + \xi/Y) = w(Y/Y + \xi/Y) = \\ &= w(Y/Y)\ w\ (\xi/Y) = w(\xi). \end{aligned} \tag{5.29}$$

Учитывая формулу (5.29), выражение для условной дифференциальной энтропии $h(Z/Y)$ можно преобразовать таким образом:

$$\begin{aligned} h(Z/Y) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(Y)\ w\ (Z/Y)\log_2 w(Z/Y)\ dY\ dZ = \\ &= -\int_{-\infty}^{\infty} w(Y)\left[\int_{-\infty}^{\infty} w(\xi)\log_2 w(\xi)\ d\xi\right]dY = \\ &= h(\xi)\int_{-\infty}^{\infty} w(Y)dY = h(\xi), \end{aligned} \tag{5.30}$$

где $h(\xi)$ - дифференциальная энтропия помехи.

Итак, в случае аддитивной помехи условная дифференциальная энтро-

пия $h(Z/Y)$ полностью определяется свойствами помехи.

Ранее было установлено, что выражение для дифференциальной энтропии сигнала, распределенного по нормальному закону, имеет вид:

$$h(\xi) = \log_2(\sqrt{2\pi e}\,\sigma_\xi), \tag{5.31}$$

где e — основа натурального логарифма, $e = 2,7$.

Подставив выражение (5.31) в формулу (5.28), получим следующее выражение для пропускной способности канала:

$$C = 2F_к \max\{h(Z) - \log_2\sqrt{2\pi e}\,\sigma_\xi\}. \tag{5.32}$$

Поскольку значение σ_ξ задано, то максимальное значение выражения (5.32) будет обеспечено при условии максимизации дифференциальной энтропии выходного сигнала $h(Z)$. Средние мощности входного сигнала $Y(t)$ и помехи $\xi(t)$ ограничены, поэтому средняя мощность выходного сигнала $Z(t)$ также ограничена. Дифференциальная энтропия $h(Z)$ будет максимальна, если $Z(t)$ характеризуется нормальным законом распределения. Если же суммарный сигнал $Z(t)$ и одна из его составляющих $\xi(t)$ распределены по нормальному закону, то и вторая составляющая, т.е. входной сигнал $Y(t)$, также должна соответствовать нормальному закону распределения.

Таким образом, дифференциальная энтропия выходного сигнала

$$h(Z) = \log_2(\sqrt{2\pi e}\,\sigma_Z) = \log_2\sqrt{(\sigma_Y^2 + \sigma_\xi^2)2\pi e}. \tag{5.33}$$

Подставляя выражение (5.31) в формулу (5.30), окончательно получаем

$$C = 2F_к\left[\log_2\sqrt{(\sigma_Y^2 + \sigma_\xi^2)2\pi e} - \log_2\sqrt{\sigma_\xi^2 2\pi e}\right] = F_к\log_2\frac{(\sigma_Y^2 + \sigma_\xi^2)}{\sigma_\xi^2} = F_к\log_2\left(1 + \frac{P_Y}{P_\xi}\right), \tag{5.34}$$

Эдмон Никола́ Лагерр (Edmond Nicolas Laguerre, 1834 - 1886), французский математик, член Парижской академии наук с 1884 г. Закончил Политехническую школу в Париже (1854). Служил офицером в артиллерии. Репетитор Политехнической школы (1864-1884); профессор в колледж де Франс (1883 - 1886). Основные работы посвящены геометрии. В своих исследованиях искал возможность конкретного изображения мысленных точек на плоскости и в пространстве. Разрабатывал аналитическую теорию функций комплексной переменной, изучал многочлены (многочлены Чебышева - Лагерра).

где $P_Y = \sigma_Y^2$ - средняя мощность полезного сигнала; $P_\xi = \sigma_\xi^2$ - средняя мощность помехи.

На основании вышеизложенного можно сделать такие выводы:

скорость передачи информации сигналами с ограниченной средней мощностью по каналу, в котором действует белый гауссов шум, оказывается максимальной в случае полного подобия сигнала и помехи;

максимальная скорость передачи информации будет обеспечена, если в качестве физического носителя информации применять стационарный случайный процесс в виде белого гауссова шума.

Как следует из формулы (5.34), пропускную способность канала можно регулировать, изменяя $F_к$ или P_Y. При этом зависимость пропускной способности канала от $F_к$ при постоянной мощности сигнала практически не линейна. Это обусловлено тем, что мощность помехи P_ξ также зависит от ширины частотного спектра. В самом деле, энергетический спектр белого шума равномерен, поэтому мощность такой помехи можно представить в виде

$$P_\xi = P_0 F_к, \tag{5.35}$$

где P_0 - мощность помехи, приходящаяся на полосу в 1 Гц (спектральная плотность мощности помехи).

Подставив формулу (5.35) в (5.34), получим выражение, определяющее настоящий характер зависимости пропускной способности канала от ширины его полосы пропускания:

$$C = F_к \log_2 \left(1 + P_Y / P_0 F_к\right). \tag{5.36}$$

Характер зависимости $C = f(F_к)$ иллюстрирует рис. 5.18.

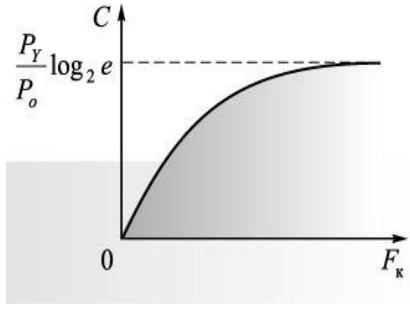


Рис. 5.18. Зависимость пропускной способности канала от ширины его полосы пропускания

Определим границу, к которой стремится пропускная способность канала при неограниченном увеличении его полосы пропускания:

$$\lim_{F_к \to \infty} C = \lim_{F_к \to \infty} \frac{\log_2 \left(1 + P_Y / P_0 F_к\right)}{1/F_k}.$$

Введя обозначение $\alpha = 1/F_к$, получим

$$\lim_{F_к \to \infty} C = \lim_{\alpha \to 0} \frac{\log_2 \left(1 + \frac{P_Y}{P_0} \alpha \right)}{\alpha}.$$

Раскрывая неопределенность, получаем предельное значение пропускной способности канала

$$\lim_{F_к \to \infty} C = \frac{P_Y}{P_0} \log_2 e. \tag{5.37}$$

Из формулы (5.37) следует, что максимальное значение, к которому стремится пропускная способность канала с ростом ширины его полосы пропускания, пропорционально отношению средней мощности сигнала к спектральной плотности мощности помехи.

Из приведенного анализа можно сделать вывод: *нет смысла чрезмерно увеличивать полосу пропускания канала, поскольку с расширением полосы пропускания возрастание пропускной способности канала замедляется, и в границе при* $F_к \to 0$ *пропускная способность приближается к постоянной величине. При этом полоса пропускания достигает значения, близкого к отношению* P_Y / P_0 .

Предельно возможное значение пропускной способности можно увеличить за счет увеличения отношения P_Y / P_0 . Сигнал ограниченной мощности передается по каналу, в котором действует аддитивная помеха в виде произвольного шума. Как уже установлено, при определенном среднеквадратичном значении (при определенной мощности) помехи наибольшую энтропию имеет помеха с нормальным законом распределения вероятностей. При любом другом законе распределения вероятностей помехи ее энтропия (при том же среднеквадратичном значении) меньше. Зависимость энтропии от вида закона распределения побудила К. Шеннона характеризовать помеху не ее настоящей мощностью, а так называемой энтропийной мощностью. Под *энтропийной мощностью* К. Шеннон понимал мощность эквивалентного белого шума, имеющего такую же продолжительность, ширину спектра и энтропию, что и данная помеха.

Таким образом, если произвольная помеха $\xi(t)$ характеризуется энтропией на один отсчет $h(t)$, то мощность эквивалентного белого шума, т.е. энтропийную мощность $\sigma^2_{\xi e}$, можно определить из условия

$$h(\xi) = \log 2\sqrt{2\pi e \sigma^2_{\xi e}},\ \sigma^2_{\xi e} = \frac{2^{p(\varepsilon)}}{2\pi e} = \frac{2^{h(\xi)-1}}{\pi e}. \tag{5.38}$$

Значение отношения средней мощности помехи σ_ξ^2 к ее энтропийной мощности $\sigma_{\xi e}^2$ определяется законом распределения помехи. Обозначив символом k_e, данное отношение, получим

$$\sigma_\xi^2 = k_e \sigma_{\xi e}^2. \tag{5.39}$$

В частности, для закона равной вероятности $k_e = 1{,}3$.

Используя понятие энтропийной мощности, можно получить выражение для пропускной способности канала, в котором действует произвольная помеха

$$\begin{aligned} C &= 2F_к[\log_2 \sqrt{(\sigma_Y^2 + \sigma_\xi^2)2\pi e} - \log_2 \sqrt{\sigma_{\xi e}^2 2\pi e}] = \\ &= F_к \log_2 (\frac{\sigma_Y^2 + \sigma_\xi^2}{\sigma_{\xi e}^2}) = F_к \log_2 (k_e \frac{\sigma_Y^2 + \sigma_\xi^2}{\sigma_\xi^2}). \end{aligned} \tag{5.40}$$

Для всех законов распределения (кроме нормального) коэффициент $k_e > 1$, поэтому пропускная способность канала, в котором действует произвольная помеха, будет всегда больше пропускной способности канала, в котором действует белый шум с такой же средней мощностью, что и произвольная помеха.

Теорема К. Шеннона для непрерывного канала с помехами: *если энтропия $\bar{H}_\xi(X)$ источника непрерывных сообщений, определяющая количество информации в единицу времени при заданной оценке g правильности воспроизведения, достаточно близка к пропускной способности канала (т.е. выполняется соотношение $\bar{H}_\alpha(X) = C - \alpha$, где α - бесконечно малая величина), то существует такой метод передачи, при котором все сообщения от источника могут быть переданы с правильностью воспроизведения как угодно близкой к g.*

Обратное утверждение этой теоремы говорит о том, что такая передача невозможна, если $\bar{H}_\alpha(X) > C$.

Теорема дает возможность находить предельно достижимую эффективность непрерывных каналов. Количество информации, которую можно передать по каналу за время его работы $T_к$ при влиянии помех типа белого шума:

$$I_T(Z,Y) = T_к F_к \log_2 \left(1 + P_Y / P_\xi\right). \tag{5.41}$$

Практически в большинстве случаев мощность полезного сигнала значительно превышает мощность помех. В этих случаях выражение (5.41) можно

с достаточным приближением представить в виде

$$I_T(Z,Y) = T_к F_к \log_2 \frac{P_Y}{P}. \qquad (5.42)$$

В формулах (5.41) и (5.42) члены $\log_2\left(1+P_Y/P_\xi\right)$, $\log_2\left(P_Y/P_\xi\right)$ выражают (с точностью до постоянного множителя) максимально возможное количество информации на один отсчет. Тогда, сравнивая формулы (5.41) и (5.42) из (5.3), можно утверждать, что емкость канала определяет максимально возможное количество информации, которую можно передать по этому каналу за время его работы. А поскольку пропускная способность выражает максимально возможное количество информации, которую можно передать по каналу за единицу времени, то связь между емкостью и пропускной способностью канала определяется зависимостью

$$V_к = T_к\ C. \qquad (5.43)$$

5.4. Синтез элементов информационных систем. Оптимальный приемник

Большинство информационных процессов связано с решением проблемы выбора. Такая же проблема решается при поиске наилучших в определенном понимании алгоритмов обработки информационных сообщений и сигналов, технической реализации этих методов. Основные алгоритмы преобразующих элементов информационных систем принимают логическое двоичное решение «да» или «нет»: найдено сообщение или не найдено, различимы сигналы на фоне помех или неразличимы, искривлено сообщение или неискривлено, доступна информация потребителю или недоступна и т.д. Эти задачи решаются при обмене данными, при формировании и принятии решений, при отображении информации и т.п.

Задача синтеза состоит *в* определении алгоритма функционирования информационных систем по заданному критерию качества и интерпретировании или внедрении этого алгоритма с помощью технических средств.

Задача анализа состоит в расчете рабочих характеристик и обобщенных структур информационных систем.

Сообщения, поступающие в информационные системы, являются случайными процессами. Сигналы как носители сообщений и помехи, действующие на них, также являются случайными процессами. Поэтому поиск алгоритмов обработки сигналов с помехами осуществляется при использовании вероятностных моделей. Оценки (решения), получаемые на основании выборок конечного размера, называют *статистическими характеристиками*. При поиске решения всегда возникает ситуация неопределенности относительно распределения вероятности сигналов и помех, их параметров и дополнительных ограничений. Если все эти данные не известны, то говорят о зада-

Альфред Хаар (Alfréd Haar, 1885 - 1933), венгерский математик. Работал в Коложваре и Сегеде. Основные работы касаются ортогональных рядов и сингулярных интегралов, дифференциальных уравнений из частными производными, вариационного исчисления, теории групп и других разделов математики. Его именем названо инвариантную меру на непрерывной группе, вейвлет и преобразование Хаара. В 1959 г. были изданы собрания сочинений ученого, куда вошли 35 основных его работ.

чах статистического синтеза в условиях априорной неопределенности.

Из-за полного отсутствия априорных данных решать задачи оптимального синтеза невозможно. Но на практике всегда удается найти какие-либо априорные сведения.

Различают такие основные типы задач статистического синтеза:

1. *Выявление сигнала на фоне помех.*
2. *Различение сигналов на фоне помех.*
3. *Выявление* (*различение*) *сигналов и оценивание их параметров на фоне помех.*
4. *Выделение сигналов на фоне помех.*

Результатом решения перечисленных задач являются соответствующие алгоритмы обработки сигнала с помехой, приводящие, как правило, к улучшению соотношения между полезным сигналом и помехой. Все эти задачи имеют не только много общего, но и свои особенности, сказывающиеся на структуре алгоритмов.

Статистические критерии выявления сигналов на фоне помех. При приеме информационных сообщений в зависимости от назначения системы и вида сигналов возникают два типа задач: выявление сигналов и распознавание сигналов.

Выявление сигналов. *Задача выявления информационного сигнала заключается в том, чтобы по результатам процесса обработки принятого сообщения, могущего быть либо помехой, либо суммой переданного сигнала и помехи, были приняты решения относительно наличия или отсутствия в этом сигнале полезной информации.*

При выявлении возможны две ошибки:

1) *ошибка 1-го рода, или* «*ошибочная тревога*» *- при отсутствии полезного сигнала выносит ошибочное решение о наличии сигнала.*

2) *ошибка 2-го рода, или* «*пропуск цели*» *- при наличии полезного сигнала в сообщении выносит ошибочное решение о его отсутствии.*

Эти ошибки количественно оцениваются условной вероятностью α о наличии сигнала при его отсутствии, и условной вероятностью β ошибочного решения об отсутствии сигнала при его наличии.

Полная вероятность ошибочного решения определяется выражением $P_0 = q\alpha + p\beta$, где q и p - априорная вероятность соответственно отсутствия и наличия полезного сигнала.

Очевидно, что потери, испытываемые потребителем информации при ошибках 1-го и 2-го рода, могут быть далеко не одинаковы.

Распознавание сигналов. ***Передачей с пассивной паузой*** *называется такой способ передачи информации, при котором наличие полезного сообщения* (*посылка*) *соответствует приему символа* 1 (*гипотеза* H_1), *а отсутствие сигнала* (*пауза*) - приему *символа* 0 (*гипотеза* H_0).

Передачей с активной паузой *называется такой способ передачи информации, при котором в процессе различения двух сигналов символу* 1 *отвечает сигнал* x_1, *а символу* 0 - *сигнал* x_0; *причем сигналы* x_0 *и* x_1 *имеют одинаковую энергию.*

При распознавании сигналов могут также появляться ошибки 1-го и 2-го рода. Пусть α - условная вероятность ошибочного решения о наличии сигнала x_1, если на самом деле принято x_2, а β - условная вероятность ошибочного решения о наличии сигнала x_2, если принято x_1. Задача распознавания, или различения, сигналов возникает в системах автоматизированного управления, радиолокации и радионавигации, телекоммуникационных систем и сетей и т.п.

Рассмотрим графики распределения плотности условной вероятности событий, которые иллюстрируют характеристику принятым сигналом состояния a_1 или a_2 (рис. 5.19). Эти графики пересекаются; причем в симметричном канале граница ab различения признаков сигналов x_1 о состоянии a_1 и сигналов x_2 о состоянии a_2 лежит по середине интервала, разделяющего $\overline{x}_1$ и $\overline{x}_2$. Кроме этого, в таком канале $\alpha = \beta$.

Если источник информации вырабатывает только два сигнала x_1 и x_2, образующих полную группу событий $\left[P(x_1) + P(x_2) = 1\right]$, то на выходе приемника возможны четыре события, которые также образуют полную группу: $P(x_1') + \alpha + P(x_2') + \beta = 1$. Указанные переходы иллюстрирует рис. 5.20. Этот тип задач является базовым и довольно распространенным в технике связи, автоматике и радиолокации, в измерениях, где широко применяются статистические критерии выявления и распознавания. Рассмотрим коротко некоторые из них.

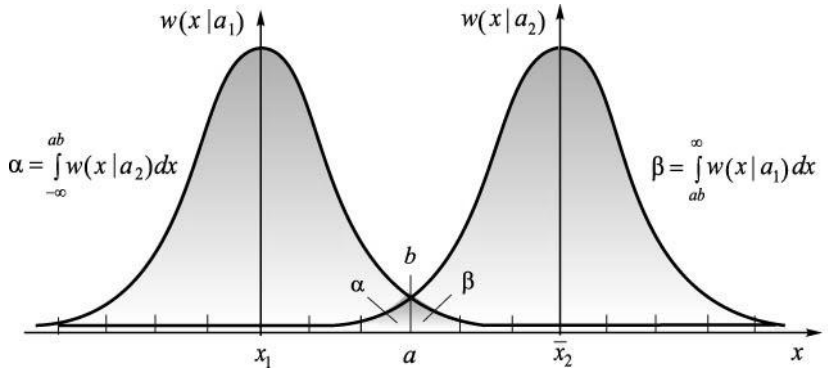


Рис. 5.19. Графики распределения плотности условной вероятности в симметричном канале

Общие критерии распознавания. На графике, изображенном на рис. 5.19, граница различения признаки *ab* проведена через точку пересечения графиков. При $w(x|a_1) = w(x|a_2)$ реализуется критерий *максимального правдоподобия* (не предоставляется преимущество ни одному из сигналов) $\alpha = \beta$.

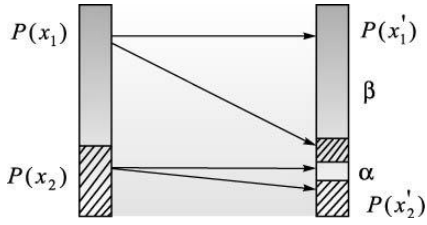


Рис. 5.20. Граф переходов вероятностей при двух сообщениях источника x_1 и x_2

Вероятность ошибочных решений зависит не только от вида, но и от априорных значений вероятности сигналов, соответствующих состоянию a_i (u=1,2) объекта. Графики $P(x_i)\, w(x|a_i)$ с учетом «веса» приведены на рис. 5.21.

Как и в предыдущем случае, граница между признаками сигналов x_1 и x_2 проводится через точку пересечения графиков. В этом случае реализуется *критерий минимума средней ошибки, или критерий идеального наблюдателя Котельникова*. Заметим, что вместе с тем реализуется и критерий максимальной апостериорной вероятности.

В тех случаях, когда кроме известной заранее вероятности сигналов x_i и x_2 известны и *потери*, испытуемые потребителем от ошибочных решений, граница различения признаков проводится с учетом относительных потерь r_{12} и r_{21} (рис. 5.22).

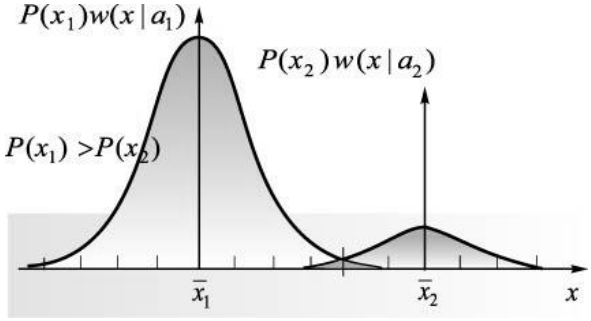


Рис. 5.21. Графики распределений плотности условной вероятности с учетом «веса»

В этом случае реализуется *критерий минимального риска*. Усредненное значение риска

$$r = P(x_i)\alpha r_{12} + P(x_2)\beta r_{21}.$$

Этот критерий целесообразно использовать при разработке таких информационных систем, в которых ошибки 1- го и 2-го рода приводят к *различным потерям*, но обе ошибки не создают опасных ситуаций.

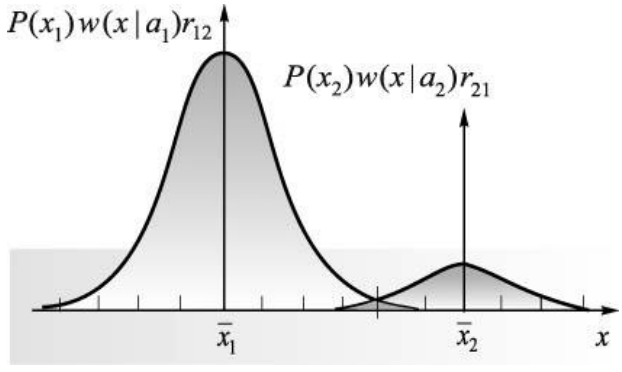


Рис. 5.22. Графики распределений плотности условной вероятности с учетом «веса» и относительных потерь

Выявление ***сигнала методом одноразового отсчета***. Все рассмотренные критерии качества приводят, в сущности, к одному правилу принятия решения. Оно состоит в определении отношения правдоподобия Λ и сравнение его с пороговым значением $\Lambda_{\text{п}}$, зависимым от применяемого критерия.

Задачу обработки реализации $y(t)$ для отыскания отношения правдопо-

добия Λ можно решить, если априорно известны хотя бы некоторые данные о полезном сигнале $s(t)$, характеристиках вероятности помехи $n(t)$ и характере взаимосвязи между полезным сигналом и помехой.

Рассмотрим простой случай выявления по методу одноразового отсчета.

Метод одноразового отсчета - это процесс обработки одного отсчета $y(t_i) = y_i$ реализации входного сигнала, взятого в некоторый момент времени t_i. На базе отсчета $y(t_i) = y_i$ выносится решение о наличии или отсутствии полезной составляющей в принятом информационном сообщении.

Мгновенное значение реализации входного сигнала как суммы полезного сигнала и помехи

$$y_i = s_i + n_i . \tag{5.44}$$

При отсутствии полезного сигнала $s_i = 0$ и $n_i \neq 0$. Тогда

$$P(y/0) = P(y_i/0) = P(n_i) = w(n_i)dx_i = w(y_i)dy_i \tag{5.45}$$

где $w(n_i)$ - одномерная плотность вероятности помехи.

Вероятность $P(y/s)$ получения реализации сигнала с помехой совпадает с вероятностью получения случайной величины $(y_i - s_i)$, которая равна n_i. Поэтому

$$P(y/s) = P(y_i/s_i) = P(y_i - s_i) = w_i(y_i, s_i)dy_i \tag{5.46}$$

$$\Lambda = P(y/s) = P(y/0) = w_i(y_i, s_i)/w(y_i). \tag{5.47}$$

Помеху можно считать стационарным нормальным случайным процессом с нулевым средним и дисперсией σ^2. Поэтому

$$w_1 = w(y_i, s_i) = \exp\{-(y_i - s_i)^2/(2\sigma^2)\};$$

$$w = w(y_i) = \exp\{-(y_i)^2/(2\sigma^2)\}/\sqrt{2\pi\sigma^2};$$

$$\Lambda = \exp\{(s_i/\sigma^2)/(y_i - 0{,}5s_i)\}. \tag{5.48}$$

Из выражения (5.48) следует, что при известных s_i и σ^2 отношение правдоподобия Λ и отсчет y_i реализации связаны между собой взаимно однозначно. Каждому отсчету y_i соответствует полностью определенное значение Λ. Поэтому достаточную сравнивать отсчеты y_i с некоторым порогом, который получаем из выражения (5.48) при $\Lambda = \Lambda_п$:

$$y_п = (\sigma^2 \ln \Lambda_п + 0{,}5s_i^2)/s_i \tag{5.49}$$

При $y_i > y_{\text{п}}$ принимается решение «да», при $y_i < y_{\text{п}}$ - решение «нет».

Основные недостатки метода, вносящие неопределенность в решение задачи:

неопределенность частоты последовательности отсчетов. При редко встречающихся отсчетах сигнал может быть пропущен;

определение значения s_i в момент отсчета. Непосредственное измерение мгновенного значения s_i по полученному мгновенному значению y_i невозможно из-за наличия случайной величины n_i.

Во избежание первого недостатка, т.е. чтобы сделать невозможным пропуск сигнала, отсчеты должны быть непрерывными. В таком случае сигнал $y(t)$ должен поступать на решающее устройство беспрерывно и сравниваться с порогом $y_{\text{п}}$ (рис. 5.23). Если $s_i = a$ и $\Lambda_{\text{п}} = 1$, то $y_{\text{п}} = 0{,}5\,a$.

Вероятность ошибочной тревоги при этом

$$P_{\text{п.т}} = P\left[(y_i > y_{\text{п}})/0\right] = \int_{y_{\text{п}}}^{\infty} w(y)dy, \tag{5.50}$$

где $y(t)$ не содержит полезного сигнала.

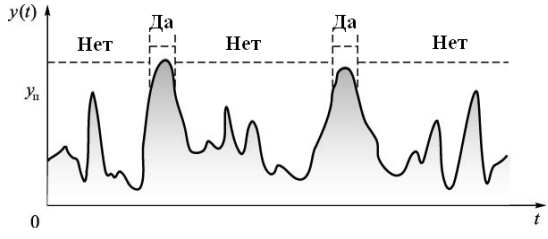


Рис. 5. 23. Схема обработки сигналов на решающем устройстве

Вероятность пропуска сигнала

$$P_{\text{прп}} = P\left[(y_i < y_{\text{п}})/s\right] = \int_{-\infty}^{y_n} w_1(y)dy\,, \tag{5.51}$$

где $y(t)$ содержит полезный сигнал.

Корреляционный метод выявления сигналов на фоне помех. Принимая решение на основании не единственного значения рассматриваемой величины, а на основании большого количества N ее значений, можно достичь более ощутимого эффекта, если различные отсчеты взаимно независимы. Для выполнения этого условия отсчеты должны отличаться друг от друга не менее чем на $\Delta\tau$ - интервал корреляции помехи. Для помехи типа белого шума $\Delta\tau \to 0$. В этом случае

$$\Lambda = \prod_{=1}^{N} w_i(y_i, s_i) / \prod_{i=1}^{N} y_i \ . \tag{5.52}$$

Для нормальных случайных процессов с нулевым средним и дисперсией σ^2 отношения правдоподобия запишется в виде

$$\Lambda = \exp\left\{\sum_{i=1}^{N}[-(y_i - s_i)^2 / 2\sigma^2]\right\} / \exp\left\{\sum_{i=1}^{N}(-(y_i^2 / 2\sigma^2)\right\}. \tag{5.53}$$

В результате преобразования формулы (5.53) получим

$$\ln\Lambda = (\sum_{i=1}^{N} s_i y_i - 0,5\sum_{i=1}^{N} s_i^2) / \sigma^2 \ . \tag{5.54}$$

Правило обработки удобнее представить в виде

$$(a_y)_N = \sum_{i=1}^{N} s_i y_i = \sigma^2 \ln\Lambda + 0,5\sum_{i=1}^{N} s_i^2, \tag{5.55}$$

откуда следует, что при обработке необходимо определить $(\sum_{i=1}^{N} s_i y_i)$ и сравнить найденное значение с порогом

$$(a_y)_N^{\text{п}} = \sigma^2 \ln\Lambda_{\text{п}} + 0,5\sum_{i=1}^{N} s_i^2, \tag{5.56}$$

который находим из формулы (5.54) при $\Lambda = \Lambda_{\text{п}}$.

При $(a_y)_N > (a_y)_N^{\text{п}}$ выносится решение «да». Если выборка взята на интервале $[0, t]$, а отсчеты в ней берутся через $\Delta\tau \to 0$, то суммы в формуле (5.56) превращаются в интегралы, а величина $\sigma^2\Delta\tau$ - в спектральную плотность мощности. Поэтому получаем

$$\int_0^{t_0} s(t)y(t)dt = \ln\Lambda + 0,5\int_0^{t_0} s^2(t)dt. \tag{5.57}$$

Процедура принятия решения согласно формуле (5.57) заключается в перемножении реализации $y(t)$ и ожидаемого сигнала $s(t)$, интегрировании полученного произведения в пределах от нуля до t_0 и сравнении результата с порогом.

Функциональная схема устройства для выявления сигналов, работающего на основе корреляционного метода, изображена на рис. 5.24.

Интеграл $\int_0^{t_0} s(t)y(t)dt$ является мерой взаимной корреляции между реализацией $y(t)$ и полезным сигналом $s(t)$. Поэтому его называют *корреляционным интегралом*, а описанную только что процедуру выявления - *корреляционным методом*.

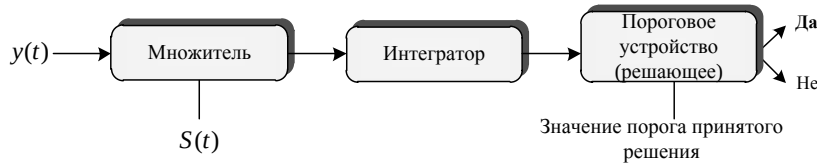


Рис. 5.24. Функциональная схема устройства для выявления сигнала корреляционным методом

Условные вероятности $P_{\text{п.т}}$ и $P_{\text{прп}}$ определяются по той же методике, что и для одноразового отсчета с предварительным уточнением закона распределения.

Синтез приемников непрерывных сигналов. ***Согласованный прием***. *Эта задача принадлежит к классу задач различения сигналов на фоне помех* (*воспроизведение сообщений*), *или задач фильтрации. Эту задачу впервые сформулировал и решил Н. Винер в 1941 г.*

Пусть на вход приемника (рис. 5.25) поступает входной сигнал с помехой $y(t) = s(t) + n(t)$. Приемник характеризуется передаточной функцией $K(i\omega)$, которую необходимо определить. Этапы поиска следующие.

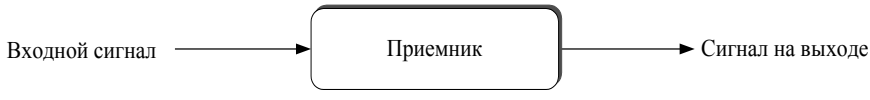


Рис. 5.25. Обработка сигнала в приемнике

Сигнал на выходе приемника

$$\gamma(t) = \int_0^{\infty} y(t-\tau)g(\tau)d\tau, \qquad (5.58)$$

где

$$g(\tau) = 1/2\pi \int_{-\infty}^{\infty} K(i\omega)e^{i\omega}d\omega$$

импульсная переходная функция.

Ошибка в воспроизведении сигнала $\nu(\tau) = \gamma(\tau) - s(t)$.

Задача ***оптимального приема*** - *найти передаточную функцию* $K(i\omega)$ *приемника, которая обеспечивает минимальное значение дисперсии* (*среднего квадрата*) *ошибки*

$$\bar{v}^2 = [\gamma(t) - s(t)]^2. \tag{5.59}$$

После подстановки в формулу (5.59) выражений для известных и искомых величин задача сводится к интегральному уравнению, решить которое можно методами вариационного исчисления. Решение, найденное Н. Винером, имеет вид

$$K(i\omega) = S(\omega) \,/\, (S(\omega) + N(\omega)), \tag{5.60}$$

где $S(\omega)$, $N(\omega)$ - энергетические спектры сигнала и помехи.

Такая постановка задачи нуждается в уточнении критерия оптимальности относительно передаточной функции приемника.

Передаточная характеристика приемника должна обеспечить максимум отношения сигнала к шуму на выходе системы. Условие оптимальности в таком смысле обеспечивает фильтр (*приемник*) *с передаточной функцией*

$$K(i\omega) = aS^*(i\omega)e^{-i\omega t_0}, \tag{5.61}$$

где a, t_0 - *постоянные значения*; $S^*(i\omega)$ - *спектр, связанный со спектром сигнала.*

Согласованной фильтрацией называется *процесс оптимизации передаточной характеристики фильтра и ее согласование со спектром входного сигнала с целью обеспечения максимума отношения полезного сигнала к шуму на выходе системы.*

На выходе приемника отношения сигнала к шуму

$$h_{\max} = \sqrt{2E / N_0},$$

где $E = 1/2\pi \int_{-\infty}^{\infty} S^2(\omega)d\omega$ - энергия сигнала на входе приемника; N_0 - энергетический спектр помех.

Заметим, что сигнал на выходе согласованного фильтра имеет такую же форму, как автокорреляционная функция входного сигнала, поскольку

$$s_{\text{вых}} = \int_{-\infty}^{\infty} s_{\text{вх}}(t-x)g(x)dx,$$

а для согласованного фильтра

$$g(t) = A_0 s_{\text{вх}}(t - t_0).$$

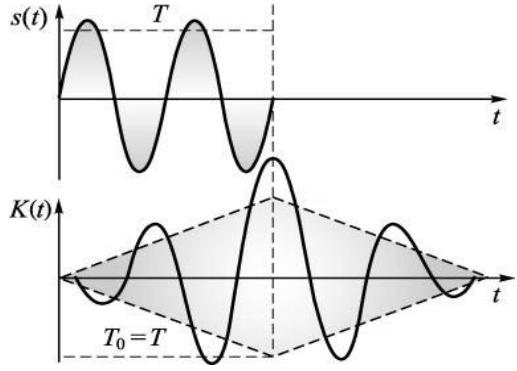


Рис. 5.26. Корреляционная функция отрезка гармонического колебания

Тогда

$$s_{\text{вых}} = A_0 \int_{-\infty}^{\infty} s_{\text{вх}}(y) s_{\text{вх}}(y-\tau) dy = A_0 K_s(\tau),$$

где $\tau = t - t_0$.

Это означает, что приемник можно строить по правилам вычисления взаимно корреляционной функции входного сигнала с ожидаемым. Корреляционная функция $K = f(t)$ отрезка гармонического колебания изображена на рис. 5.26.

Идеальный приемник Котельникова. Передача данных может осуществляться с использованием амплитудной, частотной или фазовой модуляции любых сообщений дискретного или непрерывного характера (в последнем случае - после соответствующей дискретизации согласно теореме Котельникова).

В зависимости от используемых при приеме параметров сигналов различают когерентный и некогерентный прием. В первом случае используют частоту и фазу сигнала, во втором - лишь частоту, а фазу не учитывают.

Методы передачи и приема сравнивают по их помехоустойчивости. Для ее оценки используется разработанная в 1946 г. В. А. Котельниковым теория потенциальной помехоустойчивости, согласно которой для уменьшения влияния флуктуационных помех существует наилучший (идеальный) приемник, имеющий наибольшую (потенциальную) помехоустойчивость для данного метода передачи.

Рассмотрим подходы, лежащие в основе построения идеальных приемников при когерентном приеме. Если сигналам 1 и 0 на входе когерентного приемника соответствуют сигналы $A(t)$ и $B(t)$ одинаковой продолжительности t_c, а суммой сигнала и флуктуационной помехи является $x(t)$, то согласно теории Котельникова приемник обеспечит наименьшую вероятность искажения (искажение символа), если будет выдавать сигнал A при выполне-

нии условия

$$I(A) < I(B), \tag{5.62}$$

в противоположном случае - сигнал B.

Здесь

$$I(B) = \int_0^{t_c} \left[x(t) - B(t)\right]^2 dt; \quad I(A) = \int_0^{t_c} \left[x(t) - A(t)\right]^2 dt,$$

где t_c - продолжительность двоичного сигнала 0 или 1.

Оптимальный приемник по критерию идеального наблюдателя (*идеальный приемник Котельникова*) - *это техническое средство, которое реализует алгоритм обработки входного информационного сообщения на основе сравнения принятого сигнала с помехами с неискаженными* (*идеальными*) *образцами сигнала, вычисляет энергию разности и относит принятый сигнал к тому образцу, для которого энергия разности минимальна.*

Воспользовавшись выражением (5.62), можно построить идеальный приемник Котельникова (рис. 5.27).

Схема приемника Котельникова состоит из генераторов опорных сигналов, точно повторяющих переданные сигналы *A* и *B*; вычитающего устройства; квадратирующего устройства; интегратора *I*; решающего устройства. Для генерирования опорных сигналов в приемнике должны быть известны все параметры переданных сигналов *A* и *B*. Приемник имеет две ветки, в каждой из которых вычисляется среднее квадратичное значение отклонения принятого колебания $x(t)$ от известного сигнала: в первой ветке - от сигнала $A(t)$, во второй - от сигнала $B(t)$. Решение принимается по той ветке, где это отличие меньше.

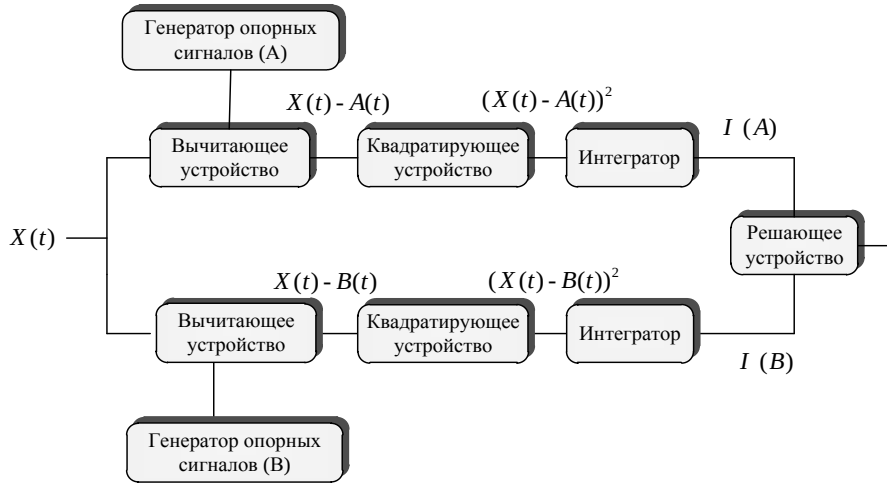


Рис. 5.27. Структурная схема идеального приемника Котельникова

Оптимальные приемники с активной и пассивной паузами. Обозначим энергию сигналов A и B (энергия на резисторе составляет 1 Ом) соответственно через E_A и E_B:

$$E_A = P_A,\ t_c = \int_0^{t_c} \left[A(t)\right]^2 dt;$$
$$E_B = P_B,\ t_c = \int_0^{t_c} \left[B(t)\right]^2 dt, \tag{5.63}$$

где P_A и P_B - соответственно удельная средняя мощность сигналов A и B (мощность, развиваемая на единичном резисторе).

Сигналом с пассивной паузой *называется сигнал, излучаемый только при передаче одного из символов* (*например, 1*), *тогда как передаче другого символа* (*например, 0*) *соответствуют паузы между сигналами.* В этом случае

$$A(t) \neq 0,\ B(t) = 0,$$
$$E_A = E,\ E_B = 0. \tag{5.64}$$

Подставив в формулу (5.62) выражения (5.63) и (5.64), получим выражение, описывающее ***алгоритм работы оптимального приемника*** *сигналов с пассивной паузой:*

$$\int_0^{t_c} x(t)A(t)dt \geq E/2 \text{ - принят сигнал А;} \tag{5.65}$$
$$\int_0^{t_c} x(t)A(t)dt \leq E/2 \text{ - принят сигнал В.}$$

Левая часть этих выражений характеризует взаимную корреляцию между $x(t)$ и $A(t)$ (корреляционный интеграл). Согласно правилу (5.65) приемник должен вычислить значение корреляционного интеграла и сравнить его с некоторым фиксированным значением - порогом. Таким образом, оптимальный приемник сигналов с пассивной паузой - это корреляционный приемник (рис. 5.28), состоящий из коррелятора и решающего устройства.

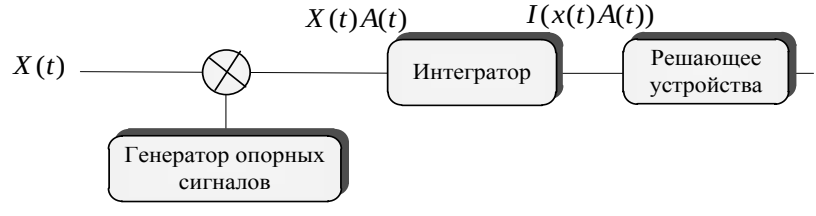


Рис. 5.28. Структурная схема оптимального корреляционного приемника с пассивной паузой

Коррелятор содержит в своей схеме множитель, блок принятия решения, интегратор и устройство снятия отсчетов (на схеме не показан) в моменты времени, кратные продолжительности сигналов $(t_k = kt_c,\ k = 1, 2, ...)$. Рассмотренный приемник называют еще *когерентным приемником с пассивной паузой.*

Если корреляционный интеграл вычислять с помощью линейного фильтра, согласованного с сигналом $A(t)$, то придем к приемнику (рис. 5.29) с оптимальным фильтром.

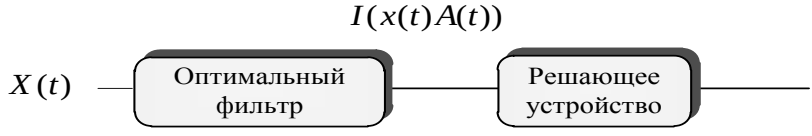


Рис. 5.29. Структурная схема с оптимальным фильтром приемника с пассивной паузой

Сигналом ***с активной паузой*** *называют сигнал, излучаемый при передаче любого символа* (0 *или* 1). *Этот сигнал должен иметь двоичную структуру для обретения двух различных значений. Чаще всего используются сигналы с одинаковыми энергиями:* $E_A = E_B = E$. Преобразуем формулу (5.62) к виду

$$2\int_0^{t_c} x(t)A(t)dt - \int_0^{t_c}\left[A(t)\right]^2 dt > 2\int_0^{t_c} x(t)B(t)dt - \int_0^{t_c} x(t)B(t)dt - \int_0^{t_c}\left[B(t)\right]^2 dt.$$

Отсюда, учитывая формулы (5.63) и (5.64), а также то, что $\int_0^{t_c}\left[A(t)\right]^2 dt = \int_0^{t_c}\left[B(t)\right]^2 dt$, получим два варианта записи.

Первый вариант:

$$\int_0^{t_c} x(t)A(t)dt - \int_0^{t_c} x(t)B(t)dt > 0 \text{ - принят сигнал } A,$$

$$\int_0^{t_c} x(t)A(t)dt - \int_0^{t_c} x(t)B(t)dt < 0 \text{ - принят сигнал } B.$$

Второй вариант:

$$\int_0^{t_c} x(t)\Delta A(t)dt > 0 \text{ - принят сигнал } A, \qquad (5.66)$$

$\int_0^{t_c} x(t)\Delta A(t)dt < 0$ - принят сигнал B, $\Delta A(t) = A(t) - B(t)$.

Полученные выражения определяют различные алгоритмы работы оптимального приемника с активной паузой.

5.5. Многоканальные сети передачи данных. Разделение информационных каналов

Разделением каналов *называется процесс размежевания информационных трактов* (*сетей*) *на фоне информационных параметров переданных сигналов, имеющий целью соединить каждый определенный источник сообщений с его приемником.*

Введем обозначения: $U_k(t)$ - сигналы датчиков, которые отображают информационные функции $x_k(t)$ и порождают модуляцию параметров a_k носителя; $u_{xk}(t)$ - сигналы на выходах передающих устройств отдельных каналов; $u(t)$ - суммарный сигнал в линии связи.

Сигнал в одном канале описывается выражением

$$u_x(t) = g[a_1,...,a_i(t),...,a_j,...,a_n],$$

где параметр $a_i(t)$ передает информацию, а параметр a_j (или несколько таких параметров) можно использовать для характеристики индивидуальных каналов. Такими параметрами могут быть: принадлежность к конкретной электрической цепи; частота или фаза носителя; положение на временной оси; форма и т.д. Каждому каналу *k* соответствует определенное значение a_{jk} или область значений Δa_{jk} параметра a_j. В случае использования этого обозначения выражение для сигнала *k*-го канала приобретет вид

$$u_{xk}(t) = g[a_1,...,a_i(t),...,a_{jk}(t),...,a_n],$$

или сокращенно: $u_{xk}(t) = g_k[a_k(t)]$.

В линию связи поступает составной сигнал, обычно представляющий собой сумму сигналов отдельных каналов

$$u(t) = \sum_k g_k[a_k(t)].$$

Процесс разделения можно рассматривать как фильтрацию, осуществляющую выделение u_{xk} или сигнала датчика U_k:

$$\Phi_k\left\{\sum_k g_k[a_k(t)]\right\} = g_k[a_k(t)].$$

В зависимости от вида оператора фильтрации Ф различают такие методы разделения каналов: пространственный (схемный), дифференциальный, частотный, временной, фазовый, кодовый, по уровню, по форме, корреляционный (рис. 5.30).

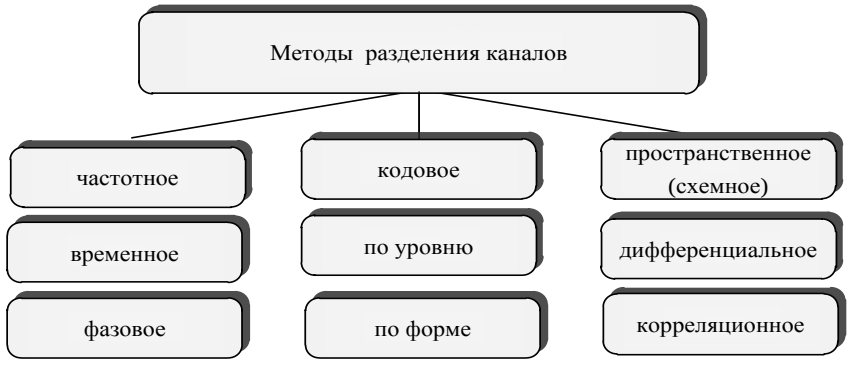


Рис. 5.30. Классификация распределения каналов передачи данных

Многоканальные системы связи с пространственным разделением. Это простейший вид разделения, при котором каждому каналу выделяется индивидуальная сеть связи (рис. 5.31). Примером является пространственное разделение сигналов с помощью матричных коммутаторов с n входами и n выходами.

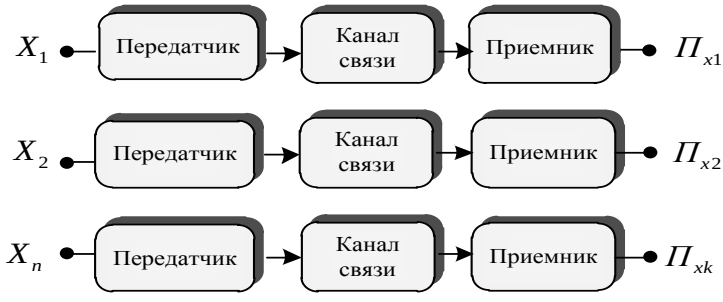


Рис. 5.31. Многоканальная система с пространственным разделением
X_k - датчик k-го канала; $П_{xk}$ - приемник информации k-го канала;

Другие формы разделения каналов допускают передачу сообщений по одной сети. Именно поэтому многоканальную передачу называют также уплотнением каналов.

Многоканальные системы связи с дифференциальным разделением. На рис. 5.32 приведена распространенная схема использования коммутатив-

ного (проводникового) канала для передачи информационных сигналов (от датчика X к приемнику). В линию на передающей и приемной стороне включаются дифференциальные трансформаторы.

Средние их точки соединяют с X и $П_x$. Таким образом, телекоммуникационные сигналы не создают помех в первичных цепях дифференциальных трансформаторов, связанных с телефонными аппаратами A. Благодаря дифференциальному включению телефонные сигналы также не создают помех.

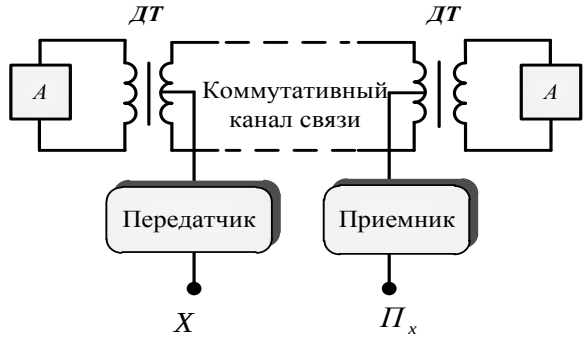
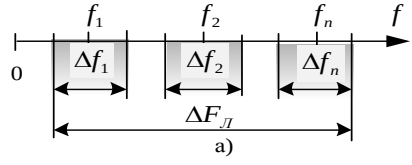


Рис. 5.32. Многоканальная система с дифференциальным разделением:
A – телефонный аппарат; $ДТ$ - дифференциальный трансформатор;
X – датчик сигналов; $П_x$ - приемник.

Многоканальные системы связи с частотным разделением. При частотном разделении для различных каналов на частотной шкале f отводятся непересекающиеся участки Δf_1, Δf_2, ..., Δf_n (рабочая ширина спектра соответствующего канала Δf_k) (рис. 5.33, *а*).



б)

Рис. 5.33. Многоканальная система с частотным разделением:
а - разделение каналов по шкале частот; *б* - схема частотного разделения

Спектры сигналов u_{kx} соответствующих каналов должны вкладываться в границы Δf_k. Полоса пропускания сети связи $\Delta F_{\text{пр}}$ определяется крайними частотами (минимальной частотой интервала Δf_1 и максимальной частотой интервала Δf_n). На рис. 5.33, *б* приведена схема многоканальной системы с частотным разделением. Низкочастотные сигналы U_k датчиков X модулируют по амплитуде или частоте высокочастотные сигналы с несущими $f_1, f_2, ..., f_n$, вырабатываемыми специальными генераторами. Сигналы на выходе модуляторов имеют спектры Δf_k, положения которых на шкале частот определяется несущими частотами f_k, а ширина зависит от ширины спектра сигналов датчиков. Полосовые фильтры передающей части служат для ограничения полосы частот своих каналов. На приемной стороне фильтры разделяют сигналы, которые, пройдя через демодуляторы, могут быть восприняты приемными устройст-вами.Важным преимуществом систем с частотным разделением является возможность одновременной передачи сигналов, относящихся к разным каналам. Еще одно их преимущество заключается в возможности передачи сигналов от рассредоточенных объектов. Недостаток таких систем - сравнительно большое взаимное влияние каналов из-за перекрытия спектров сигналов, неидеальность полосовых фильтров и появление паразитных частотных составляющих вследствие нелинейности электрических цепей (так называемая перекрестная модуляция).

Многоканальные системы связи с временным разделением. В случае временного разделения сигналы u_{xk} датчиков передаются только в отведенные для них непересекающиеся отрезки времени Δt_k (рис. 5.34, *а*).

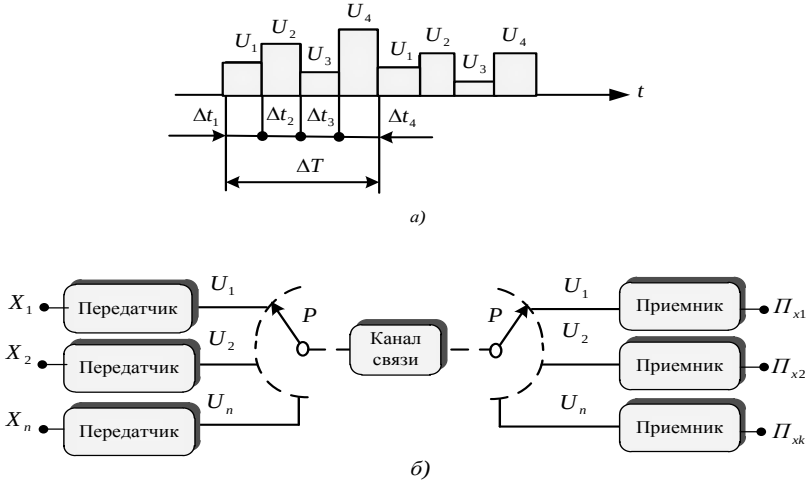


Рис. 5.34. Многоканальная система с временным разделением:
а – разделение каналов по времени; *б* – схема временного разделения

Разделение осуществляется распределителями P (см. рис. 5.34, *б)*, которые должными быть строго синхронизированными (работать с одинаковой скоростью) и синфазными (работать без сдвига). Взаимное влияние каналов при временном разделении обычно незначительно, что дает возможность строить системы с большим количеством каналов. Благодаря этому, а также простоте технических средств указанный метод используется довольно широко.

Многоканальные системы связи с фазовым разделением. Фазовое разделение применяют в двухканальной системе (рис. 5.35) с синусоидальными сигналами, фазы которых отличаются на 90°. Сигналы датчиков X_k модулируют амплитуду синусоидальных носителей, отличающихся фазой. Таким образом, сигналы u_{xk} на выходе модуляторов имеют амплитуды, обусловленные модулирующими функциями датчиков, и фазы соответственно φ_1 и $\varphi_2 = \varphi_1 + \pi/2$: $u_{x1} = U_1 \sin\omega_0 t$; $u_{x2} = U_2 \sin(\omega_0 t + \pi/2) = U_2 \cos\omega_0 t$.

Фазовые детекторы выделяют соответствующие модулирующие функции U_1 и U_2.

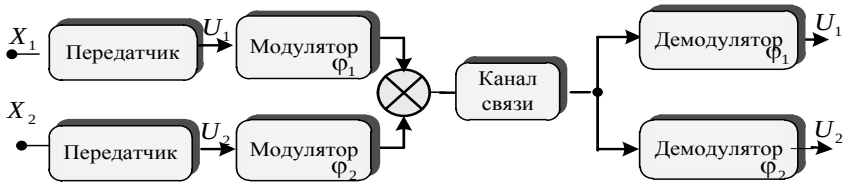


Рис. 5.35. Многоканальная система с фазовым разделением

Многоканальные системы связи с кодовым разделением. В случае кодового разделения адрес канала указывается кодированным сигналом, посылаемым в сеть связи. Разделение на приемной стороне осуществляется декодирующим устройством, направляющим сообщение по выбранному каналу. Код адреса может быть как последовательным, так и параллельным. В последнем случае используется отдельная сеть связи или индивидуальный частотный канал на каждый разряд кода. Кодовое разделение каналов дает возможность опрашивать каналы в произвольном порядке, что делает удобным использование данного разделения в системах передачи данных.

Многоканальные системы связи с разделением по форме. Для разделения сигналов, различающихся формой, используются операции, наиболее чувствительные к изменению формы, - обычное дифференцирование, интегрирование и вычитание. Рассмотрим процедуру разделения, при которой функции носителя получают посредством последовательного дифференцирования.

Пусть, например: $u_{x1}(t) = U_1$; $u_{x2}(t) = U_2 t$. В линию поступает сумма сигналов $u(t) = U_1 + U_2 t$. Процесс разделения имеет целью выделение инфор-

мационных параметров U_1 и U_2. Выделение U_2 осуществляется дифференцированием функции $u(t)$. Интегрирование U_2 восстанавливает переданный сигнал второго канала $u_{x2}(t)$. U_1 получаем вычитанием $u_{x2}(t)$ от $u(t)$.

Многоканальные системы связи с корреляционным разделением. В некоторых случаях сигналы отдельных каналов можно представить в виде

$$u_{xk}(t) = g_k[a_k(t)] = a_k(t)g_k(t) = U_k(t)g_k(t).$$

Здесь функция $g_k(t)$ описывает носитель с некоторым значением разделительного параметра a_{jk}, а информационный параметр $a_k(t)$, модулирующий функцию $g_k(t)$ по амплитуде, равен сигналу $U_k(t)$ соответствующего датчика. Этот параметр является функцией времени, изменяющегося медленнее по сравнению с $g_k(t)$, и его можно считать постоянным. Сигнал в сети является линейной комбинацией функций $u(t) = \sum_k U_k g_k(t)$.

Если функции $g(t)$ линейно независимы, они могут быть разделены линейными фильтрами. Такие многоканальные системы передачи называют *линейными*. К линейным относятся, в частности, системы с частотным, временным, фазовым разделением и разделением по форме. Важной разновидностью линейно независимых сигналов являются ортогональные сигналы, для которых существует общий метод разделения, базирующийся на применении оператора корреляционной фильтрации к поступающему из сети связи сигналу. Для геометрической интерпретации понятия ортогональности рассмотрим две функции $g_i(t)$ и $g_j(t)$, определенные на интервале $T_1 < t < T_2$. Представим их в виде совокупности дискретных отсчетов:

$$g_i^* = \left(g_i(t_1),\ g_i(t_2),\ ...,\ g_i(t_m)\right);$$

$$g_j^* = \left(g_j(t_1),\ g_j(t_2),\ ...,\ g_j(t_m)\right).$$

Эти совокупности можно рассматривать как два вектора в *m-мерном* эвклидовом пространстве с координатами $g_i(t_k)$ и $g_j(t_k)$, $k = 1, 2, ..., m$. Если векторы ортогональны, то их скалярное произведение равно нулю: $(g_i^*,\ g_j^*) = \sum_k g_i(t_k)g_j(t_k) = 0$. При бесконечном увеличении количества отсчетов m дискретные функции превращаются в непрерывные. Их также можно рассматривать как векторы, но уже в бесконечномерном пространстве. Условие ортогональности приобретает интегральный вид и определяется на заданном интервале (T_1, T_2):

$$\left(g_i(t), g_j(t)\right) = \int_{T_1}^{T_2} g_i(t)g_j(t)dt = 0.$$

Ортогональную систему образовывают такие функции:

1. *Бесконечное множество функций* $\cos k\omega t$, $\sin k\omega t$, ортогональных на интервале $0 < t < 2\pi / \omega$, где k - неотрицательное число (рис. 5.36, *а*). Носитель с такими функциями используется в устройствах с частотным и фазовым разделением.

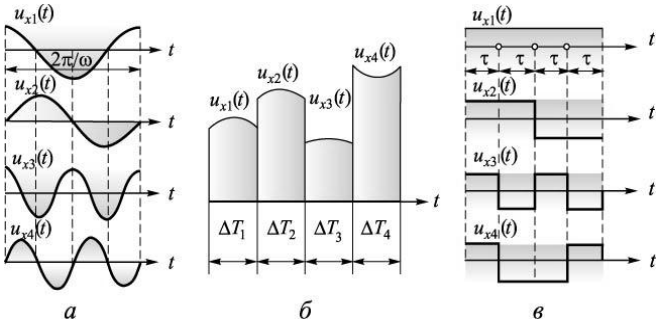


Рис. 5.36. Примеры ортогональных сигналов: *а* – на гармоническом носителе;
б – на каком-либо носителе с распределением по времени;
в – на импульсном носителе

2. *Множество произвольных функций*, определенных на непересекающихся интервалах времени, тождественно равных нулю вне этих интервалов (см. рис. 5.36, *б)*, также представляет собой ортогональную систему, поскольку при $i \neq j$ выполняется равенство $\left(g_i(t), g_j(t)\right) = 0.$

3. *Множество, состоящее из дискретных знакопеременных функций*, которые можно получить с помощью *n*-разрядного счетчика в режиме вычитания, находящегося сначала в заполненном состоянии (см. рис. 5.36, *в)*, и которые являются ортогональными на заданном интервале $T = n\tau$ (τ -продолжительность импульса младшего разряда). Коды, соответствующие этим последовательностям, принадлежат к групповым и называются кодами Рида - Мюллера 1-го порядка.

4. *Ортогональными на определенных интервалах* являются большое количество специальных функций: полиномы и функции Лежандра, Чебышева, Якоби, Эрмита, Лагерра, Хаара, Уолша, Радемахера и др.

Шарль Эрмит (Charles Hermite, 1822 - 1901), французский математик, член Парижской АН (1856), иностранный член-корреспондент (1857) и иностранный почетный член (1895) Петербургской АН. С 1848 г. работал в Политехнической школе, с 1869 г. - профессор Парижского университета. Основные работы касаются теории чисел, теории квадратичных форм, теории инвариантов, ортогональных полиномов, эллиптических функций и алгебры. Среди его учеников был Анри Пуанкаре. Эрмит первым показал, что число e (основа натурального логарифма) является трансцендентным.

Ортогональную систему удобно использовать в нормированном виде при выполнении условия

$$\int_{T_1}^{T_2} g_k^2(t)dt = 1.$$

Если $\varphi_k(t)$ - ненормированные ортогональные функции, то операция нормирования выполняется умножением на коэффициент

$$\lambda_k = 1 \Big/ \sqrt{\int_{T_1}^{T_2} \varphi_k^2(t)dt}.$$

В этом случае в сеть поступает сигнал вида $u_{xk}(t) = U_k \lambda_k \varphi_k(t) = U_k g_k(t)$, где $g_k(t)$ уже являются нормированными функциями, образующими ортонормированную систему.

Для выделения информационного параметра U_k нужно умножить принятый сигнал $u(t)$ на функцию $g_k(t)$ и проинтегрировать полученное произведение в пределах $T_1 \le t \le T_2$:

$$\int_{T_1}^{T_2} \left[\sum_i U_i g_i(t) \right] g_k(t)dt = U_k \int_{T_1}^{T_2} g^2{}_k(t)dt = U_k .$$

Умножение сигнала линии на все функции $g_k(t)$ обеспечивает полное разделение любых ортогональных сигналов.

Оператор разделения Ф, выполняющий это преобразование, определяет, в сущности, степень взаимной корреляции сигналов $u(t)$ и $g_k(t)$. Таким образом, многоканальная система (рис. 5.37) на передающей стороне содержит генераторы Γ_k ортогональных функций и модуляторы M_k с нормализаторами, а на приемной - такие же генераторы Γ_k и корреляторы K_k. Эффективность корреляционного метода разделения заключается в предоставлении возможности значительно ослабить влияние перекрестных помех, а это существенно в случае перекрываемых спектров сигналов.

Рассмотрим в качестве примера систему, использующую полиномы Лежандра 1-го рода. Эти полиномы ортогональны на интервале $-1 < t < +1$ и описываются соотношениями:

$$\varphi_0(t) = 1;\ \ldots;\ \varphi_i(t) = t;\ \ldots;\ \varphi_k(t) = \left[(2k-1) \int_{-1}^{t} \varphi_{k-1}(t)dt + \varphi_{k-2}(t) \right],\ k \ge 2.$$

Скалярное произведение двух функций:

$$\int_{-1}^{t} \varphi_i(t)\varphi_j(t)dt = \begin{cases} 0 \text{ при } i \ne j, \\ 2 / (2i+1) \text{ при } i = j. \end{cases}$$

Итак, нормировочный множитель $\lambda_k = \sqrt{(2k+1)/2}$.

Сигналы образуются на основе носителя в виде постоянного события, промодулированного по форме в соответствии с приведенными соотношени-

ями. Таким образом, форма служит параметром разделения, а для перенесения информации используется амплитуда.

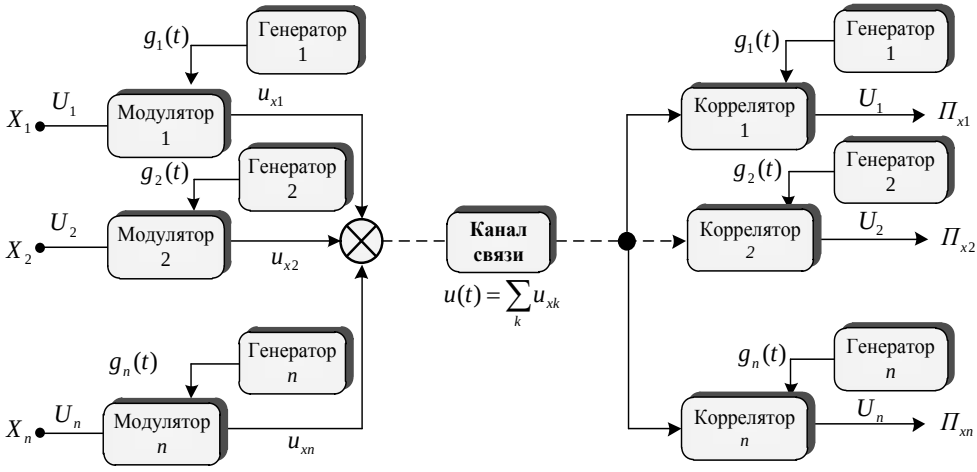


Рис. 5.37. Многоканальная система с корреляционным разделением

Функциональная схема генератора ортогональных функций $\varphi_k(t)$ приведена на рис. 5.38.

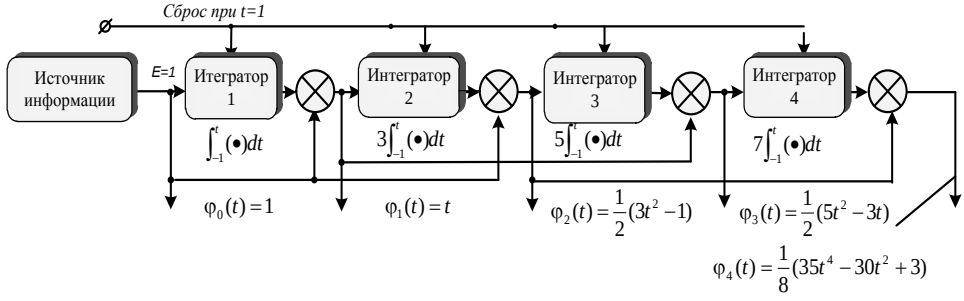


Рис. 5.38. Схема разделения каналов на фоне ортогональных функций

Источник E создает постоянный сигнал $E = 1$, играющий роль функции $\varphi_0(t)$. После прохождения через интегратор I_1 и дальнейшего вычитания $\varphi_0(t)$ от результата формируется функция $\varphi_1(t)$.

Формирование следующих функций осуществляется в соответствии с приведенными рекуррентными соотношениями последовательным интегрированием функций $\varphi_{k-1}(t)$, умножением результата на соответствующий коэффициент $2k-1$ и прибавлением функции $\varphi_{k-2}(t)$. После нормирования и модуляции образуется сигнал

$$u_{xk}(t) = U_k \lambda_k \varphi_k(t) = U_k g_k(t).$$

Полный сигнал в сети связи $u(t) = \sum_k U_k g_k(t)$.

Разделение каналов выполняется в соответствии с общей схемой, приведенной на рис. 5.38. С этой целью на приемной стороне используется аналогичный генератор полиномов, синхронизированный с первым, и коррелятор, выделяющий информационные параметры U_k.

Многоканальные системы связи с комбинированными методами разделения. Одновременное использование нескольких методов разделения предоставляет возможность увеличить количество каналов и уменьшить их взаимное влияние. Так, разделение по форме совместно с частотным или временным разделением удваивает общее количество каналов. Перекрестные искажения ограничивают емкость системы с частотным разделением каналов. Кроме этого, аппаратура с частотным разделением при большом количестве каналов усложняется из-за наличия раздельных фильтров. Комбинирование частотного и временного методов разделения устраняет указанные недостатки. При этом временное разделение осуществляется коммутацией одного или нескольких частотных каналов. Применяются также комбинации кодовых и частотных, кодовых и временных методов разделения и т.д.

5.6. Помехоустойчивость систем передачи информации

Критерии оценивания помехоустойчивости систем передачи. Основу методов повышения помехоустойчивости составляют различные способы введения избыточности. Повышение функциональной надежности или помехоустойчивости систем передачи информации достигают увеличением временной, частотной, энергетической избыточности и т.п.

Помехоустойчивость системы - *способность системы противостоять: влиянию случайных* (*или намеренных*) *помех на полезный сигнал; уменьшению* (*или невозможности*) *достоверной передачи данных от источника сообщения к потребителю.*

Поскольку в результате действия помехи принятый сигнал всегда отличается от переданного, то помехоустойчивость характеризуется степенью соответствия принятого сообщения переданному.

При сравнении различных информационных систем более помехоустойчивой будет система, которая сможет обеспечить меньшее отличие между принятым и переданным сообщениями при данном уровне помех.
Количественная мера этого соответствия выбирается по-разному в зависимости от характера сообщения.

При оценивании помехоустойчивости систем передачи непрерывных сообщений в каналах с флуктуационной помехой типа *белого шума* используется отношение h^2 средних мощностей полезного сигнала и помехи. При сравнении разных способов приема используют понятие *преимущества системы,*

где $h^2_{вх}$ и $h^2_{вых}$ - отношение средних мощностей сигнала и помехи на входе и выходе устройства.

В случае передачи дискретных сообщений в качестве критерия помехоустойчивости (надежности или правильности) используют вероятность правильного приема символа, где P_0 - вероятность ошибки при приеме символа. Значение ошибки P_0 зависит от h^2 в непрерывном канале, по которому проходят сигналы, соответствующие переданным символам.

При оценивании способов передачи оценивают помехоустойчивость различных видов модуляции в сравнении с амплитудной (см. гл. 4). Введенную избыточность проще реализовать в передатчике: повысить мощность сигнала, применить, например, более высокочастотный носитель, в конце концов, использовать корректирующее кодирование, т.е. во всех случаях увеличить объем переданного сигнала.

Этап передачи в принципе изменить сложнее. На входе приемника появляются помехи, возникшие в сети связи, исключить которые нельзя, но можно каким-либо образом уменьшить.

Любой дискретный приемник можно представить в виде структурной схемы (рис. 5.39).

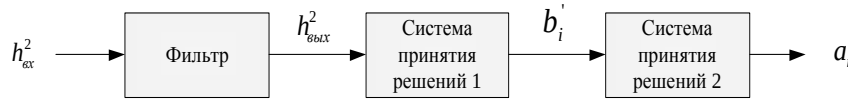


Рис. 5.39. Обобщенная структурная схема дискретного приемника

При этом выполняются такие операции:

фильтрация или «очищение» сигнала от помехи до максимально возможного значения $B = h^2_{вых} / h^2_{вх}$;

обеспечение правильного выбора символа сообщения - первая система принятия решений;

решение задачи об отождествлении сигнала, являющегося комбинацией символов, с переданным полезным сообщением - вторая система принятия решений.

Поэлементным* (*последовательным*) *приемом *называется такой прием, при котором окончательный вывод о принятом сообщении делается путем анализа каждого элемента сигнала.*

Приемом в целом *называется такой прием, при котором принятый сигнал можно не разбивать на элементы, а вывод о принятом сообщении делать, анализируя всю последовательность символов.*

В ряде случаев, когда переданное сообщение имеет избыточность, прием в целом обеспечивает *большую надежность вывода*, чем поэлементный прием. В некотором понимании такие устройства могут быть даже проще традици-

Оливер Хевисайд (Oliver Heaviside, 1850 - 1925), английский ученый-самоучка, инженер, математик и физик. Создал теорию передачи сигналов на далекие расстояния. Впервые применил комплексные числа для изучения электрических цепей, разработал технику применения преобразования Лапласа для решения дифференциальных уравнений, сформулировал уравнение Максвелла в терминах электрической и магнитной сил и потока, а также независимо от других математиков создал векторный анализ. Впервые разработал операционное исчисление, которое со временем широко применяется в физике и других науках.

онных, поскольку в них нет первой системы для принятия решения относительно символов сообщения.

Потенциальная помехоустойчивость при приеме дискретных сигналов. Если два сигнала отличаются вследствие действия помехи, то кроме правильных решений возможны и ошибочные.

Найдем вероятность искажения $P_{\text{пом}}$ с помощью идеального приемника Котельникова (см. рис. 5.27). Вероятность искажения - вероятность ошибочного решения приемника - равна вероятности невыполнения условия реализации оптимального приемника по формуле (5.62).

Искажения возникают из-за того, что в принятый сигнал входит случайная составляющая - флуктуационная помеха.

В большинстве случаев помеху считают стационарным эргодическим процессом с нормальным законом распределения мгновенных значений напряжений

$$w(U_{\text{п}}) = \exp\left[-(U_{\text{п}} - \bar{U}_{\text{п}})^2 / (2\sigma_{\text{п}}{}^2)\right] / \left[\sigma_{\text{п}}{}^2\sqrt{2\pi}\right],$$

где $U_{\text{п}}$ - мгновенное значение напряжения помехи; $\bar{U}_{\text{п}}$ - среднее значение напряжения помехи (постоянная составляющая); $\sigma_{\text{п}}^2$ - дисперсия (средний квадрат отклонения от постоянной составляющей) помехи; $w(U_{\text{п}}) = \exp\left[-U_{\text{п}}{}^2 / (2\sigma_{\text{п}}{}^2)\right] / \left[\sigma_{\text{п}}{}^2\sqrt{2\pi}\right]$ -

функция плотности распределения вероятностей согласно нормальному закону. Поскольку постоянная составляющая помехи равна нулю, то ошибка происходит при превышении мгновенным значением напряжения помехи некоторого значения x. Вероятность этого события равна площади под кривой распределения от x к ∞ (рис. 5.40):

$$P_{\text{пом}} = P(|U_{\text{п}}| > x) = \int_x^{\infty} w(U_{\text{п}})\, dU_{\text{п}} = -\Phi(x). \quad (5.67)$$

Здесь $\Phi(\alpha) = \left(\int_{-\infty}^{\alpha} \exp[-z^2/2]dz\right) / \left[\sigma_{\text{п}}{}^2\sqrt{2\pi}\right]$ -

функция Лапласа, где α - обобщенная функция, обусловленная отношениям энергий сигнала и помехи, а также степенью расхождения сигналов A и B

$$\alpha = \sqrt{(\int_0^{\alpha} [A(t) - B(t)]^2 dt) / (2N_0)}, \qquad (5.68)$$

$N_0 = U^2_{\text{п.эфф}} / \Delta F$ - спектральная плотность мощности помехи ($U^2_{\text{п.эфф}}$ - интегральная мощность помехи; ΔF - полоса пропускания приемного фильтра).

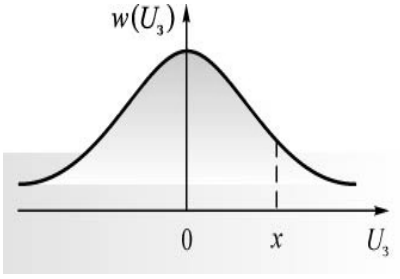


Рис. 5.40. К определению вероятности искажения при оптимальном приеме

Преобразуем выражение (5.68), для чего воспользуемся выражением (6.63), а также понятием удельной разности энергий разностного сигнала

$$\Delta E = \Delta P_{\text{с}} \cdot t_{\text{с}} = \int_0^{t_{\text{с}}} \Delta A(t)^2 dt, \qquad (5.69)$$

где $\Delta P_{\text{с}}$ - удельная средняя мощность разностного сигнала $\Delta A(t)$. Тогда

$$\alpha = \sqrt{\Delta E / (2N_0)}. \qquad (5.70)$$

Подставив это значение в выражение (5.67), получим

$$P_{\text{пом}} = 1 - \Phi\left(\sqrt{\Delta E / (2N_0)}.\right) \qquad (5.71)$$

Вероятность искажения (5.71) определяет потенциальную помехоустойчивость идеального и любого

Эндрю Джеймс Витерби (Andrew James Viterbi, 1935), итальянско-американский электротехник та бизнесмен. Родился в Бергамо (Италия) в еврейской семье, которая эмигрировала вместе с ним в 1939 году в США как беженцы. В 1952 году поступил в Массачусетский технологический институт, где изучал электротехнику и получил степень магистра наук. Получил степень доктора философии в области цифровой связи в университете Южной Калифорнии. Изобрел алгоритм, названный в его честь, который использовал для декодирования свернуто закодированных данных.

другого оптимального приемника. Уменьшить вероятность искажения можно за счет уменьшения спектральной плотности мощности помех N_0.

Потенциальная помехоустойчивость приема информационных сигналов по Котельникову. Помехоустойчивость к действию помехи типа *белого шума* называют *потенциальной* (по Котельникову), если она обеспечивает минимальную среднюю вероятность ошибки различения сигналов. Это достигается при использовании критерия *идеального наблюдателя*, который делает вывод о переданном коде x_2, сравнивая отношения правдоподобности с определенным порогом

$$\frac{w(u_y \mid x_2)}{w(u_y \mid x_1)} > \frac{p(x_1)}{p(x_2)}. \qquad (5.72)$$

Пусть переданные сигналы $u_{x1}(t)$ и $u_{x2}(t)$, соответствующие кодам x_1 и x_2, имеют произвольную форму, а принятый сигнал $u_y(t) = u_x(t) + u_\xi(t)$. Представим сигнал $u_y(t)$ в виде совокупности отсчетов u_{y1}, ..., u_{yn}, разделенных интервалами $\Delta t = 1/2 f_m$.

Распределения $w(u_y \mid x_i)$ в пространстве сигналов u_y соответствует распределение $w(u_\xi) = w(u_y - u_{xi})$ в пространстве сигналов u_ξ. Для гауссовской помехи при отсутствии корреляции между значениями шума в моменты отсчета разностная плотность

$$w(u_y - u_{xi}) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-(u_y - u_{xij})^2 / 2\sigma_\xi^2} = \frac{1}{\left(\sqrt{2\pi}\sigma_\xi\right)^n} e^{-\frac{1}{2\sigma_\xi^2}\sum_{j=1}^{n}(u_y - u_{xij})^2},$$

где $\sigma_\xi = \sqrt{D[U_\xi]}$ - среднее квадратичное отклонение помехи.

Правило для принятия решения:

$$e^{\frac{1}{2\sigma_\xi^2}\left[\sum_{j=1}^{n}(u_y - u_{x1j})^2 - \sum_{j=1}^{n}(u_y - u_{x2j})^2\right]} > p(x_1) \Big/ p(x_2),$$

или после логарифмирования:

$$\frac{1}{2\sigma_\xi^2}\left[\sum_{j=1}^{n}\left(u_y-u_{x1j}\right)^2-\sum_{j=1}^{n}\left(u_y-u_{x2j}\right)^2\right]>\ln p(x_1)-\ln p(x_2). \qquad (5.73)$$

Умножим обе части неравенства на Δt и перейдем к непрерывным во времени функциям, считая $u_x(t)$ и $u_\xi(t)$ на малых интервалах постоянными. Тогда суммы можно заменить на интегралы согласно приближенным равенствам:

$$\begin{aligned}\sum_{j=1}^{n}\left(u_y-u_{x1j}\right)^2\Delta_t &\approx \int_0^x\left[u_y(t)-u_{x1}(t)\right]^2 dt;\\ \sum_{j=1}^{n}\left(u_y-u_{x2j}\right)^2\Delta_t &\approx \int_0^x\left[u_y(t)-u_{x2}(t)\right]^2 dt.\end{aligned} \qquad (5.74)$$

Поскольку $\Delta t=\frac{1}{2f_m}$, то (5.74) приобретает вид

$$\int_0^{T_x}\left[u_y(t)-u_{x2}(t)\right]^2 dt-C_2<\int_0^{T_x}\left[u_y(t)-u_{x1}(t)\right]^2 dt-C_1;$$

$$C_k=\frac{\sigma^2{}_\xi}{f_m}\ln p(x_k),\ k=1,2.$$

Схема, реализующая это правило, получила название *оптимального приемника* (рис. 5.41).

Рассмотрим важный частный случай, когда вероятности $p(x_1)$ и $p(x_2)$ равны между собой, а функции $u_{x1}(t)$ и $u_{x2}(t)$ пронормированы таким образом, что их энергии на интервале одинаковы. Раскрыв в последнем соотношении скобки и учитывая, что $C_1=C_2$, получим

$$\int_0^{T_x}u^2{}_{x2}(t)dt-\int_0^{T_x}u_y(t)u_{x2}(t)dt<\int_0^{T_x}u^2{}_{x1}(t)dt-\int_0^{T_x}u_y(t)u_{x1}(t)dt\ .$$

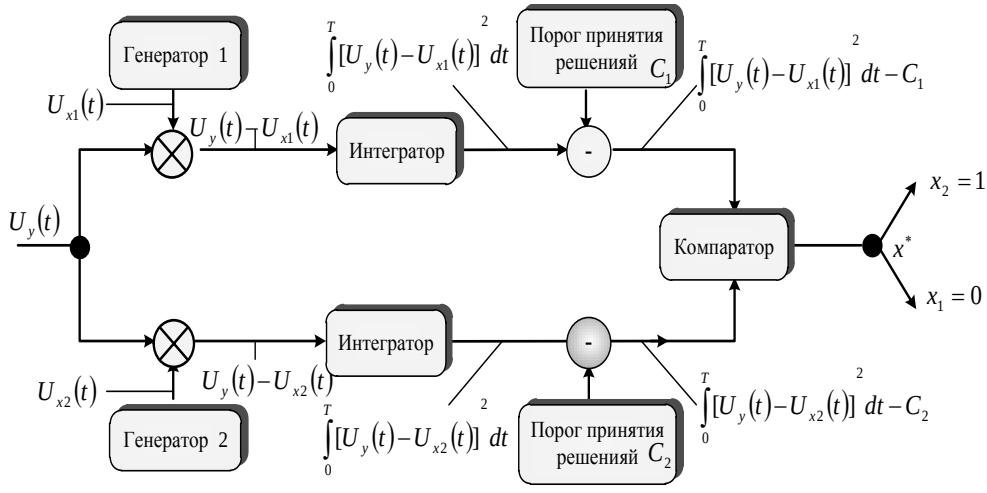


Рис. 5.41. Оптимальный приемник Котельникова с системой принятия решений на основе определенных порогов на фоне помех

Первые слагаемые каждой части неравенства, которые являются энергиями полезных сигналов, взаимно уничтожаются. Правило принятия решения в этом случае превращается в сравнение исходных сигналов взаимнокорреляционных фильтров (рис. 5.42)

$$\int_0^{T_x} u_y(t)u_{x2}(t)dt < \int_0^{T_x} u_y(t)u_{x1}(t)dt\,.$$

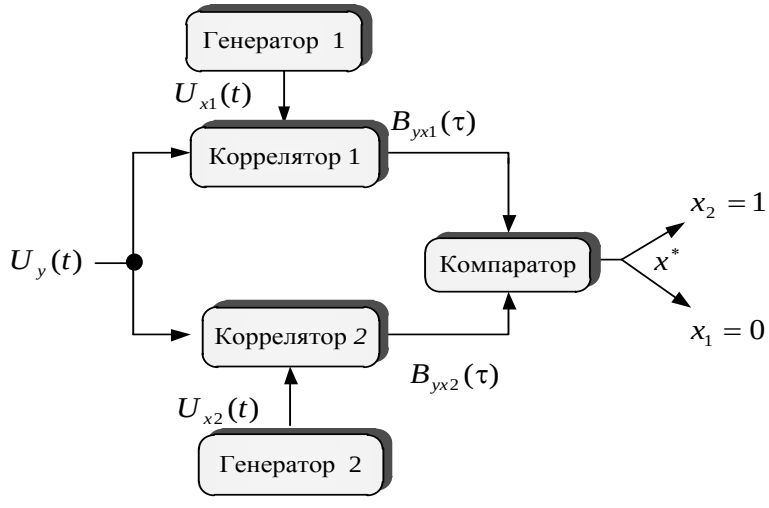


Рис. 5.42. Схема корреляционного приемника с определенным порогом принятия решений на фоне помех

Основные выводы

Любую информационную систему можно подразделить на источник, преобразователь, канал передачи, накопитель и устройство, отображающее информацию, а информационные процессы, происходящие в этих устройствах, представить в общем случае в виде процесса передачи информации по каналу связи.

Если сжатие производится так, что по сжатым данным можно абсолютно точно восстановить начальную информацию, кодирование называется неразрушающим. Неразрушающее кодирование используется при передаче (или хранении) текстовой информации, числовых данных, компьютерных файлов и т.д., т.е. в случаях, где недопустимы даже наименьшие отличия начальных и восстановленных данных.

Кодирование в канале, или помехоустойчивое кодирование, - это способ обработки переданных данных, обеспечивающий уменьшение количества ошибок, которые возникают в процессе передачи по каналу с помехами.

Обобщенные характеристики информационных каналов: время передачи сигнала; ширина его частотного спектра; энергетическая характеристика - средняя мощность.

Дискретный канал в общем виде представляется совокупностью дискретного модулятора на входе, непрерывного канала и дискретного демодулятора на выходе.

Модели дискретных каналов передачи данных подразделятся на такие виды: двоичный симметричный канал без памяти; двоичный несимметричный канал без памяти; симметричный канал без памяти.

Скорость передачи информации сигналами с ограниченной средней мощностью по каналу, в котором действует белый гауссовский шум, оказывается максимальной в случае полного сходства между сигналом и помехой.

Максимальная скорость передачи информации обеспечивается, если в качестве физического носителя информации применять стационарный случайный процесс в виде белого гауссовского шума.

Нет смысла безгранично расширять полосу пропускания канала, по-

скольку с ее расширением возрастание пропускной способности канала замедляется, и у границы при $F_к \to 0$ пропускная способность приближается к постоянной величине.

Задача синтеза элементов информационных систем состоит в определении алгоритма функционирования информационных систем по заданному критерию качества и интерпретировании этого алгоритма с помощью технических средств.

Задача анализа элементов информационных систем заключается в расчете рабочих характеристик и структурных схем информационных систем.

Различают такие основные типы задач статистического синтеза: выявление сигнала на фоне помех; различение сигналов на фоне помех; одновременное выявление (различение) сигналов и оценивание их параметров на фоне помех; выделение сигналов на фоне помех.

Теория потенциальной помехоустойчивости заключается в том, что для уменьшения влияния флуктуационных помех существует наилучший (идеальный) приемник, который имеет наибольшую (потенциальную) помехоустойчивость для данного метода передачи.

Повышение функциональной надежности, или помехоустойчивости, систем передачи информации достигают увеличением временной, частотной, энергетической избыточности.

Вопросы для самоконтроля

1. Какие основные элементы входят в модель системы передачи информации?

2. Что вы понимаете под понятием «канал связи»?

3. Раскройте понятие «объем сигнала» и «емкость канала».

4. Назовите основные характеристики дискретных каналов.

5. Поясните отличие между двоичным симметричным и несимметричным каналами без памяти.

6. Сформулируйте теорему К. Шеннона для дискретного канала с помехами и без помех.

7. Назовите статистические критерии выявления сигналов на фоне помех и укажите, в каких случаях они реализуются.

8. Приведите структурную схему идеального приемника Котельникова.

9. Укажите пути достижения увеличения помехоустойчивости систем передачи информации.

The main conclusions

Any information system can be divided into the source, the converter, the chnnel of transmission, the drive and the device which displays information: and informational processes which occur in these devices can be presented generally in the form of process of transmission of the information on a communication channel.

If compression is conducted in such way that it is possible to restore the initial information absolutely precisely using the compressed data, the coding is called nondestructive. Nondestructive coding is used during transmission (or storage) of the text information, numerical data, computer files and so on, that is where even the slightest differences of initial and restored data is inadmissible .

The coding in the channel or noiseless coding is a way of processing of the transmitted data that provides the decrease of quantity of errors which arise during transmission on a noisy channel.

Generalized characteristics of information channels are the following: the time of transmission of a signal; the width of its frequency spectrum; the energy characteristic - average power.

The discrete channel in a general view is represented as a collection of the discrete modulator on an input, the continuous channel and the digital demodulator on an output

The models of discrete data channels are divided on: the binary symmetric channel without memory; the binary asymmetric channel without memory; the symmetric channel without memory.

The information rate by the signals with the limited average power on the channel, in which white gauss noise operates, appears maximal at complete similarity between a signal and interference;

The maximal information rate will be provided if stationary stochastic process in the form of white gauss noise use as the physical medium.

There is no sense to increase a channel bandwidth enough because according to extension of a bandwidth the growth of capacity of the channel is decelerated and capacity comes nearer to the constant on boundary at $F_{\hat{E}} \to 0$.

The task of synthesis of units of information systems is determining of algorithm of operation of information systems on the set criterion of quality and interpretation of this algorithm by means of techniques.

The task of the analysis of units of information systems is calculation of performance values and the block-structure charts of information systems.

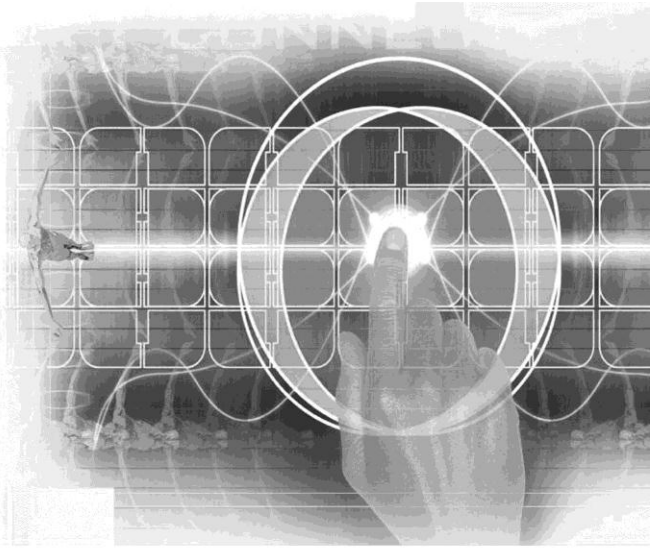
The following main types of tasks of statistical synthesis are distinguished: revealing of signal on a background of interferences; distinguishing of signals on a background of interferences; compatible revealing (distinguishing) of signals and estimation of their parameters on a background of interferences; signals extraction on a background of interferences.

The theory of potential noise immunity means that for decrease of influence of fluctuation interferences there is the best (ideal) receiver which owns the greatest (potential) noise immunity for the given method of transmission.

The rise of functional reliability or noise immunity of systems of transmission of the information is achieved at increase of time, frequency, power redundancy.

Ключевые слова

Русский	**Английский**
канал связи	communication channel
кодер источника	coder of the source
декодер источника	decoder of the source
дискретный канал	discrete channel
непрерывный канал	continuous channel
оптимальный приемник	optimum receiver



СЕТИ ПЕРЕДАЧИ ИНФОРМАЦИИ

6

6.1. Информационно-коммуникационные сети
6.2. Проектирование информационных сетей
6.3. Системы беспроводной передачи информации и защита информационных ресурсов
6.4. Спутниковые каналы
6.5. Множественный доступ к информационным ресурсам

6.1. Информационно-коммуникационные сети

Концепция информационно-коммуникационных сетей является логическим результатом развития информационных технологий и их внедрения во все сферы деятельности современного общества. Основная функция информационно-коммуникационных систем и сетей, или информационных систем передачи информации, в условиях функционирования интегрированных информационных комплексов заключается в организации оперативного и надежного обмена информацией между абонентами (пользователями), а также в сокращении расходов (экономических, технических и т.п.) на передачу данных. Главный показатель эффективности информационно-коммуникационных сетей - время доставки информации и ее количество. Зависит этот показатель от ряда факторов: структуры сети связи, пропускной способности информационной сети связи, способов соединения каналов связи между взаимодействующими абонентами, протоколов информационного обмена, методов доступа абонентов к среде передачи, методов маршрутизации пакетов и тому подобное.

Информационно-коммуникационная сеть *- это интегрированный комплекс организационно-технических мероприятий и взаимосвязанных и программных и программно-аппаратных компонентов, которые обеспечивают достоверную передачу информации от источника сообщения к потребителю.*

Изучение сети в целом нуждается в знании принципов работы и характеристик ее отдельных элементов: компьютеров, коммуникационного оборудования, операционных систем, сетевых приложений и т.п.

Многоуровневой моделью *информационно-коммуникационной сети называется полный комплекс программно-аппаратных средств сети, которые применяются с целью достоверной передачи информационных потоков от источника сообщения к потребителю.*

Первый (аппаратный) уровень *- основа любой сети, образуемая стандартизированными компьютерными платформами и используемая с целью автоматизированной обработки данных.*

Второй уровень *- коммуникационное оборудование отмеченной информационно-коммуникационной сети.*

Хотя компьютеры и являются центральными элементами обработки данных в сетях, не менее важную роль в организации сети играют коммуникационные устройства. Кабельные системы, повторители, мосты, коммутаторы, маршрутизаторы и модульные концентраторы - все эти составляющие превратились из вспомогательных компонентов сети в основные как за влиянием на характеристики сети, так и по стоимости. Сегодня коммуникационное устройство может быть сложным специализированным мультипроцессором, который необходимо конфигурировать, оптимизировать и администрировать. Для изучения принципов работы коммуникационного оборудования

необходимо знание большого количества протоколов, используемых как в локальных, так и в глобальных сетях.

Третий уровень - *операционные системы* (*ОС*), *которые создают и обеспечивают программную платформу сети.*

От того, какие концепции управления локальными и распределенными ресурсами положены в основу сетевой ОС, зависит эффективность работы всей сети. При проектировании сети важно учитывать:

оптимальность взаимодействия данной ОС с другими ОС сети;

возможности обеспечения безопасности информации и защищенности данных;

возможность наращивать количество пользователей (сложность сети);

возможность адаптации или инсталляции на другие типы вычислительных платформ и тому подобное.

Четвертый уровень - уровень *сетевых средств* - *образует сетевые приложения, такие как сетевые базы данных, почтовые системы, средства архивации данных, системы автоматизации коллективной работы и т.п.*

Важно представлять диапазон возможностей, которые предоставляются приложениями для разных сфер применения, а также знать, насколько они совместимы с другими сетевыми приложениями и ОС.

При объединении в сеть большого количества пользователей появляется целый комплекс технических вопросов. Проектирование сети прежде всего предусматривает решение задачи о ее топологии - способе организации физических связей между элементами сети.

Топологией информационно-коммуникационной сети *называется конфигурация графа, вершинам которого соответствуют пользователи (компьютеры*) *сети* (*иногда и другое оборудование, например концентраторы), а ребрам - физические связи между ними.*

Компьютеры, включенные в сеть, часто называют *станциями* или *узлами* сети.

Полносвязная топология отвечает сети, в которой каждый компьютер сети связан со всеми другими (рис. 6.1, *а*).

Ячеистая топология (см. рис. 6.1, *б)* образуется из полносвязной исключением некоторых возможных связей.

Общая шина (см. рис. 6.1, *в*) - распространенная топология для локальных сетей. В этом случае компьютеры подключаются к одному коаксиальному кабелю. Переданная информация может распространяться в обе стороны. Применение общей шины снижает стоимость проводки, унифицирует подключение разных модулей, обеспечивает возможность мгновенного широковещательного обращения ко всем станциям сети. Недостаток общей шины заключается в ее низкой надежности: любой дефект кабеля или какого-либо из многочисленных разъемов полностью парализует всю сеть. Ей присущая также невысокая производительность, поскольку при таком способе подключения в каждый момент времени только один компьютер может передавать

данные в сеть. Поэтому пропускная способность канала связи всегда распределяется здесь между всеми узлами сети.

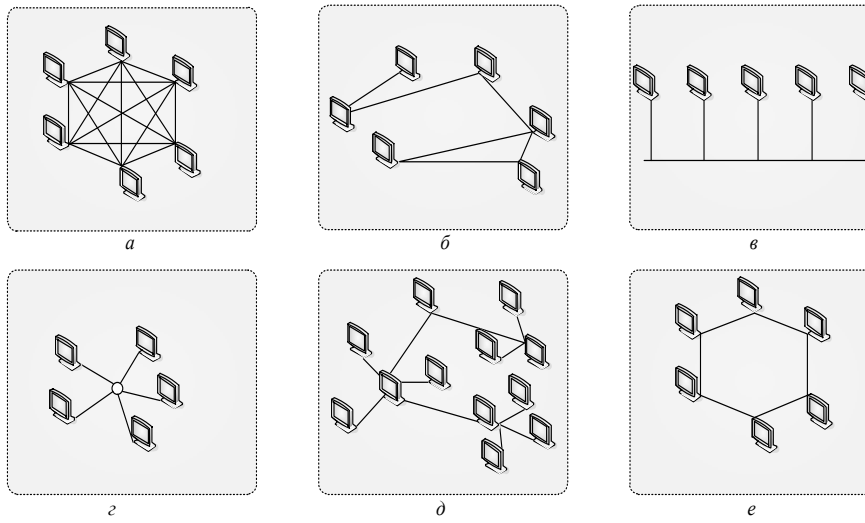


Рис. 6.1. Топологии сетей: *а* - полносвязная топология; *б* - ячеистая топология; *в* - общая шина; *г* - топология звезда; *д* - расширенная звезда; *е* - кольцевая топология

Топология звезда (см. рис. 6.1, *г*). В этом случае каждый компьютер подключается отдельным кабелем к общему устройству, которое называется *концентратором*, которое содержится в центре сети.

Расширенная звезда (***иерархическая топология***) (рис. 6.1, *д*) - топология сети с использованием нескольких концентраторов, иерархически соединенных между собой связями типа звезда. В настоящее время расширенная звезда является самым распространенным типом топологии связей как в локальных, так и в глобальных сетях.

Кольцевая топология (см. рис. 6.1, *е*) - топология сети, в которой данные передаются по кольцу от одного компьютера к другому, как правило, в одном направлении.

Небольшие сети по большей части имеют типичную топологию - звезда, кольцо или общая шина, для больших сетей характерно наличие произвольных связей между компьютерами. В таких сетях можно выделить отдельные произвольно связанные фрагменты (подсети), которые имеют типичную топологию, потому их называют ***сетями со смешанной топологией.***

Сетевая технология - *это согласованный набор стандартных протоколов и программно-аппаратных средств, которые их реализуют* (*например, сетевых адаптеров, драйверов, кабелей и разъемов), достаточных для построения и функционирования информационно-коммуникационной сети.*

Одной из наиболее развитых сетевых технологий является технология Ethernet. Протоколы, на основе которых строится сеть определенной техно-

логии (в узком смысле), специально разрабатывались для совместной работы пользователей (абонентов), поэтому от разработчика сети не требуется дополнительных усилий относительно организации ее взаимодействия. Иногда сетевые технологии называют *базовыми*, имея в виду то, что на их основе строится базис любой сети.

Главный принцип, положенный в основу Ethernet - *случайный метод доступа* к среде передачи данных и общих информационных ресурсов (CSMA/CD). В качестве среды передачи может использоваться толстый или тонкий коаксиальный кабель, витая пара, оптоволокно или радиоволны для передачи информационных потоков.

Сущность случайного метода доступа: *пользователь в сети Ethernet может передавать данные по сети только тогда, когда сеть свободна, т.е. когда никакой* другой пользователь (компьютер) в данный момент не обменивается данными. Поэтому важной частью технологии Ethernet является процедура определения доступности среды.

После того как компьютер убедился, что сеть свободна, он начинает передачу, «захватывая» при этом среду. Время монопольного использования среды одним узлом ограничивается временами передачи одного кадра.

Кадр (***фрейм***) - *это единица данных, которой обмениваются пользователи* (*компьютеры*) *в сети Ethernet. Кадр имеет фиксированный формат и вместе с полем данных содержит разную служебную информацию (например, адрес получателя и адрес отправителя).*

Иногда может возникать ситуация, когда одновременно два или больше пользователей решают, что сеть свободна, и начинают передавать информацию. Такая ситуация называется *коллизией* - препятствием достоверной передачи данных по сети. В стандарте Ethernet предусмотрен алгоритм определения и корректной обработки коллизий. Вероятность возникновения коллизии зависит от интенсивности сетевого трафика.

После выявления коллизии сетевые адаптеры, которые пытались передать свои кадры, прекращают передачу и после паузы случайной длительности пытаются опять получить доступ к среде и передать тот кадр, который вызвал коллизию.

В сетях с небольшим количеством пользователей чаще всего используется одна из базовой топологии - общая шина, кольцо, звезда или смешанная топология. Все перечисленные топологии имеют свойство *однородности, т.е. все компьютеры в такой сети имеют одинаковые права относительно доступа к другим компьютерам (за исключением центрального компьютера при соединении звезда). Такая однородность структуры упрощает процедуру наращивания количества компьютеров, облегчает обслуживание и эксплуатацию сети.*

Однако при построении больших сетей однородная структура связей превращается из преимущества в недостаток. В таких сетях использование

типичных структур порождает разные ограничения, важнейшими из которых являются:

ограничение на длину связи между узлами;

ограничение на количество узлов в сети;

ограничение на интенсивность трафика, который порождается узлами сети.

Например, технология Ethernet на тонком коаксиальном кабеле дает возможность использовать кабель длиной не более чем 185 м, к которому можно подключить не более чем 30 компьютеров.

Для снятия этих ограничений используются специальные методы структуризации сети и специальное *структурообразующее* оборудование - повторители, концентраторы, мосты, коммутаторы, маршрутизаторы. Оборудование такого рода также называется *коммуникационным*, имея в виду, что с его помощью отдельные сегменты сети взаимодействуют между собой.

Повторитель (***repeater***) - *коммуникационное устройство, используемое для физического соединения разных сегментов локальной сети с целью увеличения общей длины сети.*

Повторитель передает сигналы, которые приходят из одного сегмента сети в другие ее сегменты (рис. 6.2). Повторитель дает возможность преодолеть ограничение на длину линий связи за счет повышения качества переданного сигнала - возобновления его мощности и амплитуды, улучшения фронтов и т.п.

Концентраторы характерны для всех базовых технологий локальных сетей.

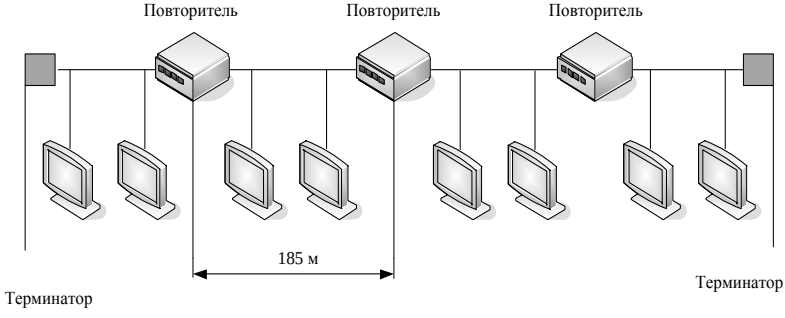


Рис. 6.2**.** Повторитель для увеличения длины сети Ethernet

Концентратор *информационно-коммуникационной сети - коммуникационное оборудование**,** функцией которого является непосредственное направление переданной пользователем информации одному или остальным компьютерам сети.*

Заметим, что в работе концентраторов любых технологий много общего - они повторяют сигналы, которые пришли из одного из их портов, на других своих портах. Разница заключается в том, на каких именно портах повторя-

ются входные сигналы. Концентратор всегда изменяет физическую топологию сети, но при этом может оставлять без изменения ее логическую топологию. Физическая структуризация сети с помощью концентраторов полезна не только для увеличения расстояния между узлами сети, но и для повышения ее надежности.

Физическая структуризация сети полезна во многих пониманиях, однако в ряде случаев, когда речь идет о сетях большого и среднего размера, невозможно обойтись без логической структуризации сети. Важнейшей проблемой, которая не развязывается путем физической структуризации, остается проблема перераспределения переданного трафика между разными физическими сегментами сети. Решение проблемы заключается в отказе от идеи единственной однородной распределенной среды.

Логическая структуризация сети *- это процесс деления сети на сегменты с локализованным трафиком.*

Для логической структуризации сети используются такие коммуникационные устройства, как мосты, коммутаторы, маршрутизаторы и шлюзы.

Мост *разделяет среду передачи сети на части* (*логические сегменты), передавая информацию из одного сегмента в другой только в том случае, если такая передача действительно необходима, т.е. если адрес компьютера назначения принадлежит другой подсети.*

Тем самым мост изолирует трафик одной подсети от трафика других, повышая общую производительность передачи данных в сети.

На рис. 6.3 изображена сеть, которая была сформирована на базе сети с центральным концентратором его заменой на мост. Сети рабочих групп 1 и 2 состоят из одиночных логических сегментов, а сеть группы 3 - из двух логических сегментов. Каждый логический сегмент, построенный на базе концентратора, имеет простую физическую структуру, образованную отрезками кабеля, которые связывают компьютеры с портами концентратора.

Мосты используют для локализации трафика и аппаратной адресации компьютеров. Это осложняет распознавание принадлежности того или другого пользователя к определенному логическому сегменту - сам адрес не содержит какой-либо информации по этому поводу.

Коммутатор (*switch, switching hub*) *по принципу обработки кадров не отличается от моста. Но он является своего рода коммуникационным мультипроцессором, поскольку каждый его порт оборудован специализированным процессором, который обрабатывает кадры согласно с алгоритмом моста независимо от процессоров других портов.*

За счет этого общая производительность коммутатора выше производительности традиционного моста, который имеет один процессорный блок. Можно сказать, что коммутаторы - это мосты нового поколения, которые обрабатывают кадры в параллельном режиме. Ограничения, связанные с применением мостов и коммутаторов за топологией связей, а также рядом других параметров, привели к тому, что в ряде коммуникационных устройств по-

явился еще один тип оборудования - *маршрутизатор* (*router*). Маршрутизаторы надежнее и эффективнее, чем мосты, изолируют трафик отдельных частей сети один от другого.

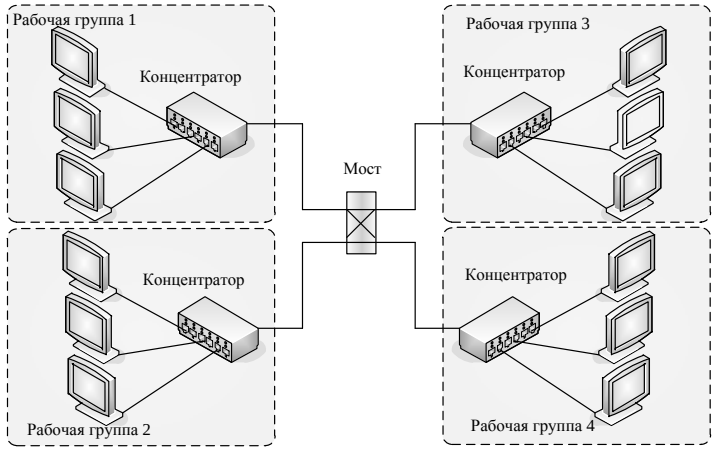


Рис. 6.3. Логическая структуризация сети с применением моста

Маршрутизатор *- это коммуникационное оборудование, которое собирает информацию о топологии межсетевых соединений и на ее основании пересылает пакеты сетевого уровня в сеть назначения.*

Маршрутизаторы образуют логические сегменты с помощью явной адресации, поскольку используют не простые аппаратные, а сложные числовые адреса. В этих адресах есть поле номера сети, так что все компьютеры, в которых значение этого поля одинаково, принадлежат к одному сегменту, который в этом случае называют *подсетью* (*subnet*). Другой очень важной характеристикой маршрутизаторов является их способность соединять в одну сеть подсети, построенные с использованием разных сетевых технологий.

Наряду с перечисленными устройствами отдельные части сети может соединять *шлюз* (*gateway*). Обычно основной причиной, по которой в сети используют шлюз, является необходимость объединить сети с разными типами системного и прикладного программного обеспечения, а не потребность локализовать трафик. Однако обеспечение шлюзом локализации трафика есть, в сущности, некоторым побочным эффектом.

Большие сети практически никогда не строятся без логической структуризации. Для отдельных сегментов и подсетей характерна типичная однородная топология базовых технологий, и для их объединения всегда используется оборудование, которое обеспечивает локализацию трафика: мосты, коммутаторы, маршрутизаторы и шлюзы.

Для упрощения структуры большинство сетей организовываются в наборы *уровней*, каждый следующий из которых возводится над предыдущим. Количество уровней, их названия, содержание и назначение отличаются

от сети к сети. Однако во всех сетях целью каждого уровня является предоставление некоторого сервиса для уровней, расположенных выше.

В начале 1980-х годов такие международные организации по стандартизации, как ISO, ITU-T и некоторые другие, разработали модель, которая сыграла значительную роль в развитии информационных сетей. Эта модель называется моделью *взаимодействия открытых систем (Open System Interconnection - OSI),* или моделью *OSI.* Модель OSI определяет разные уровни взаимодействия систем, дает им стандартные имена и указывает, какие функции должен выполнять каждый уровень.

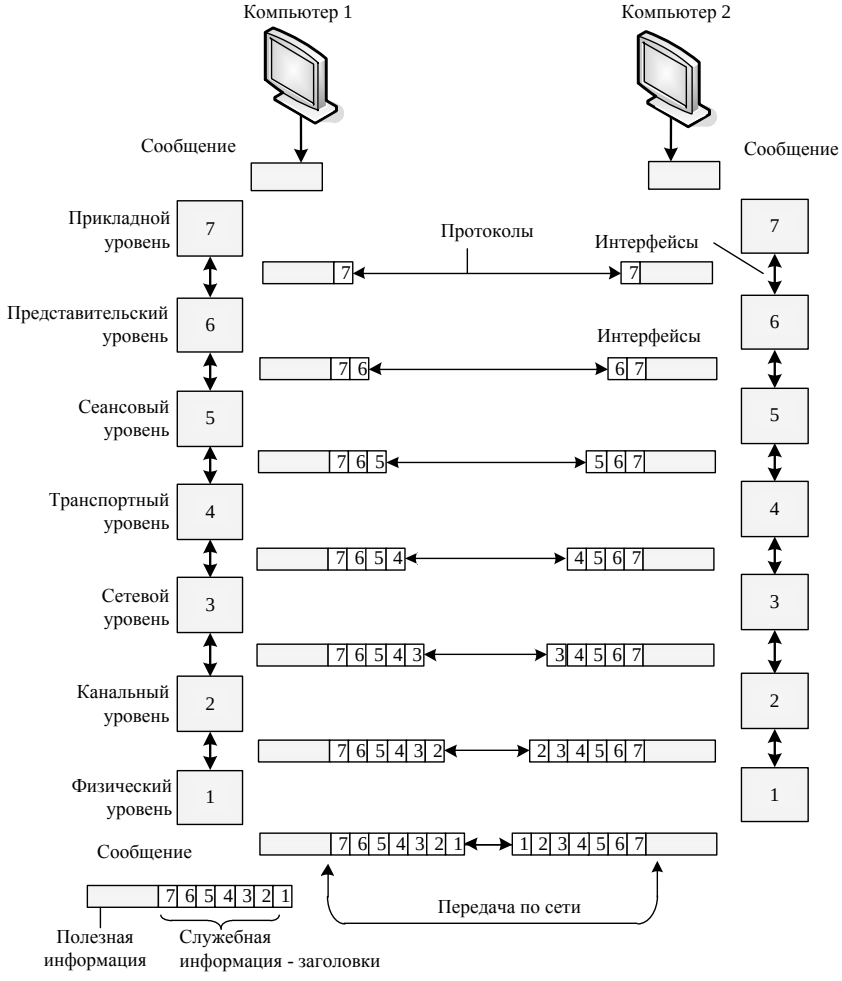


Рис. 6.4. Модель взаимодействия открытых систем ISO/OSI

В модели ISO/OSI (рис. 6.4) средства взаимодействия разделяются на семь уровней: прикладной, представительский, сеансовый, транспортный, се-

тевой, канальный и физический. Каждый уровень имеет дело с одним определенным аспектом взаимодействия сетевых устройств.

Модель OSI описывает только системные средства взаимодействия, которые реализовываются операционной системой, системными утилитами, системными аппаратными средствами. Модель не содержит средств взаимодействия приложений конечных пользователей. Свои собственные протоколы взаимодействия реализуют программные приложения, обращаясь к системным средствам. Потому необходимо различать уровень взаимодействия приложений и прикладной уровень.

Следовательно, пусть приложение обращается с запросом к прикладному уровню, например к файловой службе. На основании этого запроса программное обеспечение прикладного уровня формирует сообщение стандартного формата.

Обычное сообщение состоит из заголовка и поля данных. Заголовок содержит служебную информацию, которую необходимо передать через сеть прикладному уровню машины-адресата, чтобы сообщить ему, какую работу нужно выполнить. После формирования сообщения прикладной уровень направляет его вниз по стеку представительского уровня. Протокол представительского уровня на основании информации, полученной из заголовка прикладного уровня, выполняет необходимые действия и добавляет к сообщению собственную служебную информацию.

Полученное в результате сообщение передается вниз на сеансовый уровень, который, в свою очередь, добавляет свой заголовок и т.д.

Некоторые реализации протоколов размещают служебную информацию не только в начале сообщения в виде заголовка, но и в конце в виде так называемого окончателя. Наконец, сообщение достигает нижнего, физического уровня, который собственно и передает его по линиям связи машине-адресату. К этому моменту сообщение «обрастает» заголовками всех уровней (рис. 6.5).

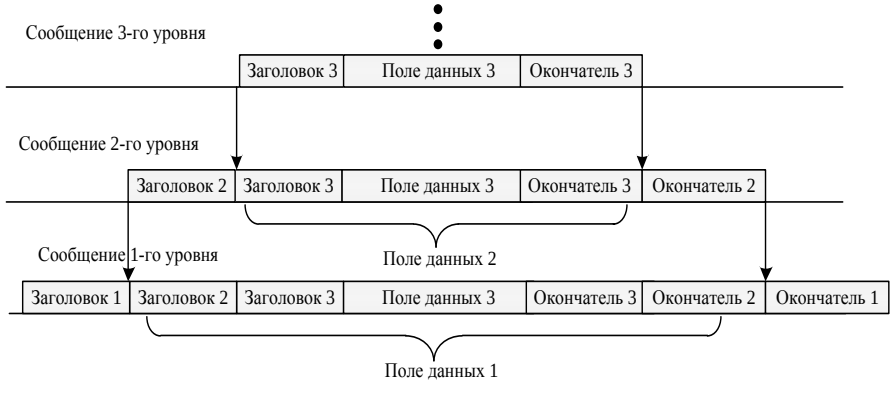


Рис. 6.5. Структура сообщений разных уровней модели ISO/OSI

Когда сообщение по сети поступает на машину-адресат, оно принимается ее физическим уровнем и последовательно перемещается вверх из уровня

на уровень. Каждый уровень анализирует и обрабатывает заголовок своего уровня, выполняя соответствующие данному уровню функции, а затем изымает этот заголовок и передает сообщение высшего уровня.

В модели OSI различают два основных типа протоколов.

Первая группа протоколов - протоколы с установлением *соединения (connection-oriented*), в которых перед обменом данными отправитель и получатель должны сначала установить соединение и, возможно, выбрать некоторые параметры протокола, какие они будут использовать при обмене данными. После завершения диалога они должны разорвать это соединение. Телефон - это пример взаимодействия, которое основывается на установлении соединения.

Вторая группа протоколов - протоколы *без предыдущего установления соединения* (*connectionless*). Такие протоколы называются также *дейтаграммными* протоколами. Отправитель просто передает сообщение, когда оно готово. При взаимодействии компьютеров используются протоколы обоих типов.

Функции уровней модели OSI. *Физический уровень (1)*. Физический уровень имеет дело с передачей битов по физическим каналам связи, таких например, как коаксиальный кабель, витая пара, оптоволоконный кабель или цифровой территориальный канал. Этого уровня касаются характеристики физических сред передачи данных, такие как полоса пропускания, помехоустойчивость, волновое сопротивление и т.п. На этом самом уровне определяются характеристики электрических сигналов, которые передают дискретную информацию, например крутизна фронтов импульсов, уровни напряжения или тока переданного сигнала, тип кодировки, скорость передачи сигналов. Кроме этого, здесь стандартизируются типы разъемов и назначения каждого контакта.

Функции физического уровня реализуются во всех устройствах, подключенных к сети. Со стороны компьютера функции физического уровня выполняются сетевым адаптером или последовательным портом.

Канальный уровень (2). На физическом уровне пересылаются биты информации. При этом не учитывается, что в некоторых сетях, в которых линии связи используются (разделяются) по времени несколькими парами взаимодействующих компьютеров физическая среда передачи может быть занятой. Поэтому одним из заданий канального уровня есть проверка доступности среды передачи. Другим заданием канального уровня является реализация механизмов определения и коррекции ошибок. Для этого на канальном уровне биты группируются в наборы, которые называют *кадрами* (*frames*). Канальный уровень обеспечивает корректность передачи каждого кадра, размещая специальную последовательность битов в начале и в конце каждого кадра для его выделения, а также вычисляет контрольную сумму, обрабатывая все байты кадра определенным способом и добавляя контрольную сумму к кадру. Когда кадр приходит по сети, получатель опять вычисляет контроль-

ную сумму полученных данных и сравнивает результат с контрольной суммой из кадра. Если они совпадают, кадр считается как действительное сообщение и принимается. Если же контрольные суммы не совпадают, то фиксируется ошибка.

Канальный уровень может не только обнаруживать ошибки, но и исправлять их за счет повторной передачи поврежденных кадров. Необходимо отметить, что функция исправления ошибок не является обязательной для канального уровня, поэтому в некоторых протоколах этого уровня она отсутствует, например в *Ethernet* и *Frame relay.*

В локальных сетях протоколы канального уровня используются компьютерами, мостами, коммутаторами и маршрутизаторами. В компьютерах функции канального уровня реализуются совместными усилиями сетевых адаптеров и их драйверов.

В глобальных сетях, которые редко имеют регулярную топологию, канальный уровень часто обеспечивает обмен сообщениями только между двумя соседними компьютерами, соединенными индивидуальной линией связи. Примерами протоколов «точка-точка» (как часто называют такие протоколы) могут служить широко распространенные протоколы PPP и LAP-B. В таких случаях для доставки сообщений между конечными узлами через всю сеть используются средства сетевого уровня.

Сетевой уровень (3). Сетевой уровень служит для образования единственной транспортной системы, объединяя несколько сетей, причем эти сети могут использовать абсолютно разные принципы передачи сообщений между конечными узлами и иметь произвольную структуру связей. Функции сетевого уровня достаточно многообразны. Начнем их рассмотрение на примере объединения локальных сетей.

Протоколы канального уровня локальных сетей обеспечивают доставку данных между любыми узлами только в сети с соответствующей типичной топологией, например топологией иерархической звезды. Это очень жесткое ограничение, которое не дает возможности строить сети с развитой структурой, например сети, которые объединяют несколько сетей предприятия в единственную сеть, или высоконадежные сети, в которых существуют избыточные связи между узлами. Можно было бы осложнять протоколы канального уровня для поддержки петлеобразных избыточных связей, но принцип деления обязанностей между уровнями приводит к другому решению. Чтобы, с одной стороны, сохранить простоту процедур передачи данных для типичной топологии, а со второй - допустить использование произвольной топологии, вводится дополнительный сетевой уровень.

На сетевом уровне сам термин *сеть* наделяют специфическим значениям. В этом случае под сетью понимается совокупность компьютеров, соединенных между собой в соответствии с одной из стандартной типичной топологий, которая использует для передачи данных один из протоколов канального уровня, определенный для этой топологии.

Внутри сети доставка данных обеспечивается соответствующим канальным уровнем, а вот о доставке данных между сетями заботится сетевой уровень, который и поддерживает возможность правильного выбора маршрута передачи сообщения даже в том случае, когда структура связей между составными сетями имеет характер, отличающийся от приемлемого в протоколах канального уровня.

Сети соединяются между собой специальными устройствами, которые называют *маршрутизаторами.*

Чтобы передать сообщение от отправителя, который находится в одной сети, получателю, который находится в другой сети, нужно сделать некоторое количество транзитных передач между сетями, или *хопов* (от *hop* - прыжок), каждый раз выбирая соответствующий маршрут. Таким образом, маршрут является последовательностью маршрутизаторов, через которые проходит пакет.

Сообщения сетевого уровня называются *пакетами* (*packets*). При организации доставки пакетов на сетевом уровне используется понятие «номер сети». В этом случае адрес получателя состоит из старшей части - номера сети и младшей - номера узла в этой сети. Все узлы одной сети должны иметь одну и ту же старшую часть адреса, поэтому термину «сеть» на сетевом уровне можно дать и другое, более формальное определение: сеть - это совокупность узлов, сетевые адреса которых содержат один и тот же номер сети.

На сетевом уровне определяются такие виды протоколов:

сетевые протоколы (*routed protocols*) - *реализуют продвижение пакетов через сеть.* Именно эти протоколы обычно имеют в виду, когда говорят о протоколах сетевого уровня;

протоколы маршрутизации (*routing protocols*) - *протоколы обмена маршрутной информацией. С помощью этих протоколов маршрутизаторы собирают информацию о топологии межсетевых соединений;*

протоколы разрешения адресов - (*Address Resolution Protocol, ARP*), *которые отвечают за отображение адреса узла локальной сети.* Иногда их относят не к сетевому, а к канальному уровню, хотя тонкости классификации не изменяют их назначения.

Протоколы сетевого уровня реализуются программными модулями операционной системы, а также программными и аппаратными средствами маршрутизаторов. Примерами протоколов сетевого уровня является протокол межсетевого взаимодействия IP стека TCP/IP и протокол межсетевого обмена пакетами IPX стека Novell.

Транспортный уровень (4). На пути от отправителя к получателю пакеты могут быть искажены или потеряны. Хотя некоторые приложения имеют собственные средства обработки ошибок, существуют и такие, которые считают лучшим сразу иметь дело с надежным соединением. Транспортный уровень обеспечивает приложениям или верхним уровням стека (прикладному и сеансовому) передачу данных с той степенью надежности, которая им необхо-

дима. Модель OSI определяет пять классов сервиса, которые предоставляются транспортным уровнем.

Эти виды сервиса отличаются качеством услуг, которые предоставляются: срочностью, возможностью возобновления прерванной связи, наличием средств мультиплексирования нескольких соединений между разными прикладными протоколами через общий транспортный протокол, а главное - способностью к определению и исправлению ошибок передачи, таких как искажение, потеря и дублирование пакетов.

Сеансовый уровень (5). Сеансовый уровень обеспечивает управление диалогом: фиксирует, какая из сторон является активной сейчас, предоставляет средства синхронизации. Последние дают возможность вставлять контрольные точки в длинные передачи, чтобы в случае отказа можно было вернуться обратно к последней контрольной точке, а не начинать все сначала. На практике немного приложений используют сеансовый уровень, и он редко реализуется в виде отдельных протоколов, хотя функции этого уровня часто объединяют с функциями прикладного уровня и реализуют в одном протоколе.

Представительский уровень (6). Представительский уровень имеет дело с формой представления информации, которая передается по сети, не изменяя при этом ее содержания. За счет уровня представления информация, которая передается прикладным уровнем одной системы, всегда понятна прикладному уровню другой системы. С помощью средств данного уровня протоколы прикладных уровней могут преодолеть синтаксические отличия в представлении данных или же отличия в кодах символов, например кодов ASCII и EBCDIC. На этом уровне может выполняться шифрование и дешифрование данных, благодаря чему секретность обмена данными обеспечивается сразу для всех прикладных служб. Примером такого протокола является протокол Secure Socket Layer (SSL), который обеспечивает секретный обмен сообщениями для протоколов прикладного уровня стека TCP/IP.

Прикладной уровень (7). Прикладной уровень - это в действительности просто набор многообразных протоколов, с помощью которых пользователи сети получают доступ к распределенным ресурсам, таким как файлы, принтеры или гипертекстовые Web-страницы, а также организовывают свою совместную работу, например с помощью протокола электронной почты. Единица данных, которой оперирует прикладной уровень, обычно называется ***сообщением* (*message*).**

Принципы адресации в информационных сетях. IP адреса. Еще одним заданием, которое нужно учитывать при использовании информационно-коммуникационных сетей, является задание *адресации.* На практике обычно используется сразу несколько схем адресации.

Наибольшее распространение получили три схемы адресации узлов информационных сетей.

Аппаратные адреса. Эти адреса предназначены для сети небольшого или среднего размера, потому они не имеют иерархической структуры. Типичным представителем адреса такого типа является адрес сетевого адаптера локальной сети. Такой адрес обычно используется только аппаратурой, поэтому ее пытаются сделать по мере сил компактной и записывают в виде двоичного или шестнадцатиричного значения. В литературе такие адреса называют MAC-адресами.

Символьные адреса, или ***имена***. Эти адреса предназначены для запоминания пользователем и поэтому обычно имеют смысловое значение. Символьные адреса легко использовать как в небольших, так и в глобальных сетях. Для работы в больших сетях символьное имя может иметь сложную иерархическую структуру, например cisco.netacad.net.

Числовые составные адреса. Символьные имена удобны для пользователя, но через переменный формат и потенциально большую длину их передача по сети не очень экономична. Поэтому часто для работы в больших сетях в качестве адресов узлов используют числовые составные адреса фиксированного и компактного форматов. Типичными представителями адресов этого типа являются *IP адреса*. У них поддерживается двухуровневая иерархия; адрес разделяется на старшую часть - номер сети и младшую - номер узла.

Для того чтобы сетевой уровень мог выполнить свое задание, ему необходима собственная система адресации, которая не зависит от способов адресации узлов в отдельных подсетях. Такая система адресации, которая дала бы возможность на сетевом уровне в универсальный и однозначный способ идентифицировать любой узел составленной сети. Естественным способом формирования сетевого адреса является уникальная нумерация всех подсетей основной сети и нумерация всех узлов в пределах каждой подсети.

Таким образом, сетевой адрес являет собой пару: *номер сети* (*подсети*) *и номер узла.*

Номером узла может быть или локальный адрес этого узла, или некоторое число, не связанное с локальной технологией, которая однозначно идентифицирует узел в пределах данной подсети.

В первом случае сетевой адрес становится зависимым от локальных технологий, и это ограничивает его применение.

Второй подход более универсален. Но в обоих случаях каждый узел составной сети имеет, рядом со своим локальным адресом, еще один - универсальный сетевой адрес.

Данные, которые поступают на сетевой уровень и которые необходимо передать через основную сеть, обеспечиваются заголовком сетевого уровня. Данные вместе с заголовком образуют пакет. Заголовок пакета сетевого уровня имеет унифицированный формат, который не зависит от форматов кадров канального уровня тех сетей, которые могут входить в объединенную сеть, и несет, рядом с другой служебной информацией, данные о номере той сети, которой назначается этот пакет.

При передаче пакета из одной подсети в другую пакет сетевого уровня, инкапсулированный в полученный канальный кадр первой подсети, освобождается от заголовка этого кадра и окружается заголовками кадра канального уровня следующей подсети. Информацией, на основе которой выполняется эта замена, являются служебные поля пакета сетевого уровня.

Следовательно, сетевой уровень использует собственную адресацию, которая обеспечивает каждому узлу подсети основной сети свой универсальный сетевой адрес, который состоит из номера сети и номера узла. Благодаря такой системе адресации сетевой уровень может пересылать информацию к узлу получателю, «не обращая внимания» на внутреннюю структуру подсетей, а в то же время для непосредственного прохождения пакетов к адресату этот уровень использует технологию передачи данных конкретной подсети, через которую идут эти пакеты.

Основной тип адресов сетевого уровня - *IP адрес. IP адрес разделяется на две части:* ***номер сети*** *и* ***номер узла.***

IP адрес имеет длину 4 байта (8 бит), это дает в совокупности 32 бита доступной информации. 32-битовая разрядность IP адреса приводит к тому, что числа выходят большими, даже если они поданы в десятичной форме исчисления. IP адрес записывается в виде четырех чисел, разделенных точками.

Для того чтобы более рационально определиться с размером сети и при этом размежевать части IP адреса, которые касаются номера сети и номера узла, используется *система классов*. Система классов использует значение первых битов адреса.

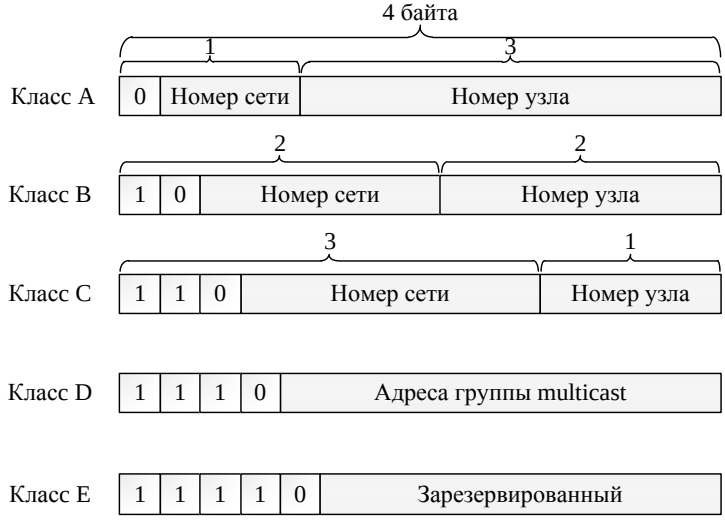


Рис. 6.6. Классы IP адресов

Значение первых битов адреса является признаком того, к какому классу принадлежит тот или другой IP адрес (рис. 6.6).

Если адрес начинается с 0, то сеть относят к *классу A***.** Номер сети класса A занимает один байт, остальные 3 байта выделяются для номеров узла в этой сети. Таким образом, сети класса A имеют номера в диапазоне от 1 до 126. (№ 0 не используется, а № 127 зарезервировано для специальных целей).

Если первые два бита адреса равняются 10, то сеть принадлежит *классу B* (если первый октет IP адреса находится в диапазоне от 128 до 191).

Если адрес начинается с последовательности 110, то это сеть *класса C* (если значение первого октета в IP адресе находится в диапазоне от 192 до 223). В этом случае под номер сети выделяется 24 бита, а под номер узла - 8 бит. Сети класса C имеют небольшое (2^8, т.е. 256) количество узлов. Следует отметить, что именно сети класса C самые распространенные.

Если адрес начинается с последовательности 1110, то она является адресом *класса D* и обозначает особенный, групповой адрес - multicast. Если в пакете как адрес назначения отмечен адрес класса D, то такой пакет должны получить все узлы, которым присвоен этот адрес.

Если адрес начинается с последовательности 11110, то это значит, что этот адрес принадлежит к *классу E. Адреса* этого класса зарезервировано для будущих применений.

Диапазоны номеров сетей и максимальное количество узлов, которые отвечают каждому классу сетей, приведены в табл. 6.1.

Таблица 6.1

Класс сети	Первые биты	Наименьший адрес сети	Наибольший адрес сети	Максимальное количество узлов
A	0	1.0.0.0	126.0.0.0	2^{24}
B	10	128.0.0.0	191.255.0.0	2^{16}
C	110	192.0.1.0	223.255.255.0	2^8
D	1110	224.0.0.0	239.255.255.255	Multicast
E	11110	240.0.0.0	247.255.255.255	зарезервированный

Таким образом, можно однозначно определить, что большие сети получают адреса класса A, средние - класса B, а малые - класса C. В зависимости от того, к какому классу (A, B, C) принадлежит адрес, номер сети может быть представлен первыми 8, 16 или 24 разрядами, а номер хоста - последними 24, 16 или 8 разрядами.

Диапазон значений первого байта для первых трех классов сетей приведен в табл. 6.2.

Таблица 6.2

Класс сети	Диапазон (десятичный) значений первого байта
A	1 до 126
B	128 до 191
C	192 до 254

6.2. Проектирование информационных сетей

Первыми шагами проектирования сети должны быть цель и задачи, которые в зависимости от конкретной организации или сложившейся ситуации предусмотрено возложить на информационную сеть.

Основные требования для большинства сетевых проектов:

Функциональность - сеть должна удовлетворять рабочие требования пользователей и обеспечивать связь пользователь-пользователь и пользователь-прикладная программа с необходимой скоростью и надежностью.

Масштабируемость - сеть должна быть способной к росту. Начальный проект должен расширяться без каких-либо серьезных изменений общего проекта.

Адаптируемость - сеть должна разрабатываться с учетом внедрения перспективных технологий. Сеть не должна содержать элементов, которые могут ограничивать реализацию новых технологий.

Управляемость - проектируемая сеть не должна осложнять решения вопросов сетевого мониторинга и управления.

При проектировании высокотехнологических информационно-коммуникационных сетей приходится решать такие базовые вопросы:

формирование и деление информационных ресурсов сети;

топология сети;

коммуникационное оборудование:

функции и размещения серверов;

домены коллизий;

сегментация сети;

широковещательные домены и тому подобное.

Серверы позволяют сетевым пользователям взаимодействовать и совместно использовать файлы, принтеры и сервисы прикладных программ. Серверы обычно не используют в качестве рабочих станций. На них запущены специализированные операционные системы, например NetWare, Windows NT, UNIX и Linux. Каждый сервер, как правило, реализует одну функцию, например электронную почту или доступ к файлам.

Файл-сервер информационно-коммуникационной сети - *узел сети, который обслуживает и предоставляет сервис другим узлам* (*пользователям*) *с помощью программного обеспечения и коммуникационного оборудования на основе деления совместно используемого информационного ресурса.*

Программное обеспечение, которое дает возможность серверу предоставлять услуги другим компьютерам, часто также называют сервером, с которым контактируют программы-клиенты, установленные на этих компьютерах.

Программное обеспечение, ориентированное на работу в сети, находясь на сетевом сервере файла, одновременно доступно группе пользователей.

Клиент-серверной информационной сетью *называется сеть, которая использует один или несколько центральных выделенных серверов.*

Серверное оборудование может выполнять свои функции в интересах всей информационно-коммуникационной сети или отдельных ее сегментов или рабочих групп.

Сервер информационно-коммуникационной сети поддерживает всех пользователей в пределах этой сети, предлагая общие сервисы, например электронную почту или деление адресации и контроль трафика. Сервер рабочей группы поддерживает специфический состав пользователей и реализует, например, текстовую обработку и файловое разделение.

Серверы целесообразно устанавливать в центральной серверной зоне (main distribution facility - MDF) - рис. 6.7.

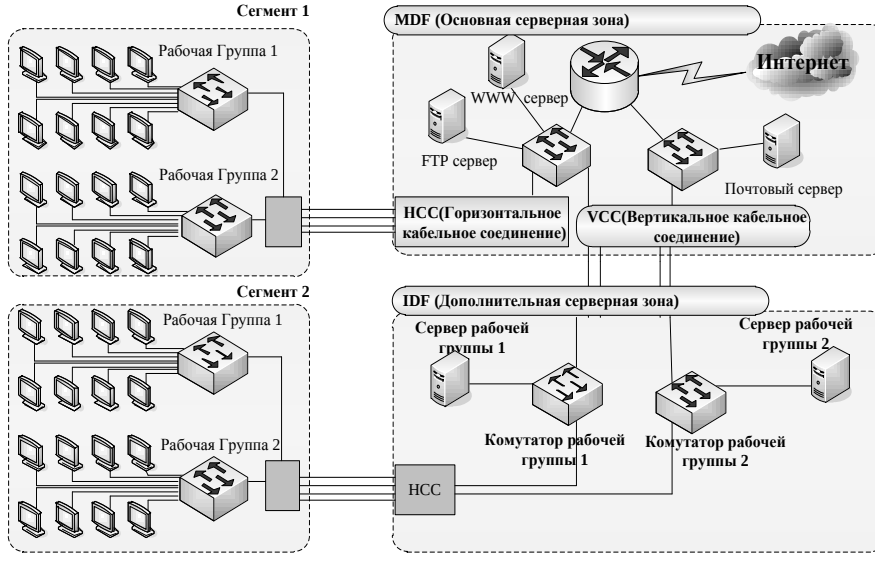


Рис. 6.7. Размещение серверов и рабочих групп информационно-коммуникационной сети

По возможности трафик к серверам информационно-коммуникационной сети должен передаваться непосредственно в MDF и не задевать других сетей. Однако некоторые сети используют маршрутизованную базовую магистраль для серверов. В этих случаях движение трафика через другие сети обычно нельзя предотвратить.

В идеале серверы рабочих групп должны устанавливаться в дополнительных серверных зонах (intermediate distribution facilities - IDF) ближе к пользователям, которые имеют доступ к прикладным программам на этих серверах. Такое размещение дает возможность трафику двигаться только через сетевую инфраструктуру дополнительных серверных зон и не влиять на дру-

гих пользователей в этом сетевом сегменте. Сетевые коммутаторы *уровня 2* модели OSI информационно-коммуникационной сети размещены в центральной серверной зоне, при этом серверы дополнительных серверных зон должны работать со скоростями 100 Мбит/с или более высокими.

Поскольку узлы Ethernet используют протокол CSMA/CD, то каждый узел сети должен «соревноваться» со всеми другими узлами относительно получения доступа к среде передачи (домену коллизий). Если два узла сети начинают одновременную передачу - происходит коллизия. При столкновении кадров переданный фрейм уничтожается, а во все узлы сегмента отсылается соответствующий сигнал. Узлы ожидают произвольный период времени, потом передают данные повторно.

Избыточное количество столкновений может снизить доступную ширину полосы частот сетевого сегмента и соответственно уменьшить производительность сети до 35 или 40 %. Решение этой проблемы - *сегментация.*

Сегментация - *деление единственного домену коллизий на несколько более мелких областей.* В доменах меньшего размера количество коллизий меньше, что дает возможность эффективнее использовать ширину полосы частот.

Для сегментации домену коллизий на *уровне 2* модели OSI могут использоваться мосты и коммутаторы. Сегментация на *уровне 3* достигается применениям маршрутизаторов.

Важным при проектировании является оценивание доступности сети. Через доступность оценивается полезность сети. На доступность влияют производительность, время отзыва и доступ к ресурсам.

Физическая топология сети определяет, каким образом связаны между собой разные компоненты информационной сети (рис. 6.8).

Логическое проектирование сети касается информационных потоков данных, а также имен и схем адресации, использованных в реализации проектного решения отмеченной сети.

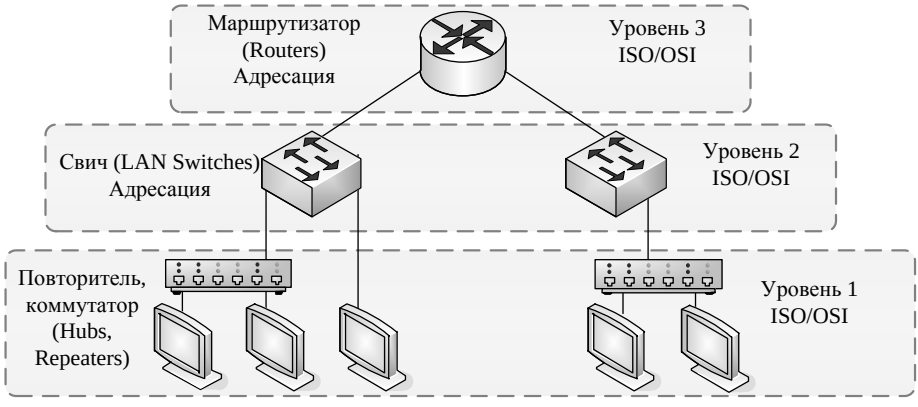


Рис. 6.8. Проектирование топологии сети по уровнях модели OSI

Топология сети согласно с уровнями модели OSI. ***Топология сети уровня 1.*** Одним из важнейших компонентов сети является кабельное оборудование. Сегодня соединение информационно-коммуникационной сети чаще всего базируется на технологии *Fast Ethernet*. Fast Ethernet - это 10 Мбит/с Ethernet, модернизируемый для скорости 100 Мбит/с и такой, который характеризуется полнодуплексной функциональностью.

Он использует стандартную Ethernet-ориентированную логическую топологию шины для адреса канального уровня (MAC-адреса) (рис. 6.9).

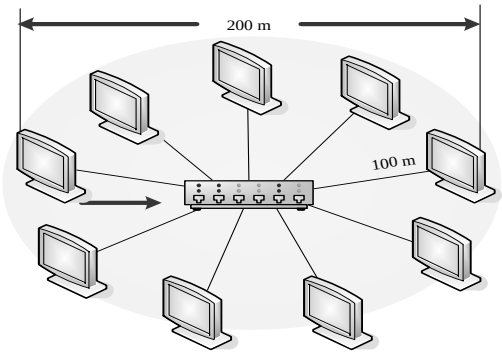


Рис. 6.9. Соединение сети по технологии Fast Ethernet

Проектные вопросы для топологии *уровня 1* содержат тип используемого для прокладки кабеля (обычно медь или волоконно-оптический) и общую структуру кабельного оборудования.

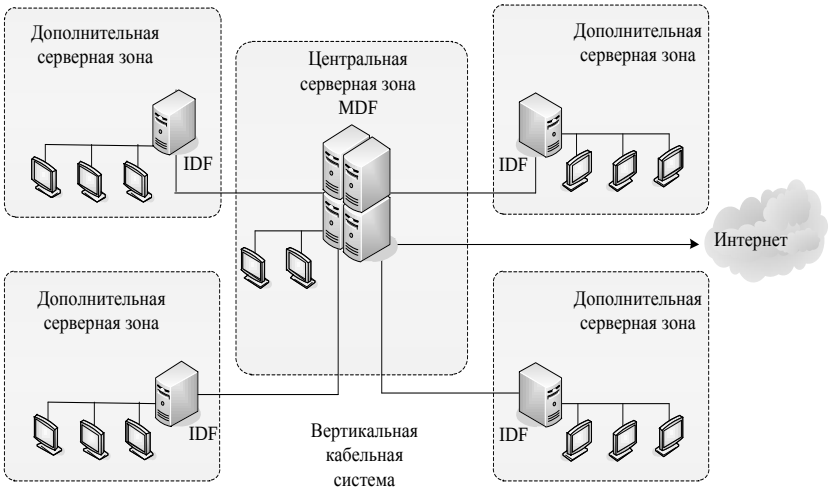


Рис. 6.10. Вертикальное соединение сети

Сильные и слабые стороны топологии должны быть тщательным образом проанализированы, поскольку эффективность сети полностью зависит от используемых кабелей.

Для магистральных линий (вертикального соединения) необходимо использовать волоконно-оптический кабель (рис. 6.10).

В больших сетях, где невозможно выполнить ограничение на 100-метровую длину кабеля для отдельного сегмента (особенно для нескольких дополнительных серверных зон), может использоваться несколько коммутационных панелей. Схема соединения для такой топологии приведена на рис. 6.11.

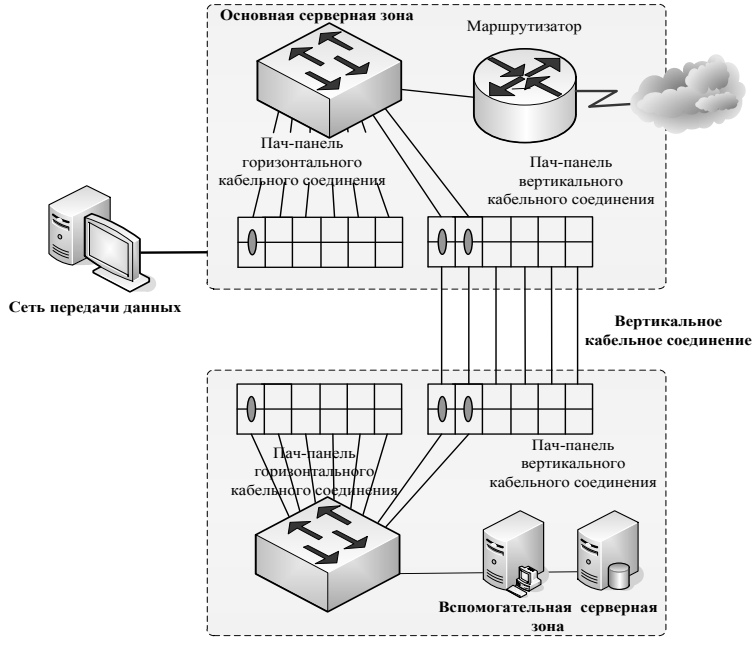


Рис. 6.11. Коммутирование основной и дополнительной серверной зоны на базе вертикальной и горизонтальной кросс-панели

Соединительные кабели (патч-кабели) служат для подключения горизонтальных сетевых кабелей *уровня 1* к портам сетевого коммутатора уровня *2*. Исходный порт коммутатора *уровня 2* соединяется патч-кабелем с Ethernet-портом маршрутизатора *уровня 3*, чем обеспечивается физическое соединение отдельной конечной хост-машины с маршрутизатором.

Для связи разных дополнительных серверных зон с центральным MDF используются кросс-панели для вертикальных кабелей (VCC). Здесь необходимо применение волоконно-оптического кабеля, поскольку вертикальные кабельные длины превышают стандартную 100-метровую границу.

Топология сети уровня 2. Назначение устройств *уровня 2* модели OSI в сети - перенаправление фреймов на основе информации из MAC-адреса узла-получателя, определение ошибок и снижение перегрузок в сети. Наиболее

распространенные сетевые устройства *уровня 2* - это мосты (bridges) и сетевые коммутаторы (switches). Устройства *уровня 2* ограничивают размер домена коллизий.

Коллизии и размер домена коллизий - два показателя, которые негативно влияют на производительность сети. Размеры домена коллизий и, соответственно, количество коллизий уменьшаются путем микросегментации сети. Принцип микросегментации иллюстрирует рис. 6.12.

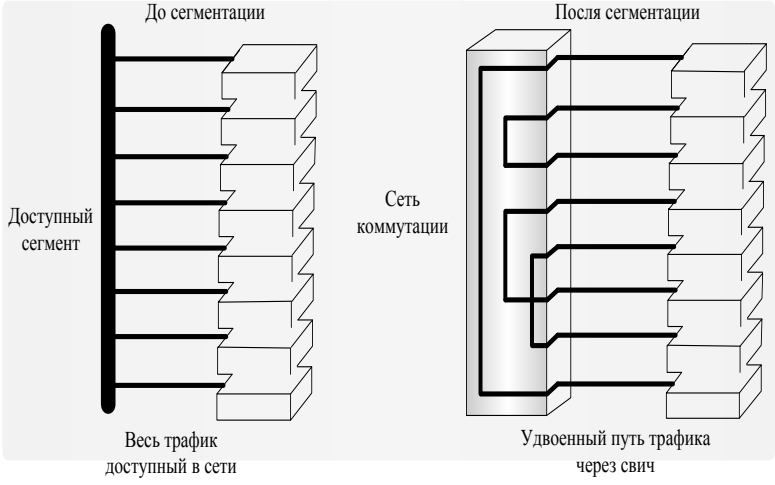


Рис. 6.12. Микросегментация

Кроме микросегментации сетевой коммутатор поддерживает так называемую асимметричную коммутацию, другими словами, пересылка фреймов между каналами с разными скоростями передачи. Эта характеристика сетевого коммутатора важна для согласования трафика между общей для всей компании частью сети и элементами сети в подразделениях компании (рис. 6.13).

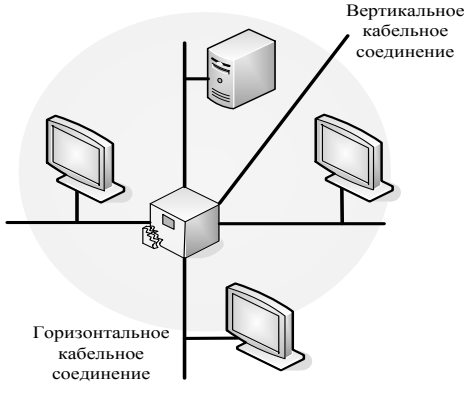


Рис. 6.13. Асимметрическая коммутация

Одна из задач проектирования сети - определение количества 10- и 100-мегабитных портов для центральной и дополнительной серверной зон. Это задание выполняется путем анализа требований пользователей относительно количества горизонтальных кабельных соединений для каждой комнаты и суммарного количества портов на все комнаты компании. Кроме того, подсчитывается количество вертикальных кабельных прогонок.

Размер домена коллизий определяется количеством хостов, которые физически подключаются к отдельному порту сетевого коммутатора. Этот размер фактически определяет доступную для отдельного пользователя часть полосы пропускания кабеля. В идеальном случае к отдельному порту сетевого коммутатора имеет возможность подключиться один пользователь. Однако в действительности к отдельному порту коммутатора через многопортовые повторители (Hubs) подключено несколько пользователей, что увеличивает размеры домена коллизий и количество коллизий, а следовательно, снижает эффективную производительность соответствующего сегмента сети и сетевую производительность отдельной хост-машины (рис. 6.14).

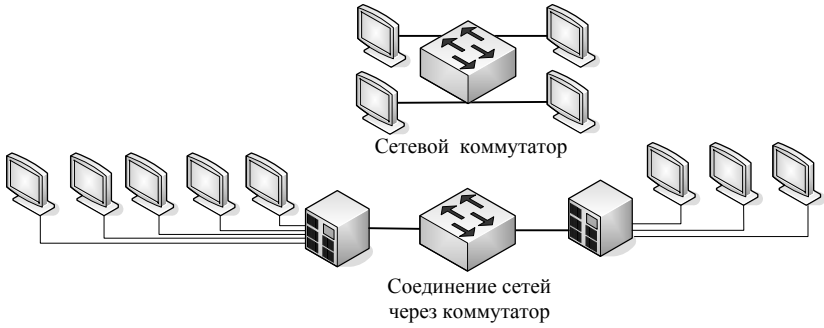


Рис. 6.14. Способы подключения к сетевым коммутаторам

В результате проектирования топологии сети *уровня 2* определяют необходимое количество сетевых коммутаторов, их тип, количество портов с учетом резерва и возможного роста сети, а также типы и необходимое количество многопортовых повторителей (Hubs).

Топология сети уровня 3. Устройства *уровня 3* используются для создания уникальных сегментов. К таким устройствам в первую очередь принадлежит маршрутизатор. Устройства обеспечивают взаимодействие между сегментами на основе адресов *уровня 3* модели OSI, таких как IP-адреса. Использование устройств *уровня 3* реализует деление сетей на уникальные физические и логические сети. Маршрутизаторы обслуживают выход в глобальные сети, например Интернет.

Маршрутизация *уровня 3* определяет движение трафика между уникальными физическими сетевыми сегментами на основе специальной системы адресации. При этом пересылаются пакеты данных на основе адреса получателя и не пересылается служебная информация сети, например, запить ARP. Следовательно, интерфейс маршрутизатора считается входом и исходной

точкой домена широковещания, который останавливает широковещание в другие сегменты информационно-коммуникационной сети.

Маршрутизаторы обеспечивают масштабируемость сети, поскольку они являются помехой широковещанию и разделяют сети на подсети на основе адресов *уровня 3* (рис. 6.15).

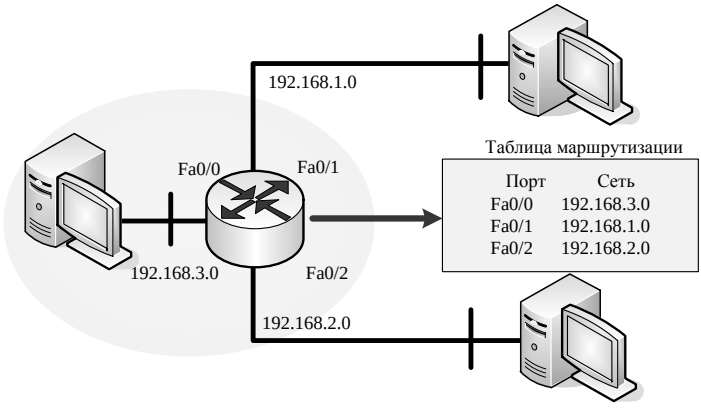


Рис. 6.15. Метод разделения информационной сети на подсети на базе маршрутизаторов

Для принятия решения об использовании маршрутизаторов или коммутаторов важно подробнее рассмотреть эту проблему. Если проблема предопределяется протоколом, а не вопросами конкуренции, то преобладают маршрутизаторы. Маршрутизаторы решают проблемы, которые касаются широковещания, плохо масштабируемых протоколов, аспектов безопасности и адресации сетевого уровня. Однако они больше стоят и сложнее конфигурируются.

Пример проекта, который содержит несколько сетей, приведен на рис. 6.16.

Весь трафик из сети 1 в сеть 2 должен пройти через маршрутизатор. В этой реализации есть два широковещательных домена. Обе сети имеют уникальную схему сетевой адресации. По такой технологии может создаваться множество физических сетей, если горизонтальные и вертикальные кабели подключены к порту соответствующего коммутатора *уровня 2*. Подключение можно выполнить патч-кабелями. Отмеченная реализация обеспечивает необходимый уровень безопасности, поскольку информационный поток, который входит и выходит из локальной сети, должен пройти через маршрутизатор.

Концепция виртуальных информационно-коммуникационных сетей. Концепция виртуальных сетей (VLAN) допускает почти полную независимость физической и логической топологии. Администраторы могут использовать средства виртуальных сетей для группирования рабочих станций даже тогда, когда они отделяются коммутаторами и размещены в других сег-

ментах информационно-коммуникационной сети. Отдельная виртуальная сеть допускает один домен коллизий и один широковещательный домен.

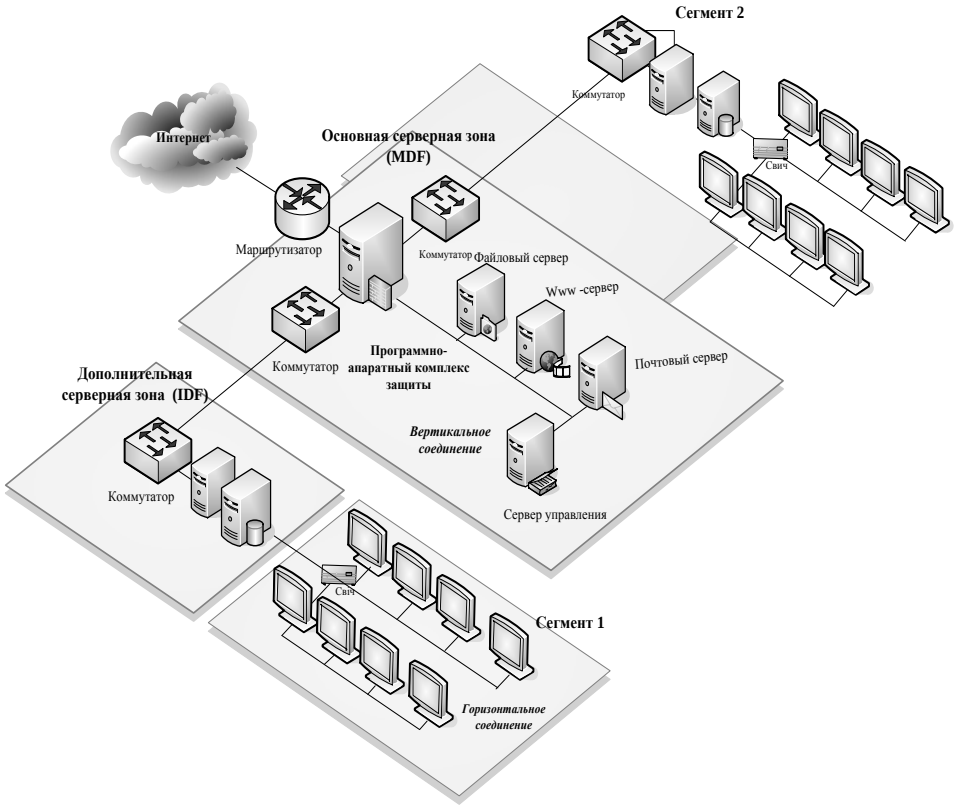


Рис.6.16. Топология разветвленной информационной сети с использованием сегментации, основных и дополнительных серверных зон и базового коммуникационного оборудования

Особенности и преимущества использования виртуальных информационно-коммуникационных сетей. Виртуальная сеть облегчает администрирование логических групп станций и серверов, которые могут взаимодействовать так, как будто они находятся в том же физическом сегменте информационно-коммуникационной сети. Виртуальные сети упрощают администрирование изменений структуры этих групп.

Виртуальные сети логично сегментируют коммутированные сети, опираясь на рабочие функции, принадлежность пользователя к конкретному отделу или сегменту независимо от физического размещения пользователей или физического подключения к сети.

Все рабочие станции и серверы, используемые конкретной рабочей группой, принадлежат одной виртуальной сети независимо от их физического размещения.

Логическая группа сетевых станций, сервисов и устройств не ограниченна физическим сегментом информационно-коммуникационной сети (рис. 6.17).

Конфигурация или деконфигурация виртуальных сетей реализуется через программное обеспечение. Следовательно, конфигурация виртуальной сети не нуждается в физическом перемещении или переключении сетевого оборудования. Функционирование рабочих станций ограничивается взаимодействием с файловыми серверами в той же виртуальной сети.

VLAN логично сегментируют сеть в разные широковещательные домены таким способом, что пакеты коммутируются только между портами, предназначенными той же виртуальной сети.

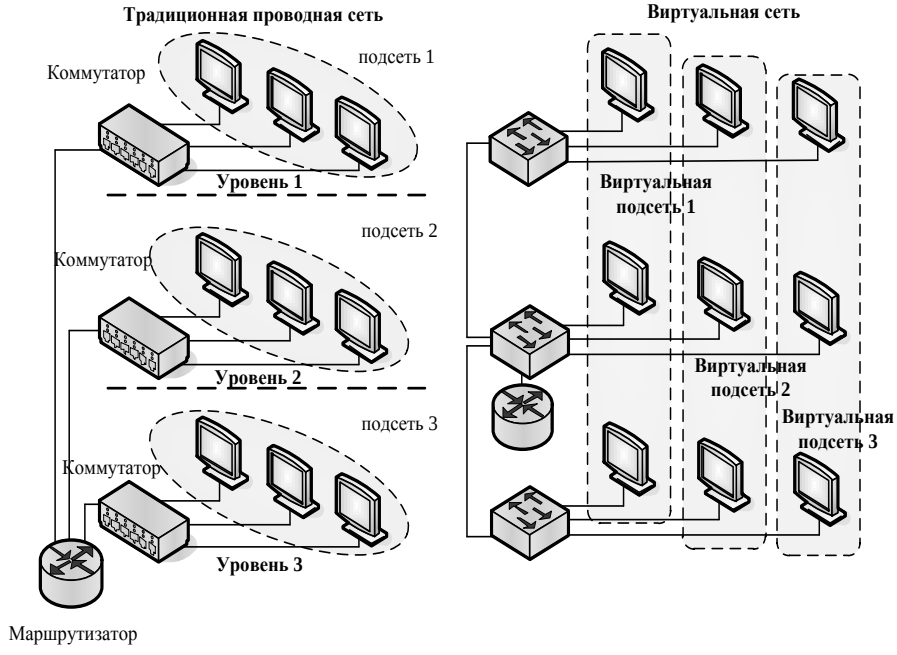


Рис. 6.17. Концепция построения виртуальных сетей

Виртуальные сети создаются для поддержки сервисов сегментации, которые традиционно реализовываются маршрутизаторами. Маршрутизаторы в топологии виртуальных сетей реализуют широковещательную фильтрацию, безопасность и управление трафиком. Коммутаторы не обеспечивают трафик между разными виртуальными сетями, поскольку это нарушает целостность широковещательного домена VLAN. Трафик между виртуальными сетями реализует маршрутизатор.

Для определения виртуальных сетей используется физическое назначение портов сетевого коммутатора. Связь между двумя или больше виртуальными сетями (VLAN1 и VLAN2, рис. 6.18) может существовать только через маршрутизатор.

Концепция виртуальных информационно-коммуникационных сетей объединяет технологии коммутации *уровня 2* и маршрутизации *уровня 3* модели OSI на основе ограничения размеров доменов коллизий и широковещательных доменов.

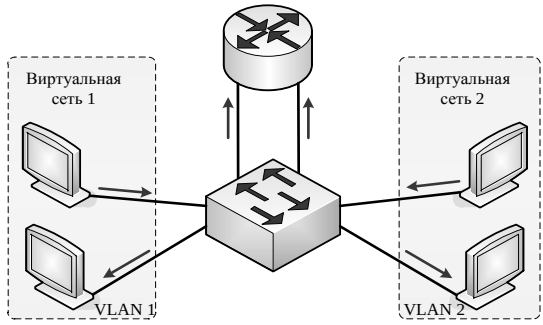


Рис.6.18. Распределение портов сетевого коммутатора
для соединения двух сетей

6.3. Системы беспроводной передачи информации и защита информационных ресурсов

Беспроводные соединения между пользователями и беспроводный доступ к глобальным информационным ресурсам является показателем абсолютного доминирования и развития информационных технологий во всех сферах жизнедеятельности общества.

Благодаря активному подключению пользователей или абонентов к информационно-коммуникационной сети из произвольной точки интегрированного офисного комплекса или территориально распределенных информационных объектов без необходимости коммутативного (проводного) соединения структура информационной сети становится более гибкой и мобильной. Эти преимущества касаются оптимальности топологии сети, использования трафика, сетевого оборудования, организации и *внедрения защиты информационных ресурсов и тому подобного*.

Распространение беспроводных сетей, развитие инфраструктуры информационно-коммуникационных сетей, появление мобильных технологий со встроенным беспроводным решением (Intel Centrino) приводит к глобальному распространению использования таких систем. Такие решения рассматриваются в первую очередь как средство развертывания мобильных и стационарных беспроводных локальных сетей и обеспечения оперативного доступа к глобальным информационным ресурсам.

Технология расширения спектра. В основу всех беспроводных протоколов положена технология расширения спектра (Spread Spectrum, SS). Эта технология предусматривает, что узкополосный полезный информационный сигнал при передаче преобразуется таким образом, что его спектр оказывается

значительно шире спектра начального сигнала. То есть спектр сигнала «расплывается» по частотному диапазону.

Одновременно с расширением спектра сигнала происходит и перераспределение спектральной энергетической плотности сигнала - энергия сигнала также «расплывается» по широкой полосе частот. В итоге максимальная мощность преобразованного сигнала оказывается значительно ниже мощности исходного сигнала. При этом уровень полезного информационного сигнала уравнивается с уровнем помех канала связи. В результате сигнал становится в прямом смысле невидимым - он теряется на уровне помех по всему диапазону переданных частот.

Технологией расширенного спектра (***Spread Spectrum - SS***) *называется технология, которая базируется на изменении спектральной энергетической плотности узкополосного полезного информационного сигнала с целью обеспечения сосуществования нескольких пользователей в одном частотном диапазоне передачи данных.*

В передаче данных по беспроводной сети принимают участие три элемента: радиосигналы, формат данных и структура сети. Каждый из этих элементов не зависит от двух других. С точки зрения эталонной модели ISO/OSI радиосигналы действуют на физическом уровне, а формат данных руководит несколькими из верхних уровней.

Сети беспроводной связи работают в специальном диапазоне радиочастот 2,4 ГГц. Этот диапазон зарезервирован в большинстве стран мира для нелицензионного вида деятельности радиослужб соединений «точка-точка» (свободное применение радиочастотного ресурса пользователями без получения лицензии на использование отмеченных частот в собственных или коммерческих потребностях) на базе частотного деления спектра информационных сообщений.

Радиослужба соединения «точка-точка» управляет коммуникационным каналом, который переносит информацию от передатчика к отдельному приемнику. Противоположностью такому соединению есть широковещательная (brodcast) служба (например, радио- или телевизионная станция), которая отправляет тот же сигнал большому количеству приемников одновременно.

Передачей с расширенным спектром *называется ряд способов передачи отдельного радиосигнала с использованием широкого сегмента радиоспектра.*

В беспроводных сетях Ethernet используются две разные системы радиопередачи с расширенным спектром:

частотное расширение спектра (*FSSS*);

расширение спектра с прямой последовательностью (*DSSS*).

Сравнительно с другими типами сигналов, которые используют отдельный узкий канал, радиосвязь с расширенным спектром обеспечивает такие преимущества:

расширенного спектра вполне достаточно для передачи дополнительной энергии, поэтому радиопередатчики могут работать на очень малой мощности;

поскольку они действуют в относительно широком диапазоне частот, то менее чувствительны к помехам от других радиосигналов и электрического шума.

Это значит, что сигналы можно использовать в средах, где традиционный узкополосный тип информационного сигнала принять и распознать невозможно. Неавторизированному (несанкционированному) абоненту тяжело перехватить и декодировать содержание информационного сообщения, поскольку сигнал с частотным расширением спектра перемещается по множеству каналов.

Частотное расширение спектра (***FHSS***). ***Технология FHSS*** *- это процесс, который дает возможность разделить информационный сигнал на малые энергетические сегменты и на протяжении незначительных интервалов времени (частицы секунды) многократно изменить одну частоту радиосигнала на другую во время передачи данных этих сегментов.*

Передатчик и приемник используют синхронизированную модель сдвига, которая определяет порядок использования разных подканалов (рис. 6.19, где t_d- длительность сигнала).

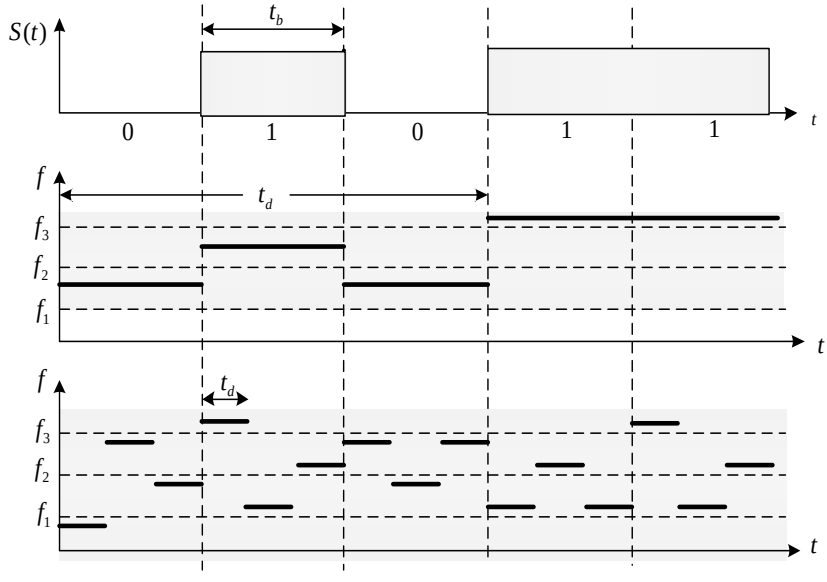


Рис. 6.19. Изменение частот подканалов при передаче полезного сигнала по технологии частотного расширения спектра

Системы на базе частотного расширения спектра маскируют помехи от других пользователей, используя узкополосный сигнал несущей, который многократно изменяет частоту на протяжении каждой секунды. Дополни-

тельные пары передатчиков и приемников одновременно могут использовать разные модели сдвига в том же наборе подканалов. В любой отдельно взятый момент времени каждая система передачи использует свой подканал, поэтому помех между сигналами не возникает. При возникновении конфликта система повторно отправляет тот же пакет до тех пор, пока приемник не получит достоверную копию и не подтвердит прием передающей станции.

Для беспроводных служб передачи данных нелицензированный диапазон 2,4 ГГц разделяется на 75 подканалов шириной в 75 МГц. Поскольку каждый частотный «прыжок» будет задержкой для передачи потока данных, обмен данными на основе FHSS осуществляется сравнительно медленно.

Расширение спектра с прямой последовательностью (DSSS). ***Технология DSSS*** *- это процесс, который дает возможность осуществлять передачу информационного радиосигнала по одному каналу с определенной шириной спектра без изменения частот на основе использования 11-символьного алгоритма последовательности Баркера.*

Каждая связь с применением расширения спектра с прямой последовательностью использует только один канал передачи данных без использования методов перехода между частотами диапазона. Как показано на рис. 6.20, при DSSS-передаче задействовано большую полосу частот, но используется меньшая мощность, чем при традиционном сигнале.

При потенциальной кодировке информационные биты - логические нули и единицы - передаются на базе прямоугольных импульсов напряжения. Прямоугольный импульс длительностью T имеет спектр, ширина которого обратно пропорциональна к длительности импульса. Таким образом, чем меньше длительность информационного бита, тем большую ширину спектра занимает такой сигнал. Для расширения спектра узкополосного сигнала в каждый переданный информационный бит (логический нуль или единица) в буквальном понимании встраивается последовательность так называемых *чипов*. Если информационные биты - логические нули или единицы, образуемые последовательностью прямоугольных импульсов, то каждый отдельный чип - это тоже прямоугольный импульс, но его длительность в несколько раз меньше длительности информационного бита.

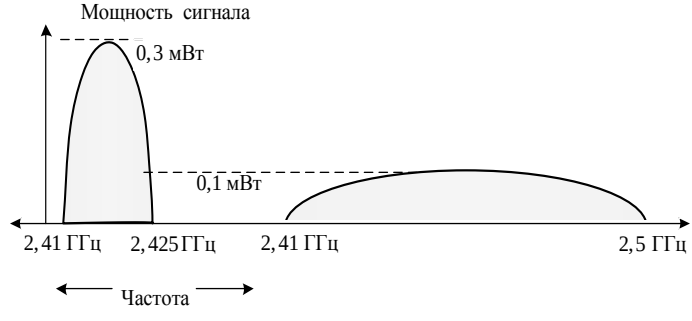


Рис. 6.20. Распределение частотного диапазона при DSSS-передаче

Последовательность чипов представляет собой последовательность прямоугольных импульсов, т.е. нулей и единиц, однако эти нули и единицы не являются информационными. Поскольку длительность одного чипа в n раз меньше длительности информационного бита, то и ширина спектра преобразованного сигнала будет в n раз больше ширины спектра начального сигнала. При этом амплитуда переданного сигнала уменьшится в n раз. Чиповые последовательности, которые встраиваются в информационные биты, называются *помехоподобными кодами* (PN-последовательности). Этим отмечается, что результирующий сигнал становится помехоподобным и его тяжело отличить от естественной помехи.

Используемые для расширения спектра сигнала чиповые последовательности должны удовлетворять определенные требования автокорреляции. Если подобрать такую чиповую последовательность, для которой функция автокорреляции будет иметь резко выраженный пик лишь для одного момента времени, то такой информационный сигнал можно будет выделить на уровне шума. Для этого в приемнике полученный сигнал перемножается на ту же чиповую последовательность, т.е. вычисляется автокорреляционная функция сигнала. В итоге сигнал становится опять узкополосным, поэтому его фильтруют в узкой полосе частот, причем любая помеха, которая попадает в полосу исходного узкополосного сигнала, обрезается фильтрами. В узкую информационную полосу попадает лишь часть помехи, мощность которой значительно меньше чем мощность помехи, которая действует на входе приемника (рис. 6.21).

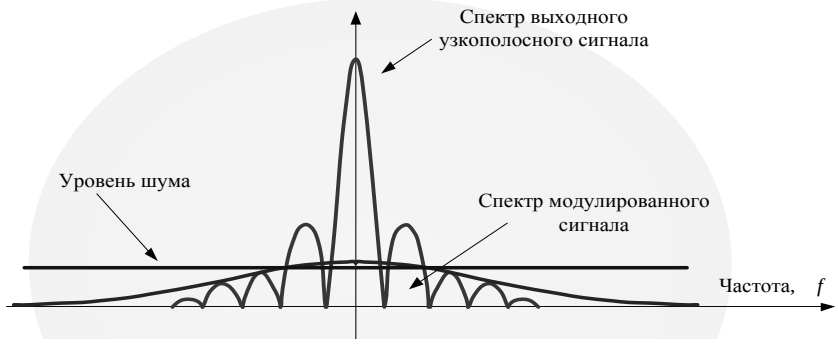


Рис. 6.21. Использование технологии расширения спектра позволяет передавать данные на уровне природного шума

Коды Баркера. Чиповых последовательностей, которые отвечают отмеченным требованиям автокорреляции, существует достаточно много, но для нас особый интерес представляют так называемые *коды Баркера*, поскольку именно они используются для формирования последовательностей чипов.

Коды Баркера имеют наилучшие среди псевдослучайных последовательностей свойства помехоустойчивости, чем и предопределено их широкое применение.

Модуляция в каналах беспроводной связи. Для модуляции синусоидальной несущей сигнала используется относительная двоичная фазовая модуляция (Differential Binary Phase Shift Key - DBPSK). При этом кодировка информации происходит за счет сдвига фазы синусоидального сигнала относительно предыдущего состояния сигнала. Двоичная фазовая модуляция предусматривает два возможных значения сдвига фазы - 0 и π. Тогда логический ноль может передаваться синфазным сигналом (сдвиг по фазе равняется нолю), а единица - сигналом, сдвинутым по фазе на π. (*Подробнее алгоритмы и процессы, связанные с фазовой модуляцией,* рассмотрены в гл. 4.)

Кодирование в каналах беспроводной связи. В системах и сетях беспроводной связи используется сверточное кодирование при котором входная последовательность информационных битов преобразуется в специальном сверточном кодере таким образом, чтобы каждому входному биту отвечал более чем один выходной. Т.е. сверточный кодер добавляет определенную избыточную информацию к исходной последовательности.

Любой сверточный кодер строится на основе нескольких последовательно связанных ячеек с памятью и логических элементов, которые связывают эти ячейки между собой. Количеством ячеек определяется количество возможных состояний кодера.

Дело в том, что *в случае избыточности кодировки даже при возникновении ошибок приема исходную последовательность битов можно безошибочно возобновить. Для возобновления исходной последовательности битов на стороне приемника применяется декодер Витерби.*

Подробнее алгоритмы сверточного кодирования и декодирования на базе алгоритмов Витерби рассмотрены в гл. 5.

Многоканальные системы беспроводной связи. Разделение каналов. При разработке стандартов для систем и сетей беспроводной связи используются две базовые технологии: *метод ортогонального частотного разделения и метод пакетного сверточного кодирования.*

Ортогональное частотное разделение каналов. Распространение сигналов в открытой среде сопровождается возникновениям разного рода помех, источником которых являются сами распространяемые сигналы. Классический пример такого рода помех - эффект многолучевой интерференции сигналов.

Следствием многолучевой интерференции является искажение принятого сигнала. Многолучевая интерференция присуща любому типу сигналов, но особенно негативно она отражается на широкополосных сигналах. При использовании широкополосного сигнала в результате интерференции определенные частоты добавляются синфазно, что приводит к увеличению сигнала, а некоторые, напротив - противофазно, вызывая ослабление сигнала на данной частоте.

Говоря о многолучевой интерференции, которая возникает при передаче сигналов, различают два крайних случая.

Максимальная задержка между разными сигналами не превышает длительности одного символа и интерференция возникает в пределах одного переданного символа.

Максимальная задержка между разными сигналами больше длительности одного символа. В результате интерференции добавляются сигналы, которые подают разные символы, и возникает так называемая межсимвольная интерференция (*Inter Symbol Interference - ISI*).

Наиболее негативно на искажение сигнала влияет межсимвольная интерференция. Поскольку символ - это дискретное состояние сигнала, который характеризуется значениями частоты несущей, амплитуды и фазы, а для разных символов амплитуда и фаза сигнала изменяются, то возобновить исходный сигнал крайне сложно.

Чтобы частично компенсировать эффект многолучевого распространения, используются частотные полосовые фильтры. Однако с ростом скорости передачи данных или символьной скорости, а также с осложнением схемы кодировки эффективность использования частотных полосовых фильтров снижается.

В случае высоких скоростей передачи применяется принципиально другой метод кодировки данных - ортогональное частотное разделение каналов с мультиплексированием (Orthogonal Frequency Division Multiplexing - OFDM).

Частотное разделение каналов с мультиплексированием (*FDM*) - *это метод деления и параллельной передачи информационного потока данных на определенных частотных подканалах.*

При этом высокая скорость передачи достигается именно за счет одновременной передачи данных по всем каналам, а скорость передачи в отдельном подканале может быть невысокой.

Поскольку в каждом из частотных подканалов скорость передачи данных можно сделать не очень высокой, то это создает предпосылки для эффективного притеснения межсимвольной интерференции.

При частотном разделении каналов необходимо, чтобы ширина отдельного канала была *достаточно узкой для минимизации искажения сигнала в пределах отдельного канала, а в то же время - достаточно широкой для обеспечения необходимой скорости передачи.*

Кроме того, для расчетливого использования всей полосы канала, разделенного на подканалы, желательно как можно плотнее расположить частотные подканалы, избежав при этом межканальной интерференции, чтобы обеспечить полную независимость каналов один от другого. Частотные каналы, которые удовлетворяют перечисленные требования, называются *ортогональными.*

Ортогональными частотными каналами **(*OFC*)** *называются каналы при условии, что несущие сигналы всех частотных подканалов* (*а точнее, функции, которые описывают эти сигналы*) *ортогональны друг к другу.*

Хотя сами частотные подканалы могут частично перекрывать друг друга, ортогональная несущая сигналов гарантирует частотную независимость каналов один от другого, а следовательно, и отсутствие межканальной интерференции (рис. 6.22).

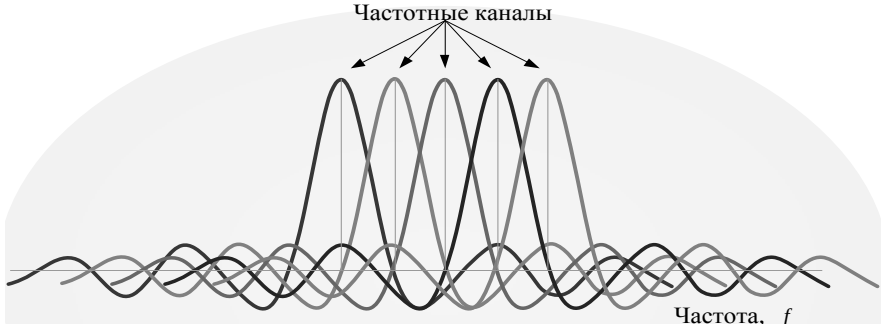


Рис. 6.22. Пример перекрывающихся частотных каналов с ортогональными несущими

Ортогональным частотным разделением каналов с мультиплексированием (OFDM) *называется метод деления широкополосного канала и процесс параллельной передачи информационного потока по ортогональным частотным подканалам данных.*

Одним из ключевых преимуществ метода OFDM есть сочетание высокой скорости передачи с эффективным предотвращением многолучевому распространению. Сама по себе технология OFDM не устраняет многолучевого распространения, но создает предпосылки для устранения эффекта межсимвольной интерференции. Дело в том, что неотъемлемой частью технологии OFDM является охранный интервал.

Охранным интервалом **(*Guard Interval* - *GI*)** *называется циклическое повторение окончания символа, который добавляется к началу символа (рис. 6.23).*

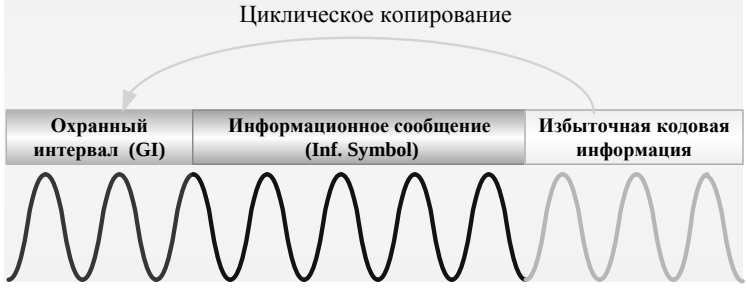


Рис. 6.23. Охранный интервал GI

Охранный интервал является избыточной информацией и поэтому снижает полезную (информационную) скорость передачи, но именно он обеспечивает защиту от возникновения межсимвольной интерференции. Эта избыточная информация добавляется к переданному символу в передатчике и отбрасывается при принятии символа в приемнике.

Наличие охранного интервала создает временные паузы между отдельными символами, и если длительность охранного интервала превышает максимальное время задержки сигнала в результате многолучевого распространения, то межсимвольной интерференции не возникает (рис. 6.24).

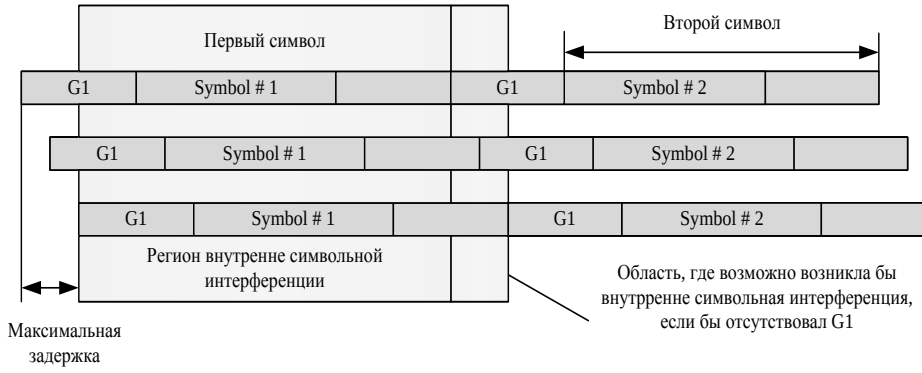


Рис. 6.24. Предотвращение межсимвольной интерференции за счет использования охранных интервалов

При использовании технологии OFDM длительность охранного интервала составляет одну четвертую длительности самого символа.

Для передачи на высших скоростях используется квадратурная амплитудная модуляция QAM (Quadrature Amplitude Modulation), при которой информация кодируется за счет изменения фазы и амплитуды сигнала (гл. 4).

До сих пор мы рассматривали лишь физический (PHY) уровень протоколов. На физическом уровне определяются механизмы, которые используются для превращения данных и обеспечения необходимой скорости передачи в зависимости от среды передачи данных. Следовательно, физический уровень определяет методы кодирования/ декодирования и модуляции/демодуляции сигнала при его передаче и принятии.

В то же время такие вопросы, как регулирование общего использования среды передачи данных, определяются на высшем уровне - уровне доступа к среде передачи данных.

Сетевые устройства и точки доступа систем беспроводной связи. Информационные сети беспроводной связи содержат две категории сетевого оборудования*: станции и точки доступа*.

Станция информационной сети беспроводной связи - *это компьютер или другое периферийное устройство, сетевое оборудование, подключенное к*

беспроводной сети через внутренний или внешний беспроводный адаптер сетевого интерфейса.

Точка доступа информационной сети беспроводной связи *- это базовая станция беспроводной сети и городов между беспроводной и традиционной коммутативной* (*проводной) сетью.*

Сетевой адаптер представляет собой интерфейс между компьютером (пользователем) и сетью. В беспроводной сети адаптер содержит радиопередатчик, который отправляет данные из компьютера в сеть, и приемник, который детектирует входные радиосигналы с данными из сети, передавая их на компьютер. В компьютерной операционной системе беспроводный адаптер имеет тот же внешний вид, что и любой другой сетевой интерфейс.

Маршрутизаторы (***Routers***). Беспроводные маршрутизаторы работают, как «ворота», между внешней глобальной сетью и внутренними локальными или корпоративными сетями. Они распределяют трафик между определенной коммутативной сетью и глобальной. Большинство из них имеют встроенный адаптер, который автоматически назначает IP-адрес каждому компьютеру внутри определенной сети. Отдельный сетевой вход соединяет маршрутизатор с внешней сетью, давая возможность пользователям разделять соединение с глобальной сетью. Большинство маршрутизаторов имеют встроенную аппаратно-программную систему защиты информации (firewall) и другие специальные функции для повышения безопасности беспроводной сети.

Коммутаторы беспроводной сети (***Switches and hubs***). Ethernet-технология - это основа каждой информационной сети. При подключении дополнительных элементов сети используются обычные Ethernet-коммутаторы. Практически каждый современный компьютер имеет встроенный Ethernet-адаптер. Соответственно, большинство беспроводных маршрутизаторов имеют встроенные три или четыре Ethernet-порта. Switches и hubs могут поддерживать множество из них без значительного уменьшения скорости связи для каждого из них. Много точек доступа и большинство беспроводных сетевых адаптеров имеют встроенные всенаправленные антенны. В большинстве ситуаций эти встроенные антенны служат для осуществления процесса передачи и приема информационного потока данных между точкой доступа и ближайшим пользователем (рис. 6.25).

При связи между базовой станцией и беспроводным сетевым адаптером принимают участие две антенны - по одной на каждом конце. На каждом из них антенна с высоким коэффициентом усиления влияет на качество сигнала, поэтому эффективной может быть замена стандартной антенны на направленную в точке доступа или на сетевом интерфейсе.

Направленные антенны могут обеспечить значительное улучшение качества сигнала в узкой зоне покрытия. Кроме того, они способны уменьшить влияние помех от боковых зон за пределами данного лепестка направленности антенны.

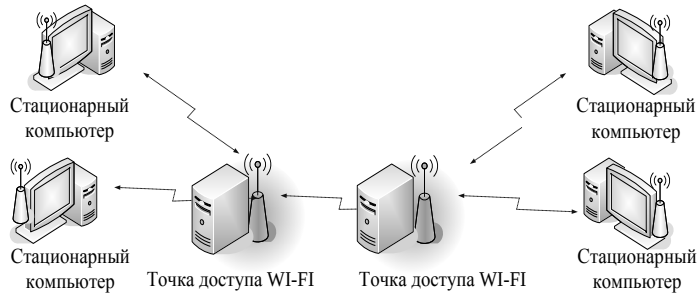


Рис. 6.25. Методика обмена данными через точку доступа в беспроводной сети

Рабочие режимы систем и сетей беспроводной связи. Уровень доступа к среде передачи данных называют MAC-уровнем (или подуровнем — Media Access Control) канального уровня. Именно на MAC-уровне устанавливаются правила общего использования среды передачи данных одновременно несколькими узлами беспроводной сети.

На MAC-уровне определяются два основных типа архитектуры и режимов работы беспроводных сетей:

локальные беспроводные Ad-Hoc-сети;

инфраструктурные сети (Infrastructure Mode).

Локальной беспроводной сетью (*Ad-Hoc*) *называется сеть, которая представляет собой автономную группу станций, которая работает без подключенияя к глобальным сетям передачи данных.*

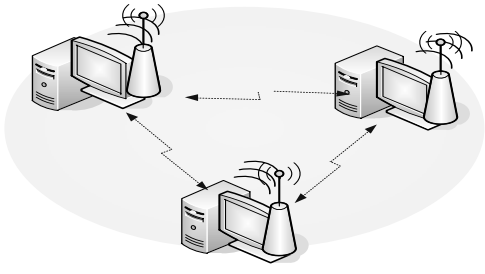


Рис.6.26. Локальная беспроводная сеть с *Ad-Hoc* режимом

Она содержит две или больше беспроводных станций без использования точек доступа. Ad-Hoc-сети также называются *одноранговыми и независимыми базовыми наборами служб. В режиме Ad-Hoc* (рис. 6.26), который называют также *независимым базовым обслуживанием (Independent Basic Service SET-IBSS), или режимом типа «точка-точка (Point to Point)», станции взаимодействуют непосредственно.*

Ad-Hoc-режим: *одноранговое взаимодействие типа «точка-точка». Пользователи (компьютеры*) *осуществляют обмен информационными сообщениями без применения точек доступа.*

Для этого режима нужен минимум оборудования: каждая станция должна быть оснащена беспроводным адаптером. При такой конфигурации нет потребности создавать сетевую инфраструктуру. Основными недостатками режима Ad-Hoc является ограниченный диапазон действия возможной сети и невозможность подключения к внешней сети (например, к Интернету).

Режим Infrastructure Mode. ***Инфраструктурной беспроводной сетью*** *называется сеть, которая представляет собой группу станций, которая работает с подключением к глобальным сетям передачи данных и информационных ресурсов с использованием одной или более точек доступа.*

Каждая беспроводная станция обменивается сообщениями и данными с точкой доступа, которая передает их на другие узлы в проводной сети. Любая сеть, которая требует проводного подключения через точку доступа к периферийному оборудованию, файловому серверу или интернет-шлюзу, является инфраструктурной. Инфраструктурная сеть изображена на рис. 6.27.

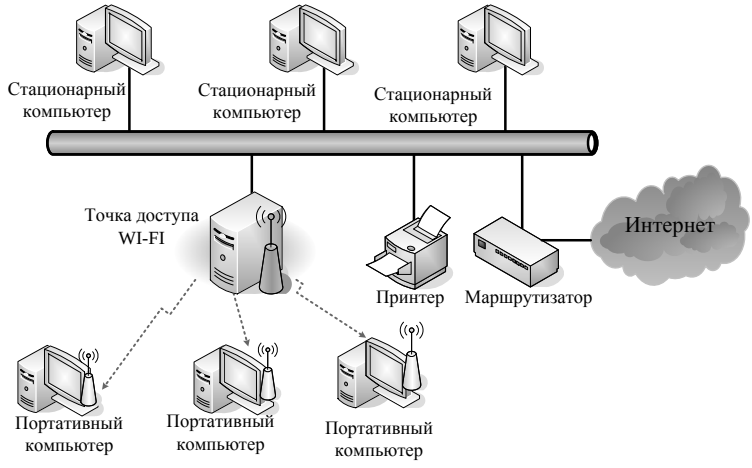


Рис. 6.27. Инфраструктурная беспроводная сеть с коммутацией через точку доступа к стандартной сети

Базовым набором служб *называется инфраструктурная сеть, которая использует только одну базовую станцию для обеспечения приема и передачи информационных потоков данных.*

Расширенным набором служб *называется инфраструктурная беспроводная сеть, которая использует две или более точек доступа для обеспечения приема и передачи информационных потоков данных.*

Кроме двух разных режимов функционирования беспроводных сетей на канальном уровне ISO/OSI определяются правила коллективного доступа к среде передачи данных. Необходимость существования таких регламентирующих правил вполне очевидна при передаче каждому узлу беспроводной сети одновременно. В результате интерференции нескольких таких сигналов узлы, которым назначалась отправленная информация, не смогли бы не толь-

ко ее получить, но и понять, что данная информация адресованная им. Именно поэтому необходимо существование твердых регламентирующих правил, которые определяли бы коллективный доступ к среде передачи данных.

На канальном уровне определяются два типа коллективного доступа к среде передачи данных:

функция распределенной координации (*Distributed Coordination Function - DCF*);

функция централизованной координации (*Point Coordination function - PCF*).

Функция распределенной координации (*DCF*) - *метод, который дает возможность снизить вероятность возникновения коллизий и одновременно гарантирует всем узлам сети равноправный доступ к среде передачи данных.*

Эта функция основывается на методе коллективного доступа с делением несущей частоты информационного сигнала при использовании алгоритма предотвращения коллизий (Carrier Sense Multiple Access/Collision Avoidance, CSMA/CA). При такой организации каждый узел прежде чем начать передачу, «прослушивает» среду, пытаясь обнаружить несущий сигнал, и только при условии, что среда свободна, начинает передачу данных.

Однако в этом случае больше вероятность возникновения коллизий: когда два или больше узлов сети одновременно (или почти одновременно) решат, что среда свободна и начнут передавать данные. Для снижения вероятности возникновения таких ситуаций используется механизм предотвращения коллизий (Collision Avoidance - CA).

Суть этого механизма заключается вот в чем. Каждый узел сети, убедившись, что среда свободна, прежде чем начать передачу, ожидает на протяжении определенного промежутка времени. Этот промежуток является случайным и состоит из двух составляющих: обязательного промежутка (DCF Interframe Space) и выбранного в случайный способ промежутка обратного отсчета. В итоге каждый узел сети перед началом передачи ожидает на протяжении случайного промежутка времени, что, естественно, значительно снижает вероятность возникновения коллизий, поскольку вероятность того, что два узла сети будут ожидать на протяжении того же промежутка времени, чрезвычайно мала.

Для того чтобы гарантировать всем узлам сети равноправный доступ к среде передачи данных, необходимо соответствующим образом определить алгоритм выбора длительности промежутка обратного отсчета (backoff time). Промежуток обратного отсчета хоть и является случайным, но определяется на основании некоторых дискретных промежутков времени, то есть равняется целому числу - количеству элементарных часовых промежутков, названных тайм-слотами (Time Slot). Для выбора промежутка обратного отсчета каждый узел сети формирует так называемое *окно конкурентного доступа* (Contention Window - CW), которое используется для определения количества

тайм-слотов, на протяжении которых станция ожидала перед передачей. Фактически окно CW - это диапазон для выбора количества тайм-слотов, причем минимальный размер окна определяется в 31 тайм-слот, а максимальный размер - в 1023 тайм-слота. Промежуток обратного отсчета определяется как количество тайм-слотов, зависимое от размера окна CW:

$$\text{Backoff time}\,[CW_{min}, CW_{max}]\times\text{Slot time}.$$

Когда узел сети пытается получить доступ к среде передачи данных, то после непременного промежутка ожидания запускается процедура обратного отсчета, т.е. включается обратный отсчет счетчика тайм-слотов начиная от выбранного значения окна CW. Если на протяжении всего промежутка ожидания среда остается свободной (счетчик обратного отсчета равняется нулю), то узел начинает передачу.

После успешной передачи окно CW формируется опять. Если за время ожидания передачу начал другой узел сети, то обратный отсчет счетчика прекращается, а передача данных откладывается. После того как среда станет свободной, этот узел опять начнет процедуру обратного отсчета, но уже с меньшим размером окна CW, который определяется предыдущим значением счетчика обратного отсчета и, соответственно, меньшим значением времени ожидания. При этом очевидно, что чем больше количество раз узел откладывает передачу через занятость среды, тем выше вероятность того, что в следующий раз он получит доступ к среде передачи данных (рис. 6.28). Рассмотренный алгоритм реализации коллективного доступа к среде передачи данных гарантирует равноправный доступ всех узлов сети к среде. Однако при таком подходе вероятность возникновения коллизий хотя и мала, но все-таки существует. Понятно, что снизить вероятность возникновения коллизий можно, увеличив максимальный размер формируемого окна CW. Однако в этом случае увеличивается время задержек при передаче и тем самым снижается производительность беспроводной сети.

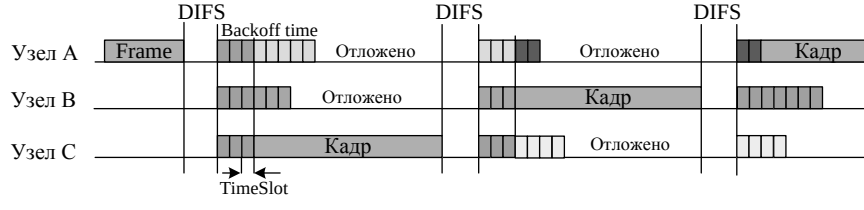


Рис. 6.28. Реализация равноправного доступа к среде передачи данных в методе DCF

Интеграция беспроводного сегмента. Беспроводная информационная сеть нуждается в не таком наборе компонентов коммуникационного оборудования, как традиционная коммутативная (проводная) сеть. Наиболее существенное отличие заключается в отсутствии коммутативного соединения между сетевым сервером, компьютерами беспроводных клиентов и других

устройств, которые формируют инфраструктуру сети. Интерфейс беспроводной сети между проводной и беспроводной частями должен использовать радиопередатчики и приемники для обеспечения прохождения информационных потоков данных.

Вопрос выбора конфигурации. Для выбора конфигурации беспроводной сети прежде всего необходимо измерить уровень сигнала на определенной территории и учитывая уже имеющуюся сетевую инфраструктуру составить проект будущей сети. При наличии стандартной проводной сети с выходом к глобальной сети и функционированием серверов, рабочих станций, сетевых принтеров и т.д. рекомендовано сохранить соответствующую инфраструктуру, лишь дополнив ее беспроводными возможностями.

В таком случае применяется одна или несколько точек доступа, которые работают в режиме инфраструктурной беспроводной сети (Infrastructure Mode), используемом для объединения всех беспроводных устройств с образованием прозрачного моста между беспроводным и проводным сегментами сети. При этом каждая точка должна подключатся к порту проводного коммутатора/маршрутизатора.

Если используются несколько не связанных между собой и расположенных на разных территориях сегментов проводной сети, их можно связать с помощью дополнительных точек доступа, которые работают в режиме мультиплексорного моста «Wireless/Multiple Bridge» (рис. 6.29), предназначенном для объединения двух и больше сегментов проводных сетей.

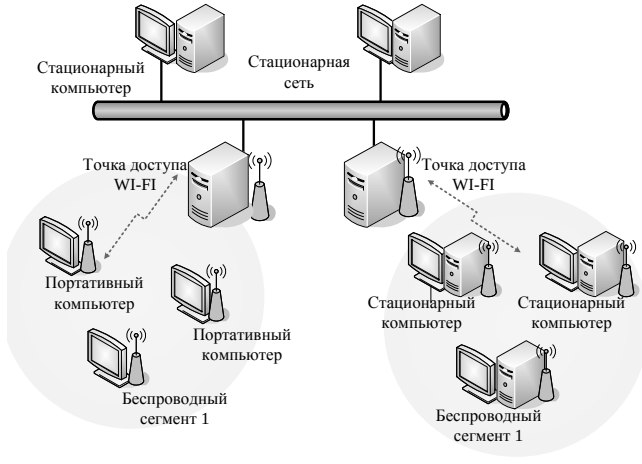


Рис. 6.29. Режим мультиплексорного моста для объединения двух и более сегментов проводных сетей

Для улучшения качества сигнала можно воспользоваться внешними дополнительными антеннами: узконаправленной для соединения в зоне прямого виденья (когда необходимо, чтобы сигнал распространялся в одном

направлении) и всенаправленной (когда необходимо увеличить зону покрытия в помещении).

Большинство беспроводных сетевых интерфейсных адаптеров выполняют только одну функцию: они осуществляют обмен данными между компьютером и сетью.

Беспроводный доступ к проводной сети. Любая точка доступа может выступать как базовая станция, пополняя имеющуюся проводную сеть беспроводной связью (рис. 6.30).

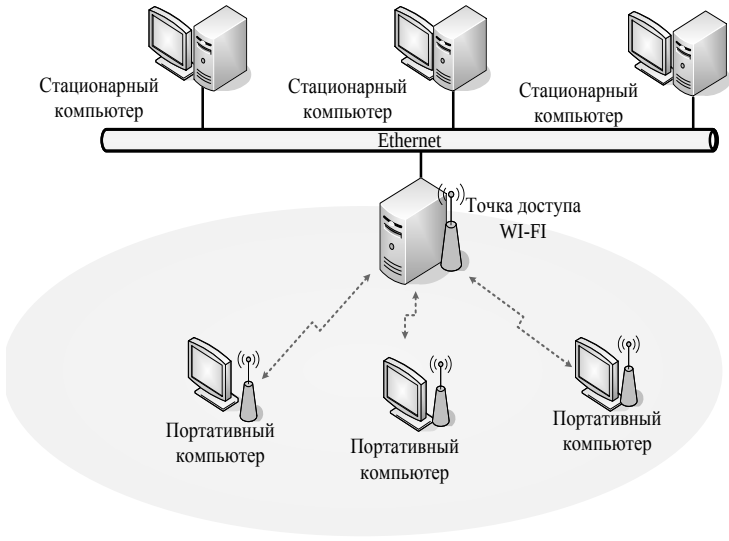


Рис. 6.30. Интеграция беспроводных сегментов к стандартной сети

Точка доступа обеспечивает такой же внешний вид другой части сети, как и дополнительный коммутатор (хаб или свитч), который подсоединяет к ней проводные узлы. В такой разновидности интегрированной сети каждое устройство может осуществлять обмен данными с любым другим сетевым узлом независимо от способа подключения. При этом не столь важно, каким образом - с помощью проводной или радиосвязи - подсоединено конкретное коммуникационное оборудование к сети.

В смешанной сети, которая содержит как проводные подключения, так и беспроводные связи, наилучшим способом функционирования системы может стать использование отдельных устройств, которые совмещают функции беспроводной точки доступа с проводным хабом или свитчем (рис 6.31). Такой тип точки доступа иногда описывается как широкополосный маршрутизатор.

Главными преимуществами комбинирования точки доступа и хаба является удобство их использования для малых локальных беспроводных сетей. Комбинированная единица может также стать самым быстрым способом

расширения имеющейся сети с добавлением как проводных, так и беспроводных узлов в отдаленном месте расположения.

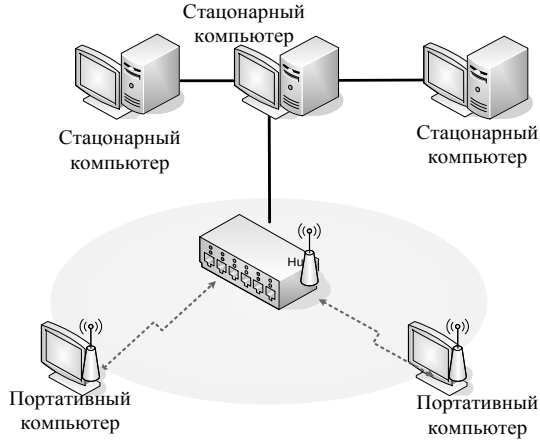


Рис. 6.31. Использование типа точки доступа – широкополосный маршрутизатор

Широкополосный шлюз *представляет собой точку доступа, которая содержит и использует порт для непосредственного подключения к DSL-передатчику или кабельного модема, что поддерживает высокоскоростной доступ к глобальной сети* (рис. 6.32).

Некоторые шлюзовые устройства содержат несколько Ethernet-портов для проводных подключений к локальным компьютерам. Такой подход особенно эффективен для наиболее разветвленной комбинированной информационной сети.

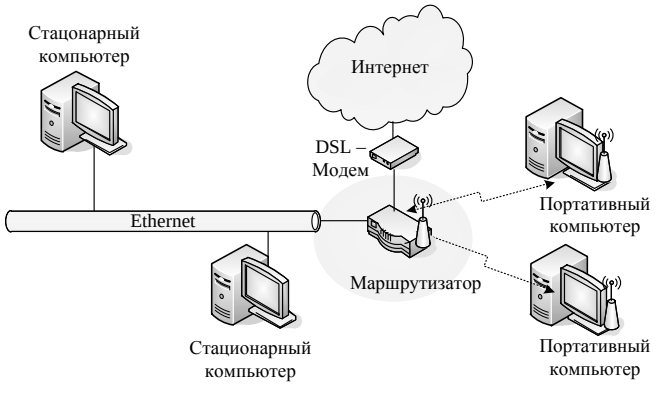


Рис. 6.32. Режим - широкополосный шлюз для смешанной сети

6.4. Системы спутниковой связи

Системы спутниковой связи **(*ССЗ*)** используются для передачи разного типа информационных потоков и данных на значительные расстояния. С момента своего появления спутниковая связь стремительно развивалась, и с

нагромождением опыта, усовершенствованием аппаратуры, развитием методов передачи сигналов происходил переход от отдельных линий спутниковой связи к локальным и глобальным системам.

Такие темпы развития спутниковых систем связи объясняются рядом преимуществ, которыми они характеризуются. К ним, в частности, принадлежат большая пропускная способность каналов, неограниченное перекрытие пространства, высокое качество и надежность каналов. Благодаря преимуществам, которые определяют широкие возможности спутниковой связи, она становится уникальным и эффективным средством информационного обмена данными. Спутниковая связь в это время является основным видом международной и национальной системы передачи данных на большие и средние расстояния.

Все системы можно разделить на системы двух видов: спутники, которые работают на негеостационарных и на геостационарных орбитах.

Негеостационарные спутники используются в основном для военных, научных и метеорологических исследований. Их главная особенность - невозможность поддержки круглосуточной связи между спутником и базовыми станциями. Однако перемещаясь по заданной орбите относительно поверхности Земли, они могут собирать данные с большой площади земной поверхности.

Геостационарные спутники выводятся на такую орбиту в плоскости экватора, при которой их угловая скорость совпадает со скоростью вращения Земли вокруг своей оси. Высота над поверхностью Земли, где выполняются условия постоянства скоростей и равенства центробежной и гравитационной сил, составляет 36 тыс. км. Теоретически один расположенный таким способом спутник может обеспечить высококачественную связь для трети земной поверхности. В действительности обслуживаются существенно меньшие территории. Особенностью спутников на геостационарных орбитах является значительная часовая задержка (около 240 мс) в спутниковом канале, вызванная необходимостью дважды преодолевать расстояние в 36 тыс. км от наземной станции к спутнику.

Дальше будем рассматривать системы, где применяются спутники связи, которые вращаются на орбитах синхронно с вращением Земли. Это дает возможность существенно упростить систему связи. В таком случае каждая земная станция работает непрерывно с одним и тем самым спутником. Раньше при использовании несинхронных спутников существовала необходимость периодического переключения антенной системы каждой земной станции из одного спутника на другой, что, естественно, вызывало перерывы при передаче информационных потоков. К тому же значительную часть стоимости спутниковых систем связи составляла не очень надежная аппаратура наблюдения. Использование стационарных спутников обеспечивает бесперебойную связь, но нуждается в дополнительном запасе рабочего ресурса времени для проведения многократных коррекций орбиты спутника. Считается, что этот

дополнительный запас рабочего ресурса времени для коррекции орбиты является сравнительно небольшой платой за простоту эксплуатации системы и отсутствие перерывов связи. Земные станции при использовании стационарных спутников упрощаются за счет отказа от сложной и дорогой системы наблюдения.

Спутниковые системы связи могут отличаться также и типом передаваемого сигнала, который может быть цифровым или аналоговым. Передача информации в цифровой форме имеет ряд преимуществ сравнительно с другими методами передачи.

***Преимущества систем цифровой спутниковой связи*:**

простота и эффективность объединения многих независимых сигналов и превращение цифровых сообщений в «пакеты» для удобства коммутации;

меньшие энергозатраты сравнительно с передачей аналогового сигнала;

относительная нечувствительность цифровых каналов к эффекту нагромождения искажений при ретрансляциях, что обычно составляет серьезную проблему в аналоговых системах связи;

потенциальная возможность получения очень малых вероятностей ошибок передачи и достижения высокой точности воссоздания переданных данных путем определения и исправления ошибок;

конфиденциальность связи;

гибкость реализации цифровой аппаратуры, которая допускает использование микропроцессоров, цифровую коммутацию и применение микросхем с большей степенью интеграции компонентов.

Сегменты систем спутниковой связи. Большинство систем спутниковой связи состоят из нескольких сегментов (рис. 6.33).

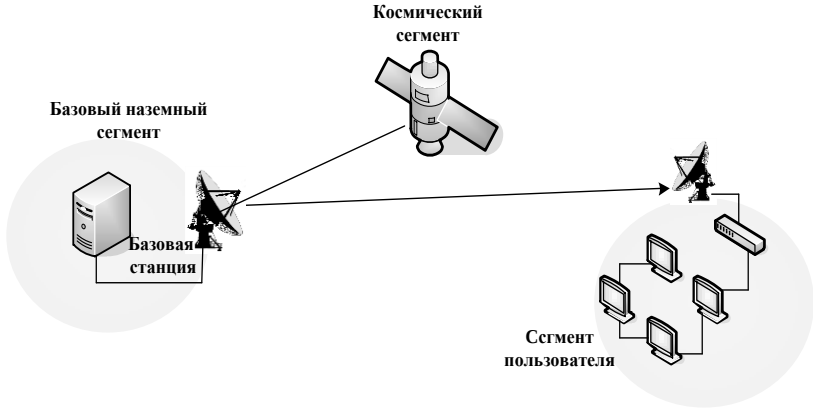


Рис. 6.33. Сегменты систем спутниковой святи

Космический сегмент. Космический сегмент систем спутниковой связи состоит из геостационарных спутников, которые находятся на высоте до 35600 км.

Геостационарные спутники вращаются вокруг Земли с той же скоростью, что и сама Земля, поэтому кажутся нам неподвижными.

Связь с такими спутниками намного более стойкая, чем со спутниками на других орбитах, поскольку:

на протяжении всего сеанса связи между пользователем и геостационарным спутником устанавливается постоянный канал обмена данными. На протяжении всего сеанса пользователь не изменяет источника информации;

сеанс связи между пользователем и геостационарным спутником может поддерживаться на протяжении неограниченного интервала времени;

геостационарный спутник имеет стационарное место на орбите, что обеспечивает достоверную и высококачественную связь.

Система спутниковой связи может состоять из трех или большего количества основных и нескольких запасных спутников, которые находятся на орбите. Основные спутники получают свои названия в соответствии с территориями, над которыми они располагаются. Трех геостационарных спутников, расположенных равномерно по всей длине экватора, достаточно для покрытия 98 % поверхности Земли их глобальными лучами. Вне зоны обслуживания остаются только околополярные области.

Антенны спутников кроме глобальных лучей диаграмм направленности формируют зональные (локальные) лучи, в направлении которых концентрируется излучаемая мощность систем передачи информации. На карте зоны обслуживания глобальный луч каждого спутника изображают в виде овала, центр которого совпадает с данным спутником, а зональные лучи - в виде кругов (рис. 6.34).

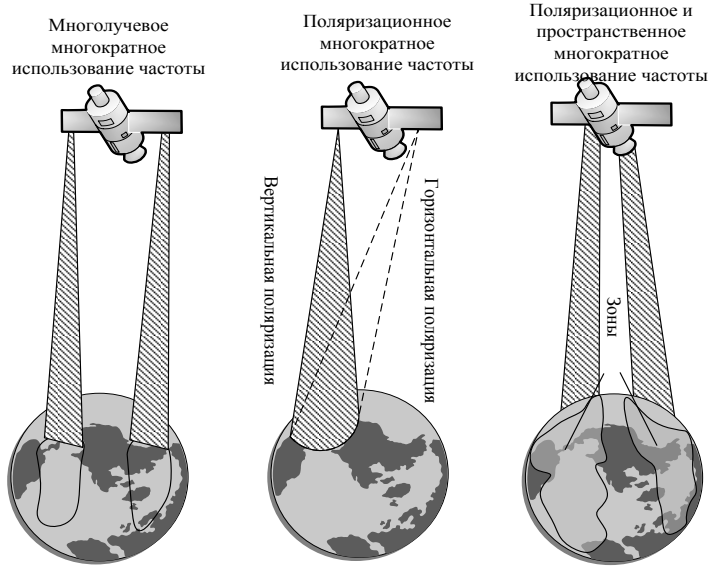


Рис. 6.34. Зональные лучи диаграмм направленности при разных методах поляризации и типах антенн

Технология зональных лучей дает возможность значительно уменьшить массу и габаритные размеры предназначенных для пользователя спутниковых терминалов, не ухудшая при этом качества их работы. Зональными лучами покрывается вся суша и основные морские пути. Спутник может перераспределять энергию между своими зональными лучами, увеличивая пропускную способность одного из них за счет ее уменьшения в другом луче. Таким способом спутник динамически изменяет нагрузку в соответствии с интенсивностью спутникового траффика, который поступает из определенного региона.

Наземный сегмент. Наземный сегмент систем спутниковой связи состоит из таких компонентов:

спутникового центра управления (SCC-Satellite Control Center);

сети береговых и наземных станций (LES-Land Earth Station);

сетевых координирующих станций (NCS-Network Coordination Station);

сетевого операционного центра (NOC-Network Operation Center).

К заданиям спутникового центра управления принадлежат поддержка спутников в заданных позициях над экватором и непрерывное наблюдение за исправностью всех бортовых систем. Данные о состоянии спутников поступают из четырех станций наблюдения, телеметрии и контроля.

Спутник принимает сигналы от предназначенных для пользователя спутниковых телефонов и передает их на береговую станцию, которая служит шлюзом между космическим сегментом и наземными сетями. Все информационные каналы контролируются сетевым операционным центром, который, в свою очередь, опирается на сетевые координирующие станции.

Предназначенный для пользователя сегмент - это оборудование спутниковой связи, серверы, телефоны и терминалы, которыми непосредственно пользуется абонент. Современные спутниковые телефоны имеют достаточно широкие возможности и в то же время простые в эксплуатации. К спутниковому коммутационному оборудованию пользователя можно легко подключить компьютер, факсимильный аппарат, видеотерминал или другие устройства - в зависимости от заданий.

Станции спутниковой связи. *Станция спутниковой связи используется с целью обмена информацией между наземными информационными объектами, а также в системах сбора и деления данных. Спутниковая станция с сетью земных станций обеспечивает систему телекоммуникационной связи и передачи информационных потоков данных* (*в частности, цифровой низкочастотной языковой информации*).

При передаче телекоммуникационного трафика спутниковые системы образуют групповые тракты.

Групповые тракты станций спутниковой связи *- совокупность программно-аппаратных и технических методов и средств, которые обеспечивают прохождение группового сигнала на основе организации нескольких те-*

лекоммуникационных подканалов, которые совмещаются в один спутниковый канал передачи данных.

Групповые каналы передачи станций спутниковой связи *- совокупность программно-аппаратных и технических методов и средств, которые обеспечивают передачу информационных сигналов из одной точки пространства в другую.*

Каналы и групповые тракты спутниковых систем широко используются на участках магистральных и внутризональных телекоммуникационных сетей.

В частности, на местных линиях связи спутниковые системы дают возможность:

организовывать прямые закрепленные каналы и тракты между любыми пунктами связи в зоне обслуживания спутника;

работать в режиме незакрепленных каналов, при котором спутниковые каналы и тракты могут оперативно переключаться из одних направлений на другие при изменении потребностей трафика сети;

использование наиболее эффективных методов обработки информационных потоков (пакетная обработка).

Кроме систем с закрепленным каналом, эффективных при постоянной передаче информации на высоких скоростях (10 кбит/с и более высоких), существуют системы, которые используют часовое, частотное, кодовое или комбинированное разделение канала между многими абонентскими станциями спутника.

Еще одной характеристикой, которая дает возможность классифицировать спутниковые системы, является использование протокола. Первые спутниковые системы были беспротокольными и предлагали пользователю прозрачный канал. Недостатком таких систем была, например, передача информации пользователя, как правило, без подтверждения принимающей стороной ее доставки. Иначе говоря, в таких системах нет предостережений относительно правила диалога между участниками обмена информацией. В этом случае качество переданной информации определяется качеством спутникового канала. При типичных значениях вероятности ошибки на символ в пределах $10–6...10^{-7}$ передача больших объемов информации через спутниковые системы, даже с использованием разных помехоустойчивых кодов, очень осложняется, и даже делается невозможным. Современные спутниковые системы используют протокол, который повышает надежность связи с сохранением высокой скорости обмена информацией между абонентами.

Составляющие и технические характеристики спутниковых систем. Спутниковая станция по конструктивному признаку состоит из таких компонентов: высокочастотного модуля (ODU); низкочастотного модуля (IDU). Высокочастотный модуль состоит из антенны и приемника-передатчика, в котором установлены низкочастотные модули, которые, в

свою очередь, состоят из модема и мультиплексора (каналообразующие оборудование).

Стандартный вариант комплектации предусматривает наличие параболической антенны небольшого диаметра и приемника-передатчика. В зависимости от места расположения спутниковой станции относительно центра зоны освещения спутника и скорости передачи в канале используются более мощные передатчики или антенны большего диаметра. На наземном сегменте устанавливается модем и мультиплексор.Высоко- и низкочастотные модули соединены между собой радиочастотными (RF) кабелями коммутации. По них идет сигнал промежуточной частоты (IF). По функциональным назначениям наземный сегмент разделяется на базовый комплект, который обеспечивает передачу самого канала, и на дополнительное оборудование, которое обеспечивает направленную передачу этого канала.

Однозеркальная антенна обычно выполняется по схеме офсет (со смещенным центром). Схема офсет дает возможность снизить уровень боковых лепестков, которые идут параллельно, создавая максимальные помехи. Эта схема дает также возможность избежать нагромождения атмосферных осадков на поверхности рефлектора.

Спутниковый модем - *радиотехническое средство подсистемы модулятора, предназначенное для кодировки переданного цифрового потока, который пришел из мультиплексора.*

Спутниковый мультиплексор - *радиотехническое средство, предназначенное для обеспечения направленной передачи низкочастотной телекоммуникационной информации и других информационных потоков данных*.

Спутниковый мультиплексор дает возможность скомбинировать телекоммуникационные сообщения с синхронной и асинхронной передачей данных по одному каналу локальной наземной или спутниковой сети связи. Это дает возможность снизить телекоммуникационные расходы путем увеличения возможностей передачи важной информации и одновременного уменьшения пропускной способности канала.

Спутниковый шлюз - *радиотехническое программно-аппаратное средство, которое обеспечивает выход на информационно-коммуникационные сети наземных телекоммуникаций*.

Шлюз может обеспечивать:

выход на телекоммуникационные сети;

услуги междугородной связи с выходом на информационные сети общего пользования;

услуги международной телекоммуникационной связи;

выход на специальные телекоммуникационные сети;

выход на глобальные информационно-коммуникационные сети передачи данных (POCHET, INTERNET, RELCOM и т. др.).

Ресурс системы спутниковой связи. ***Ресурс связи*** (Communications Resource - CR) - это время (длительность сеанса связи) и ширина полосы ча-

стот, доступных для передачи полезного сигнала в определенную информационно-коммуникационную систему.

Для создания эффективной системы связи необходимо спланировать деление ресурса между пользователями системы, чтобы соотношение время / частота было оптимальным. Результат такого планирования - равноправный доступ пользователей к ресурсу.

С проблемой общего использования ресурса связи связаны термины «*уплотнение» и «множественный доступ*». При использовании термина «уплотнение» требования пользователя к общему использованию ресурса связи постоянные или (чаще всего) изменяются незначительно. Деление ресурса выполняется априорно, а общее использование ресурса обычно привязано к локальному устройству.

Применение множественного доступа, как правило, нуждается в отдаленном общем использовании ресурса, как, например, в случае спутниковой связи. При динамической схеме множественного доступа контролер системы должен учитывать потребности каждого пользователя ресурса связи. Время, необходимое для передачи соответствующей управляющей информации, устанавливает верхнюю границу эффективного использования ресурса связи.

Существуют три основных способа увеличения пропускной способности (общей скорости передачи данных) ресурса связи:

увеличение эффективной изотропно-излучаемой мощности (effective isotropic radiated power - EIRP) передатчика или снижения потерь системы, что в каждом случае приведет к увеличению отношения сигнал / шум $\left(E / N_0\right)$.

увеличение ширины полосы канала передачи данных;

повышение эффективности деления ресурса связи на базе множественного доступа.

Основные способы разделения ресурса связи таковы.

Частотное разделение (frequency division - FD). Разделяются определенные поддиапазоны используемой полосы частот.

Часовое разделение (time division - TD). Пользователям выделяются периодические часовые интервалы. В некоторых системах им предоставляется ограниченное время для связи. В других случаях время доступа пользователей к ресурсу определяется динамически.

Кодовое разделение (code division - CD). Выделяются определенные элементы набора ортогонально (или почти ортогонально) разделенных спектральных кодов, каждый из которых использует весь диапазон частот.

Пространственное разделение (space division - SD), *или многолучевое многократное использование частоты.* С помощью точечных лучевых антенн радиосигналы разделяются и направляются в разные стороны. Этот метод допускает многократное использование одного частотного диапазона.

Рис.6.35. Классификация методов и средств построения интегрированных информационно-коммуникационных систем спутниковой связи

Поляризационное разделение (polarization division - PD), или *двойное поляризационное многократное использование частоты.* Для разделения сигналов применяется ортогональная поляризация, которая дает возможность использовать один частотный диапазон.

Методы и принципы, которые используются при построении интегрированных информационно-коммуникационных систем спутниковой связи, приведены на рис. 6.35.

Ключевым моментом всех схем уплотнения и множественного доступа является то, что при использовании ресурса разными сигналами интерференция не предоставляет неуправляемых взаимных помех, которые делают невозможным процесс детектирования. Интерференция допустима до тех пор, пока сигналы одного канала увеличивают вероятность появления ошибок в другом канале. Избежать взаимных помех между разными пользователями дает возможность применение в разных каналах ортогональных сигналов.

Напомним, что сигналы $x_i(t)$ и $x_j(t)$ $\left(i,\ j=1,\ 2,...,\ N\right)$ являются ортогональными, если в часовой области выполняется условие

$$\int_{-\infty}^{\infty} x_i(t)\ x_j(t)\ dt = \begin{cases} K \text{ при } i = j, \\ 0 \text{ при } i \neq j, \end{cases} \tag{6.1}$$

где K - ненулевая константа. Аналогично сигналы ортогональны, если в частотной области выполняется условие

$$\int_{-\infty}^{\infty} X_i(f)\ X_j(f)\ df = \begin{cases} K \text{ при } i = j, \\ 0 \text{ при } i \neq j, \end{cases} \tag{6.2}$$

где функции $X_i(f)$ - Фурье-образы сигналов $x_i(t)$.

$$\int_{-\infty}^{\infty} X_i(f)\ X_j(f)\ df = \begin{cases} K \text{ при } i = j, \\ 0 \text{ при } i \neq j, \end{cases} \tag{6.2}$$

где функции $X_i(f)$ - Фурье-образы сигналов $x_i(t)$.

Методы кодировки в системах спутниковой связи. Поскольку каналы и тракты спутниковых линий входят в национальные и международные сети общего пользования, к их качественным показателям выдвигаются достаточно жесткие требования.

Поэтому при передаче цифровых сигналов в спутниковых системах применяют помехоустойчивое кодирование, которое называют также *прямым исправлением ошибок* (FEC в англоязычной литературе).

Сегодня широко применяются хорошо разработанные коды с прямым исправлением ошибок двух основных классов.

Блочные коды. Последовательность данных разделяется на блоки из n символов (обычно от 1 до 7); каждому блоку ставится в соответствие кодовая комбинация из n символов $\left(n > k\right)$, что передается по каналу связи; прибавленные $r = n - k$ символы называются проверочными; код характеризуется кодовой скоростью $R = k/n$ и максимальным количеством L ошибок в кодовой комбинации, которые он может исправить.

Сверточное кодирование. Применение сверточного кодирования дает возможность не только повысить вероятность переданной информации, но и получить энергетический выигрыш $h_{\text{е.в}}$, что равняется уменьшению мощности передатчика. Платой за этот выигрыш является расширение полосы частот, которую занимает радиосигнал, учитывая необходимость передачи избыточных проверяющих символов. Величина выигрыша зависит от кодовой скорости R, способа кодировки и алгоритма декодирования. Для декодирования используют алгоритм, предложенный А. Витерби. При этом энергетический выигрыш достигает 5...6 дБ при $R = 1/2$. Как первый (внешний) код используют блочный код (обычно код Рида - Соломона). Дальше символы образованных кодовых комбинаций по большей части перемешивают (переставляют в определенном порядке) и подают на второй (внутренний) кодер, конечно - сверточный. Декодирование осуществляется в обратном порядке: сначала декодируют внутренний код, а дальше символы декодированного сигнала поддают обратному перемешиванию (переставляют на исходящие позиции), в результате чего пакеты ошибок «разбивают» на одиночные ошибки (какие легче исправить) и потом декодируют внешний код.

6.5. Множественный доступ к информационным ресурсам

Множественный доступ с частотным распределением в спутниковых системах. Большинство спутников связи находятся на геостационарной геосинхронной орбите. Это значит, что спутник находится на круговой орбите, которая лежит в плоскости земного экватора. При этом спутник находится на такой высоте над уровнем моря (приблизительно 35 830 км), на которой период его вращения вокруг Земли равняется периоду вращения самой Земли вокруг Солнца. Поскольку при наблюдении с Земли такие объекты кажутся неподвижными, то три спутника, расположенных через 120° один от другого, дают возможность охватить территорию всего земного шара (кроме полярных областей). Большинство спутниковых систем связи используют нерегенеративные ретрансляторы или транспондеры. Нерегенеративный значит, что сигналы «Земля-спутник» усиливаются, сдвигаются по частоте и ретранслируются на Землю без обработки сигнала, демодуляции или повторной модуляции.

Самым широко используемым диапазоном в коммерческих системах спутниковой связи есть так называемая *полоса* С (*C-band).* В этом диапазоне для передачи сигнала «Земля-спутник» применяются несущая частота 6 ГГц и частота 4 ГГц передачи сигнала «спутник-Земля». В соответствии с международными соглашениями для систем передачи в полосе разрешено использовать любой спутник, который работает в спектральном диапазоне шириной в 500 МГц. По большей части такой спутник имеет 12 транспондеров с шириной полосы 36 МГц каждый. Наиболее распространенные транспондеры

работают в режиме FDM/FM/FDMA (уплотнение с частотным распределением, частотная модуляция, множественный доступ с частотным распределением).

Рассмотрим составляющие отмеченных режимов.

FDM *- сигналы, подобные телекоммуникационным, имеющие одиночную боковую полосу шириной 4 кГц и обрабатывающиеся с использованием частотного распределения каналов, в результате чего формируется многоканальный сигнал.*

FM *- составной сигнал модулирует несущую на базе метода фазовой модуляции и передается на спутник.*

FDMA *- поддиапазоны полосы транспондера (36 МГц) могут распределяться между разными пользователями. Каждому пользователю выделяется определенная полоса, на которой он получает доступ к транспондеру.*

Таким образом, составные каналы с частотным распределением модулируются методом фазовой модуляции, после чего информация, разделенная за разными полосами в соответствии с системой FDMA, передается на спутник. Основным преимуществом технологии FDMA сравнительно с TDMA является простота. Каналы FDMA не требуют синхронизации или централизованного разделения времени. Каждый из каналов независимый от других.

Эта технология эквивалентна методам разделения частотных каналов для систем беспроводной связи.

Уплотнение/множественный доступ с временным распределением. На рис. 6.36 приведен случай, когда ресурс связи разделен путем предоставления каждому из M сигналов (или пользователей) всего спектра на протяжении небольшого отрезка времени, который называется *временным интервалом* (time slot).

Промежутки времени, которые разделяют используемые интервалы, называются *защитными интервалами* (guard time). Защитный интервал создает некоторую временную неопределенность между соседними сигналами, играя роль буфера и снижая тем самым интерференцию.

Рис. 6.36. Уплотнение с временным распределением

Пример использования технологии TDMA в спутниковой связи приведен на рис. 6.37.

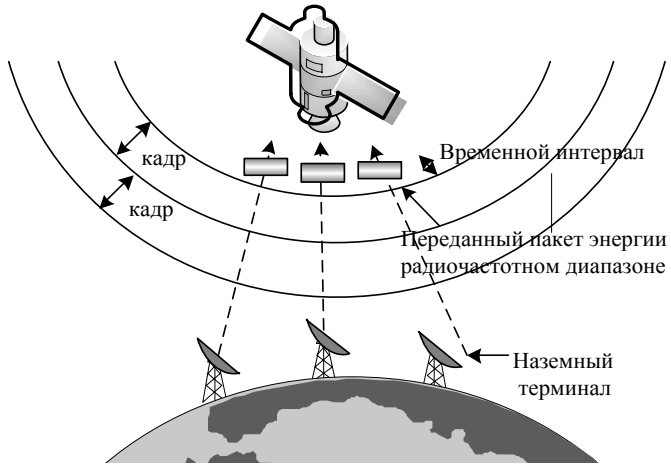


Рис. 6.37. Базовая конфигурация TDMA

Время разбивается на интервалы, которые называют *кадрами* (frame). Каждый кадр разделяется на временные интервалы, которые можно распределить между пользователями. Общая структура кадров периодически повторяется, так что передача данных за схемой TDMA - это один или более временных интервалов, которые периодически повторяются на протяжении каждого кадра.

Каждая наземная передаточная станция транслирует информацию в виде пакетов таким способом, чтобы они поступали на спутник согласно с установленным расписанием. После принятия транспондером такие пакеты ретранслируются на Землю вместе с информацией от других передаточных станций. Приемная станция детектируется и разуплотняет уплотненные данные соответствующего пакета, после чего информация поступает к соответствующим пользователям.

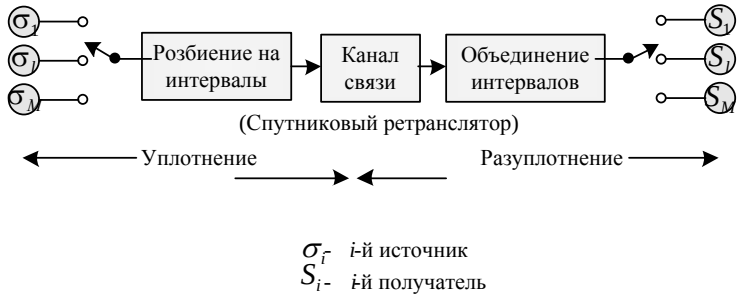


Рис. 6.38. TDM с фиксацией уплотнения

Самой простой среди схем TDM/TDMA является схема с фиксированным разделением. При использовании такой схемы M временных интервалов, образующих кадр, предварительно распределенны между источниками сиг-

нала на достаточно длительный промежуток времени. Работу такой системы показывает наглядно схема, изображенная на рис. 6.38.

Операция уплотнения заключается в предоставлении каждому источнику возможности использовать один или больше интервалов. Разуплотнение - это распознавание интервалов с последующим делением данных между соответствующими пользователями. Два коммутирующих ключа на рис. 6.39 должны быть синхронизированны таким образом, чтобы сообщение, которое отвечает источнику 1, попадало на выход канала 1, и т.д.

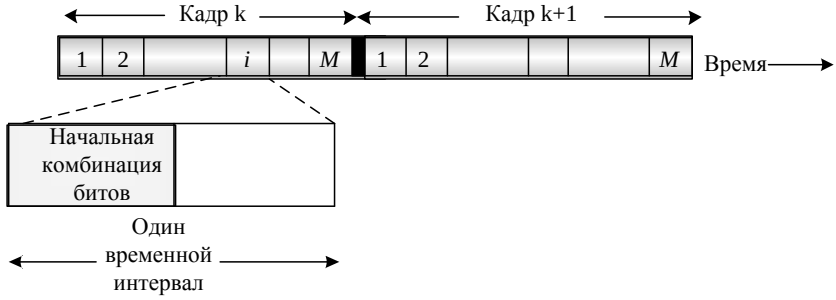


Рис. 6.39. TDM с фиксацией уплотнения

Именно сообщение в общем случае состоит из начальной комбинации битов (preamble) и собственно информационной части. Начальная комбинация обычно состоит из элементов, которые отвечают за синхронизацию, адресацию и защиту от ошибок.

Схема TDM/TDMA с фиксированным распределением чрезвычайно эффективна, когда требования пользователя можно предусмотреть, а поток данных значительный (то есть временные интервалы практически всегда заполнены). Однако в случае пульсирующего или случайного потока данных отмеченный метод себя не оправдывает.

Рассмотрим простой пример (рис. 6.40). Здесь кадр образуют четыре интервала, каждый из которых закреплен за пользователями *A*, *B*, *C* и *D*. Схемы активности четырех пользователей приведены на рис. 6.40, *а*.

На протяжении первого интервала передачи кадра пользователь *C* не отправляет данные, пользователь *B* не передает данных на протяжении второго интервала, а пользователь *A* - на протяжении третьего.

В случае использования TDMA с фиксированным разделением все интервалы кадра распределены предварительно. Если «владелец» интервала не передает данные на протяжении отмеченного промежутка времени, этот интервал не используется. На рис. 6.40, *б* изображен поток данных и неиспользованные интервалы. Если требования пользователей не предусмотрены, как в приведенном примере, то должны применяться более эффективные методы с применением динамического разделения интервалов. Именно о таких методах идет речь при применении систем с коммутацией пакетов, статистиче-

ских мультиплексоров или концентраторов. Соответствующие системы дают возможность достичь результата, изображенного на рис. 6.40, *в*, где пропускная способность системы остается постоянной благодаря использованию всех доступных временных интервалов.

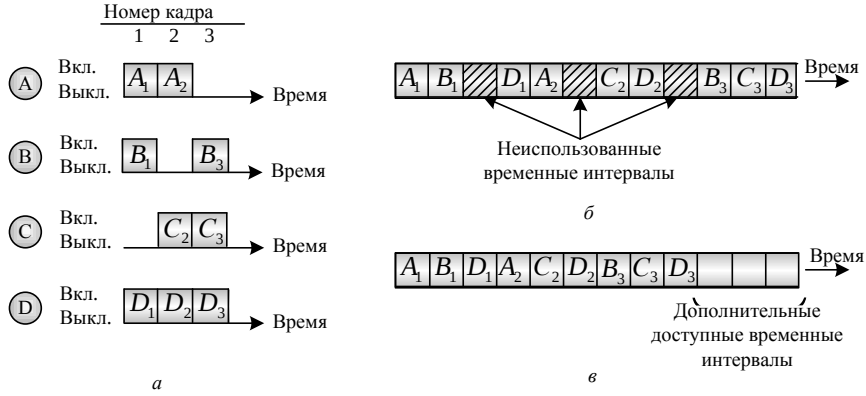


Рис. 6.40. TDM с фиксированным распределением и система с коммутацией пакетов: *а* - схема активности пользователей; *б* - TDM с фиксированным распределением; *в* - коммутация пакетов с временным распределением

Многоканальное деление ресурса связи. Общий способ управления ресурсом связи, который дает возможность разделять частотные диапазоны на предварительно определенный период времени. Такую систему множественного доступа называют *комбинированными FDMA/TDMA.*

Для определения деления частотных диапазонов рассмотрим случай равномерного пропорционального деления полосы шириной W между M группами (или классами) пользователей. Соответственно будем считать, что частотный диапазон разбит на полосы шириной W / M Гц, которые будут постоянно доступны соответствующим группам. Аналогично для назначения деления временных интервалов ось времени разбивается на интервалы длительностью T. В свою очередь каждый из кадров разбивается на N интервалов длительностью T/N каждый. Допустим, что активность пользователей синхронизирована по времени, а распределенные интервалы периодически расположены в кадрах. Каждый пользователь может передавать данные, когда начинается его интервал времени, причем на протяжении этого интервала пользователь может использовать выделенную полосу частот. Временной интервал однозначно задается как m-й интервал кадра n. Обратившись к рис. 6.41, можно описать интервал (n, m) таким способом: область сигнала является пересечением временного интервала (n, m) и частотного диапазона j. Допустим, что система модуляции / кодирования выбрана таким образом, что полная полоса W ресурса связи может поддерживать скорость передачи данных

R бит/с. Для любого частотного диапазона, который содержит полосу W/M Гц, соответствующая скорость передачи данных составляет R/M бит/с.

Временной интервал $(n,\ m)$: $nT + \dfrac{(m-1)T}{N} \le t \le nT + \dfrac{mT}{N},\ n = 0,1...,$ $m = 1,2...,\ N.$

Технология FDMA дает возможность использовать M диапазонов с шириной, которая равняется полной ширине полосы ресурса связи, а TDMA - полный диапазон частот для каждого из N интервалов времени, при этом длительность каждого интервала составляет $1/N$ длительности кадра.

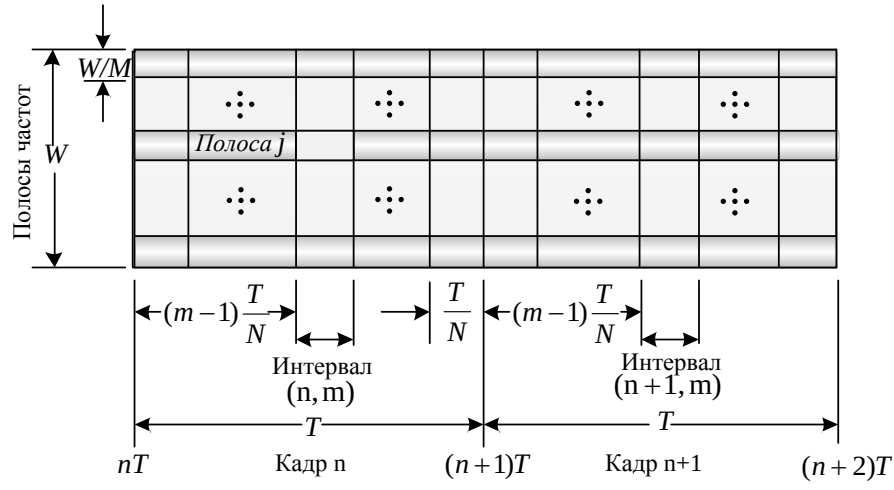


Рис. 6.41. Ресурсы связи: временно-частотное распределение по каналам

Сравнение производительности FDMA и TDMA. Скорость передачи данных FDMA и TDMA. Основные отличия между системами FDMA и TDMA для ресурса связи, который поддерживает скорость передачи данных R бит/с, иллюстрирует рис. 6.42. В случае FDMA полоса системы разделена на M ортогональных полос частот. Следовательно, все M источников $\sigma\ (1 \le m \le M)$ могут одновременно передавать данные со скоростью R/M бит/с каждое (см. рис. 6.42, *а*).

Согласно TDMA время разделено на M ортогональных временных интервалов - один пакет на интервале времени (рис. 6.42, *б).*

Таким образом, каждое M источников передает данные со скоростью R бит/с, что в M раз больше скорости передачи от пользователя FDMA за время $1/M$.

В обоих случаях источник передает информацию со средней скоростью R/M бит/с. Пусть информация, переданная каждым источником (см. рис. 6.42, *а*, *б),* собирается в b-битовые группы или пакеты. В случае FDMA b-битовые пакеты передаются за T секунд по каждому из M непересекаемых

каналов. Таким образом, полную скорость передачи данных можно подать в таком виде:

$$R_{FD} = M\frac{b}{T} \text{ бит/с.} \tag{6.3}$$

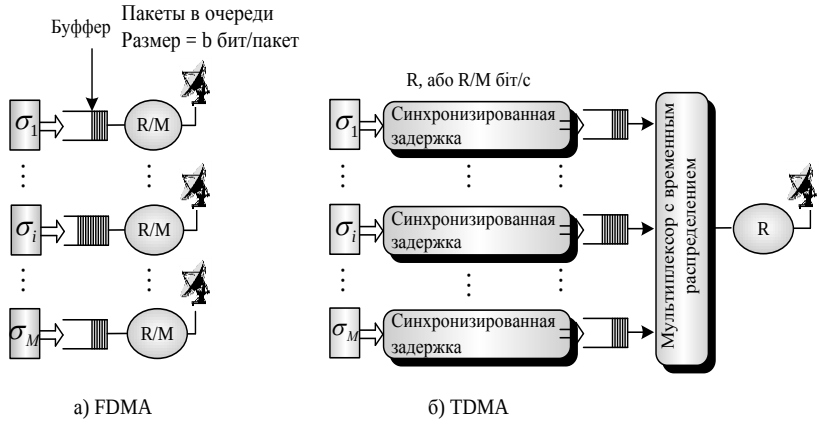


Рис. 6.42. Сравнительное представление технологии FDMA/TDMA:
а - FDMA: частота делится на M ортогональных частотных диапазонов;
б - TDMA: время делится на M ортогональных временных интервалов
(один пакет на интервале времени)

При использовании TDMA каждым источником при T / M секундах передается b бит. Следовательно, необходимая скорость передачи данных

$$R_{TD} = \frac{b}{T/M} \text{ бит/с.} \tag{6.4}$$

Поскольку равенства (6.3) и (6.4) идентичны, то

$$R_{FD} = R_{TD} = R = \frac{Mb}{T} \text{ бит/с.} \tag{6.5}$$

Следовательно, обе системы требуют одинаковой скорости передачи данных, а именно R бит/с.

Задержка сообщений в системах FDMA и TDMA. На основании только что изложенного приходим к выводу, что, несмотря на некоторые разногласия, FDMA и TDMA не отличаются производительностью. Однако отличие между ними становится очевидным, если за единицу измерения производительности взять время средней задержки пакета. TDMA имеет значительное преимущество по сравнению с FDMA по этому параметру, поскольку среднее время задержки пакета при использовании первой схемы меньше, чем при использовании второй.

Как и раньше, допустим, что при FDMA диапазон частот системы разбит на M ортогональных полос. При использовании TDMA кадр разделен на M

ортогональных временных интервалов. Для анализа времени задержки сообщения рассмотрим самый простой случай детерминированных источников данных. Допустим, что ресурс связи используется на 100 %. Тогда все частотные диапазоны при FDMA и все временные интервалы при TDMA будут заполнены пакетами данных. Для упрощения будем считать, что отсутствуют дополнительные расходы, связанные с защитными полосами или интервалами. Тогда время задержки сообщения

$$D = w + \tau \tag{6.6}$$

где w - среднее время ожидания пакета (до передачи); τ - время передачи пакета. При FDMA каждый пакет пересылается на протяжении T секунд; тогда время передачи пакета

$$\tau_{FD} = T. \tag{6.7}$$

В случае использования TDMA каждый пакет пересылается на протяжении часового интервала T/M секунд. С помощью уравнения (6.7) время передачи пакета

$$\tau_{TD} = \frac{T}{M} = \frac{b}{R}. \tag{6.8}$$

Поскольку каналы FDMA доступны постоянно, а пакеты пересылаются непосредственно после создания, время ожидания

$$w_{FD} = 0. \tag{6.9}$$

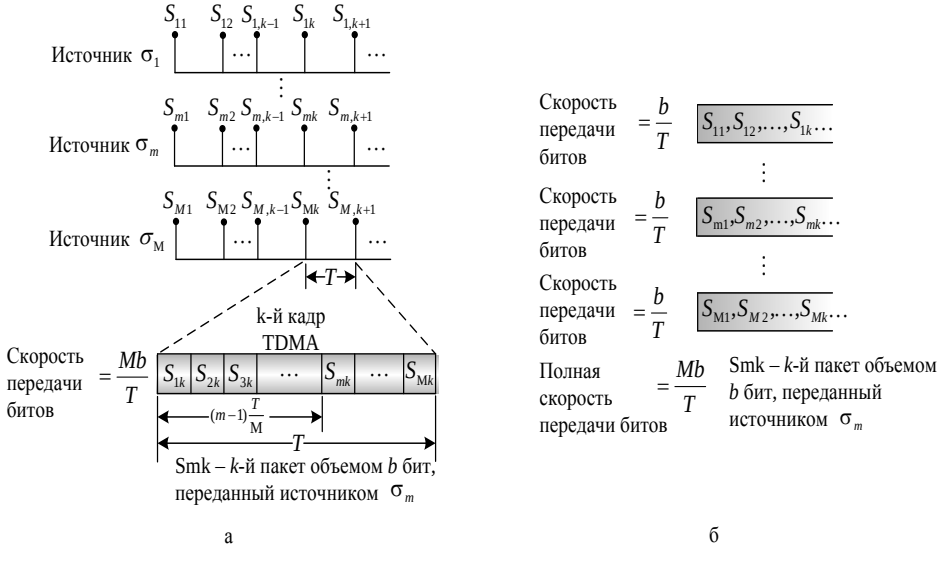


Рис. 6.43. Распределение по каналам: *а* - TDMA; *б* – FDMA

На рис. 6.43 сравниваются потоки данных для схем FDMA и TDMA. Как показано на рис. 6.43, *а*, при использовании TDMA временные интервалы

пользователей начинаются в разных точках кадра длиной T секунд. Пакет S_{mk} начинает отправляться по окончании $(m-1)\,T/M$ секунд $(1 \le m \le M)$ после создания пакета. Таким образом, для TDMA среднее время ожидания пакета перед отправлением

$$w_{TD} = \frac{1}{M}\sum_{m=1}^{M}(m-1)\frac{T}{M} = \frac{\mathrm{T}}{\mathrm{M}^2}\sum_{n=0}^{M-1} n = \frac{T}{M^2}\frac{(M-1)M}{2} = \frac{T}{2}\left(1-\frac{1}{M}\right). \qquad (6.10)$$

Максимальное время ожидания пакета перед отправлением составляет $(M-1)\,T/M$ секунд, а среднее время задержки пакета:

$$1/2(M-1)\,T/M = (T/2)\,(1-1/M). \qquad (6.11)$$

Сравним среднее время задержки D_{FD} и D_{TD} при использовании FDMA и TDMA:

$$D_{FD} = T, \qquad (6.12)$$

$$D_{TD} = \frac{T}{2}\left(1-\frac{1}{M}\right)+\frac{T}{M} = D_{FD} - \frac{T}{2}\left(1-\frac{1}{M}\right). \qquad (6.13)$$

С помощью уравнения (6.7) формулу (6.13) можно записать в виде

$$D_{TD} = D_{FD} - \frac{b}{2R}(M-1). \qquad (6.14)$$

Результат показывает, что FDMA значительно уступает TDMA по времени задержки сообщения. Несмотря на то, что уравнение (6.3) выполняется для источника детерминированных данных, малые задержки передачи сообщений для TDMA сохраняются для любого независимого процесса получения данных.

Множественный доступ с кодовым разделением. В случае FDMA плоскость ресурса связи была разделена на горизонтальные отрезки, которые соответствуют частотным диапазонам. Ту же плоскость на рис. 6.38 было разбито по вертикалям на временные интервалы TDMA. Эти два подхода являются самыми распространенными в применении множественного доступа. Метод множественного доступа, который является результатом сочетания FDMA и TDMA, иллюстрирует рис. 6.44.

Этот метод называется ***множественным доступом с кодовым разделением*** (code-division multiple access - CDMA).

Множественный доступом с кодовым разделением является практическим воплощением методов расширения спектра (spread-spectrum - SS), которые можно разделить на две основные категории:

расширение спектра методом прямой последовательности (direct sequence - DSSS);

расширение спектра методом скачкообразной перестройки частоты (frequency hopping - FHSS).

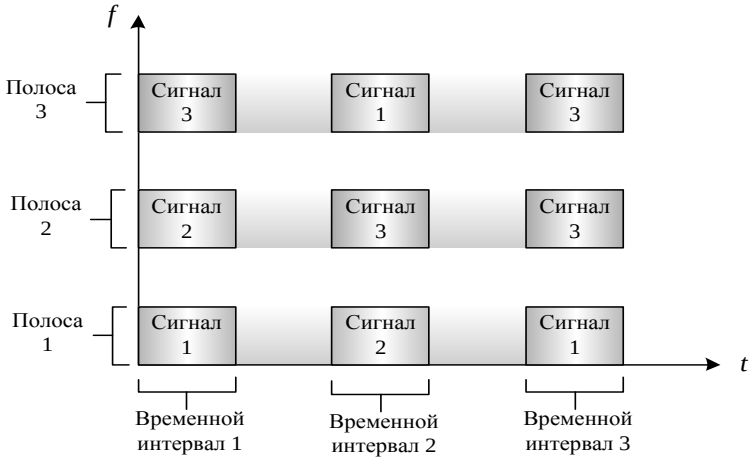


Рис.6.44. Уплотнение с кодовым распределением

Эти методы разделения каналов на базе расширения спектра передаваемого сигнала вполне идентичны относительно организации разных режимов работы беспроводных сетей и детально рассмотрены в предыдущем разделе.

Процесс модуляции с использованием перестройки частоты иллюстрирует рис. 6.45.

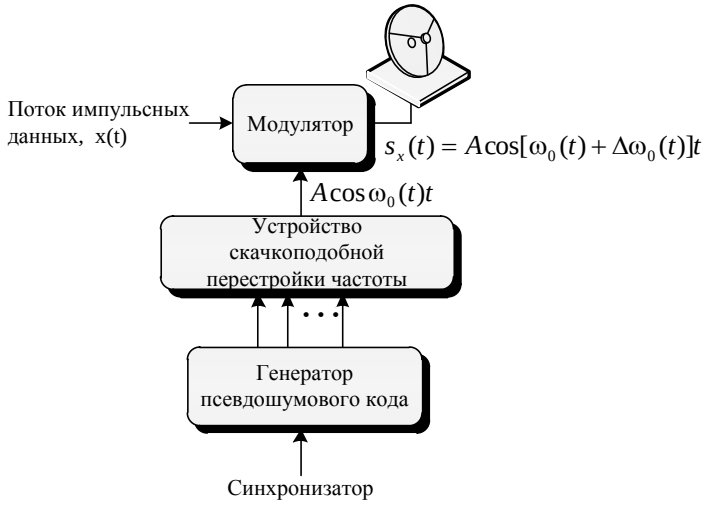


Рис. 6.45. Процесс модуляции схемы FH-CDMA

Во время каждого изменения частоты генератор псевдошумовой последовательности направляет кодовую последовательность на устройство скачкообразной перестройки частоты. Это устройство выдает одну из допустимых для

прыжка частоту. Будем считать, что используется *М*-я частотная манипуляция (M frequency shift keying - MFSK). В случае обычной системы MSFK данные модулируют несущую волну с фиксированной частотой. В случае MFSK с перестройкой частоты (FH-MFSK) частота несущей смещается по всему диапазону частот передачи. FH-модуляцию (см. рис. 6.45) можно представить как процесс, который состоит из модуляции данных и модуляции перестройки частоты. Отмеченные действия можно совместить. Тогда модулятор на основе псевдошумового кода и собственно данных будет генерировать тона передачи.

Конфиденциальность передачи *информационных потоков данных в случае использования смешанного метода CDMA является основным и уникальным преимуществом этого метода множественного доступа.*

Если код группы пользователей известен лишь разрешенным членам этой группы, CDMA обеспечивает конфиденциальность связи, поскольку несанкционированные лица, которые не имеют кода, не могут получить доступ к переданной информации.

Каналы с замираниями. Для определенной части используемого спектра характерны замирания, сигналы в этой части диапазона будут ослабленными. В случае применения схемы FDMA пользователь части спектра может испытывать постоянные трудности со связью. Согласно схеме FH-CDMA пользователь будет испытывать аналогичные трудности только при смещении частоты в соответствующую часть спектра. Таким образом, возможные проблемы со связью равномерно распределяются между всеми пользователями.

Информационный поток в системах множественного доступа. На рис. 6.46 представлена блок-схема потока данных относительно реализации алгоритма множественного доступа (multiple access algorithm - MAA), между контролером и наземной станцией связи.

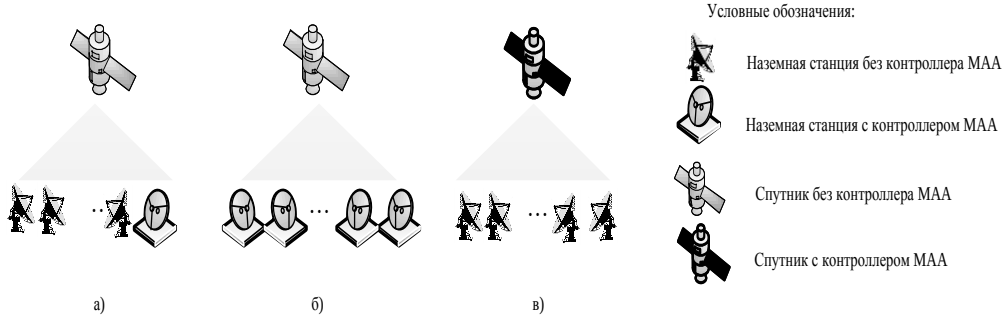


Рис. 6.46. Архитектура спутниковой системы множественного доступа:
а - управление осуществляет одна базовая наземная станция;
б - управление распределено между всеми наземными станциями;
в - управление осуществляет спутник

Направлять и руководить информационными потоками может непосредственно спутник или одна наземная базовая станция. Управление может быть распределено также между всеми наземными станциями. Порядок передачи данных приведен на рис. 6.47.

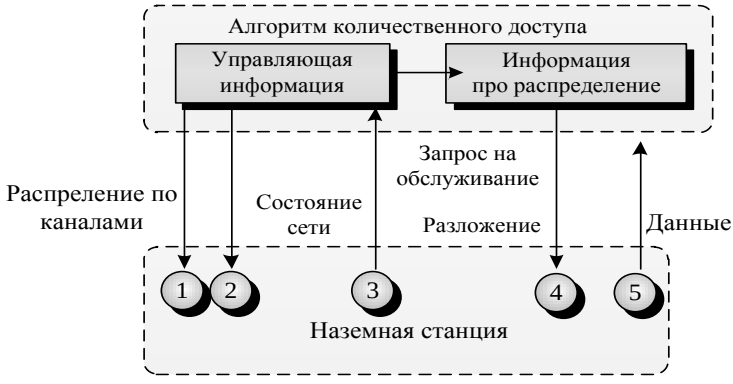


Рис. 6.47 Информационный поток в системах количественного доступа

Деление информационного потока по каналам - *деление информационных ресурсов и данных, осуществляемое при обмене информацией между спутником и пользователем* (*например, каналы* N *могут быть предоставлены пользователю* $\tilde{O}$, *а каналы* $(N+1) = M$ - *пользователю* Y).

Эта информация изменяется редко и может распространяться между наземными станциями без использования системы связи, например с помощью информационного бюллетеня деления.

Состояние сети (*network state - NS*) - это *состояние ресурса связи, то есть время и ширина полосы частот, доступных для передачи информационного сигнала в определенную информационно-коммуникационную систему.*

Наземная станция получает указания относительно доступности ресурса связи, а также о том, как стоит использовать время, частоту, кодовые позиции ресурса для передачи запроса на обслуживание.

Запросом на обслуживание *называется переданное станцией кодовое сообщение (запрос) на выделение ресурса связи для передачи т сегментов информационного сигнала.*

После получения запроса (запросов) на обслуживание контроллер передает станции расписание, в соответствии с которым данные должны распределяться в ресурсе связи.

Станция передает данные в соответствии с отмеченным расписанием.

Множественный доступ с предоставлением каналов по требованию.

Спутниковыми системами с фиксированным делением *называются системы множественного доступа, которые дают возможность передаточ-*

ной станции периодически получать доступ к информационному каналу независимо от реальных потребностей пользователя или системы.

Спутниковыми системами с предоставлением каналов по требованию (*demand-assignment multiple access - DAMA*) *называются системы множественного доступа с динамическим делением, которые предоставляют доступ к информационному каналу только на соответствующий запрос передаточной станции.*

Если передача данных станцией связи ведется нерегулярно или скачкообразно, схема DAMA может быть значительно эффективнее схемы фиксированного деления. Полезность схемы DAMA объясняется тем, что фактическая потребность в ресурсах редко совпадает с максимальным спросом. Однако система с низшей пропускной способностью, использующая буферизацию и схему DAMA, может успешно поддерживать скачкообразный процесс обмена данными, хотя в этом случае все-таки возможны некоторые задержки передачи данных. На рис. 6.48 обобщаются основные разногласия между системой с фиксированным делением, пропускная способность которой равняется сумме требований всех пользователей, и динамической системой, пропускная способность которой определяется средними требованиями пользователей.

Использование методов спутниковой коммутации информационных потоков. Современные спутники связи обычно используют несколько лучей, которые обеспечивают покрытие в определенном регионе. Примером является спутник, который находится над Атлантическим океаном, отдельные лучи направляются в Северную Америку, Европу, Южную Америку и Африку. Для взаимосвязи станций разных регионов используются коммутаторы.

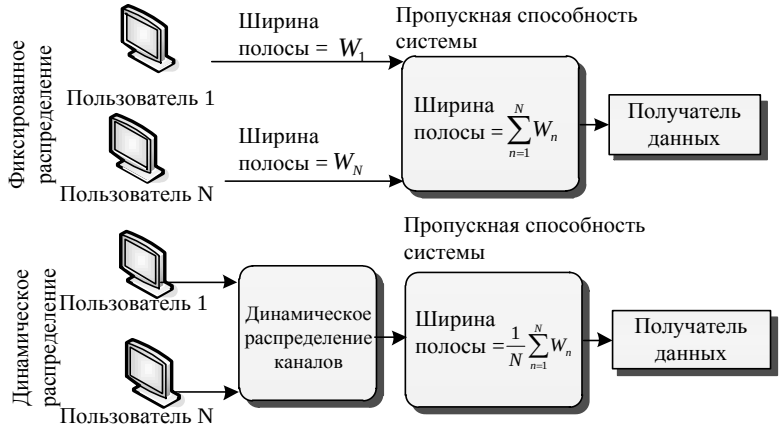


Рис. 6.48. Уменьшение ширины полосы для систем с динамическим распределением каналов

Спутниковая коммутация (*satellite-switched TDMA - SS/TDMA*) - *это процесс обеспечения эффективной циклической взаимосвязи обмена инфор-*

мационными потоками и данными (*с временным разделением доступа к информационным ресурсам) с зонами перекрытия диаграмм направленности лучей разных спутников.*

Основой системы является размещенная на спутнике микроволновая матрица коммутации, программируемая с помощью наземного управления на циклическую смену состояний. Таким образом, в каждый момент коммутации связываются раздельные лучи каналов Земля - спутник. Наземная станция может связаться со станциями, которые используют другой луч, посылая пакеты TDMA во время соответствующих выделенных интервалов времени. Схема коммутации состояний выбирается так, чтобы максимально увеличить пропускную способность системы с учетом имеющихся ограничений на обмен данными. Для достижения полной взаимосвязи *N* лучей нужно *N!* разных состояний, или режимов, спутника. Шесть режимов, необходимых для полной взаимосвязи трехлучевой системы, приведены в табл. 6.3.

Таблица 6.3

Вход	Выход					
	Режим 1	Режим 2	Режим 3	Режим 4	Режим 5	Режим 6
A	*A*	*A*	*B*	*B*	*C*	*C*
B	*B*	*C*	*A*	*C*	*A*	*B*
C	*C*	*B*	*C*	*A*	*B*	*A*

В режиме 1 приемники спутника на лучах *A*, *B* и *C* соединены с передатчиками для лучей *A*, *B* и *C*. *Наземная* станция, которая использует один из этих лучей, может связаться с другой станцией, которая использует тот же луч. Такой луч называют *самоориентированным.*

Пример трехлучевой (лучи *A*, *B* и *C*) системы SS/TDMA приведен на рис. 6.49. Микроволновая матрица коммутации для данного спутника является *координатной.* Такую матрицу можно представить как набор продольных и поперечных линий. При активизации линий (одной продольной и одной поперечной) возникает контакт на их пересечении. Координатный коммутатор дает возможность одновременно устанавливать связь только между двумя компонентами матрицы - одним продольным и одним поперечным. Если канал станции A_U связан с каналом станции B_D, ни один из этих каналов не может быть одновременно связан с любым другим каналом. Три схемы процедуры обмена данными на протяжении интервалов времени T_1 T_2 и T_3 при существовании трех состояний коммутации S_1 S_2 и S_3 изображены на рис. 6.49. На протяжении интервала T_1 имеем режим S_1: лучи самоориентированы. На протяжении интервала T_2 режим S_2 дает возможность передавать сигналы из станций A_U, B_U и C_U на станции B_D, C_D и A_D. На протяжении интервала T_3 (режим S_3) каналы передачи так же связываются с каналами приема, обеспечивая доставку данных необходимому адресату.

Схемы процедуры обмена данными, а также их длительность выбираются с целью оптимизации пропускной способности спутника и максимально

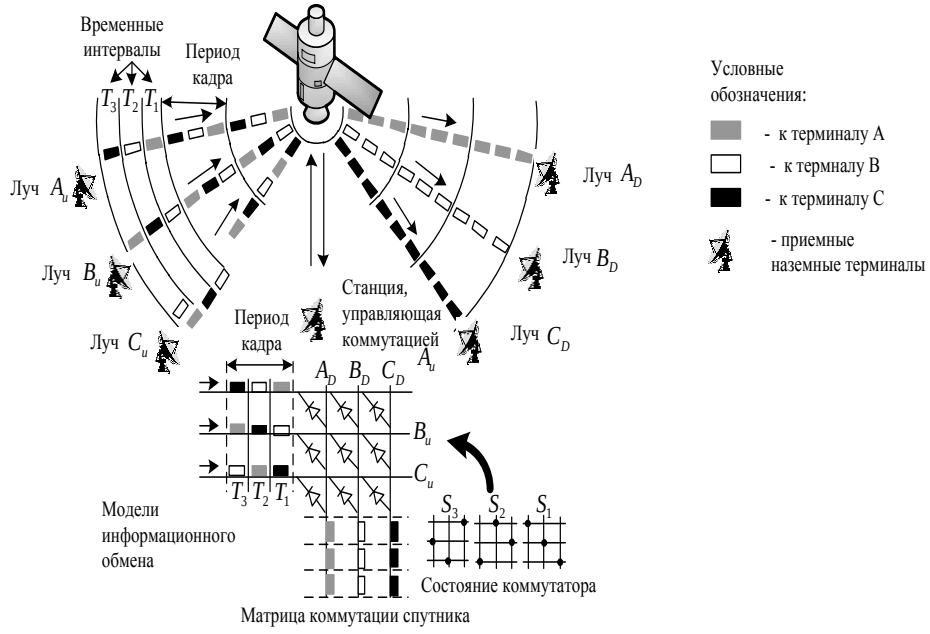


Рис. 6.49. TDMA со спутниковой коммутацией (satellite switched TDMA-SS/TDMA)

эффективного обслуживания пользователей. Для учета изменений в информационном потоке циклическая схема в случае потребности может изменяться наземной станцией.

Матрица информационного обмена. Матрица, которая характеризует обмен данными между N областями, которые обслуживаются сфокусированным лучом, приведена на рис. 6.50. Промежуточная сумма

$$S_i = \sum_{j=1}^{N} t_{ij} \tag{6.15}$$

является полным информационным потоком от *i*-го луча наземной станции, а

$$R_j = \sum_{i=1}^{N} t_{ij} \tag{6.16}$$

- полным информационным потоком к *j*-му лучу наземной станции; t_{ij} - объем информационного потока от луча i к j.

Если обменом данными системы SS/TDMA руководит коммутатор, который не блокирует (позволяет передачу всех сообщений без выдачи любого аналога сигнала «занято»), каждому каналу в кадре TDMA назначается временной интервал длительностью k секунд. Для эффективного использования

ресурса связи полный информационный обмен на рис. 6.50 должен состояться на протяжении времени кадра T, который должен быть малейшим. Минимальное время передачи кадра для обеспечения такой неблокировочной связности

$$T_{\min} = k \max\left(\{S_i\},\{R_j\}\right). \tag{6.17}$$

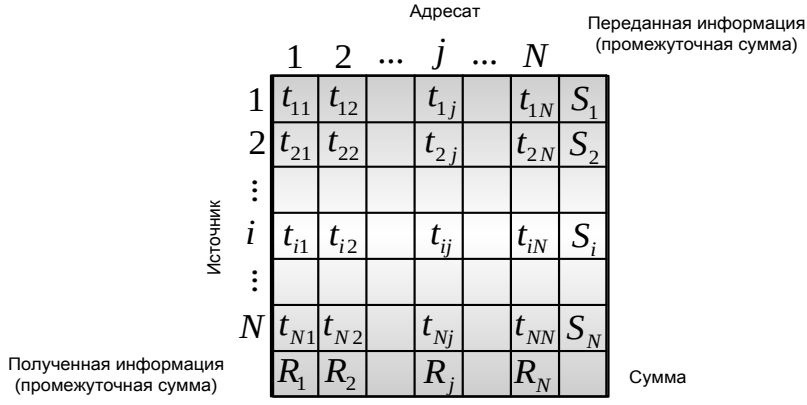


Рис. 6.50. Матрица информационного обмена

Здесь $\max\left(\{S_i\},\{R_j\}\right)$ - максимальное значение, взятое из всех возможных $\{S_i\}$ и $\{R_j\}$. Выражение (6.17) описывает минимальное время, необходимое для передачи *всех* данных *всем* адресатам (оба условия отмечены в матрице информационного обмена), если все каналы имеют полосы одинаковой ширины.

Системы связи множественного доступа и их архитектура. Информация об использовании времени, частоты и кодовых функций, необходимая пользователям для общения между собой с помощью спутника, содержится в протоколе или ***алгоритме множественного доступа*** (multiple access algorithm - MAA).

Система множественного доступа - *это система, которая объединяет аппаратное и аппаратно-программное обеспечение, которое поддерживает алгоритм множественного доступа (протокол доступа) с целью своевременной упорядоченности и эффективной передачи информационных потоков и данных между станцией спутниковой связи и пользователем.*

Основные структуры спутниковых систем связи множественного доступу приведены на рис. 6.46. В условных обозначениях даны символы, используемые для наземных станций, которые имеют или не имеют контроллера алгоритма доступа. Система, в которой одна из наземных станций определяется как основная (контроллер), приведена на рис. 6.46, *а*. На этой станции размещают сервер, который реагирует на запросы относительно обслужива-

ния, которые поступают от всех других пользователей. Запрос влечет передачу данных от контроллера к спутнику и назад. Реакция контроллера приводит к другой передаче с помощью спутника. Таким образом, каждая услуга требует двух сеансов передачи данных из Земли на спутник и назад. Случай деления управления согласно с алгоритмом доступа между всеми наземными станциями (выделенного контролера не существует) иллюстрирует рис. 6.46, *б*. Все наземные станции используют одинаковый алгоритм и имеют в своем распоряжении идентичные знания о запросе относительно доступа и алгоритма деления доступа. Следовательно, каждая услуга в этом случае нуждается в одном цикле связи «станция - спутник - станция». Контроллер алгоритма доступа, который находится непосредственно на спутнике, изображен на рис. 6.46, *в*. Запрос пользователя поступает на спутник, который может немедленно послать соответствующий сигнал. Таким образом, в этой системе для предоставления услуги связи достаточно одного цикла связи.

Система наземных станций - SCPC-система (рис. 6.51). Большинство сетей наземных станций работают согласно принципу стандартного варианта связи, где используется связь по схеме «point-to-point» («точка-точка»), - две наземные станции, соединенные спутниковым каналом и расположенные у пользователей. При наличии такого канала пользователи могут устанавливать связь друг с другом в любой момент. Чаще приходится иметь дело с конфигурацией сети типа «звезда» (принцип «центр с каждым»). В случае использования этой схемы возможна организация потоков цифровой информации со скоростью от 32 кбит/с до 8 Мбит/с и обеспечение телефонной, телефаксовой связи между центром и периферией.

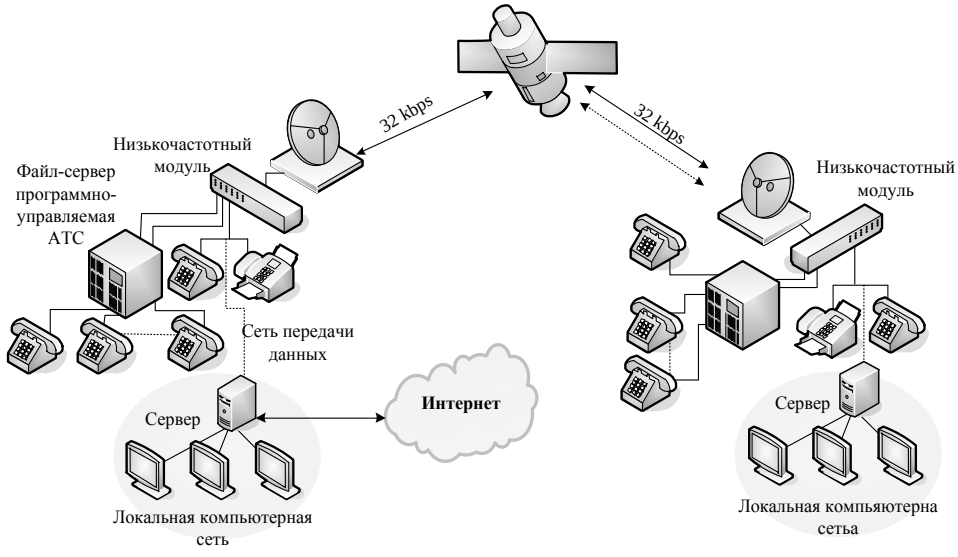


Рис. 6.51. Схема работы SCPC - системы

Такая система имеет одно из основных преимуществ - передачу больших объемов информации с высокой скоростью. Благодаря использованию спутниковых цифровых каналов она является некритической относительно дальности и достаточно помехозащищенная.

Наземные системы полного доступа (TES-система). ***Наземная система полного доступа*** *(Telecommunications Earth STATION-TES) предназначена для обмена телекоммуникационной и цифровой информацией в информационных сетях спутниковой связи, построенных по принципу «каждый с каждым» («mesh») или в сетях с полным доступом* (рис. 6.52).

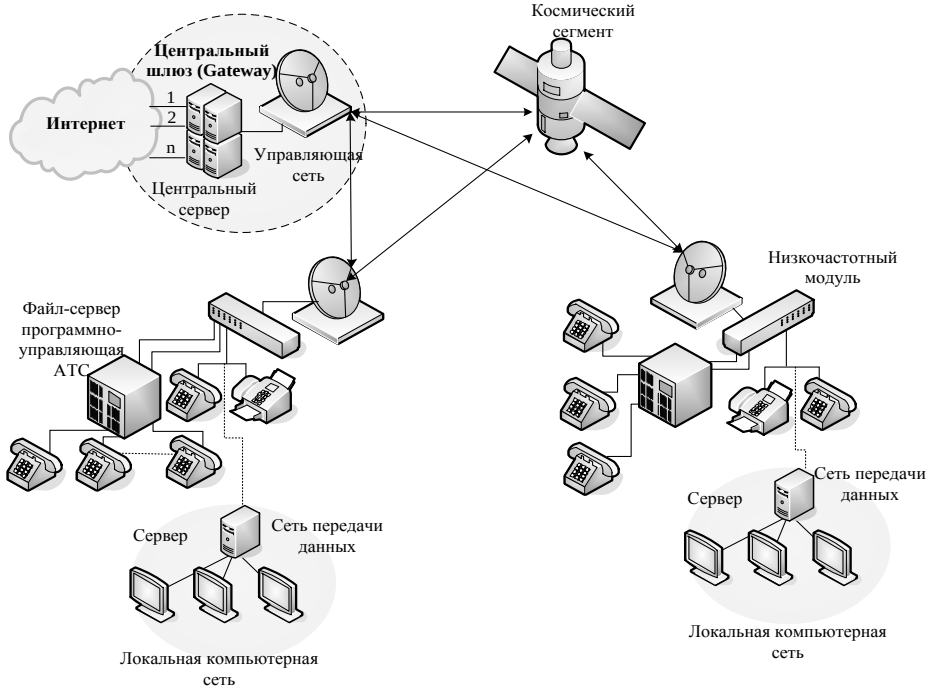


Рис. 6.52. Схема работы TES - системы

Эта система дает возможность устанавливать телекоммуникационную связь между любыми двумя абонентами сети. Кроме того, абонентам обеспечивается выход в международные глобальные информационно-коммуникационные сети общего пользования через ***телепорт*** (*Gateway*). В самой простой конфигурации обеспечивается связь по одному телефонному или факсимильному каналу.

Абоненту предоставляется дополнительная возможность организации передачи цифровой информации между двумя станциями, которые входят в сеть. Сеть работает по принципу, согласно которому абонент не имеет жестко закрепленного за ним спутникового канала (DAMA), а этот канал предостав-

ляется ему по первому требованию (запрос), причем с высокой (свыше 99 %) вероятностью коммутации.

Этот способ дает возможность уменьшить количество задействованных спутниковых каналов. В целом использование именно TES-системы является наиболее оперативным и действенным способом доступа в международные телекоммуникационные сети, а также оптимальным средством связи с теми территориальными областями, которые имеют неразвитую инфраструктуру связи или вообще не имеют ее.

Система персональных наземных станций (PES). Наземная система имеет в своем составе центральную станцию (HUB station) со многими периферийными (PES) или отдаленными станциями. Эффективность излучения, большая мощность и высокое качество приема центральной станции делает возможным применение на периферийных станциях малых антенн диаметром 0,5…1,8 м и маломощных (0,5...2 Вт) передатчиков. Это значительно снижает стоимость абонентской системы связи. В отличие от других упомянутых систем в этой системе передача информации всегда происходит через HUB-станцию. С точки зрения энергетики системы и ее стоимости (соответственно и стоимости предлагаемых услуг) оптимальным является расположение центральной базовой стации в центре зоны освещения спутника.

Система персональных наземных станций (*Personal Earth STATION-PES*) - *спутниковая диалоговая пакетно-коммутированная сеть, предназначенная для обмена телекоммуникационной и цифровой информацией в рамках системы спутниковой связи с топологией типа «звезда» с возможностью полного дуплекса* (рис. 6.53).

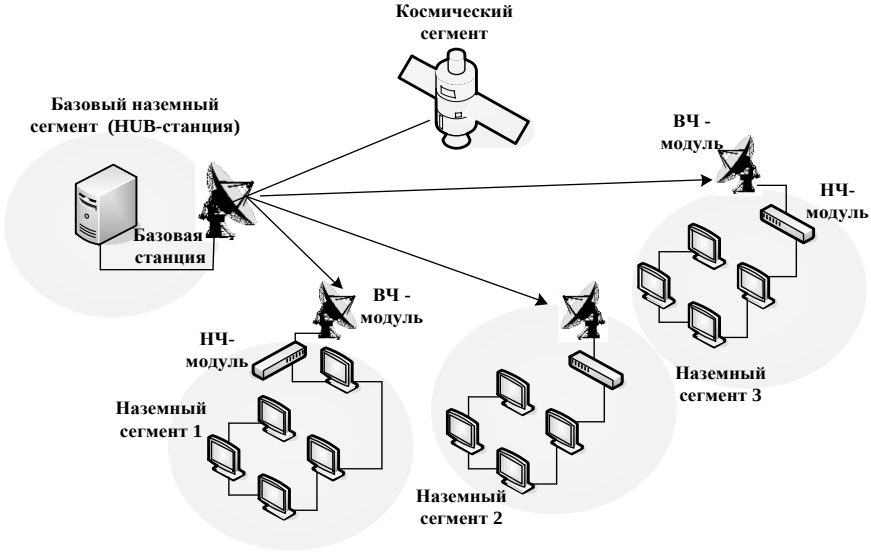


Рис. 6.53. Структура PES - системы

Основные выводы

Информационно-коммуникационная сеть - это интегрированный комплекс взаимосвязанных и согласованно функционирующих программных и аппаратных компонентов, которые обеспечивают достоверную передачу информации от источника сообщения к потребителю.

Многослойной моделью информационно-коммуникационной сети называется полный комплекс программно-аппаратных средств сети, который применяется с целью достоверной передачи информационных потоков от источника сообщения к потребителю.

Топологией информационно-коммуникационной сети называется конфигурация графа, вершинам которого соответствуют компьютеры сети (иногда и другое оборудование, например концентраторы), а ребрам - физические связи между ними.

Сетевая технология - это согласованный набор стандартных протоколов и программно-аппаратных средств, которые их реализуют (например, сетевых адаптеров, драйверов, кабелей и разъемов), достаточных для построения и функционирования информационно-коммуникационной сети.

Для снятия ограничений при построении больших сетей используются специальные методы структуризации сети и специальное структурообразовывающее (сетевое) оборудование - повторители, концентраторы, мосты, коммутаторы, маршрутизаторы. Оборудование такого рода также называют коммуникационным, имея в виду, что с его помощью отдельные сегменты сети взаимодействуют между собой.

Наибольшее распространение получили три схемы адресации узлов информационных сетей: аппаратные адреса; символьные адреса, или имена; числовые составные адреса.

Концепция виртуальных информационно-коммуникационных сетей с целью ограничения размеров доменов коллизий и широковещательных доменов объединяет технологии коммутации уровня 2 и маршрутизации уровня 3 модели OSI.

Передачей с расширенным спектром называется ряд способов передачи отдельного радиосигнала с использованием широкого сегмента радиоспектра. Используются две разные системы радиопередачи с расширенным спектром:

частотное расширение спектра;

расширение спектра с прямой последовательностью.

Информационные сети беспроводной связи включают две категории сетевого оборудования*:* станции и точки доступа.

Станция информационной сети беспроводной связи - это компьютер или другое периферийное устройство, сетевое оборудование, подключенное к

беспроводной сети через внутренний или внешний беспроводный адаптер сетевого интерфейса.

Точка доступа информационной сети беспроводной связи - это базовая станция беспроводной сети и городов между беспроводной и традиционной коммутативной (проводной) сетью.

Существуют два типа беспроводных информационных сетей:

Ad-Hoc-сети;

инфраструктурные сети.

Локальной беспроводной сетью (Ad-Hoc) называется сеть, которая представляет собой автономную группу станций, которая работает без подключения к глобальным сетям передачи данных.

Инфраструктурной беспроводной сетью называется сеть, которая представляет собой группу станций, которая работает с подключением к глобальным сетям передачи данных и информационных ресурсов с использованием одной или большего количества точек доступа.

Станции спутниковой связи используются с целью обмена информацией между наземными информационными объектами, а также в системах сбора и распределения данных. Спутниковая станция с сетью земных станций обеспечивают систему телекоммуникационной связи и передачу информационных потоков данных (в том числе цифровой низкочастотной языковой информации).

Спутниковая станция по конструктивному признаку состоит из таких компонентов:

высокочастотного модуля (ODU);

низкочастотного модуля (IDU).

Существует три основных способа увеличения пропускной способности (общей скорости передачи данных) ресурса связи:

увеличение эффективной изотропно-излучаемой мощности (effective isotropic radiated power - EIRP) передатчика или снижения потерь системы, которая всегда приводит к увеличению отношения сигнал/шум $\left(E / N_0\right)$.

увеличение ширины полосы канала передачи данных;

повышение эффективности разделения ресурса связи на базе множественного доступа.

Разделение в каналах базируется на двух методах:

уплотнение с временным разделением (time-division multiplexing-TDM) или множественным доступом с временным разделением (time-division multiple access-TDMA);

уплотнение с частотным разделением (frequency-division multiplexing - FDM) или множественным доступом с частотным разделением (frequency-division multiple access - FDMA).

Множественный доступ с кодовым разделением является практическим воплощением методов расширения спектра, которые можно разделить на две основные категории:

расширение спектра методом прямой последовательности (direct sequence - DSSS);

расширение спектра методом скачкообразной перестройки частоты (frequency hopping - FHSS).

Конфиденциальность передачи информационных потоков данных в использовании смешанного метода CDMA является основным и уникальным преимуществом этого метода множественного доступа.

Система множественного доступа - это система, которая объединяет аппаратное и аппаратно-программное обеспечение, которое поддерживает алгоритм множественного доступа (протокол доступа) с целью своевременного упорядочения и эффективной передачи информационных потоков и данных между станцией спутниковой связи и пользователем.

Наземная система полного доступа (TES) предназначена для обмена телекоммуникационной и цифровой информацией в информационных сетях спутниковой связи, построенных по принципу «каждый с каждым» («mesh») или в сетях с полным доступом.

Система персональных наземных станций (PES) - спутниковая диалоговая пакетно-коммутированная сеть, предназначенная для обмена телекоммуникационной и цифровой информацией в рамках системы спутниковой связи с топологией типа «звезда» с возможностью полного дуплекса.

Вопросы для самоконтроля

1. Назовите основную топологию и технологии проектирования информационно-коммуникационных сетей.

2. В чем заключается отличие между физической и логической структуризацией сети?

3. Какие основные типы протоколов используются в модели OSI?

4. Перечислите основные функции сетевого и транспортного уровней модели OSI.

5. Что такое IP адрес?

6. Назовите основные требования для проектирования большинства сетевых проектов.

7. Раскройте понятие «сегментация сети».

8. В каких случаях проектирования топологии сети используется вертикальное или горизонтальное кабелирование?

9. Назовите основные особенности и преимущества использования виртуальных информационно-коммуникационных сетей.

10. Назовите основные преимущества применения беспроводных локальных сетей.

11. В чем заключается технология частотного расширения спектра?

12. Какие рабочие режимы систем и сетей беспроводной связи существуют?

13. Какую функцию выполняют беспроводные сетевые интерфейсные адаптеры?

14. Какое назначение широкополосного шлюза?

15. Дайте определение понятия «системы спутниковой связи».

16. Назовите основные преимущества систем цифровой спутниковой связи.

17. Назовите составляющие наземного сегмента систем спутниковой связи.

18. Основные методы кодирования в системах спутниковой связи.

19. Особенности множественного доступа в спутниковых системах.

20. Сравните производительность FDMA и TDMA.

21. Назовите особенности деления информационного потока по каналам связи.

22. Раскройте содержание методов спутниковой коммутации информационных потоков.

23. Назовите основные характеристики системы связи множественного доступа и ее архитектуры.

The main conclusions

The information-communication network is integrated complex of the interconnected and in coordination-functioning software and hardware components that provide authentic transmission of the information from a source of a message to the consumer.

The complete complex of firmware of a network which is used with the purpose of authentic transmission of informational streams from a source of the message to the consumer is called multilayer model of an information-communication network.

The topology of an information-communication network is the configuration of the graph points of which are conformed to computers of a network (sometimes and other equipment, for example concentrators) and ribs are conformed to physical ties between them.

The network technology is a consistent set of standard protocols and firmware that realize them (for example, network adapters, drivers, cables and plugs), sufficient for construction and functioning of an information-communication network.

Special methods of structurization of a network and the special structurally-formative (network) equipment such as repeaters, concentrators, bridges, commutators, routers are used for removal of limitations at construction of great networks.

The equipment of such sort is also called communication, meaning, that separate segments of a network interact among themselves with its help.

Three schemes of addressing of nodes of information networks have got the greatest distribution:

the hardware addresses;

character addresses or names;

numerically made addresses.

The conception of virtual information-communication networks unites technologies of commutation of the Level 2 and routing of the Level 3 of OSI model with the purpose of limitation of the sizes of domains of collisions and broadcasting domains.

Transmission with the extended spectrum is the number of ways of transmission of a separate radio signal with the use of a wide segment of a radiospectrum. Two different systems of a radio transmission with the extended spectrum are used:

the frequency extension of a spectrum;

the extension of a spectrum with direct sequence.

Information networks of wireless link include two categories of the network equipment: stations and points of access.

The station of an information network of wireless link is a computer or other peripheral unit, the network equipment connected to wireless network through the internal or external wireless adapter of the network interface.

The point of access of an information network of wireless link is the base station of a wireless network and bridge between a wireless and traditional commutative (wire) network.

There are two types of wireless information networks:

Ad-Hoc networks;

infrastructure networks.

The local wireless network (Ad-Hoc) is the network that is the autonomous group of stations that works without connection to the global networks of data transmission.

The infrastructure wireless network is the network that is the group of stations that works with connection to the global networks of data transmission and informational resources with the use of one or more points of access.

The station of a satellite communication is used with the purpose of information interchange between ground information objects and also in systems of collection and division of data. The satellite station with a network of earth stations provides a system of telecommunication link and transmission of informational dataflows (including the digital low-frequency language information).

According to a constructive feature the satellite station consists of:

high-frequency module (ODU);

low-frequency module (IDU).

There are three main ways of increasing the capacity (the general data transfer rate) of a resource of connection:

increasing of effective isotropic radiated power (EIRP) of the transmitter or lowering of losses of system that in any case will lead to increasing the relation signal/noise $\left(E/N_0\right)$;

increasing of width of a band of a data communication;

rise of efficiency of division of resource of connection on the basis of multiple access.

The division in channels is based on two methods:

time-division multiplexing (TDM) or time-division multiple access (TDMA);

frequency-division multiplexing (FDM) or frequency-division multiple access (FDMA).

Code-division multiple access is practical addition of methods of the extension of a spectrum that are possible to be divided into two main categories:

the extension of a spectrum by a method of direct sequence (direct sequence - DSSS);

the extension of a spectrum by a method of frequency hopping (frequency hopping - FHSS).

Confidentiality of transmission of information dataflows in the use of mixed method CDMA is the main and unique advantage of the given method of multiple access.

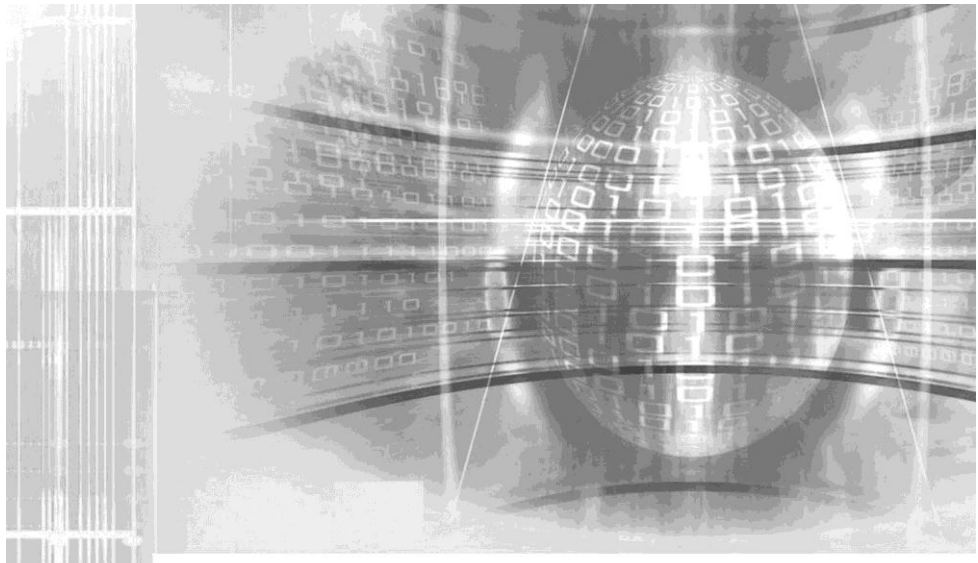
The system of multiple access is a system that unites hardware and a firmware which supports the algorithm of multiple access (the access protocol) with the purpose of timely ordering and effective transmission of information streams and data between the station of a satellite communication and the user.

Total access earth systems (TES) is intended for an exchange of the telecommunication and digital information in information networks of a satellite communication that are built on a principle “everyone with everyone” (“mesh”) or in networks with total access.

Personal earth stations system (PES) is satellite dialogue packet-switching network that is intended for an exchange of the telecommunication and digital information within the system of a satellite communication with topology of "star" type with possibility of a full duplex.

Ключевые слова

Русский	**Английский**
топология информационно-коммуникационной сети	topology of information communication network
сетевая технология	network technology
виртуальная информационно-коммуникационная сеть	virtual information communication network
беспроводная сеть	wireless network



КОДИРОВАНИЕ ИНФОРМАЦИИ

7

7.1. Кодирование источника сообщения и сжатие данных
7.2. Помехоустойчивое кодирование
7.3. Блочное помехоустойчивое кодирование
7.4. Свёрточное помехоустойчивое кодирование

7.1. Кодирование источника сообщений и передача данных

Под *кодированием в широком смысле* понимают процесс преобразования сообщений в сигнал. Как при передаче, так и при хранении и обработке информации значительные преимущества дает дискретная форма представления сигналов. Поэтому в случае, когда начальные сигналы являются непрерывными, происходит, как правило, предыдущее преобразование их в дискретные сигналы. В связи с этим термин «кодирование» относят по обыкновению к дискретным сигналам и под *кодированием в узком смысле* понимают отображение дискретных сообщений сигналами в виде определенных комбинаций символов. Совокупность правил, согласно которым происходят эти операции, называют *кодом*.

Под ***кодированием*** *в общем случае понимают преобразование алфавита сообщения* $\mathrm{A}\{\lambda_i\}$ $(i=1, 2\ldots, K)$ *в алфавит определенным способом выбранных кодовых символов* $R\{x_j\}$ $(j=1, 2\ldots, N)$. *По обыкновению* (*но не обязательно*) *размер алфавита кодовых символов* $\dim\{x_j\}$ *меньший, чем размер алфавита источника* $\dim A\{\lambda_i\}$.

Кодирование сообщений может иметь разные цели, но в информационных сетях передачи данных существует последовательность использования того или другого метода кодирования в зависимости от поставленных перед системой заданий и технических требований. Рассмотрим информационную систему передачи данных (рис. 7.1).

Источник информации или сообщение *- это физический объект, система или явление, которые формируют переданное сообщение.*

Как правило, начальные сообщения - язык, музыка, изображение, результаты измерения параметров окружающей среды и т.п. - представляют собой функции времени, например $s(t)$, или других аргументов неэлектрической природы (акустическое давление, температура, распределение яркости на некоторой плоскости и т.д.), например, $s(x, y, z)$. С целью передачи по каналу связи эти сообщения обычно превращаются в электрический сигнал, изменения которого во времени $s(t)$ отображают переданную информацию. Такие сообщения называются *непрерывными*, или *аналоговыми*, сообщениями (сигналами), и для них выполняются условия

$$s(t) \in (s_{\min}, s_{\max}),\ t \in (0, t). \qquad (7.1)$$

Т.е. значения и функции, и аргумента для таких сообщений - непрерывные или определенные для любого значения непрерывного интервала, как по функции s, так и по времени t.

Форматирование источника сообщений *- это процесс аналого-цифрового преобразования информационного сигнала источника сообщения*

$\tilde{x}(t)$ в цифровой сигнал $x_{ц}(k\Delta t)$. Такое преобразование базируется на процедурах дискретизации и квантовании сигнала $\tilde{x}(t)$ и его представлении в двоичной системе исчисления.

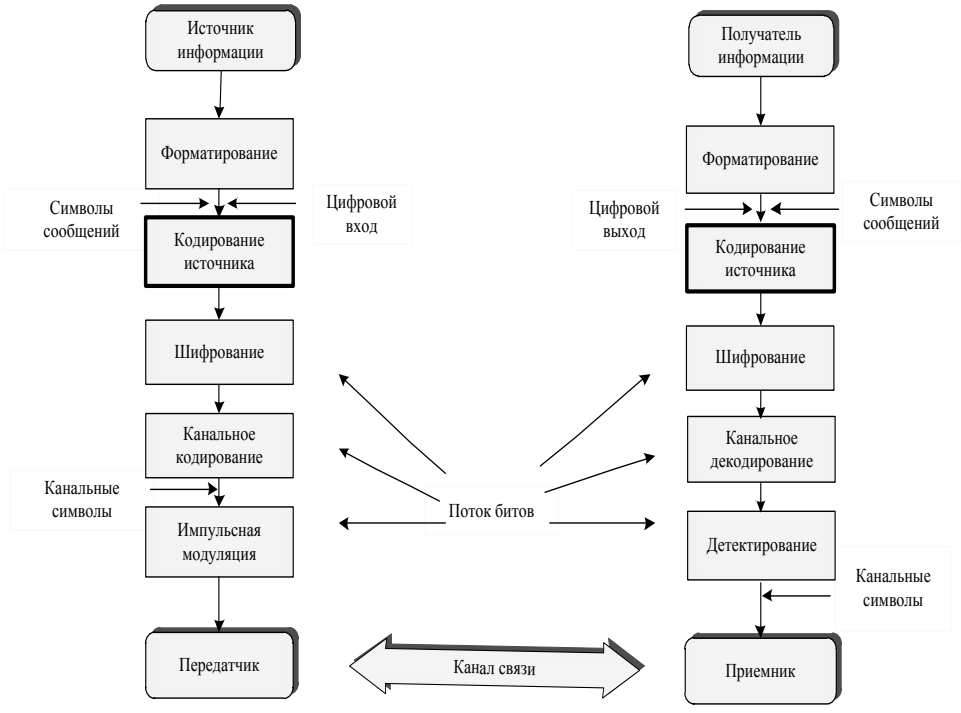


Рис. 7.1. Структурная схема информационной сети передачи данных

Аналого-цифровой преобразователь осуществляет одновременно процедуру дискретизации сигнала по времени, квантование по уровню и формирует значение исходного сигнала, подавая его в цифровом виде.

Передача и хранение информации нуждаются в довольно больших затратах. Часть данных, которые нужно передавать по каналам связи и сохранять, имеет не самое компактное представление. Чаще всего эти данные сохраняются в форме, которая обеспечивает их простейшее использование, например обычные книжные тексты, ASCII коды текстовых редакторов, двоичные коды данных ЭВМ, отдельные отсчеты сигналов в системах сбора данных и т.д. Тем не менее, такое наипростейшее в использовании представление данных заставляет тратить вдвое, втрое, а иногда и в сотне раз больше места для их хранения и намного более широкую полосу частот для их передачи, чем нужно на самом деле. Поэтому сжатие данных - это один из наиболее актуальных направлений современной теории информации.

Кодирование источника сообщений *проводится с целью обеспечения компактного представления данных, сокращения объема информации, кото-*

рая вырабатывается источником, и с целью повышения скорости передачи информационных сообщений по каналам связи.

Такое кодирование называют *экономным, безызбыточным, эффективным кодированием, или сжатием данных* (рис. 7.2).

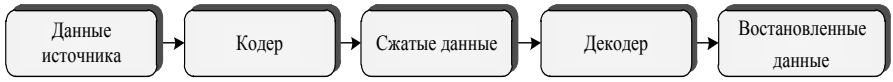


Рис. 7.2. Блок-схема алгоритма сжатия данных

В этой схеме данные, вырабатываемые источником, определим как данные источника, а их компактное представление - как сжатые данные. Система сжатия данных состоит из кодера и декодера источника. Кодер превращает данные источника в сжатые данные, а декодер предназначен для восстановления данных источника со сжатых данных. Восстановленные данные, которые вырабатываются декодером, могут или абсолютно точно совпадать с начальными данными источника, или почти не отличаться от них.

Существуют два типа систем сжатия данных:

системы сжатия без потерь информации (неразрушительное сжатие);

системы сжатия с потерями информации (разрушительное сжатие).

Общую классификацию методов кодирования источника сообщения приведены на рис. 7.3.

Рис. 7.3. Классификация методов кодирования источника сообщений

Сжатие без потерь информации. *Системой сжатия без потерь* называется система, реализующая процесс восстановления сообщений источника таким способом, в котором определенная процедура является неразрушаю-

щей относительно структуры и значений вектора входных данных $X = (x_1, x_2, ..., x_n)$, которые подлежали сжатию или кодированию.

Структура системы сжатия изображена на рис. 7.4. Вектор данных источника X, подлежащего сжатию, - это последовательность $X = (x_1, x_2, ..., x_n)$ конечной длины - результат форматирования аналогового информационного сигнала. Отсчеты x_i - компоненты вектора X - выбрано с конечного *алфавита данных* A. При этом размер n вектора данных ограничен, но может быть довольно большим. Таким образом, источник на своем выходе формирует как вектор данных X последовательность длиной n из алфавита A.

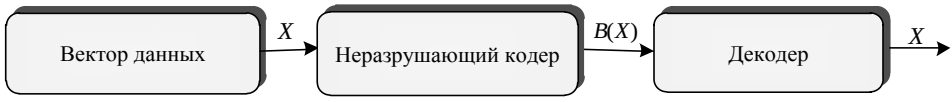


Рис. 7.4. Блок-схема алгоритма кодирования без потерь

Выход кодера - сжатые данные, которые отвечают входному вектору X, дадим в виде двоичной последовательности $B(X) = (b_1, b_2, \ldots, b_k)$, размер которой k. Назовем $B(X)$ кодовым словом, присвоенным вектору X кодером (или кодовым словом, в которое вектор X преобразован кодером). Поскольку система сжатия неразрушительная, одинаковым векторам $X_l = X_m$ должны отвечать одинаковые кодовые слова $B(X_l) = B(X_m)$.

При решении задачи сжатия важным является вопрос, насколько эффективна та или другая система сжатия. Поскольку, как уже отмечалось, в основном используется двоичное кодирование, то мерой эффективности системы может быть *коэффициент сжатия* r, определенный как отношение размера данных источника в битах к размеру k сжатых данных в битах

$$r = \frac{n \log_2 (\dim A)}{k}, \tag{7.2}$$

где $\dim A$ - размер алфавита данных A.

Таким образом, коэффициент сжатия $r = 2$ означает, что объем сжатых данных, представляет половину объема данных источника. Чем больше коэффициент сжатия r, тем лучше работает система сжатия данных. Рядом с коэффициентом сжатия r эффективность системы сжатия можно характеризировать *скоростью сжатия* R, которая определяется как отношение

$$R = k/n \tag{7.3}$$

и измеряется количеством кодовых бит, которые отвечают отсчету данных источника. Система, имеющая *больший* коэффициент сжатия, обеспечивает *меньшую* скорость сжатия.

Сжатие с потерей информации. *Системой сжатия с потерями* называется система, реализующая процесс восстановления сообщений источника таким образом, что определенная процедура восстановления является разрушительной относительно структуры и значений вектора входных данных $X = (x_1, x_2, ..., x_n)$, которые подлежали сжатию или кодированию. Структура системы сжатия изображена на рис. 7.5.

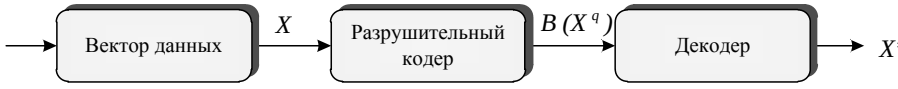


Рис. 7.5. Блок-схема алгоритма кодирования с потерями

Как и в предыдущей схеме, $X = (x_1, x_2, ..., x_n)$ - вектор данных, который подлежит сжатию. Восстановленный вектор обозначим как $X^* = \left(x_1^*, x_2^* ... x_n^*\right)$. Отметим наличие в этой схеме сжатия элемента, которого не было при неразрушительном сжатии, - *разрушительного кодера.*

Кодер подвергает разрушительному сжатию вектор квантованных данных $X = (x_1, x_2, ..., x_n)$ и таким образом обеспечивается соответствие между X и $B\left(X^q\right)$ (возможно невыполнение условия $X = X^q$). Система в целом остается разрушительной, поскольку двум разным векторам X^* может отвечать один и тот же вектор X. Разрушительный кодер относительно вектора входных данных X формирует вектор X^q, достаточно близкий к X в смысле среднеквадратичного расстояния.

Разрушительный кодер характеризуется двумя параметрами - скоростью сжатия R и значением искажений D, определяемыми выражениями:

$$R = k/n, \quad D = (1/n) \sum (x_i - x_i^*)^2. \qquad (7.4)$$

Параметр R характеризует скорость сжатия в битах на один отсчет источника, а значение величины D является мерой среднеквадратичного отличия между X^* и X.

Если существует система разрушительного сжатия со скоростью R_1 и искажениями D_1 и вторая система со скоростью R_2 и искажениями D_2, то первая из них лучшая, если $R_1 < R_2$ и $D_1 < D_2$. Тем не менее, к сожалению, невозможно построить систему разрушительного сжатия, которое обеспечивает одновременное снижение скорости R и уменьшение искажений D, поскольку эти два параметра связанные обратной зависимостью. Поэтому целью оптимизации системы сжатия с потерями может быть минимизация или скорости при заданной величине искажений или искажений при заданной скорости сжатия.

Выбор системы неразрушительного или разрушительного сжатия зависит от типа данных, которые подлежат сжатию. При сжатии текстовых данных, компьютерных программ, документов, черчений и т.д. нужно применять неразрушительные методы, поскольку необходимо абсолютно точное восстановление начальной информации после ее сжатия. При сжатии языка, музыкальных данных и изображений, наоборот, чаще используется разрушительное сжатие, поскольку при практически незаметных искажениях оно обеспечивает на порядок меньшую скорость *R*. В общем случае разрушительное сжатие обеспечивает, как правило, существенным образом высшие коэффициенты сжатия, чем неразрушительное.

К сожалению, неразрушительное сжатие при всей привлекательности перспективы получения *абсолютного совпадения* начальных и восстановленных данных имеет невысокую эффективность - коэффициенты неразрушительного сжатия редко превышают 1-2 (за исключением случаев кодирования данных с высокой степенью повторяемости одинаковых участков). Тем не менее, зачастую нет потребности в абсолютной точности передачи начальных данных потребителю. Речь идет, в частности, о таких случаях.

Источники данных имеют ограниченный динамический диапазон и вырабатывают начальные сообщения с определенным уровнем искажений и ошибок. Этот уровень может быть большим или меньшим, но абсолютной точности воспроизведения достичь невозможно.

Передача данных по каналам связи и их хранение всегда происходят при наличии разного рода помех. Поэтому принятое (*воспроизведенное*) *сообщение всегда определенной мерой отличается от переданного, т.е. на практике невозможна абсолютно точная передача при наличии помех в канале связи* (*в системе хранения*).

Заметим, что сообщения передаются и сохраняются для их восприятия и использования получателем. Получатели информации - органы чувств человека, исполнительные механизмы и т.д. - также имеют конечную раздельную способность, т.е. не замечают незначительной разности между *абсолютно*

Элиаким Гастингс Мур (Eliakim Hastings Moore, 1862 - 1932), американский математик. Сначала работал в области абстрактной алгебры, доказал (1893), что любое финитное поле является полем Галуа. Переформулировал аксиомы Гилберта для геометрии так, что только точки были первоначальным понятием, переведя линии и плоскости из первоначальных понятий Гильберта в обозначенные понятия. Больше того, показал (1902), что некоторые из аксиом Гильберта избыточны. Разрабатывал систему аксиом с точки зрения метаматематики и теории моделей. После 1906 г. вернулся к основам анализа.

точным и *приближенным* значениями воспроизведенного сообщения. Порог чувствительности к искажениям также может быть разным, но он всегда существует.

Кодирование с разрушением учитывает эти аргументы в пользу приближенного восстановления данных и дает возможность получить за счет некоторой контролируемой по размеру ошибки коэффициенты сжатия, которые в десятки раз превышают степень сжатия для неразрушительных методов.

Большинство методов разрушительного сжатия основываются на кодировании не самих данных, а некоторых линейных преобразований от них, например коэффициентов дискретного преобразования Фурье, коэффициентов косинусного преобразования, преобразований Хаара, Уолша и т.д.

Постановка задачи посимвольного кодирования. Предположим, что для некоторого дискретного источника X с известным распределением вероятностей $\{p(x), x \in X\}$ нужно построить эффективный неравномерный двоичный код с алфавитом $A = \{a\}$. Поскольку на практике для кодирования источников используются большей частью лишь двоичные коды, допустим, что $A = \{0,1\}$.

Требованием однозначного декодирования *называют необходимость существования такого двоичного кода, который допускает однозначное распределение последовательности кодовых слов на отдельные кодовые слова без использования любых дополнительных символов.*

Неравномерный посимвольный код $C = \{c\}$ *объемом* $|C| = M$ *над алфавитом A определяется как произвольное множество последовательностей одинаковой или разной длины из элементов алфавита A. Код является однозначно декодированным, если любая последовательность символов с A единым способом разбивается на отдельные кодовые слова.*

Префиксным кодом *называется такой код, для которого ни одно кодовое слово не является началом другого. Префиксные коды являются однозначно декодированными.*

Пример 7.1. Для источника $X = \{0, 1, 2, 3\}$ среди четырех кодов:

а) $C_1 = \{00, 01, 10, 11\}$;

б) $C_2 = \{1, 01, 001, 000\}$;

в) $C_3 = \{1, 10, 100, 000\}$;

г) $C_4 = \{0, 1, 10, 01\}$ —

первые три кода однозначно декодированные, последний код - нет.

Первый код этого примера - равномерный код. Понятно, что любой равномерный код может быть однозначно декодированный.

Для декодирования второго кода можно применить такую стратегию. Декодер считывает символ за символом, и каждый раз проверяет, не совпадает ли полученная последовательность с одним из кодовых слов. В случае успеха соответствующее сообщение выдается получателю, и декодер присту-

пает к декодированию следующего сообщения. В случае кода C_2 неоднозначности не может быть, поскольку ни одно слово не является продолжением другого.

Код C_3, очевидно, не префиксный. Тем не менее, мы утверждаем, что он однозначно декодирован. Каждое слово кода C_3 получено переписыванием в обратном порядке соответствующего слова кода C_2 . Для декодирования последовательности кодовых слов кода C_3 можно переписать принятую последовательность в обратном порядке и для декодирования использовать декодер кода C_2.

Префиксность *- достаточное, но не необходимое условие однозначной декодированности.*.

Графически удобно изображать префиксные коды в виде кодовых деревьев. Кодовое дерево кода C_2 из примера 7.1 приведено на рис. 7.6.

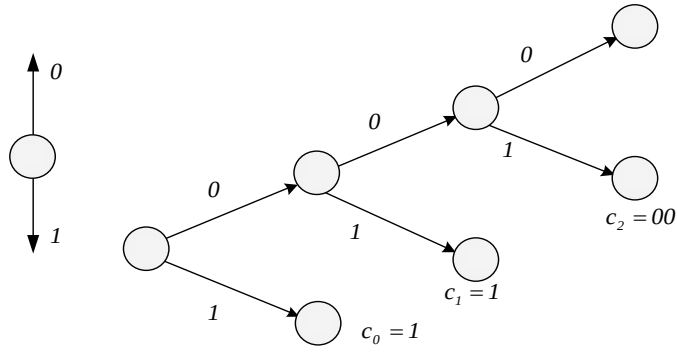


Рис. 7.6. Кодовое дерево кода C_2

Узлы дерева размещаются на ярусах. На начальном (нулевом) ярусе расположен один узел, который называется *корнем дерева.* Узлы следующих ярусов связаны с узлами предыдущих ярусов ребрами. В случае двоичного кода из каждого узла выходит не более чем два ребра. Ребрам приписаны кодовые символы. В этом примере считаем, что ребру, направленному вверх, приписывается символ 0, а ребру, направленному вниз - символ 1. Таким образом, каждой вершине дерева отвечает последовательность, которая считывается вдоль пути, который связывает данный узел с корнем дерева.

Конечным узлом *называют узел, из которого не выходит ни одного ребра.*

Древовидным кодом *называют такой код, который содержит только такие кодовые слова, которые отвечают конечным вершинам кодового дерева.*

Якоб Зив (Jacob Ziv, 1931), родился в Израиле. В 1962 г. получил степень доктора философии в Массачусетском технологическом институте (Кембридж, США). Его исследовательские интересы охватывают сжатие данных, теорию информации и статистическую теорию связи. Зив вместе с Абрахомом Лемпелем разработал алгоритм сжатия данных без потерь LZ77. В разные года работал старшим инженером в Министерстве обороны Израиля, где занимался исследованием и развитием систем коммуникаций.

Древоподобность кода и префиксность - синонимы в том понимании, что каждый древовидный код является префиксным и каждый префиксный код можно подать с помощью кодового дерева. Дальше будем рассматривать лишь префиксные (однозначно декодированные) коды.

Критерием качества кода *относительно кодирования источника сообщения есть средняя длина кодовых слов.*

Рассмотрим источник $X = \{1,..., M\}$, который порождает буквы с вероятностями $\{p_1,..., p_M\}$. Предположим, что для кодирования букв источника взят код $C = \{c_1,..., c_M\}$ с длинами кодовых слов $\text{length}(c_1) = l_1, ..., \text{length}(c_M) = l_M$.

Средней длиной кодовых слов *является*

$$\overline{l} = M[l_i] = \sum_{i=1}^{M} p_i l_i.$$

Еще один важный аспект, который нужно учитывать при сравнении способов неравномерного кодирования, - это сложность реализации кодирования и декодирования. Как увидим дальше, символьные коды часто используются как составная часть более сложных алгоритмов. В этих случаях играет роль не только сложность кодирования и декодирования символа алфавита, а и сложность построения или модификации кода в случае изменения статистических данных об источнике.

Задача посимвольного неравномерного кодирования *- построение однозначно декодированного кода с наименьшей средней длиной кодовых слов при заданных ограничений на сложность.*

Неравенство Крафта для префиксного кодирования. Требование префиксности накладывает жесткие ограничения на множество длин кодовых слов и не дает возможности выбирать кодовые слова слишком короткими. Формально эти ограничения записываются в виде неравенства, которое называется *неравенством Крафта*.

Необходимым и достаточным условием существования префиксного кода *объемом* M *с длинами кодовых слов* $l_1,... l_M$ *является выполнение неравенства*

$$\sum_{i=1}^{M} 2^{-l_i} \le 1. \qquad (7.5)$$

Необходимое условие префиксности. Убедимся в том, что неравенство (7.5) правильно для любого префиксного кода.

Рассмотрим двоичное кодовое дерево произвольного префиксного кода объемом M с длинами кодовых слов $l_1,..., l_M$. Выберем целое число L, чтобы $L \ge \max l_i$. Продолжим все пути в дереве к ярусу с номером L. На последнем ярусе получим 2^L вершин. Заметим, что конечная вершина исходного дерева, размещенная на глубине l_i, имеет следующие 2 узла на глубине $l_i + l$, 4 узла на глубине $l_i + 2$, и т.д. На глубине L будет 2^{L-l_i} узлов этой вершины. Множества узлов разных конечных вершин не пересекаются, поэтому суммарное количество узлов не превышает общего количества вершин на ярусе L. Получим неравенство:

$$\sum_{i=1}^{M} 2^{L-l_i} \le 2^L.$$

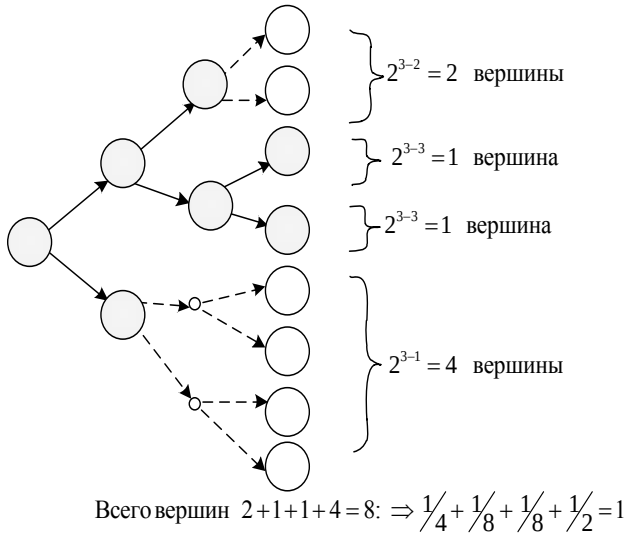


Рис. 7.7. Дерево выполнения условий Крафта для префиксного кода

Поделив обе части на 2^L получим искомый результат. Доведение необходимости выполнения неравенства Крафта на примере кода из 4 слов иллюстрирует рис. 7.7. В этом примере $l_1 = 2,\ l_2 = l_3 = 3,\ l_4 = 1,\ L = 4$. Кодовое дерево на рис. 7.7 изображено сплошными линиями, а ребра, которые появились при продолжении дерева к ярусу с номером L, - пунктиром.

Достаточное условие префиксности. Покажем, что из формулы (7.5) вытекает существование кода с заданным набором длин кодовых слов. По-

строим такой код. Без потери всеобщности можем считать числа l_i упорядоченными по росту.

Из общего количества 2^{l_1} вершин на ярусе l_1 выберем любую, сделаем ее конечной и закрепим за первым кодовым словом. Продолжим вершины, которые остались, к ярусу l_2. Из общего количества возможных вершин нужно исключить $2^{l_2-l_1}$ вершин, которые принадлежат поддереву, что начинается в узле, который отвечает первому слову. На ярусе l_2 останется $2^{l_2}-2^{l_2-l_1}\geq 1$ вершин.

Последнее неравенство вытекает из неравенства (7.5), в чем нетрудно убедиться, поделив его правую и левую части на 2^{l_2}. Сделаем одну из них конечной и закрепим ее за вторым словом. Аналогично для третьего слова получим множество из выражения $2^{l_3}-2^{l_3-l_2}-2^{l_3-l_1}\geq 1$ вершин. Согласно формуле (7.5) всегда обнаружится одна вершина для третьего слова. Продолжая построение, на последнем ярусе с номером l_M получим вершин: $2^{l_M}-2^{l_{\tilde{N}}-l_{M-1}}-2^{l_M-l_{M-2}}-\ldots-2^{l_M-l_l}$.

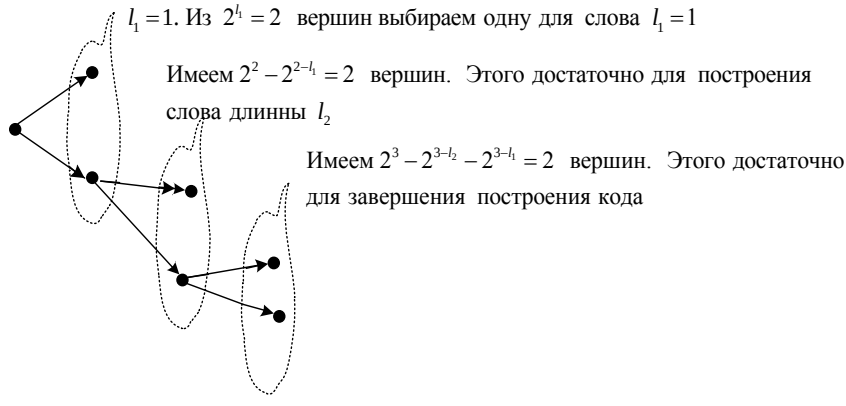


Рис. 7.8. Построение префиксного кода с длинами слов, которые удовлетворяют неравенству Крафта

Простые выкладки показывают, что это число не меньше 1, если неравенство (7.5) правильное. Выбрав эту вершину для последнего слова, закончим построение префиксного кода. Процесс построения кодового дерева для набора длин кодовых слов $l_1=l, l_2=2, l_3=l_4=3$ иллюстрирует рис. 7.8. В рассмотренном примере неравенство Крафта превращается в равенство. Для достижения равенства в формуле (7.5) кодовое дерево должно быть полным, т.е. каждая промежуточная вершина дерева должна иметь ровно 2 узла и всем конечным вершинам должны быть сопоставленные кодовые слова.

Неравенство Крафта ограничивает снизу длину кодовых слов префиксного кода заданного объема M.

В связи с этим важно быть уверенным, что это неравенство выполняется не только для древовидных (префиксных) кодов, а и для любых других однозначно декодированных кодов.

Для любого однозначно декодированного двоичного кода объемом M с длинами кодовых слов $l_1, ..., l_M$ выполняется неравенство

$$\sum_{i=1}^{M} 2^{-l_1} \le 1. \qquad (7.6)$$

Коды без памяти. Коды Хаффмана. Рассмотрим один из наиболее распространенных методов сжатия данных - *код Хаффмана*, или *минимально избыточный префиксный код*, предложенный в 1952 г. Дэвидом Хаффманом. Идея, положенная в основу кода Хаффмана, довольно простая. Вместо того чтобы кодировать все символы одинаковым количеством бит (как это сделано, например, в ASCII кодировании, где каждому символу отводится ровно по 8 бит), будем кодировать символы, которые встречаются чаще, меньшим количеством бит, чем те, которые случаются реже. Более того, необходимо, чтобы код был оптимальным, или, другими словами, минимально избыточным. Предложенный Хаффманом алгоритм построения оптимальных неравномерных кодов - одно из важнейших достижений теории информации как из теоретического, так и с прикладной точки зрения.

Рассмотрим ансамбль сообщений $X=\{1, ..., M\}$ с вероятностями сообщений $\delta_1, \delta_2, ... \delta_M$ и упорядочим его по спаданию вероятностей, т.е. $p_1 < p_2 < ... < p_M$. *Задача* состоит в построении оптимального кода, т.е. кода с наименьшей возможной средней длиной кодовых слов. Понятно, что при заданных вероятностях такой код может не быть единым, возможно существование семьи оптимальных кодов. Выясним некоторые свойства всех кодов этой семьи. На основании этих свойств найдем и обоснуем один из оптимальных кодов.

Пусть двоичный код $C=\{c_1, ..., c_M\}$ с длинами кодовых слов $\{l_1, ..., l_M\}$ оптимальный для рассмотренного ансамбля сообщений.

Свойства оптимальных префиксных кодов. 1. *Если* $p_i < p_j$, *то* $l_i > l_j$.

Свойство легко доказывается методом от противоположного. Предположим, что $l_i < l_j$. Рассмотрим другой код C', в котором сообщению x_i отвечает слово c_j, а сообщению x_j - слово c_i. Нетрудно убедиться в том, что средняя длина кодовых слов для кода C' меньшая, чем для кода C, что противоречит предположению об оптимальности кода C.

2. *Не менее чем два кодовых слова имеют одинаковую длину* $l_M = \max l_m$.

Предположив, что существует только одно слово максимальной длины, приходим к выводу, что соответствующее кодовое дерево будет неполным.

Очевидно, слово максимальной длины можно будет сделать менее коротким, по меньшей мере, на один символ. При этом уменьшится средняя длина кодовых слов, которая противоречит предположению об оптимальности кода.

3. *Среди кодовых слов длиной* $l_M = \max l_m$ *обнаруживаются два слова, которые отличаются только одним последним символом.*

В соответствии с предыдущим свойством два слова длиной l_M существуют в любом оптимальном коде. Рассмотрим конечный узел, который отвечает одному из слов максимальной длины. Чтобы дерево было полным, должны существовать узел, который является общим с предыдущим узлом. Кодовые слова, которые отвечают двум конечным вершинам, имеют одинаковую длину l_M и отличаются одним последним символом.

Введем обозначения. Для рассмотренного ансамбля $X = \{1, ..., M\}$ и некоторого кода C, удовлетворяющий свойству 1 - 3, введем вспомогательный ансамбль $X = \{1, ..., M-1\}$, сообщение которого поставим в соответствие вероятности $\{p'_1, ..., p'_{M-1}\}$ таким образом: $p'_1 = p_1, ..., p'_{i-2} = p_{i-2}, p'_{i-1} = p_{i-1} + p_i$. Из кода C построим код C' для ансамбля X', приписав сообщению $x'_1, ... x'_{M-2}$ те самые кодовые слова, что и в коде C, т.е. $c'_i = c_i$, $i = 1, ..., M-2$, а сообщению x'_{M-1} - слово c'_{M-1}, которое представляет собой общую часть слов c_{M-1} и c_M (соответственно свойству 3 эти два кодовых слова отличаются только одним последним символом).

4. *Если код* C' *для* X' *оптимальный, то код* C *оптимальный для* X.

Длины кодовых слов кодов C и C' связаны соотношением

$$l_m = \begin{cases} l'_m \text{ при } m \le M-2; \\ l'_{M-1} + 1 \text{ при } m \le M-1. \end{cases}$$

Отсюда

$$\bar{l} = \sum_{m=1}^{M} p_m l_m = \sum_{m=1}^{M-2} p_m l_m + p_{M-1} l_{M-1} + p_M l_M = \sum_{m=1}^{M} p_m l_m + (p_{M-1} + p_M)(l'_{M-1} + 1) =$$

$$= \sum_{m=1}^{M-2} p'_m l'_m + p'_{M-1} l'_{M-1} + p_{M-1} + p_M = \sum_{m=1}^{M-1} p'_m l'_m + p_{M-1} + p_M = \bar{l}' + p_{M-1} + p_M.$$

Последние два слагаемых в правой части не зависят от кода, поэтому код, который минимизирует $\bar{l}'$, одновременно обеспечивает минимум для $\bar{l}$.

Итак, сформулированные свойства оптимальных префиксных кодов сводят задачи построения кода объемом M к задаче построения кодов объемом $M' = M - 1$. Это означает, что мы получили рекуррентное правило построения кодового дерева оптимального неравномерного кода.

Пример 7.2. Рассмотрим источник и коды из примера 7.1. Вычислим среднюю длину кодовых слов кодов C_1 и C_2 для двух распределений вероятностей на буквах источника:

а) $p_0 = p_1 = p_2 = p_3 = \frac{1}{4}$;

б) $p_0 = \frac{1}{2}$; $p_1 = \frac{1}{4}$; $p_3 = p_4 = \frac{1}{8}$;

в) $p_0 = p_1 = \frac{1}{8}$; $p_2 = \frac{1}{4}$; $p_3 = \frac{1}{2}$.

Результаты расчетов показывают, что не всегда и не любой неравномерный код эффективный.

Пример 7.3. Пусть имеем источник данных, который передает символы $\{\alpha_1, \alpha_2, ..., \alpha_n\}$ с разной вероятностью, т.е. каждому α_i отвечает своя вероятность (или частота) $P_i(\alpha_i)$, причем существует хотя бы одна пара α_i и α_j, $i \neq j$, для которых $P_i(\alpha_i) \neq P_j(\alpha_j)$. Таким образом, образовывается набор частот $\{P_1(\alpha_1), P_2(\alpha_2), ..., P_n(\alpha_n)\}$, причем $\sum_{i=1}^{n} P_i(\alpha_i) \equiv 1$, поскольку передатчик не передает больше никаких символов, кроме $\{\alpha_1, \alpha_2, ..., \alpha_n\}$.

Наша задача - подобрать такие кодовые символы $\{b_1, b_2, ..., b_n\}$ с длинами $\{L_1(b_1), L_2(b_2), ..., L_n(b_n)\}$, чтобы средняя длина кодового символа не превышала средней длины исходного символа. При этом нужно учитывать, что если $P_i(\alpha_i) > P_j(\alpha_j)$ и $i \neq j$, то $L_i(b_i) \leq L_j(b_j)$.

Хаффман предложил строить дерево, в котором узлы с наибольшей вероятностью наименее отдалены от корня. Отсюда и вытекает сам способ построения дерева.

1. Выбрать два символа α_i и α_j $(i \neq j)$, так, чтобы $P_i(\alpha_i)$ и $P_j(\alpha_j)$ из всего списка $\{P_1(\alpha_1), P_2(\alpha_2), ..., P_n(\alpha_n)\}$ были минимальными.

2. Свести ветки дерева от этих двух элементов в

Абрахам Лемпель (Abraham Lempel, 1936), ученый-компьютерщик и один из родителей семьи алгоритмов сжатия данных без потерь LZ, по происхождению поляк. Ныне проживает в Израиле. Его исторически важные работы берут начало из презентации алгоритма LZ77 в статье "A Universal Algorithm for Sequential Data Compression" (1977 г.). Эта работа написана в соавторстве с Якобом Зивом. Немало алгоритмов с буквой L в названии указывают на Лемпеля: LZ77, LZ78, LZW, LZR, LZS, LZO и LZMA. Названия LZX, LHA (LHarc) и LZH ссылаются также на Лемпеля. Его работы заложили основу для таких графических форматов сжатия, как GI, TIFF и JPEG.

одну точку с вероятностью $P = P_i(\alpha_i) + P_j(\alpha_j)$, обозначив одну ветку нулем, а другую - единицей.

3. Повторить п. 1 с учетом новой точки вместо α_i и α_j, если количество образованных точек больше, чем единица. В противоположном случае мы достигли корня дерева.

Теперь попробуем воспользоваться изложенной теорией и закодировать информацию, переданную источником, на примере семи символов.

a_i	$P_i\ (a_i)$
a_1	0.01
a_2	0.40
a_3	0.08
a_4	0.02
a_5	0.10
a_6	0.35
a_7	0.04

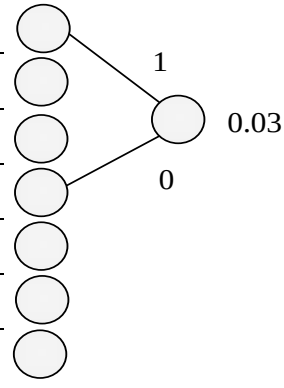


Рис. 7.9. Таблица соответствия символов и их вероятностей

Рассмотрим подробно первый цикл. На рис. 7.9 изображена таблица, в которой каждому символу α_i соответствует своя вероятность (частота) $P_i(\alpha_i)$. Соответственно п. 1 выбираем два символа из таблицы с наименьшей вероятностью. Это α_1 и α_4. Соответственно п. 2 сводим ветки дерева от α_1 и α_4 в одну точку и обозначаем ветку, которая ведет к α_1, единицей, а ветку, которая ведет к α_4, - нулем. Возле новой точки записываем ее вероятность (в этом случае 0,03). В дальнейшем действия повторяются уже с учетом новой точки, но без учета α_1 и α_4.

После многоразового повторения указанных действий выстраивается дерево, приведенное на рис. 7.10.

По построенному дереву можно найти значение кодов $\{b_1, b_2, ..., b_n\}$, осуществив спуск от корня до соответствующего элемента α_i. При прохождении каждой ветки приписываем к создаваемой последовательности нуль или единицу (в зависимости от того, как именуется соответствующая ветка). Значения искомых кодов приведены в табл. 7.10.

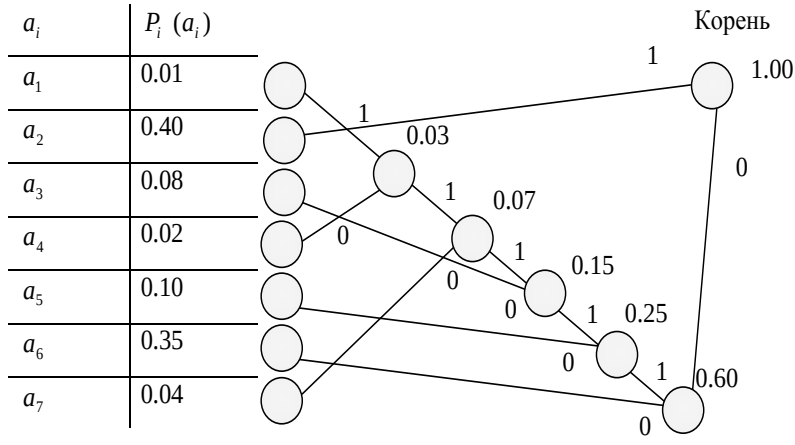


Рис. 7.10. Таблица символов и вероятностей и соответствующее им дерево Хаффмана

Теперь закодируем последовательность из символов. Предположим например, что символа α_i отвечает число i. Пусть есть последовательность 12672262. Нужно найти результирующий двоичный код.

Для кодирования можно использовать уже имеющуюся таблицу кодовых символов b_i, взяв во внимание, что b_i отвечает символу α_i. В таком случае код для цифры 1 будет представлять собой последовательность 011111, для цифры 2 будет 1, для цифры 6 последовательность 00, для цифры 7 - последовательность 01110.

Таблица 7.1

i	b_i	$L_i(b_i)$
1	011111	6
2	1	1
3	0110	4
4	011110	6
5	010	3
6	00	2
7	01110	5

Искомый результат приведен в табл. 7.2.

Таблица 7.2

Данные	12672262	Длина кода
Исходные	001 010 110 111 010 010 110 010	24 бит
Кодированные	011111 1 00 01110 1 1 00 1	19 бит

В результате кодирования мы выиграли 5 бит и записали последовательность с помощью 19 бит вместо 24. Однако это не дает полной оценки сжатия данных. Оценим степень сжатия кода. Для этого понадобится энтропическая оценка. Математически энтропия подается как сумма произведений вероятностей разных состояний системы на логарифмы этих вероятностей, взятых с противоположным знаком:

$$H(X) = -\sum_{i=1}^{n} P_i \log_d P_i,$$

где X - случайная величина (в этом примере - кодовый символ), а d - произвольная основа, больше единицы. Выбор основы равносильный выбору определенной единицы измерения энтропии. Поскольку рассматриваются двоичные цифры, то за основу целесообразно взять $d = 2$.

Таким образом, энтропию для нашего случая можно записать как

$$H(b) = -\sum_{i=1}^{n} P_i(\alpha_i) \log_2 P_i(\alpha_i).$$

Энтропия равна минимально допустимой средней длине кодового символа $\bar{L}_{\min}$ в битах, а самая средняя длина кодового символа вычисляется по формуле

$$L(b) = \sum_{i=1}^{n} P_i(\alpha_i) L_i(b_i).$$

Подставляя соответствующие значения в формулы для $H(b)$ и $\bar{L}(b)$, получим

$$H(b) = 2,048,$$
$$\bar{L}(b) = 2,100.$$

Значения $H(b)$ и $\bar{L}(b)$ очень близкие, что говорит о реальном выигрыше в выборе алгоритма. Теперь сравним среднюю длину выходного символа и среднюю длину кодового символа:

$$\frac{\bar{L}_{\text{âèõ.ñ}}}{L(b)} = \frac{3}{2,1} = 1,429.$$

Таким образом, получили сжатие в соотношении 1:1,429.

Избыточность кода Хаффмана. Для посимвольных по алгоритму Хаффмана кодов средняя длинна кодовых слов удовлетворяет неравенству

$$\bar{l} \le H + 1, \qquad (7.7)$$

где H - энтропия ансамбля.

Избыточностью неравномерного кода Хаффмана называется разность

$$r = \overline{l} - H.$$

Она показывает степень «несовершенства» кода в том понимании, что при кодировании с избыточностью r на каждое сообщение тратится на r бит больше, чем в принципе можно было бы израсходовать, если использовать теоретически наилучший (возможно, нереализованный) способ кодирования.

Итак, из формулы (7.7) вытекает, что для кода Хаффмана избыточность $r \le 1$. Однако при решении практических задач избыточность существенным образом меньше единицы, поэтому нужно найти самую точную оценку средней длины кодовых слов. Этого нельзя сделать, не ограничив множества рассматриваемых источников.

Код Шеннона. Рассмотренный раньше префиксный код Хаффмана является оптимальным неравномерным кодом. Во время рассмотрения прямой теоремы посимвольного кодирования уже говорилось о том, что избыточность 1 бит на букву не такая уже и большая при большом значение энтропии.

Рассмотрим источник, который выбирает сообщение из множества $X = \{1, ..., M\}$ с вероятностями $\{p_1, ..., p_M\}$. Считаем, что символы упорядочены по спаданию вероятностей, т.е. $p_1 \ge p_2 \ge ... \ge p_M$. Поставим в соответствие, кроме того, каждой букве так называемую *кумулятивную вероятность* по правилу

$$q_1 = 0,\ q_2 = p_1, ...,\ q_M = \sum_{i=1}^{M-1} p_i.$$

Кодовым словом Шеннона для сообщения с номером m *является двоичная последовательность, которая представляет собой первые* $l_m = \left[-\log p_m\right]$ *разрядов после запятой в двоичной записи числа* q_m.

Пример 7.4. Рассмотрим тот самый ансамбль, что и в предыдущем примере. В табл. 7.1 приведены промежуточные вычисления и результат построения кода Шеннона. Средняя длина кодовых слов $\overline{l} = 2{,}95$. В этом случае избыточность кода Шеннона оказалась на 0,5 бит большей, чем избыточность кода Хаффмана. Кодовое дерево кода показано на рис. 7.11, из которого вытекает, почему код неэффективный (или неоптимальный). Кодовые слова для символов *b*, *d*, *e*, *f* можно укоротить на 1 бит без потери свойства однозначной декодированности.

Вспомним, что длина слова и его вероятность связаны соотношением

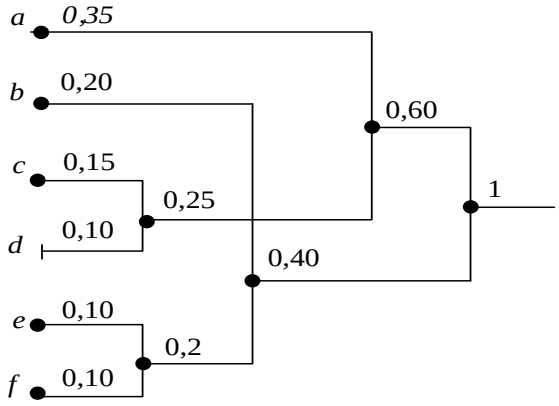
$$l_i = \left[-\log_2 p_i\right] \ge -\log_2 p_i.$$

Поэтому

$$p_i \ge 2^{-l_i}.$$

С учетом этого неравенства

$$q_j - q_i \geq 2^{-l_i}.$$



Буква	Кодовое слово
a	00
b	10
c	010
d	011
e	110
f	111

Рис. 7.11. Пример построения дерева кода Хаффмана

Построение кода Шеннона для этого примера иллюстрирует табл. 7.3.

Таблица 7.3

Буква	Вероятность p_m	Кумулятивная вероятность q_m	Длина кодового слова I_m	Двоичная запись $[q]_2$	Кодовое слово c_m
a	0,35	0,00	2	0,00...	00
b	0,20	0,35	3	0,0101...	010
c	0,15	0,55	3	0,10001...	100
d	0,10	0,70	4	0,10110...	1011
e	0,10	0,80	4	0,11001...	1100
f	0,10	0,90	4	0,11100..	1110

Соответствующее дерево иллюстрирует рис. 7.12.

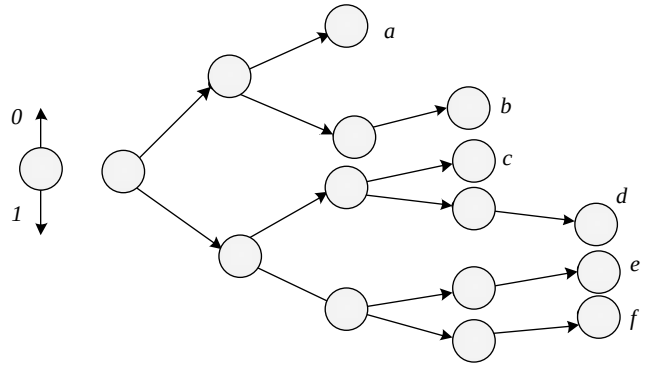


Рис. 7.12. Дерево Шеннона для источника табл. 7.3

В двоичной записи числа в правой части имеем после запятой $l_i - 1$ нулей и единицу в позиции с номером l_i. Это означает, что по крайней мере в одном из l_i разрядов слова c_i и c_j отличаются, а значит, c_i не является префиксом для c_j. Поскольку это правильно для любой пары слов, то код является префиксным.

Длины кодовых слов в коде Шеннона точно такие же, какие были взяты при доказательстве прямой теоремы кодирования. Повторяя выкладки, получаем уже известную оценку для средней длины кодовых слов: $\overline{l} \le H + 1$.

При построении кода Шеннона мы выбрали длины кодовых слов так, чтобы они приближенно равнялись (были немного большими) собственной информации соответствующих сообщений. В результате средняя длина кодовых слов оказалась такой, что приближенно равняется (немного больше) энтропии ансамбля.

Пример 7.5. Рассмотрим еще одну графическую интерпретацию процесса кодирования. Она будет полезной в дальнейшем при обсуждении арифметического кодирования.

Рассмотрим числовой отрезок [0,1], на котором разместим один за другим отрезки длиной $p_1, ..., p_m$. Пример разбивки отрезка [0,1] для случая $M = 3,\ p_1 = 0{,}6;\ p_2 = 0{,}3;\ p_3 = 0{,}1$ приведен на рис. 7.13. Как вытекает из рис. 7.13, *а*, кумулятивные вероятности $q_1 = 0;\ q_2 = 0{,}6;\ q_3 = 0{,}9$ отвечают началам отрезков. Эти точки идентифицируют сообщение источника.

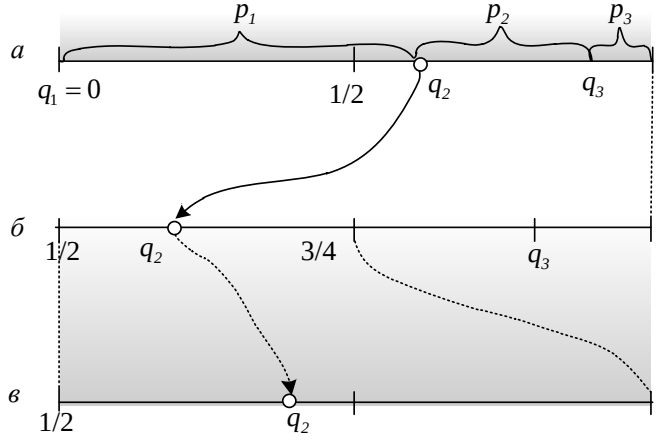


Рис. 7.13. Графическая интерпретация кода Шеннона

Предположим, что нужно закодировать сообщение с номером $m = 2$. Соответствующая ему точка обозначена на рис. 7.13, *в* кружочком. На первом шаге передачи кодового символа 0 или 1 кодер указывает, в которой (левой или правой) половине отрезка $[0,1]$ находится начало соответствующего

сообщению отрезка. В этом случае передается 1, и тем самым область возможных положений переданной точки уменьшается вдвое, что и показано на рис. 7.13, *б.* На следующем шаге передается символ 0, поскольку точка находится в левой половине интервала, то длина интервала неопределенности уменьшается до 1/4.

Заметим, что после второго шага в интервале неопределенности осталась только одна точка, а потому передачу можно закончить. Это состоялось не случайно. Дело в том, что длина интервала равняется 1/4, что меньше длины кратчайшего близлежащего отрезка, которая равна 0,3. Именно поэтому гарантируется единственность точки, а соответственно, однозначность декодирования.

В общем случае после передачи l_m двоичных символов длина интервала неопределенности равна 2^{-l_m}. Декодирование будет однозначным, если $2^{-lm} \le p_m$ или $l_m \ge -\log_2 p_m$. Это условие полностью совпадает с правилом выбора длин кодовых слов в коде Шеннона.

При выборе длины кодовых слов мы ориентировались только на отрезок, который лежит справа от точки q_m. Упорядоченность букв по спаданию вероятностей гарантирует, что левый отрезок всегда будет длинней, чем правый.

Код Гильберта - Мура. При построении кода Шеннона требовалась благоустроенность сообщений по спаданию вероятностей. В алгоритме построения кода Шеннона сортировка букв или символов входного алфавита - наиболее трудоемкая часть. Упростим построение кода, модифицировав кодирование так, чтобы упорядоченность не требовалась. Графическая интерпретация кода Шеннона (рис. 7.14) подсказывает путь к выполнению этой задачи.

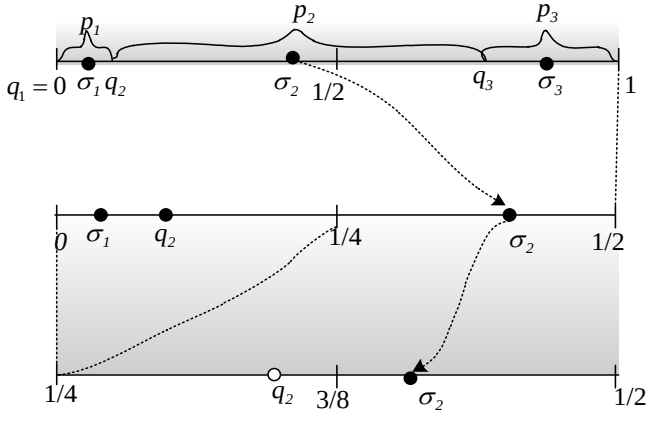


Рис. 7.14. Графическая интерпретация кода Гильберта - Мура

Предположим, что вероятности не благоустроенные. Тогда при кодировании сообщения с номером m нужно учитывать не только вероятность

(длину отрезка) p_m, а и длину предыдущего отрезка p_{m-1}, которая может быть очень малой, почти нулевой, и тогда длина слова будет больше даже в случае, если вероятность p большая. Соответствующую сообщению точку смещаем с начала отрезка (точка q_m) в его середину (точка $q_m + p_m / 2$), а длину кодового слова выбираем так, чтобы до конца передачи кодового слова длина интервала неопределенности была не больше, чем $p_m / 2$. Это и будет кодовое слово кода Гильберта - Мура.

Рассмотрим источник, который выбирает буквы из алфавита $X = \{1, ..., M\}$ с вероятностями $\{p_1, ..., p_M\}$. Поставим в соответствие каждой букве $m = 1, ..., M$ кумулятивную вероятность $q_m = \sum_{i=1}^{m-1} p_i$ и вычислим для каждой буквы значения σ_m по формуле

$$\sigma_m = q_m + \frac{p_m}{2}.$$

Кодовым словом Гильберта - Мура *для* x_m *является двоичная последовательность, которая представляет собой первые* $l_m = \left[-\log_2 (p_m/2)\right]$ *разрядов после запятой в двоичной записи числа* σ_m.

Пример 7.6. Рассмотрим источник с распределением вероятностей $p_1 = 0,1;\ p_2 = 0,6;\ p_3 = 0,3$. Вычисления, связанные с построением кода Гильберта - Мура для этого источника, приведены в табл. 7.4.

Представление числа *a* в двоичной форме обозначается записью [*a*]. В последнем столбце табл. 7.4 показано, какой вид имели бы кодовые слова соответствующего кода Шеннона, если бы буквы не были упорядочены по вероятностям. Код вышел непрефиксным.

Таблица 7.4

Буква x_m	Вероятность p_m	Кумулятивная вероятность q_m	σ_m	Длина кодового слова I_m	Кодовое слово Гильберта - Мура	Кодовое слово Шеннона
1	0,1	0,0 = (0,00000...)	0,05 = (0,00001)	5	00001	0000
2	0,6	0,1 = (0,00011...)	0,40 = (0,01100)	2	01	0
3	0	0,7 = (0,0 = 10110...)	0,85 = (0,11011)	3	110	10

Код Гильберта - Мура в общем случае является префиксным и однозначно декодированным. Все слова кода Гильберта - Мура не более чем на единицу длиннее слов кода Шеннона и имеют такую оценку средней длины кодовых слов: $\overline{l} \le H + 2$.

Рассмотрим код Гильберта - Мура на базе его графической интерпретации, приведенной на рис. 7.14, для источника из примера 7.6. Предположим, что передается буква с номером 2. Ей отвечает точка σ_2. По построению соседние точки σ отдалены от нее на расстояние по меньшей мере $p_2/2$. Кодер передает бит за битом, и при этом каждый раз интервал неопределенности суживается вдвое. Передачу можно закончить, когда длина интервала неопределенности будет не больше $p_2/2$. В этом примере достаточно передать 2 бита.

Алгоритм декодирования кода Гильберта - Мура формируется согласно длине кодовых слов, которые выбираются так, что в результате округления к l_m разрядам значения $\sigma_m = q_m + p_m/2$ уменьшаются не больше, чем на $p_m/2$ (погрешность округления не превышает $2^{-l_m} \le p_m/2$). Поэтому в неравенстве

$$q_m \le \hat{\sigma} < q_{m+1}.$$

В этом алгоритме при декодировании используются только значение q_i, $i \le m$ и не используются значение $\hat{\sigma}_i$. Это важно, когда речь идет о декодировании арифметического кода. Также при декодировании арифметического кода для полученного из канала округленного значения F мы будем рекуррентно вычислять наиболее близкое к $\hat{F}$, но не больше F значение кумулятивной вероятности $q(x)$. Результатом декодирования будет соответствующая последовательность сообщений x.

Итак, декодеру известен алфавит $X = \{1, ..., M\}$, вероятности $\{p_1, ..., p_M\}$, кумулятивные вероятности $\{q_1, ..., q_M\}$, длина последовательности сообщений n и полученное из канала значение $\hat{F}$. Задача состоит в вычислении последовательности сообщений x.

Арифметическое кодирование. *Метод арифметического кодирования дает возможность эффективно кодировать блоки длины* n *с избыточностью порядка* $2/n$ *и со сложностью, которая возрастает только пропорционально квадрату длины блока* n.

За счет малого проигрыша в скорости кода можно достичь даже линейной по длине кода сложности. Арифметическое кодирование все шире применяется в разнообразных системах обработки информации.

Рассмотрим для упрощения соображений дискретный постоянный источник, который выбирает сообщение из множества $X = \{1, ..., M\}$ с вероятностями $\{p_1, ..., p_M\}$. Обозначим через $\{q_1, .., q_i\}$ кумулятивные вероятности сообщений. Наша задача состоит в кодировании последовательностей множества $X^n = \{x\}$. Описывая алгоритм кодирования, будем использовать обо-

значение x_i^j для короткой записи подпоследовательности $(x_i, ..., x_j)$ последовательности $x = (x_1, ..., x_n)$.

Применим к ансамблю $X^n = \{x\}$ довольно простой и эффективный посимвольный код. Упрощение заключается в том, что ни кодер, ни декодер не сохраняет и не строит всего множества из $\left|X^n\right|$ кодовых слов. Вместо этого при передаче конкретной последовательности x кодером вычисляется кодовое слово $c(x)$ только для данной последовательности x. Правило кодирования обычно известно декодеру, и он восстанавливает x по $c(x)$, не имея полного списка кодовых слов.

Возможными алгоритмами для использования в такой схеме можно рассматривать код Шеннона и код Гильберта - Мура. Тем не менее, использование кода Шеннона допускает благоустроенность сообщений по спаданию вероятности. При больших n сложность приведения в порядок становится недопустимо большой, поэтому единым претендентом остается код Гильберта - Мура.

Соответственно правилу построения кода Гильберта - Мура кодовое слово формируется по вероятности $p(x)$ и кумулятивной вероятности $q(x)$ как первые $l(x) = \left[-\log p(x) + 1\right]$ разрядов после точки в двоичной записи числа $\sigma(x) = q(x) + p(x)/2$. Для того чтобы вычислить $q(x)$, нужно договориться о некоторой нумерации последовательностей из X^n. Наиболее естественный способ нумерации последовательностей - использование лексикографической благоустроенности. Лексикографический порядок на последовательностях обозначают знаком «$\prec$». Запись $y \prec x$ означает, что y лексикографически передает x. Лексикографический порядок - это порядок, который обычно используется при составлении словарей. Итак, основная задача состоит в вычислении кумулятивной вероятности

$$q(x) = \sum_{y \prec x} p(y), \tag{7.8}$$

поскольку для источника без памяти вероятности последовательностей $p(x)$ вычисляются довольно просто по формуле

$$p(x) = \prod_{i=1}^{n} p(x_i).$$

Алгоритм арифметического кодирования

1. По вероятностям $\{p_1, .., p_M\}$ сообщений источника исчисляются кумулятивные вероятности $q_1 = 0$ ï ðè $j = 2, ..., M$, $q_j = q_{j-1} + p_{j-1}$.

Ричард Весли Хэмминг (Richard Wesley Hamming, 1915 - 1998), американский математик, которого можно назвать гением одной идеи. Он сформулировал ее в 1950 г. в своей научной статье, посвященной кодам для корректирования ошибок. Речь шла о конструкции блочного кода, который корректирует одиночные ошибки, возникающие при передаче сообщений. Одним из важнейших разделов теории информации является теория кодирования, основы которой заложил Хемминг. Пионерскую работу Хемминга было отмечено многими наградами. В 1996 г. в Мюнхене за исследование кодов, которые корректируют ошибки, Хемминг был удостоен престижной премии Эдуарда Рейма.

Устанавливаются начальные значения вспомогательных сменных $F = 0,\ G = 1$. Принимается от источника последовательность сообщений $x = (x_1, ..., x_n)$.

2. Для $i = 1, ..., n$ выполняются такие вычисления:

$$F \leftarrow F + q(x_i)G,$$
$$G \leftarrow p(x_i)G.$$

3. Кодовое слово для x формируется как первые $[-\log G] + 1$ разрядов после запятой в двоичной записи числа $(F + G/2)$.

Графическая интерпретация процесса кодирования аналогична интерпретации кодов Шеннона и Гильберта - Мура (см. рис. 7.13 и 7.14).

Пример 7.7. Рассмотрим источник $X = \{a, b, c\}$ из примера 7.5, характеризующийся распределением вероятностей $p_a = 0{,}1;\ p_b = 0{,}6;\ p_c = 0{,}3$. Вычисление, выполняемые арифметическим кодером при кодировании последовательности $x = (bcbab)$ длины $n = 5$, приведено в табл. 7.5, где $\hat{F}$ обозначает число $(F + G/2)$, округленное с недостаточной точностью до $[-\log G + 1] = 9$ двоичных разрядов.

Как показано на рис. 7.15, на каждом шаге кодирования вычисляют начальную точку F и длину G отрезка, которому будет принадлежать число, которое отвечает кодовой последовательности для заданной последовательности сообщений. Так, после первого шага $(x_1 = b)$ мы знаем, что точка будет подлежать отрезку [0,1; 0,7]. Более детально этот отрезок показан на рис. 7.13, *б*. Поскольку вторая буква $x_2 = c$, начальная точка перемещается в точку 0,52, а длина интервала уменьшается до 0,18 и т.д. После 5-го шага $F = 0{,}5391$. Прибавив сдвиг $G/2$ и округлив до девяти двоичных знаков, получим $\hat{F} = 0{,}541$.

Тем самым вся последовательность сообщений отображается в одну точку интервала [0,1]. Эта точка на рис. 7.15 обозначена кружочком. Как было показа-

но раньше, девяти разрядов достаточно для того, чтобы последовательность восстановить однозначно, т.е. ближайшая возможная точка, которая отвечает другой последовательности сообщений, отдалена от $\widehat{F}$ на расстояние не менее чем $1/2^9 = 1/512$.

Таблица 7.5

Шаг i	x_i	$p(x_i)$	$q(x_i)$	F	G
0	-	-	-	0,0000	1,0000
1	*b*	0,6	0,1	0,1000	0,6000
2	*c*	0,3	0,7	0,5200	0,1800
3	*b*	0,6	0,1	0,5380	0,1080
4	*a*	0,1	0,0	0,5380	0,0108
5	*b*	0,6	0,1	0,5391	0,0065
6	Длина кодового слова		Кодовое слово		
	$\lceil -\log G + 1 \rceil = 9$		$F + G/2 = 0.5423... \to \hat{F} = 0.541 \to 100010101$		

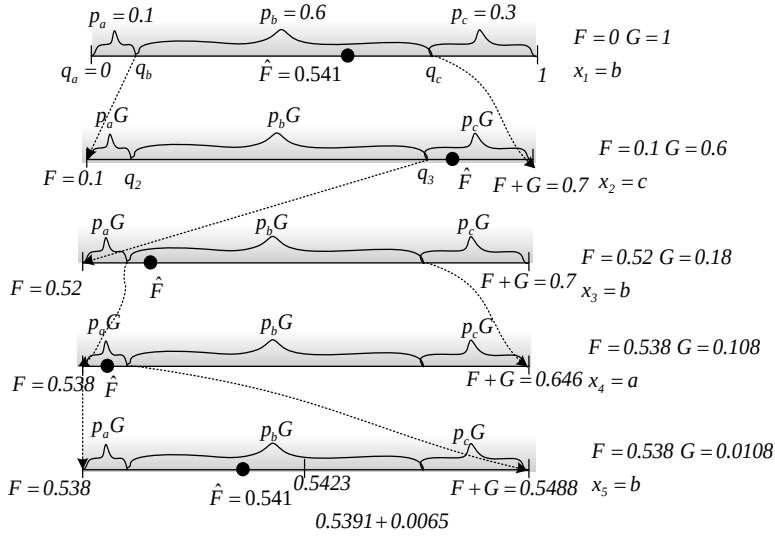


Рис. 7.15. Графическая интерпретация арифметического кодирования

Некоторые аспекты практической реализации арифметического кодирования. Как вытекает из описания арифметического кодирования, в случае его практической реализации для кодирования последовательностей большой длины n возникают такие проблемы: арифметическое кодирование нуждается в большой (в границе бесконечной) точности вычислений, которая приводит к недопустимо высокой сложности реализации; для формирования кодового слова формально необходима вся последовательность сообщений, которая приводит к недопустимо большой задержке кодирования, которая равняется длине кодированной последовательности сообщений.

Марсель Голей (Marcel J.E. Golay, 1902 - 1989), математик, физик, специалист в области теории информации, который применял математику к реальным практическим военным и промышленным задачам. Родился в Швейцарии. Изучал электротехнику в Швейцарском федеральном технологическом институте в Цюрихе. Работал над решением многих задач, включая газовую хроматографию и оптическую спектроскопию. Его достижение: разработка кодов Голея, обобщение абсолютных бинарных кодов Хэмминга к небинарным кодам, открытие комплементарных последовательностей.

Обе проблемы можно преодолеть. Решение заключается в том, что ту часть данных, которая не принимает участия в дальнейших вычислениях и уже не влияет на окончательный результат, можно изъять из вычислений и выдать на выход кодера. Тем самым уменьшается сложность вычислений и задержка кодирования. Чтобы объяснить, как это делается, рассмотрим работу декодера.

Начнем из анализа работы декодера кода Гильберта - Мура. Пусть для источника $X = \{1, ..., M\}$ известные вероятности $\{p_1, ..., p_i\}$, по которым вычислены $q_m = \sum_{j=1}^{m-1} q_j$, $\sigma_m = q_m + p_m / 2$, длины слов $l_m = \lceil \log(p_m / 2) \rceil$ и кодовые слова длиной l_m, полученные округлением значения σ_m. Округленные к l_m разрядам после запятой числа σ_m обозначим через $\hat{\sigma}_m$. Задача декодера состоит в восстановлении сообщения m по соответствующему округленному значению $\sigma_m = \hat{\sigma}_m$. Эта задача выполняется с помощью приведенного дальше алгоритма.

Пример 7.8. Рассмотрим источник $X = \{a, b, c\}$ с распределением вероятностей $p_a = 0,1$; $p_b = 0,6$; $p_c = 0,3$. Допустим, что на вход декодера поступила двоичная последовательность 0100010101. За этой последовательностью нужно восстановить последовательность закодированных сообщений. Проще возвратиться на несколько страниц назад, ведь в примере 7.7 мы получили именно эту последовательность на выходе кодера. Промежуточные результаты вычислений, выполненных согласно описанному алгоритму декодирования арифметического кода, приведены в табл. 7.6, из которой вытекает, что последовательность сообщения восстановлена правильно.

Возвратимся к вопросу о вычислительной сложности. Сущность проблемы заключается в том, что все вычисления, как в кодере, так и в декодере выполняются с переменными, разрядность которых равняется длине закодированной последовательности сообщений.

Таблица 7.6

Шаг	S	G	Гипотеза x	$q(x)$	$S+qG$	Решение x_i	$p(x)$
$100010101 \to 0.100010101 \to \widehat{F} = 0.541$							
1	0,0000	1,0000	*a*	0,0	$0.0000 < \widehat{F}$	*b*	0,6
			b	0,1	$0.1000 < \widehat{F}$		
			c	0,7	$0.7000 < \widehat{F}$		
2	0,1000	0,6000	*a*	0,0	$0.1000 < \widehat{F}$	*c*	0,3
			b	0,1	$0.1600 < \widehat{F}$		
			c	0,7	$0.5200 < \widehat{F}$		
3	0,5200	0,1800	*a*	0,0	$0.5200 < \widehat{F}$	*b*	0,6
			b	0,1	$0.5380 < \widehat{F}$		
			c	0,7	$0.6460 > \widehat{F}$		
4	0,5380	0,1080	*a*	0,0	$0.5380 < \widehat{F}$	*a*	0,1
			b	0,1	$0.5488 > \widehat{F}$		
5	0,5380	0,0108	*a*	0,0	$0.5380 < \widehat{F}$	*b*	0,6
			b	0,1	$0.5391 < \widehat{F}$		
			c	0,7	$0.5456 < \widehat{F}$		

Возвратимся к вопросу о вычислительной сложности. Сущность проблемы заключается в том, что все вычисления, как в кодере, так и в декодере выполняются с переменными, разрядность которых равняется длине закодированной последовательности сообщений.

Рассмотрим подробнее работу кодера арифметического кода при кодировании последовательности из предыдущего примера. Обратимся к табл. 7.5. Видим, что после 3-го шага кодирования значение F уже не станет меньше 0,5 и больше 0,75 независимо от того, какими будут следующие сообщения источника. Отсюда вытекает, что первые два символа числа F уже не изменятся и могут быть переданы по каналу. После этого можно выполнить нормирование $F \leftarrow 2(F - 0{,}5)$, $G \leftarrow 2G$ и продолжить кодирование. Точно такое же нормирование выполняет и декодер. Тем самым разрядность сменных сократится без потери в точности вычислений. Нетрудно сформулировать общее правило выполнения такого рода нормирования. Трудности возникают в том случае, когда двоичное представление F содержит серию единиц после нуля (число F близкое к 0,5, но меньше, чем 0,5), т.е. $F = 0{,}01111...$. Неопределенность устраняется одним из двух способов. Или появится нулевой разряд $(F = 0{,}01111...10...)$, или состоится перенесение в одном из младших разрядов $(F = 0{,}10000...0...)$. Если состояние неопреде-

ленности будет длиться долго, то разрядности ячеек, используемых для хранения F и G, может оказаться недостаточно.

Выход из положения заключается в том, что кодер, соприкоснувшись с такой ситуацией, изменяет форму хранения числа F. Количество разрядов уменьшается за счет того, что в памяти сохраняется не сама серия единиц, а ее длина. В момент появления нуля на выход выдается последовательность вида 0111...1. Если же неопределенность завершилась перенесением, то на выход поступает 1000...0.

Аналогичные приемы применяются при реализации декодера. Кроме того, важно, что для вынесения решений относительно переданных символов не нужна вся последовательность $\hat{F}$, решения о сообщении можно выдать получателю, проанализировав часть кодовой последовательности. На примере 7.9 можно отследить, какой будет задержка принятия решений для каждого из пяти сообщений. Результаты вычисления приведены в табл. 7.7.

Таблица 7.7

Шаг i	Двоичный символ	Нижняя граница $\hat{F}$	F	G
1	*1*	0,5000	0,1000	-
2	*0*	0,5000	0,7500	*b*
3	*0*	0,5000	0,6250	
4	*0*	0,5000	0,5625	
5	*1*	0,5313	0,5625	*c,d*
6	*0*	0,5313	0,5469	*a*
7	*1*	0,5390	0,5469	
8	*0*	0,5390	0,5430	*b*
9	*1*	0,5410	0,5430	

Выяснилось, что в этом случае первый символ можно выдать получателю после получения из канала двух двоичных символов. По первым пяти символам можно вычислить три первых сообщения, а для однозначного восстановления всех пяти сообщений достаточно 8 бит. Последний 9-й бит в этом случае оказался неиспользованным.

Коды с памятью. По обыкновению рассматривают два типа кодов с памятью: блочные коды и блочные коды с конечной памятью.

Блочный код *- это код, который разделяет вектор данных на блоки заданной длины и каждый указанный блок последовательно заменяет кодовым словом из префиксного множества двоичных слов. Образованную последовательность кодовых слов объединяют в результирующую двоичную строку на выходе кодера.*

О блочном коде говорят, что он представляет собой блочный код k-го порядка, если все блоки имеют длину, равную k.

Пример 7.9. Сожмем вектор данных $X = (0,1,1,0,1,0,0,1,1,1,1,0,0,1,0,1)$, воспользовавшись блочным кодом 2-го порядка. Сначала разобьем

вектор X на блоки длиной 2:01, 10, 10, 01, 11, 10, 01, 01. Будем рассматривать эти блоки как элементы нового «гипералфавита» $\{01, 10, 10\}$. Чтобы определить, какой код назначить этому или другому символу этого нового алфавита, определим вектор частот появлений элементов дополнительного алфавита в последовательности блоков. Получим вектор частот $(4, 3, 1)$, т.е. блок 01, что случается чаще всего - четыре раза, следующий за частотой появления блок 10 - он случается трижды, и блок 11, что случается наиболее реже - лишь один раз.

Оптимальный вектор Крафта для вектора частот (4, 3, 1) - это вектор (1, 2, 2). Таким образом, кодер для оптимального блочного кода 2- го порядка относительно заданного вектора данных X определяется табл. 7.8.

Заменяя каждый блок присвоенным ему кодовым словом из таблицы кодера, получаем последовательность кодовых слов: 0, 10, 10, 0, 11, 10, 0, 0.

Таблица 7.8

Таблица кодера	
Блок	Кодовое слово
01	0
10	10
11	11

Объединяя эти кодовые слова вместе, имеем исходную последовательность кодера:

$$B(X) = (0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0).$$

Можно проверить, что энтропия $H(X)$ начального вектора X равна $0{,}9887$ бит/букву, а степень сжатия, который получаем в результате использования блочного кода $R(X) = 12/16 = 0{,}75$ бит на выборку данных вышел меньшим нижней границы энтропии.

Полученный результат, казалось бы, противоречит теореме Шеннона, которая утверждает, что невозможно достичь средней длины кода меньше, чем энтропия источника (см. гл. 5). Тем не менее, на самом деле такого разногласия нет. Если рассмотреть вектор данных X, то можно заметить закономерность: за символом 0 чаще всего идет 1, т.е. условная вероятность $P(1/0)$ существенным образом большая, чем просто $P(0)$. Итак, энтропию этого источника нужно вычислять как энтропию сложного сообщения, а она, как известно, всегда меньшая, чем для источника простых сообщений.

Код с конечной памятью *- это такой код, который при кодировании вектора данных* $(X_1, X_2, ..., X_n)$ *использует кодовую книгу* (*словарь*), *который состоит из нескольких разных кодов без памяти. Каждая выборка данных кодируется таким кодом без памяти из кодовой книги, который определяется значением некоторого количества предыдущих выборок данных.*

Кодер с памятью *- это такой кодер, который при кодировании текущего символа учитывает значение предыдущего символа.*

Таким образом, кодовое слово для текущего символа A будет разным в комбинациях $RA, DA^3 CA$ (другими словами, код имеет память в один символ источника) (табл. 7.9).

Результат кодирования – вектор $B(X) = (10111011111011)$, имеющий длину 11 бит и скорость сжатия $R = 13/11 = 1{,}18$ бит на символ данных, тогда как при кодировании равномерным триразрядным кодом из $R = 3$ понадобилось бы 33 бит.

Таблица 7.9

Кодер	
Символ, предыдущий символ	Кодовое слово
(*A*, -)	1
(*B*, *A*)	0
(*C*, *A*)	10
(*D*, *A*)	11
(*A*, *R*)	1
(*R*, *B*)	1
(*A*, *C*)	1
(*A*, *B*)	1

Метод Зива - Лемпела. Практически все словарные методы кодирования принадлежат к алгоритмам, предложенным в работе двух израильских ученых Абрама Лемпела и Якоба Зива, опубликованной в 1977 г.

Суть указанных алгоритмов заключается в том, что фразы в тексте, который сжимается, заменяют показателем в том месте, где они в этом тексте уже раньше появлялись.

Эту семью алгоритмов называют *методом Зива - Лемпела* и обозначают как *LZ*-сжатие. Этот метод быстро приспосабливается к структуре текста и дает возможность кодировать короткие функциональные слова, поскольку они очень часто в нем появляются. Новые слова и фразы могут также формироваться из частей слова, которые раньше случались.

Декодирование сжатого текста осуществляется непосредственно - происходит простая замена показателя готовой фразой со словаря, на которую тот указывает. На практике *LZ*-метод дает результативное сжатие, его важным свойством является быстрая работа декодера. (Когда мы говорим о тексте, то допускаем, что кодированию подвергается некоторый вектор данных с конечным дискретным алфавитом, причем это не обязательно текст в буквальном понимании этого слова.)

Большинство словарных методов кодирования названо в честь авторов идеи метода Зива и Лемпела, и часто считают, что все методы используют

один и тот же самый алгоритм кодирования. На самом деле разные представители этой семьи алгоритмов отличаются в деталях работы.

Словарные методы кодирования можно разбить на две группы.

1. ***Методы первой группы*** *- находя в кодированной последовательности цепочки символов, которые раньше уже встречались, вместо того, чтобы повторять эти цепочки, заменяют их показателями на предыдущие повторения.*

Словарь в этой группе алгоритмов в неявном виде содержится в обрабатываемых данных, а сохраняются лишь показатели на цепочки символов, которые повторяются.

Все методы этой группы базируются на алгоритме, разработанным Лемпелем и Зивом, который получил название *LZ*77. Наиболее совершенным представителем этой группы, которая воплощает в себе все соответствующие научные достижения, есть алгоритм, опубликованный в 1982 г. Сторером и Шиманским и известный как *LZSS*.

Процедуру такого кодирования иллюстрирует рис. 7.16.

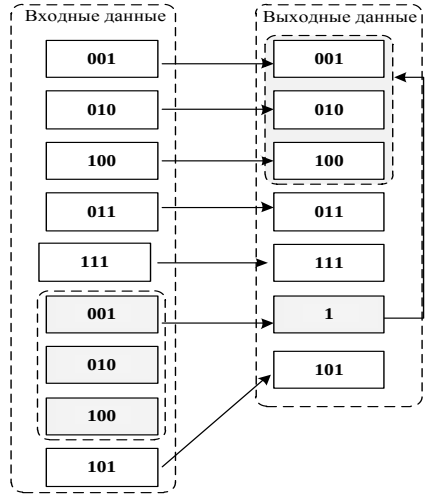


Рис. 7.16. Процедура кодирования соответственно алгоритмам первой группы

2. ***Методы второй группы*** *- в дополнение к начальному словарю источника создают словарь фраз, которые являются повторными комбинациями символов начального словаря, которые встречаются во входных данных.*

При этом размер словаря источника возрастает и для его кодирования понадобится большее количество бит, но значительную часть этого словаря будут представлять уже не отдельные буквы, а целые слова. Когда кодер обнаруживает фразу, которая раньше уже встречалась, он заменяет ее индексом словаря, который содержит эту фразу. При этом длина кода индекса выходит намного меньшей чем длина кода фразы.

Все методы этой группы базируются на алгоритме, разработанным и опубликованным Лемпелем и Зивом в 1978 г., - *LZ*78. Наиболее совершенным на данный момент представителем этой группы словарных методов есть алгоритм *LZW*, разработанный 1984 года Тери Велчем.

Идею алгоритмов этой группы делает наглядным рис. 7.17.

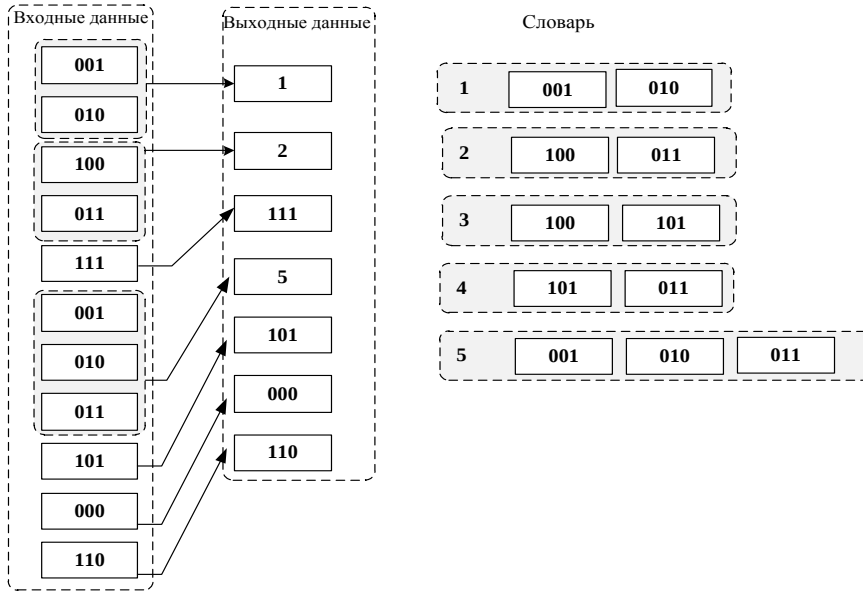


Рис. 7.17. Процедура кодирования соответственно алгоритмам второй группы

Принцип работы алгоритмов второй группы объяснить намного проще, чем принцип алгоритмов первой группы, поэтому начнем рассмотрение принципа действия *LZ*-кодеров из алгоритма *LZW*.

Процесс сжатия приобретает довольно простой вид. Последовательно прочитываем символы входного потока (строка) и проверяем, есть ли в уже созданной таблице такая строка. Если строка есть, то считываем следующий символ, а если такой строки нет, ставим в исходный поток код для предыдущего найденного ряда, заносим его в таблицу и начинаем поиск снова.

LZW-декодер, обрабатывая входной поток закодированных данных, возобновляет из него начальные данные. Равно как и алгоритм сжатия, декодер прибавляет красные строки к словарю каждый раз, когда находит во входном потоке новый код. Все, что ему остается сделать, - это превратить входной код в исходную строку символов и отдать его на выход кодера.

Преимуществом этого способа сжатия есть то, что весь словарь новых символов передается декодеру собственно без передачи. В конце процесса декодирования декодер имеет такой же словарь новых символов, какой в

процессе кодирования был накоплен кодером, при этом его создание было частью процесса декодирования.

Кодирование длин повторений. Кодирование длин участков (или повторений) может быть достаточно эффективным при сжатии двоичных данных, например черно-белых изображений, которые содержат множество прямых линий и однородных участков, схем и т.д.

Идея сжатия данных на основе кодирования длин повторений заключается в том, что вместо кодирования самих данных прибегают к кодированию чисел, которые отвечают длинам участков, на которых данные сохраняют неизменное значение.

Предположим, что нужно закодировать двоичное (двухцветное) изображение размером 8×8 элементов. Просканировав это изображение по строкам (двум объектам на изображении будут соответствовать 0 и 1), в результате получим двоичный вектор данных, например:

$$X = (011100001111000000100000001000000010000000111100001111101111101111)$$

длиной 64 бит (скорость начального кода составляет 1 бит на элемент изображения).

Выделим в векторе X участки, на которых данные сохраняют неизменное значение, и определим их длины. Результирующая последовательность длин участков - положительных целых чисел, отвечающих начальному вектору данных X, - будет иметь вид

$$r = (1, 3, 4, 4, 7, 1, 7, 1, 7,1, 7, 4, 3, 4, 1, 4, 1, 4).$$

Теперь эту последовательность, в которой заметна определенная повторяемость, можно закодировать некоторым статистическим кодом, например кодом Хаффмана без памяти (табл. 7.10).

Таблица 7.10

Кодер	
Длина участка	Кодовое слово
4	0
1	10
7	110
3	111

Ирвинг Рид (Irving S. Reed, 1923), математик и инженер. Родился в Сиетле. Получила степень доктора философии в области математики в Калифорнийском технологическом институте (1949). Член Национальной инженерной академии и Института инженеров по электротехнике и электронике, а также лауреат премии имени Клода Элвуда Шеннона. Главное достижение - разработка вместе с Густавом Соломоном алгебраических кодов выявления и исправление ошибок, известных как коды Рида-Соломона. Также является соизобретателем кодов Рида-Мюллера. Сделал значительный вклад в электротехнику, в частности радиолокацию, сигналов и изображений.

Для того, чтобы указать, что кодированная последовательность начинается с нуля, прибавим в начале кодового слова префиксный символ 0. В результате получим кодовое слово

$$B(r) = (0100011010110101101011001110100100)$$

длиной 34 бит. Тогда результирующая скорость R будет составлять 34/64 или немного больше, чем 0,5 бит на элемент изображения. При сжатии изображений большего размера и повторении элементов, содержащих множество, эффективность сжатия может быть существенно выше.

Дальше приведен другой пример использования кодирования длин повторений, когда в цифровых данных случаются участки с большим количеством нулевых значений. Каждый раз, когда в потоке данных случается «*нуль*», он кодируется двумя числами. Первое - нуль, который является показателем начала кодирования длинны потока нулей, и второе - равняется количеству нулей в очередной группе. Если среднее количество нулей в группе больше двух, имеем сжатие. Тем не менее, большое количество отдельных нулей может привести даже к увеличению размера кодированного файла (рис. 7.18).

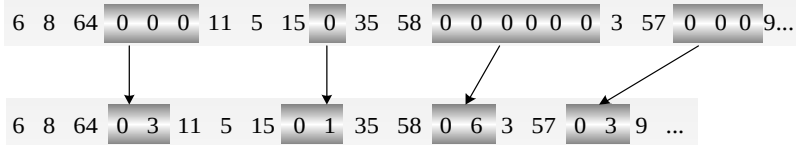


Рис. 7.18 . Графическое изображение алгоритма кодирования по длине повторов нулей

Еще одним простым и широко используемым для сжатия изображений и звуковых сигналов методом неразрушительного кодирования является метод дифференциального кодирования.

Дифференциальное кодирование. Работа дифференциального кодера основывается на том, что для многих типов данных разность между соседними отсчетами сравнительно небольшая, даже если сами данные имеют большие значения. Например, нельзя ожидать большой разности между соседними пикселями цифрового изображения.

Приведенный далее простой пример показывает, какое преимущество может иметь дифференциальное кодирование (кодирование разности между соседними отсчетами) сравнительно с простым кодированием без памяти (кодированием отсчетов независимо друг от друга).

Просканируем 8-битовое (256-уровневое) цифровое изображение, при этом 10 последовательных пикселей будут равняться 144, 147, 150, 146, 141, 142, 138, 143, 145, 142.

Закодировав эти данные последовательно пиксель за пикселем некоторым кодом без памяти, который использует 8 бит на пиксель изображения, получим кодовое слово, которое содержит 80 бит.

Предположим теперь, что прежде чем подвергать отсчеты изображения кодированию, мы вычислим разности между соседними пикселями. Эта процедура даст нам последовательность, приведенную на рис. 7.19.

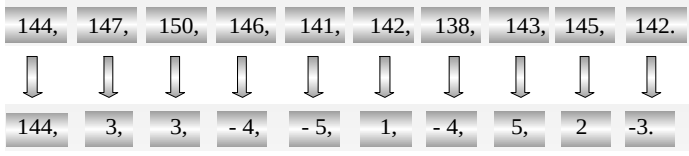


Рис. 7.19. Алгоритм дифференциального кодирования

Начальную последовательность можно легко восстановить с разностной простым суммированием (дискретной интеграцией), как это изображено на рис. 7.20.

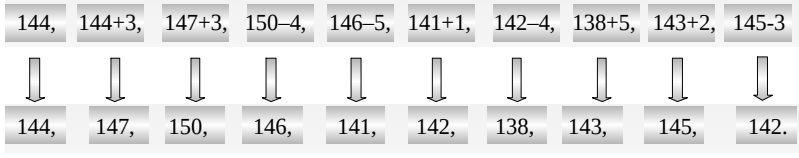


Рис. 7.20. Алгоритм дифференциального декодирования

Для кодирования первого числа из полученной последовательности разностей отсчетов, как и раньше, понадобится 8 бит, остаток чисел можно закодировать 4-битовыми словами (1 знаковый бит и 3 бита на кодирование модуля числа). Таким образом, в результате кодирования получим кодовое слово длиной $8 + 9 \cdot 4 = 44$ бит, т.е. почти вдвое короче, чем при индивидуальном кодировании отсчетов.

Метод дифференциального кодирования очень широко используется в тех случаях, когда природа данных такая, что их соседние значения немного отличаются одно от другого, но сами значения могут быть какими угодно большими. Это касается звуковых сигналов, особенно языка, изображений, соседние пиксели которых имеют практически одинаковую яркость и цвет. Тем не менее, этот метод абсолютно непригоден для кодирования текстов, черчений или любых цифровых данных с независимыми соседними значениями.

7.2. Помехоустойчивое кодирование

Искажение информации, т.е. нарушение ее целостности, возможно на любом этапе ее циркуляции в информационно-телекоммуникационных сетях: при хранении, передаче или обработке.

Под *целостностью информации* понимается ее свойство, которое заключается в том, что информация не может быть модифицирована неавтори-

зованным пользователем или процессом. Другими словами, под целостностью информации понимается отсутствие в ней любых искажений (модификаций), которые не были санкционированы ее владельцем, независимо от причин или источников возникновения таких искривлений.

Причины таких искажений могут быть случайными или намеренными. В свою очередь, случайные искажения могут быть как естественными, связанными с действием естественных факторов, так и искусственными. Случайные искусственные искажения связанны с деятельностью людей - со случайными ошибками персонала. Намеренные искажения всегда связанные с целенаправленными действиями нарушителей. И те, и другие действия имеют своим следствием искажения некоторого количества символов в цифровом представлении информации (независимо от используемой системы исчисления и формы представление информации) и в этом понимании являются угрозами функциональным свойствам защищенности информационных ресурсов - их целостности и доступности. Поэтому задача обеспечить целостность и доступность информационных ресурсов является одной из наиболее актуальных при разработке и эксплуатации информационных систем и их элементов. Эта актуальность подтверждается и требованиями относительно допустимой вероятности $D_{\text{ї}}$ ошибок в сообщениях, которую следует понимать как вероятность нарушения целостности информационных объектов, которые обрабатываются (если передача и обработка информации осуществляются в виде сообщений). Для обеспечения контроля и возобновления целостности информационных объектов, в частности для восстановления разрушенной информации, в состав информации включают избыточную информацию - признак целостности или контрольный, процедура формирования которой известна и принадлежит к помехоустойчивым методам кодирования.

Задача помехоустойчивого корректирующего кодирования *- обеспечить целостность информационных сообщений с применением помехоустойчивых корректирующих кодов.*

Помехоустойчивым корректирующим кодированием *называется такой вид кодирования, который дает возможность реализовывать программные, аппаратные или программно-аппаратные средства выявления и устранения искажений в информационных сообщениях.*

Кодирование с исправлением ошибок является, в сущности, методом обработки сигналов, предназначенным для увеличения надежности передачи информационных потоков по цифровым каналам связи. Хотя разные схемы кодирования очень отличаются одна от другой и основываются на разных математических теориях, всем им присущие два общих свойства.

1. *Использование избыточности информации относительно кодового слова.* Закодированные цифровые сообщения всегда содержат дополнительные (избыточные) символы. Их используют для того, чтобы сделать более выразительным индивидуальность каждого сообщения, устанавливая правило или алгоритм восстановления целостности информации. Эти символы выби-

рают так, чтобы сделать маловероятной потерю сообщением его индивидуальности через искажение вследствие влияния помех.

2. *Усреднение влияния помех.* Эффект усреднения достигается за счет того, что избыточные символы зависят от нескольких информационных символов. В каналах с помехами эффективным средством повышения достоверности передачи сообщений есть помехоустойчивое кодирование. Оно основывается на применении специальных кодов, которые корректируют ошибки, вызванные действием помех.

Корректирующим *называется такой код, который дает возможность обнаруживать или и обнаруживать, и исправлять ошибки при приеме сообщений.* Существует много корректирующих кодов, которые различаются как по принципам построения, так и основными характеристиками (рис. 7.21).

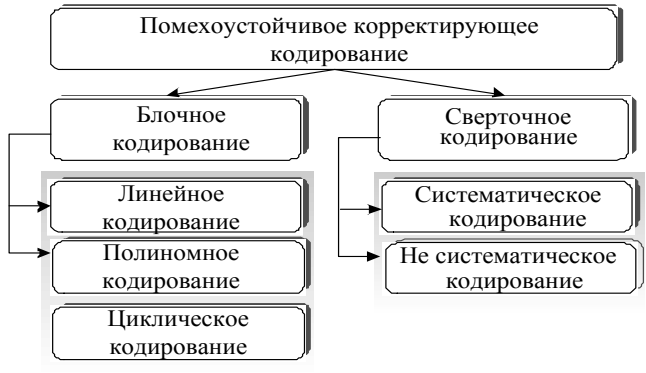


Рис. 7.21. Классификация помехоустойчивого кодирования

Обнаруживающим *называется такой код, с помощью которого только обнаруживаются ошибки.*

Исправление ошибки при таком кодировании по обыкновению происходит путем повторения обезображенных сообщений. Запрос о повторении передается по каналу обратной связи.

Исправляющим *называется такой код, который фиксирует не только сам факт наличия ошибок, а и устанавливает, какие кодовые символы приняты с ошибкой, также дает возможность их исправить без повторной передачи.*

Известны также коды, в которых исправляется только часть выявленных ошибок, а сдача ошибочных комбинаций передается повторно.

Для того чтобы код имел корректирующие свойства, в кодовой последовательности должны содержаться дополнительные (избыточные) символы, предназначенные для корректирования ошибок. Чем больше избыточность кода, тем выше его корректирующая способность.

Базовым алгоритмом для всех корректирующих кодов является использование избыточности для исправления ошибок, которые могут возникнуть в

процессе передачи или хранения информации. Согласно основной схеме, чрезмерные символы дописываются вслед за информационными, образовывая кодовую последовательность или кодовое слово.

Кодовое слово, сформированное за процедурой блочного кодирования, изображено на рис. 7.22.

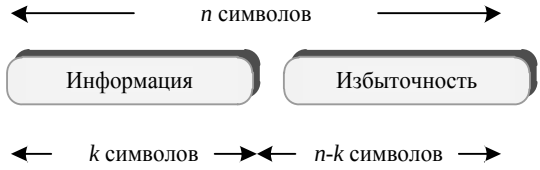


Рис. 7.22. Систематическое блочное кодирование для исправления ошибок

Такое кодирование называют *систематическим*. Это означает, что информационные символы всегда появляются на первых позициях кодового слова. Символы на *n-k* позициях, которые остались, являются разными функциями от информационных символов, которые обеспечивают тем самым чрезмерность, необходимую для выявления или исправления ошибок. Множество всех кодовых последовательностей называют *кодом, который исправляет ошибки*.

Блоковые и сверточные коды. Соответственно тому, как вводится чрезмерность в сообщение, корректирующие коды можно поделить на два класса: блоки и свертку кода. Обе схемы кодирования нашли практическое применение.

Блоковым кодированием *называется алгоритм кодирования, согласно которому каждый блок информационных символов обрабатывается независимо от других блоков информационного сообщения. Т.е. блочное кодирование является* ***операцией без памяти*** *в том смысле, что кодовые слова не зависят один от другого.*

Сверточным кодированием (***циклическо-замкнутым***) *называется алгоритм кодирования, согласно которому кодер зависит не только от информационных символов в данный момент, а и от предыдущих символов на его входе или выходе* (***кодирование с памятью***).

Чтобы упростить объяснение, мы начнем с изучения структурных свойств блочных кодов. Множество свойств является общим для обоих типов кодов. Заметим, что на самом деле блочные коды имеют память, если рассматривать кодирование как побитовый процесс в пределах кодового слова.

Принципы помехоустойчивого кодирования. В теории помехоустойчивого кодирования важным является вопрос об использовании чрезмерности для корректирования ошибок, которые возникают при передаче.

Для равномерных кодов, которые являются основными для помехоустойчивого кодирования, количество возможных комбинаций $M = 2^n$, где n - значительность кода. В обычном некорректирующем коде без чрезмерности

(например, в коде Бодо) количество комбинаций M выбирается таким, которое равняется количеству сообщений алфавита источника M_0, причем все комбинации используются для передачи информации. Корректирующие коды строятся так, чтобы количество комбинаций M превышало количество сообщений источника M_0. Тем не менее в этом случае лишь M_0 комбинаций из общего их количества используется для передачи информации. Эти комбинации называются *разрешенными*, а $M - M_0$ комбинаций - *запрещенными*. На приемном конце в декодирующем устройстве известно, какие комбинации являются разрешенными, а какие запрещенными. Поэтому если в результате ошибки передачи разрешенная комбинация превратится в некоторую запрещенную комбинацию, то такую ошибку будет обнаружено, а при определенных условиях и исправлено. Естественно, ошибки, которые приводят к образованию другой разрешенной комбинации, не обнаруживаются.

Расстоянием между комбинациями *называют количество символов, которыми различаются переданные комбинации равномерного кода. Расстояние* d_{ij} *между двумя комбинациями* A_i *и* A_j *определяется количеством единиц в сумме комбинаций по модулю два.* Например:

$$\begin{array}{r} 110011\ A_i \\ \underline{010110\ A_j} \\ 100101\ d_{ij} = 3 \end{array}$$

Кодовым расстоянием d *для любого кода* $d_{ij} \le n$ *называется минимальное расстояние между разрешенными комбинациями в этом коде.*

Расстояние между комбинациями A_i и A_j условно обозначено на рис. 7.23, *а*, где изображены промежуточные комбинации, которые отличаются одна от другой одним символом.

В общем случае некоторая пара разрешенных комбинаций A_{p1} и A_{p2}, разделенных кодовым расстоянием d, изображается на прямой (см. рис. 7.23, *б)*, где точками обозначаются запрещенные комбинации. Для того чтобы в результате ошибки комбинация A_{p1} превратилась в другую разрешенную комбинацию A_{p2}, должны исказиться d символов.

В случае искажения меньшего количества символов комбинация A_{p1} перейдет в запрещенную комбинацию и ошибку будет обнаружено.

Ошибка всегда обнаруживается, если ее кратность, т.е. количество искаженных символов в кодовой комбинации

$$g \le d - 1. \qquad (7.9)$$

Если $g > d$, то некоторые ошибки также обнаруживаются. Тем не менее, полной гарантии обнаружения ошибок здесь нет, поскольку ошибочная ком-

бинация в этом случае может совпасть с какой-нибудь разрешенной комбинацией.

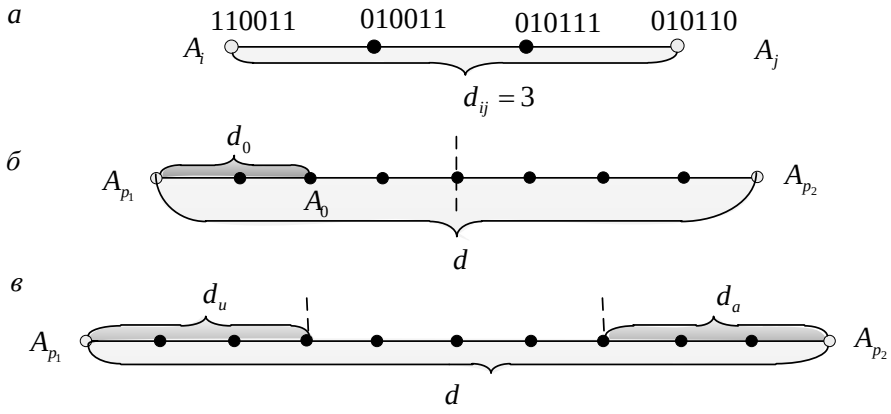


Рис. 7.23. Геометрическое представление разрешенных и запрещенных кодовых комбинаций

***Минимальное кодовое расстояние**, при котором обнаруживаются любые одиночные ошибки,* $d=2$.

Процедура исправления ошибок в процессе декодирования сводится к определению переданной комбинации при известной принятой. Расстояние между переданной разрешенной комбинацией и принятой запрещенной комбинацией d_0 равняется кратности ошибок g. Если ошибки в символах комбинации случаются независимо одна от другой, то вероятность искажения некоторых g символов в *n*-значной комбинации

$$P_0^g(1-P_0)^{n-g}, \tag{7.10}$$

где P_0 - вероятность искажения одного символа. Поскольку по обыкновению $P_0 << 1$, то вероятность многократных ошибок уменьшается с увеличением их кратности, при этом вероятнее меньшее расстояние d_1 чи d.

При этих условиях исправление ошибок может происходить по такому правилу: *если принята запрещенная комбинация, то считается переданной ближайшая разрешенная комбинация*.

Например: пусть образовалась запрещенная комбинация A_0 (рис. 7.23, *б)*, тогда принимается решение, что была передана комбинация A_1. Это правило декодирования для указанного распределения ошибок является оптимальным, поскольку обеспечивает исправление максимального количества ошибок. Напомним, что аналогичное правило используется в теории потенциальной помехоустойчивости при оптимальном приеме дискретных сигналов, когда решение сводится к выбору того переданного сигнала, который

меньше всего отличается от принятого (см. гл. 6). При таком правиле декодирования будут исправлены все кратные ошибки:

$$g \le \frac{d-1}{2}. \tag{7.11}$$

Минимальное значение d, при котором еще возможно исправление любых одиночных ошибок, равно 3.

Возможно также построение кодов, в которых часть ошибок исправляется, а часть только обнаруживается. Так, согласно рис. 7.23, *в* ошибки, кратные $g \le d_{\text{âèï}}$, исправляются, а ошибки, кратность которых лежит в пределах $d_{\text{âèï}} \le i \le d - d_{\text{âèÿâ}}$ только обнаруживаются, тем не менее, при их исправлении принимается ошибочное решение - считается переданной комбинация A_{p1} вместо A_{p2}, или наоборот.

Исправление ошибок - сложная задача, практическое выполнение которой связано с осложнением кодирующих и декодирующих устройств. Поэтому коды по обыкновению используются для корректирования ошибок малой кратности.

Корректирующая способность кода возрастает с увеличением d. При фиксированном количестве M_0 разрешенных комбинаций увеличить d возможно лишь за счет роста количества запрещенных комбинаций

$$M - M_0 = 2^n - 2^k, \tag{7.12}$$

что, в свою очередь, нуждается в чрезмерном количестве символов $r = n - k$, где k - количество символов в комбинации кода без избыточности. Можно ввести понятие избыточности кода, количественно определив ее как

$$\chi = \frac{n-k}{n} = 1 - \frac{\log_2 M_0}{\log_2 M}. \tag{7.13}$$

В случае независимых ошибок вероятность появления g ошибочных символов в *n*-значной кодовой комбинации выражается формулой (7.10), а количество всех возможных появлений g ошибочных символов в *n*-значной комбинации зависит от ее длины и определяется известной комбинаторной формулой

$$C_n^g = \frac{n!}{g!\,(n-g)!}.$$

Отсюда полная вероятность ошибки кратности g, которая учитывает все возможные появления ошибочных символов, описывается выражением

$$P_{0g} = C_n^g P_0^g = \left(1 - P_0\right)^{n-g}. \tag{7.14}$$

Используя формулу (7.14), можно определить вероятность отсутствия ошибок в кодовой комбинации, т.е. вероятность правильного приема

$$P_{\text{пр}} = (1-P_0)^n,$$

и вероятность правильного корректирования ошибок

$$P_{\text{кор}} = \sum_g P_{0g} = \sum_g C_n^g P_0^g (1-P_0)^{n-g}.$$

Здесь суммирование выполняется по всем значениям кратности ошибок g, которые обнаруживаются и исправляются. Таким образом, вероятность некорректированных ошибок

$$P_{\text{ош}} = 1 - P_{\text{пр}} - P_{\text{кор}} = 1-(1-P_0)^n - \sum_g C_n^g P_0^g (1-P_0)^{n-g}. \qquad (7.15)$$

Анализ выражения (7.15) показывает, что при малом значении P_0 и сравнительно небольших значениях n наиболее возможные ошибки малой кратности, которые и необходимо корректировать прежде всего.

Вероятность $P_{\text{пом}}$, избыточность χ и количество n символов являются основными характеристиками корректирующего кода, которые определяют уровень помехоустойчивости передачи дискретных сообщений и меру цены ее достижения.

Общая задача, которая возникает при создании кода, заключается в достижении наименьших значений $P_{\text{пом}}$ и χ. Целесообразность применения того или другого кода зависит также от сложности кодирующих и декодирующих устройств, которая, в свою очередь, зависит от χ. Во многих практических случаях эта сторона вопроса являются решающей. Часто, например, используются коды с большой избыточностью, но такие, которые имеют простые правила кодирования и декодирования.

Соответственно общему принципу корректирования ошибок, который основывается на использовании разрешенных и запрещенных комбинаций, необходимо сравнивать принятую комбинацию со всеми комбинациями данного кода. В результате M сравнений принимается решение о переданной комбинации. Этот способ декодирования логически является наиболее простым, тем не менее он нуждается в сложных устройствах, в которых должны запоминаться все комбинации кода. Поэтому на практике чаще всего используются коды, которые дают возможность с помощью ограниченного количества преобразований принятых кодовых символов получить из них всю информацию о корректированных ошибках. Изучению таких кодов и посвящены следующие подразделы.

Расстояние Хэмминга и свойства корректирования. Рассмотрим двоичный код C, который исправляет ошибки. Если не все из 2^n возможных двоичных векторов длины n будут передавать по каналу связи, то этот код может иметь свойство помехоустойчивости. Действительно, код C является подмножеством n-измеримого двоичного векторного пространства

$V_2 = \{0,1\}^n$, таким, при котором его элементы максимально отдалены один от другого.

Кодовое расстояние *в двоичном пространстве* V_2 *определяется как количество позиций, на котором два кодовых вектора* $x_1 = \left(x_{1,0}, x_{1,1}, ..., x_{1,\,n-1}\right)$ и $x_2 = \left(x_{2,0}, x_{2,1}, ..., x_{2,\,n-1}\right)$ *в этом пространстве не сохраняются.*

Расстояние Хэмминга *между векторами* x_1 *и* x_2, *обозначаемое как* $d_H(x_1, x_2)$, *определяется выражением*

$$d_H(x_1, x_2) = \left|\left\{i : x_{1,i} \neq x_{2,i},\ 0 \le i \le n\right\}\right|, \tag{7.16}$$

где вертикальные черточки, которые охватывают запись множества в правой части неравенства (*7.16*), *означают, что берется количество элементов в этом множестве.*

Минимальное расстояние *Хэмминга для заданного кода* C *определяется как минимум расстояния Хэмминга по всем возможным парам разных кодовых слов*:

$$d_{\min} = \min_{v_1, v_2 \in C} \left\{d_H(v_1, v_2) \middle| v_1 \neq v_2\right\}. \tag{7.17}$$

Запись $(n, k, d_{\min})$ означает, что рассматривается блочный код длинны n, который используется для кодирования сообщений длины k и имеет минимальное расстояние Хэмминга $d_{\min}$. Считается, что количество кодовых слов этого кода равняется $|C| = 2^k$.

Пример 7.10. Простым примером является код-повторение длины 3. Каждый информационный бит повторяется трижды. Таким образом, сообщение «0» кодируется вектором (000), а сообщение «1» — вектором (111). Поскольку эти два вектора различаются тремя позициями, хэмминговое расстояние между ними равно 3. Графическое изображение этого кода приведено на рис. 7.24. Трехмерное двоичное векторное пространство соответствует $2^3 = 8$ вершинам трехмерного единичного куба.

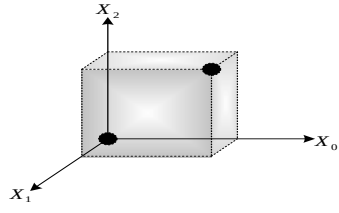


Рис. 7.24. (3,1,3) код-повторение в трехмерном векторном пространстве

Хэммингово расстояние между кодовыми словами (000) и (111), равно количеству вершин, через которые проходит совмещающий их путь, т.е. количество координат, которые необходимо изменить, чтобы превратить (000) в

(111), и наоборот. Итак, $d_H = \left[(000),(111)\right] = 3$. Поскольку в этом коде только два кодовых слова, то $d_{\min} = 3$.

Двоичное векторное пространство V_2 обычно называют *хэмминговым пространством*. Пусть v - кодовое слово кода C. *Хэмминговой сферой* $S_t(v)$ радиуса t с центром в точке v есть множество векторов (точек) в V_2, которые находятся от этого центра на расстоянии, меньшем или равным t:

$$S_t(v) = \left\{x \in C \middle| d_H(x,v) \le t\right\}. \tag{7.18}$$

Заметим, что количество слов (векторов) в $S_t(v)$:

$$\left|S_t(v)\right| = \sum_{i=0}^{t} \binom{n}{i}. \tag{7.19}$$

Пример 7.11. Хэмминговы сферы радиуса $t = 1$, которые окружают кодовые слова (3, 1, 3) двоичного кода-повторения, изображены на рис. 7.25.

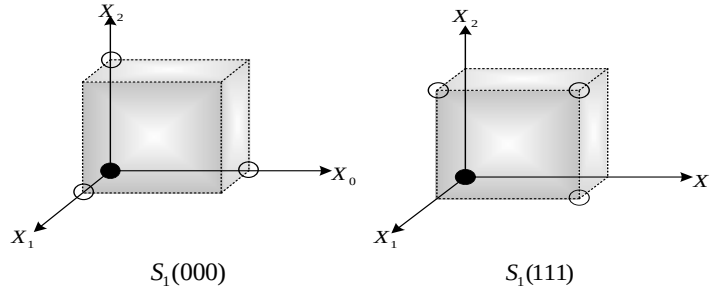


Рис. 7.25. Хэмминговы сферы радиуса $t = 1$, окружающие кодовые слова (3,1,3) двоичного кода-повторения

Укажем, что Хэмминговы сферы для этого кода не пересекаются, т.е. в пространстве V_2 нет векторов (или вершин в единичном трехмерном кубе), принадлежащих одновременно $S_1(000)$ и $S_1(111)$. Соответственно, если изменить любую одну позицию кодового слова v, то образуется вектор, который останется внутри хэмминговой сферы с центром в v. Эта идея принципиальна для понимания и определения корректирующей способности кода.

Корректирующей способностью t ***кода*** C *называют наибольший радиус хэмминговой сферы* $S_t(v)$ *для всех кодовых слов* $v \subseteq C$, *такой, при котором для любых разных пар* v_i, $v_j \subseteq C$, *соответствующие им хэмминговы сферы не пересекаются, т.е.*

$$t = \max\left\{l \middle| S_l(v_i) \bigcap S_l(v_j) = 0, v_i \ne v_j\right\}. \tag{7.20}$$

Это отвечает более распространенному определению

$$t = \left[\left(d_{\min} - 1\right)/2\right], \tag{7.21}$$

где $\lfloor x \rfloor$ - целая часть x, т.е. целое число, которое не превышает x.

Заметим, что для определения минимального кодового расстояния произвольного блочного кода C необходимо вычислить все $2^k(2^k-1)$ расстояния между разными парами кодовых слов. Это практически невозможно даже для сравнительно коротких кодов, например из $k=50$. Одной из важных преимуществ линейных блочных кодов является то, что для вычисления $d_{\min}$ достаточно знать только *хэмминговы весы* 2^k-1 ненулевых кодовых слов.

7.3. Блочное помехоустойчивое кодирование

Линейное блочное кодирование. Построение оптимального кода означает поиск в V_2 подмножества элементов, наиболее отдаленных один от другого. Это сильно сложная задача. Более того, если даже это сделано, то остается невыясненным, как назначить кодовые слова информационным сообщениям. Приведем общую классификацию блочного помехоустойчивого кодирования (рис. 7.26).

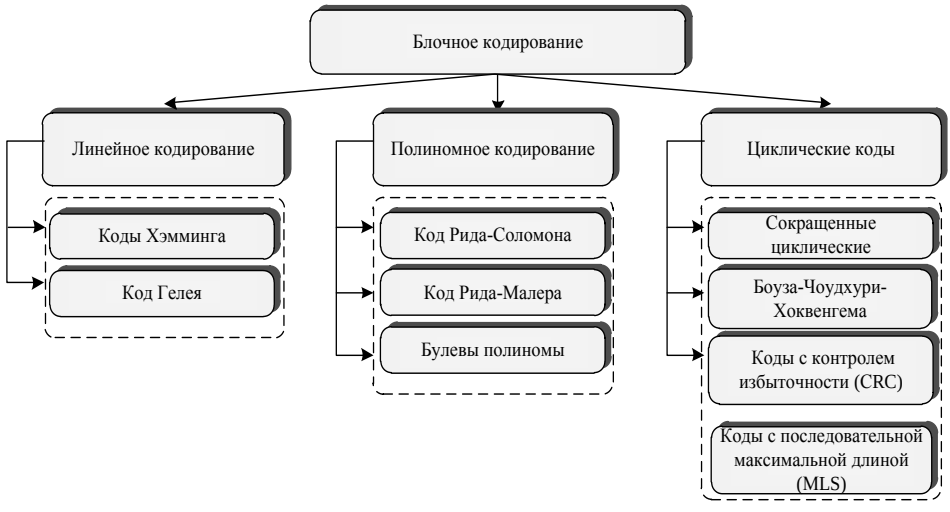


Рис. 7.26. Классификация блочного помехоустойчивого кодирования

Линейный код (множество кодовых слов) является векторным подпространством в пространстве V_2. Это означает, что операция кодирования может быть умножением на матрицу. Правила добавления и умножения двоичных чисел приведены в табл. 7.11.

Пусть C - двоичный линейный код $(n, k, d_{\min})$. Поскольку C представляет собой k-измеримое подпространство, то он имеет базис, например

$(v_0, v_1, ..., v_{k-1})$, такой, при котором любое кодовое слово $v \in C$ можно записать как линейную комбинацию элементов этого базиса

$$v = u_0 v_0 + u_1 v_1 + ... + u_{k-1} v_{k-1}, \tag{7.22}$$

где $u_i \in \{0,1\}$, $0 \le i < k$. Уравнение (7.22) можно дать в матричной форме через *породную матрицу* G и *вектор-сообщение* $u = (u_0, u_1, ..., u_{k-1})$:

$$v = uG, \tag{7.23}$$

где

$$G = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{k-1} \end{pmatrix} = \begin{pmatrix} v_{0,0} & v_{0,1} & \cdots & v_{0,k-1} \\ v_{1,0} & v_{1,1} & \cdots & v_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k-1,0} & v_{k-1,1} & \cdots & v_{k-1,k-1} \end{pmatrix}. \tag{7.24}$$

Таблица 7.11

a	b	$a + b$	ab
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	1

Поскольку C является k-измеримым векторным пространством в V_2, то существует $(n-k)$-измеримое дуальное пространство $C^{\perp}$, которое порождается строками матрицы H - так называемой *проверочной матрицы*, где $GH^{\mathrm{T}} = 0$, где H^T - транспонированная матрица H. Укажем, в частности, что любое кодовое слово $v \in C$ удовлетворяет условию

$$vH^T = 0. \tag{7.25}$$

Уравнение (7.25) является фундаментальным для декодирования линейных кодов.

Линейный код $C^{\perp}$, который генерируется матрицей H, является двоичным линейным кодом $(n, n-k, d^T_{\min})$, который называют *дуальным кодом* C.

Как уже отмечалось, линейные коды отличаются тем, что для определения минимального расстояния кода достаточно знать минимум хэмминговых весов ненулевых кодовых слов. Дальше этот факт будет доказан. Определим хэммингов вес $wt_H(x)$ вектора $x \in V_2$ как количество ненулевых элементов в x. Из определение хэммингового расстояния вытекает, что $wt_H(x) = d_H(x, 0)$. Для двоичного линейного кода C получаем

$$d_H(v_1, v_2) = d_H(v_1 + v_2, 0) = wt_H(v_1 + v_2). \tag{7.26}$$

Наконец, из свойства линейности кода имеем $v_1 + v_2 \in C$. Отсюда вытекает, что минимальное расстояние кода C можно вычислить как минимальный вес по всем $2^k - 1$ ненулевым кодовым словам. Эта задача существенно проще, чем полный перебор по всем парам кодовых слов, хотя и остается очень сложной даже для кодов среднего размера (или размерности k).

Кодирование и декодирование линейных блочных кодов. Равенство (7.23) определяет правило кодирования для линейного блочного кода, которым можно воспользоваться непосредственно. Если кодирование должно быть систематическим, то произвольную породную матрицу G линейного блочного $(n, k, d_{\min})$ кода C можно превратить в систематическую (каноническую) форму G_{sys} с помощью элементарных операций и перестановок столбцов матрицы. Матрица G_{sys} состоит из двух подматриц: единичной матрицы размера $k \times k$, что обозначается I_k, и проверочной подматрицей P размера $k \times (n - k)$. Таким образом,

$$G_{sys} = (I_k \,|\, P), \tag{7.27}$$

где

$$P = \begin{pmatrix} P_{0,0} & P_{0,1} & \cdots & P_{0,n-k-1} \\ P_{1,0} & P_{1,1} & \cdots & P_{1,n-k-1} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k-1,0} & P_{k-1,1} & \cdots & P_{k-1,n-k-1} \end{pmatrix}. \tag{7.28}$$

Поскольку $GH^T = 0$, то отсюда вытекает, что систематическая форма проверочной матрицы имеет вид

$$H_{sys} = (P^{\perp} \,|\, I_{n-k}). \tag{7.29}$$

Пример 7.12. Рассмотрим двоичный линейный (4, 2, 2) код с порождающей матрицей

$$G = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Перестановками второго и четвертого столбцов превратим эту матрицу в систематическую форму

$$G_{sys} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

Таким образом, проверочная подматрица $P = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$.

В этом случае выполняется соотношение $P = P^{\perp}$. Из формулы (7.29) вытекает, что систематическая форма проверочной матрицы имеет вид

$$H_{sys} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

В дальнейшем будут использованы обозначения $u = (u_0, u_1, ..., u_{k-1})$ для информационного сообщения и обозначения $v = (v_0, v_1, ..., v_{n-1})$ для соответствующего кодового слова кода *C*.

Если параметры *C* такие, что $k > (n-k)$, т.е. скорость кода $k/n < 1/2$, то кодирование с помощью порождающей матрицы нуждается в меньшем количестве логических операций. В этом случае

$$v = uG_{sys} = (u, v_p), \tag{7.30}$$

где $v_p = uP = (v_k, v_{k+1}, ..., v_{n-1})$ - проверочная часть кодового слова.

Тем не менее, если $k > (n-k)$ или $k/n > 1/2$, то кодирование с помощью проверочной матрицы *H* нуждается в меньшем количестве вычислений. Этот вариант кодирования обосновывается уравнением (7.25):

$$(u, v_p)H^T = 0.$$

Проверочные позиции $(v_k, v_{k+1}, ..., v_{n-1})$ вычисляются как

$$v_j = u_0\rho_{0,j} + u_1\rho_{1,j} + ... + u_{k-1}\rho_{k-1} \quad k \le j < n. \tag{7.31}$$

Можно сказать, что элементами систематической формы проверочной матрицы являются коэффициенты проверочных уравнений, из которых вычисляются проверочные символы.

Пример 7.13. Рассмотрим двоичный линейный (4, 2, 2) код из примера 7.11. Пусть сообщение и кодовые слова обозначены соответственно $u = (u_0, u_1)$ и $v = (v_0, v_1, v_2, v_3)$. Из уравнение (7.31) получаем

$$v_2 = u_0 = u_1, \; v_3 = u_0. \tag{7.32}$$

Соответствие между $2^2 = 4$ двухбитовыми сообщениями и кодовыми словами имеет вид

$$\begin{array}{ll} (00) & (0000) \\ (01) & (0110) \\ (10) & (1011) \\ (11) & (1110) \end{array} \tag{7.33}$$

Декодирование по стандартной таблице. Процедура декодирования находит кодовое слово *v*, самое ближнее к принятому с искажениями слову $r = v + e$, где *вектор ошибок* $e \in \{0,1\}^n$ образовывается двоичным симметричным каналом (см. гл. 6) в процессе передачи кодового слова. Модель изображена на рис. 7.27. По предположению переходная вероятность $p \le 1/2$.

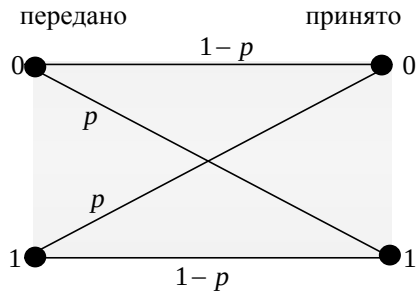


Рис. 7.27. Модель двоичного симметрического канала

Стандартной проверочной таблицей **(*или стандартной расстановкой)*** *для двоичного линейного* $(n, k, d_{\min})$ *кода C называется таблица всех возможных принятых из канала векторов r, организованная таким способом, который может быть наиболее близок к r кодовому слову v.*

Стандартная таблица содержит 2^{n-k} строк и 2^{k+1} столбцов (табл. 7.12). Размещенные по правую сторону 2^k столбцы таблицы содержат все векторы из пространства $V_2 = \{0,1\}^n$.

Таблица 7.12

s	$u_0 = 0$	u_1		u_{2k-1}
0	v_0	v_1	…	$v_{2^k\_1}$
s_1	e_1	$e_1 + v_1$	…	$e_1 + v_{2^k\_1}$
s_2	e_2	$e_2 + v_1$	…	$e_2 + v_{2^k\_1}$
⋮	⋮	⋮	⋱	⋮
$s_{2^{n-k}\_1}$	$e_{2^{n-k}\_1}$	$e_{2^{n-k}\_1} + u_{2^k\_1}$	…	$e_{2^{n-k}\_1} + v_{2^k\_1}$

Для описания процедуры декодирования необходимо ввести понятие синдрома. ***Синдромом произвольного кодового слова*** V_2 *называется произведение искаженного слова* $r = v + e$ *и транспонированной проверочной матрицы H*:

$$s = rH^T, \tag{7.34}$$

где H - проверочная матрица кода C.

Покажем, что синдром является индикатором вектора ошибок. Предположим, что кодовое слово $v \in C$, передано по ДСК, принято как $r = v + e$. Синдром принятого слова

$$s = rH^T = (v + e)H^T = eH^T. \tag{7.35}$$

Таким образом, вычисление синдрома можно рассматривать как линейное преобразование вектора ошибок.

Пример 7.14. Имеем стандартную таблицу двоичного линейного (4, 2, 2) кода (табл. 7.13).

Таблица 7.13

s	00	01	10	11
00	0000	0110	1011	1101
11	1000	1110	0011	0101
10	0100	0010	1111	1001
01	0001	0111	1010	1100

Декодирование с помощью стандартной таблицы выполняется таким образом. Пусть $r = v + e$ - принятое слово. Найдем это слово в таблице и возьмем как результат декодирования сообщения *u*, записанное в верхний (первой) ячейке того столбца, в котором лежит принятое слово *r*. В сущности, этот процесс предусматривает хранение в памяти всей таблицы и поиска в ней заданного слова. Тем не менее, можно упростить процедуру декодирования, если заметить, что все элементы одной и той самой строки имеют один и тот самый синдром. Каждая строка $\mathrm{Row}_i, 0 \le i < 2^{n-k}$, этой таблицы представляет собой *смежный класс* кода *C*, а именно $\mathrm{Row}_i = \{e_i + v | v \in C\}$. Вектор e_i называется *лидером смежного класс*а. Синдром всех элементов *i*-й строки

$$s_i = (e_i + v)H^T = e_i H^T \qquad (7.36)$$

не зависит от конкретного значения кодового слова $v \in C$. Упрощенная процедура декодирования состоит в выполнении таких действий. Вычислить синдром принятого слова $r = e_j + v$:

$$s_j = (e_j + v)H^T = e_j H^T,$$

найти его в левом столбце стандартной таблицы; взять лидера смежного класса e'_j из второго столбца той же строки и прибавить его к принятому слову, взяв наиболее ближнее к принятому $r = e'_j + v'$ кодовое слово v'. Итак, вместо таблицы $n \times 2^n$ бит для декодирования достаточно использовать таблицу лидеров смежного классов $n \times 2^{n-k}$ бит.

Алгоритмы помехоустойчивого блочного кодирования. Коды Хэмминга представляют, наверно, наиболее известный класс блочных кодов, за исключением, возможно, только кодов Рида - Соломона. Как уже отмечалось, коды Хэмминга являются оптимальными в том смысле, что они требуют минимальной чрезмерности при заданной длине блока для исправления одной ошибки. Двоичные коды Голея - это единственный нетривиальный пример оптимального кода, который исправляет тройные ошибки (другими примерами оптимальных кодов являются коды-повторения и коды с одной проверкой

на парность). Коды Рида - Маллера - очень элегантная комбинаторная конструкция с простым декодированием.

Коды Хэмминга. Напомним, что любое кодовое слово v линейного $(n, k, d_{\min})$ кода C удовлетворяет уравнению

$$vH^T = 0. \tag{7.37}$$

Полезная интерпретация этого уравнения заключается в том, что максимальное количество линейно независимых столбцов проверочной матрицы H кода C равняется $d_{\min} - 1$.

В двоичном случае для $d_{\min} = 3$ из формулы (7.37) вытекает, что сумма любых двух столбцов проверочной матрицы не равна нулевому вектору. Пусть столбцы H являются двоичными векторами длины m. Существует всего $2^m - 1$ ненулевых разных столбцов. Итак, длина двоичного кода, который исправляет одиночную ошибку, удовлетворяет условию

$$n \le 2^m - 1.$$

Эта неравенство точно совпадает с границей Хэмминга для кода длины n с $n - k = m$ проверками и исправлением $t = 1$ ошибок. Соответственно код, который удовлетворяет приведенному условию со знаком равенства, известный как *код Хэмминга*.

Пример 7.15. Для $m = 3$ получаем (7, 4, 3) код Хэмминга с проверочной матрицей

$$H = \begin{pmatrix} 1110100 \\ 0111010 \\ 1101001 \end{pmatrix}.$$

Как уже подчеркивалось, проверочная матрица кода Хэмминга имеет свойство, что все ее столбцы разные. Если возникает одиночная ошибка на j-й позиции, $1 \le j \le n$, то синдром искаженного принятого слова равняется j-му столбцу матрицы H. Обозначим e вектор ошибок, добавленный к кодовому слову в процессе его передачи по ДСК и допустим, что все его компоненты равняются нолю за исключением j-й позиции, $e_j = 1$. Тогда синдром принятого слова

$$s = rH^T = eH^T = h_j, \tag{7.38}$$

где h_j - j-й столбец матрицы H.

Процедуры кодирования и декодирования. Из уравнения (7.38) вытекает, что когда столбцы проверочной матрицы рассматривать как двоичное представление целых чисел, то значение синдрома равно номеру искаженной (ошибочной) позиции. Эта идея положена в основу алгоритмов кодирования и декодирования, приведенных дальше.

Запишем столбцы проверочной матрицы в виде двоичного представления номера (от 1 до n) позиции кодового слова в возрастающем порядке. Обозначим эту матрицу через H. Очевидно, что матрице H^* отвечает эквивалентный код Хэмминга с точностью до перестановки позиций кодового слова.

Напомним, что проверочная матрица в систематической форме содержит единичную подматрицу I_{n-k} размера $(n-k)\times(n-k)$ Очевидно, что в матрице H^* столбцы единичной подматрицы I_{n-k} (т.е. столбцы веса один) размещаются на позициях с номерами, которые равняются степени 2, т.е.: $2^e, l = 0,1,...,m.$

Пример 7.16. Пусть $m = 3.$ Тогда систематическую (каноническую) проверочную матрицу можно задать в виде

$$H = \begin{pmatrix} 1101100 \\ 1011010 \\ 0111001 \end{pmatrix}.$$

Матрица H^*, заданная двоичным представлением целых чисел от 1 до 7 (младший разряд записывается в верхней строке), имеет вид

$$H^* = \begin{pmatrix} 1010101 \\ 0110011 \\ 0001111 \end{pmatrix},$$

где матрица I_3 содержится в первом, втором и четвертом столбцах.

Вообще для $(2^m - 1, 2^m - 1 - m)$ кода Хэмминга и данного (арифметического) порядка столбцов единичная матрица I_m содержится в столбцах проверочной матрицы с номерами $1, 2, 4,..., 2^{m-1}$.

Кодирование. При вычислении проверочных символов p_j для всех $1 \le j \le m$ проверяют номера столбцов и те столбцы, номера которых не является степенью 2, ставятся в соответствие информационным позициям слова. Соответствующие информационные символы включаются в процесс вычисления проверок. Такая процедура кодирования в чем-то сложнее обычной процедуры для систематического (канонического) кода Хэмминга. Тем не менее, соответствующая ей процедура декодирования очень простая. Для некоторых применений этот подход может быть более привлекательным, поскольку обычное декодирование должно выполняться довольно быстро.

Декодирование. Если кодирование выполнялось соответственно матрице H^*, то декодирование оказывается очень простым. Синдром (7.38) равен номеру позиции, в которой случилась ошибка. После вычисления синдрома s, что рассматривается как целое число, ошибка исправляется по правилу

$$v_s = v_s + 1, \qquad (7.39)$$

где выполняется сложение по модулю 2 (0 + 0 = 0, 1 + 1 = 0, 0 + 1 = 1, 1 + 0 = 1).

Двоичный код Голея. Голей установил, что $\sum_{i=0}^{3}\binom{23}{i} = 2^{11}$.

Это равенство дает основание предположить, что может существовать совершенный двоичный (23, 12, 7) код при $t = 3$, т.е. код способен исправлять до трех ошибок в словах длиной 23 символа. В своей статье Голей привел порождающую матрицу такого двоичного кода, который исправляет до трех ошибок.

Учитывая сравнительно небольшую длину (23) и размерность (12), а также небольшое количество проверок (11) кодирование и декодирование двоичного (23, 12, 7) кода Голея можно выполнить табличным методом.

Кодирование. Табличное (LUT, look-up-table) кодирование реализуется с помощью просмотров таблицы, которая содержит список всех $2^{12} = 4096$ кодовых слов, пронумерованных непосредственно информационными символами. Пусть u - информационный вектор размерности 12 бит и v - соответствующее кодовое слово (23 бит). Табличный кодер использует таблицу, в которой для каждого информационного вектора (12 бит) вычислен и записан синдром (11 бит). Синдром берется из таблицы и приписывается по правую сторону к информационному вектору.

Операция LUT - это взаимно однозначное отображение из множества векторов u на множество векторов v, которое можно записать в виде

$$v = \mathrm{LUT}(u) = (u, \mathrm{get\_syndrome}(u, 0)). \qquad (7.40)$$

В реализации табличного кодера учтены упрощения, которые вытекают из циклической природы кода Голея. Его *породный полином* имеет вид

$$g(x) = x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1, \qquad (7.41)$$

или в шестнадцатиричной системе исчисления C75. Этот полином используется в процедуре «get_syndrome», заданной уравнением (7.41).

Декодирование. Напомним, что задача декодера заключается в оценивании наиболее возможного (такого, который имеет минимальный хэммингов вес) вектора ошибок e по принятому вектору r.

Процедура построения табличного LUT-декодера состоит из таких действий:

1) выписать все возможные векторы ошибок e, хэммингов вес которых не превышает три;

2) для каждого вектора ошибок вычислить соответствующий синдром $s = \mathrm{get\_syndrome}(e)$;

3) записать в таблицу для каждого значения s соответствующий ему вектор e, при котором $\mathrm{LUT}(s) = e$.

Исправление до трех ошибок в принятом искаженном слове r с помощью LUT-декодера можно записать как

$$v'' = r \oplus \mathrm{LUT}(\mathrm{get\_syndrome}(r)),$$

где v'' - исправленное кодовое слово.

Полиномное помехоустойчивое кодирование. Двоичные коды Рида - Маллера (РМ) образовывают семью кодов, которые исправляют ошибки, с простым декодированием, которое обосновывается на *мажоритарной логике*.

Известное определение двоичных кодов РМ обосновывается на двоичных полиномах (или булева функция). Согласно этому определению коды РМ становятся близкими к кодам, которые входят в класс *полиномных кодов.*

Булевые полиномные коды РМ. Обозначим $f(x_1, x_2, ..., x_m)$ булевую функцию от m двоичных сменных $x_1, x_2, ..., x_m$. Известно, что такие функции легко дать с помощью *таблицы истинности*. Таблица истинности содержит список значений функции f для всех 2^m комбинаций значений ее аргументов. Все обычные булевые операции (такие как «и», «или») можно дать как булевые функции.

Пример 7.17. Рассмотрим функцию, заданную такой таблицей истинности:

x_2	0	0	1	1
x_1	0	1	0	1
$f(x_1 x_2)$	0	1	1	0

Тогда имеем: $f(x_1, x_2) = (x_1 \,\&\, NOT(x_2)) \bigcup (NOT(x_1) \,\&\, x_2)$.

Ассоциируем с каждой булевой функцией f двоичный вектор f длины 2^m, составленный из значений этой функции для всех возможных комбинаций значений m ее аргументов. В примере 7.17 сделано предположение о лексикографическом (арифметическом) упорядочении значений аргументов функции, т.е. x_1 представляет собой младший разряд, а x_m - старший разряд.

Заметим, что булевую функцию можно записать прямо по таблице истинности, воспользовавшись дизъюнктивной нормальной формой (ДНФ). В терминах ДНФ любую булевую функцию можно записать как сумму 2^m элементарных функций $1, x_1, x_2, ..., x_m, x_1x_2, ..., x_mx_1, x_2, ..., x_m$:

$$f = 1 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_m x_m + \alpha_{12} x_1 x_2 + ... + \alpha_{12...m} x_1 x_2 ... x_m, \qquad (7.42)$$

где вектор 1 включен для того, чтобы вычислить составляющую (нулевой степени).

В примере 7.17. $f = x_1 + x_2$. Двоичный $(2^m, k, 2^{m-2})$ код РМ, обозначенный $PM_{r,m}$, определяется как множество векторов, которые ассоциируются со всеми булевыми функциями степени к r включительно от m сменных. Код

$PM_{r,m}$ называют также кодом РМ r-го порядка длины 2^m. Размерность $PM_{r,m}$ кода

$$k = \sum_{i=0}^{r} \binom{m}{i}. \tag{7.43}$$

Это число равно количеству способов, которыми можно построить полиномы степени, не выше r, от m сменных.

С учетом уравнения (7.43) строками порождающей матрицы $PM_{r,m}$ кода являются векторы, ассоциированные с k булевыми функциями, которые можно записать как полиномы степени, не высшего за r, от m сменных.

Пример 7.18. Код РМ $PM_{1,3}$ первого порядка длины 8 является двоичным (8, 4, 4) кодом, который можно построить из булевых функций первой степени от трех сменных: $\{1, x_1, x_2, x_3\}$. Таким образом, имеем

$$\begin{array}{rcccccccc} 1 = & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 = & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ x_2 = & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ x_3 = & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}$$

Порождающая матрица $PM_{1,3}$ кода

$$G = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}. \tag{7.44}$$

Заметим, что код $PM_{1,3}$ можно построить также и из кода Хэмминга (7, 4, 3) сложением общей проверки на парность. Расширенный код Хэмминга и $PM_{1,3}$ код могут отличаться только порядком позиций (столбцов).

Дуальные коды кодов РМ. Можно показать, что $PM_{m-r-1,m}$ дуальный код $PM_{r,m}$. Другими словами, порождающая матрица $PM_{m-r-1,m}$ кода, может использоваться как проверочная матрица $PM_{r,m}$ кода.

Определение кодов РМ можно дать и в терминах *конечной геометрии*. *Геометрия Евклида* $EG(m, 2)$ размерности m над $GF(2)$ содержит 2^m точек, которые представляют собой все двоичные векторы длины m. Заметим, что столбцы матрицы, образованной последними тремя строками порождающей матрицы $PM_{1,3}$ кода, представляют собой 8 точек $EG(3,2)$. Изъятием нулевой точки это множество точек превращается в *проективную геометрию*

$PG(m-1,2)$. Коды конечной геометрии являются, в сущности, обобщением кодов PM.

Связь между кодами и конечной геометрией можно объяснить. Возьмем $EG(m,2)$. Столбцы матрицы $(x_1^T x_2^T ... x_m^T)$ (где T - операция транспонирования матрицы) рассматриваются как координаты точек геометрии $EG(m,2)$. Тогда существует взаимно однозначное соответствие между компонентами двоичного вектора длины 2^m и точками $EG(m,2)$. В частности, подмножество $EG(m,2)$ можно ассоциировать с двоичным вектором $w=(w_1,w_2,...,w_n)$ длины $n=2^m$, если интерпретировать значение его координат $w_i=1$ как выбор точки. Другими словами, w является *вектором инцидентности* (совпадений).

Кодовыми словами $PM_{r,m}$ ***кода PM*** *являются векторы инцидентности всех подпространств* (*линейных комбинаций точек*) *размерности* $m-r$ *в* $EG(m,2)$. *Из этого определения следует, что количества кодовых слов минимального веса* $PM_{r,m}$ *кода*

$$A_{2^{m-r}} = 2^r \prod_{i=0}^{m-r-1} \frac{2^{m-i}-1}{2^{m-r-i}-1}. \qquad (7.45)$$

Код, который образовывается после изъятия координат, которые отвечают условию $x_1 = x_2 = ... = x_m = 0$, из всех кодовых слов $PM_{r,m}$ кода и являются двоичным циклическим $PM^*_{r,m}$ кодом. Количество слов минимального веса циклического кода *PM*

$$A^*_{2^{m-r-1}} = \prod_{i=0}^{m-r-1} \frac{2^{m-i}-1}{2^{m-r-i}-1}. \qquad (7.46)$$

Декодирование РМ кодов можно выполнить на основе мажоритарной логики (МЛ). Идея мажоритарного декодирования состоит вот в чем. Как известно, проверочная матрица порождает 2^{n-k} проверочных уравнений. Построение МЛ декодера сводится к выбору такого подмножества проверочных уравнений, чтобы решение о значении кодового символа на определенной позиции формировалось по большинству «голосов», причем каждый «голос» связан с одним из проверочных уравнений.

Коды Рида - Соломона. При построении кодов Рида - Соломона (РС-коды), исправляющих пакетные ошибки с разрядностью *b*, пакет разрядов слова рассматривается как *b*-значный разряд, приобретающий одно из $s=2^b$ значений (от 0 до $s-1$). В отличие от двоичных циклических кодов в этом случае в *H*-матрице символами являются не 1 и 0, а подматрицы 0, *I*, h^β, где h^β определяется выражением

$$h^\beta = \left\| f^{\beta+b-1} \; f^{\beta+b-2} \; ... \; f^{\beta+b-b} \right\|, \qquad (7.47)$$

в котором $f^{\beta-b-i}$ - столбец, который отвечает остатку от деления $x^{\beta+b-1}$ на многочлен $G(x)$ степени b; u - номер столбца в подматрице h^{β}; b - показатель степени матрицы, причем $1 \le \beta \le 2^b - 1$.

Как образующий полином используется первоначальный полином степени b, обеспечивающий максимальное количество разных матриц h^{β} и равный $2^b - 1$. Значение полиномов для типичных пакетов искажения приведены в табл. 7.14.

Таблица 7.14

Разрядность пакета искажений	Образующий многочлен матриц h^{β}
2	$x^2 + x + 1$
4	$x^4 + x + 1$
8	$x^8 + x^7 + x^6 + x + 1$

Полученные матрицы вместе с нулевой (все ее элементы равняются нулю) образовывают поле матриц и так же, как и столбцы *H*-матриц двоичных кодов, могут складываться и делиться. При сложении указанных матриц результат суммирования соответствующих символов вычисляется по модулю два. При сложении показатели степени матриц подытоживаются, а при обратной процедуре - отнимаются по модулю $2^b - 1$.

Например, *H*-матрица РС-кода, который исправляет одиночные и находит двойные пакеты искажений, имеет такой вид:

$$\begin{Vmatrix} I & I & \dots & I & \dots & I & I & & 0 & 0 \\ I & h^1 & \dots & h^j & \dots & h^{2^b-2} & 0 & & I & 0 \\ I & h^2 & \dots & h^{2j} & \dots & h^{2\left(2^b-2\right)} & 0 & & 0 & I \end{Vmatrix}. \qquad (7.48)$$

В этой матрице вторая строка (кроме трех последних подматриц) содержит подматрицы $b \times b$ вида (7.48) всех степеней от трех до $(2^b - 2)$.

Символы третьей строки (кроме трех последних) равняются квадратам соответствующих символов второй строки.

Длина РС-кодов - это количество двоичных информационных и контрольных символов, которая определяет количество символов в каждой строке *H*-матрицы и определяется выражением

$$n = b(2^b + 2).$$

При этом количество информационных разрядов $K = b(2^b - 1)$, а количество контрольных разрядов $k = 3b$ в двоичных символах или $N = (2b + 2)$, или $K = 2^b - 1$, или $K = 3$ в обобщенных *b-разрядных* символах. Причем, как и раньше, величина $n(N)$ определяет количество столбцов, а $k(K)$ — количество строк соответствующей *H*-матрицы.

Например, при $b=4$ имеем $n=72,\ k=12,\ K=60$. При $b=2$ имеем $N=12,\ k=6$. В последнем случае *H*-матрица имеет вид (в обобщенных символах)

$$\left\|\begin{matrix} I & I & I & I & 0 & 0 \\ I & h^1 & h^2 & 0 & I & 0 \\ I & h^2 & h^4 & 0 & 0 & I \end{matrix}\right\| = \left\|\begin{matrix} I & I & I & I & 0 & 0 \\ I & h^1 & h^2 & 0 & I & 0 \\ I & h^2 & h^1 & 0 & 0 & I \end{matrix}\right\|, \qquad (7.49)$$

поскольку первая (нижняя) строка представляет собой квадрат второго, а максимальная степень подматриц вида (7.48) не должна превышать 2^b-1, откуда $4 \bmod (2^b-1) = 4 \bmod 3 = 1$.

Полученная *H*-матрица размером в 12 столбцов и 6 строк (в двоичных символах) или в 6 столбцов и 3 строки в подматрице содержит

$$0=\begin{vmatrix}0 & 0\\ 0 & 0\end{vmatrix},\ I=\begin{vmatrix}1 & 0\\ 0 & 1\end{vmatrix},\ h^1=\begin{vmatrix}1 & 1\\ 0 & 1\end{vmatrix},\ h^1=\begin{vmatrix}0 & 1\\ 1 & 1\end{vmatrix},\ h^4=h^1=\begin{vmatrix}1 & 1\\ 0 & 1\end{vmatrix}.$$

Для построения матрицы h^i использовано соотношение (7.47), т.е.

$$h^i = \left| f^{1+2+1} \quad f^{1+2-1} \right| = \left| f^2 \quad f^1 \right|,$$

где f_1 - остаток от деления x_i на многочлен x_2, что порождает $(x+1)$:

$$\begin{array}{l} x^2 \\ \underline{x^2+x+1} \\ x+1 \end{array} \left| \begin{array}{l} x^2+x+1 \\ \hline 1 \end{array} \right.,$$

т.е. остаток от деления x^2 равен $(x+1)$, а остаток от деления x^1 равен x, откуда

$$f^2=\begin{vmatrix}1\\ 1\end{vmatrix},\ f^1=\begin{vmatrix}1\\ 0\end{vmatrix},\ \text{а}\ h^1=\begin{vmatrix}1 & 1\\ 1 & 0\end{vmatrix}. \qquad (7.50)$$

Аналогично:

$$h^2 = \left| f^{2+2-1} \quad f^{2+2-1} \right| = \left| f^3 \quad f^2 \right|.$$

Столбец f^2 вычислен раньше, а столбец f^3, как нетрудно показать, равен $\left\|\begin{matrix}0\\ 1\end{matrix}\right\|$, поскольку при делении x^3 на x^2+x+1 получим остаток, равный 1, или полином вида $(0 \cdot x+1)$.

При кодировании начальной числовой последовательности контрольные разряды, как и для двоичных кодов, вычисляются суммированием по модулю 2 информационных разрядов, которые соответствуют единице в соответ-

ствующей строке двоичной *H*-матрицы или (для обобщенных символов) определяются выражением

$$\alpha_{k+j}^T = C_j = \sum_{i=1}^{k} h^{ij}\alpha_i^T \,(\mathrm{mod}\,2), \quad j = 1,...,K,$$

где j - номер обобщенного контрольного символа; α_i - значение обобщенного информационного символа, записанное в виде матрицы-столбца; h_{ij} - символ подматрицы проверочной *H*-матрицы, которая содержится в i-м столбце и j-й (если считать сверху вниз) строке.

Пример 7.19. Вычислим контрольный признак для шестиразрядного $(n = 6)$ начального числа, считая, что код должен находить и исправлять двухразрядные $(b = 2)$ пакеты искажений, а начальное число $A = \alpha_1\alpha_2\alpha_3 = 11\,01\,10.$

Параметры такого кода, как доказано раньше, $N = 6,\ M = 3,\ k = 3,$ а проверочная матрица имеет вид формулы (7.48). Тогда

$$\alpha_{k+j}^T = C_j = \sum_{i=1}^{3} h^{ij}\alpha_i^T \,(\mathrm{mod}\,2),\ j = 1,\ 2,\ 3,$$

а именно:

$$\alpha_4^T = C_1 = h^{11}\cdot\alpha_1^T + h^{21}\cdot\alpha_2^T + h^{31}\cdot\alpha_3^T\,(\mathrm{mod}\,2) = I\cdot\alpha_1^T + I\cdot\alpha_2^T +$$

$$+ I\cdot\alpha_3^T\,(\mathrm{mod}\,2) = \alpha_1^T + \alpha_2^T + \alpha_3^T\,(\mathrm{mod}\,2) = \begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}0\\1\end{vmatrix} + \begin{vmatrix}1\\0\end{vmatrix}(\mathrm{mod}\,2) = \begin{vmatrix}0\\0\end{vmatrix};$$

$$\alpha_5^T = C_2 = h^{12}\cdot\alpha_1^T + h^{22}\cdot\alpha_2^T + h^{32}\cdot\alpha_3^T\,(\mathrm{mod}\,2) = I\cdot\alpha_1^T + h^1\cdot\alpha_2^T +$$

$$+ h^2\cdot\alpha_3^T\,(\mathrm{mod}\,2) = \begin{vmatrix}1 & 0\\0 & 1\end{vmatrix}\cdot\begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}1 & 1\\1 & 0\end{vmatrix}\cdot\begin{vmatrix}0\\1\end{vmatrix} + \begin{vmatrix}0 & 1\\1 & 1\end{vmatrix}\cdot\begin{vmatrix}1\\0\end{vmatrix}(\mathrm{mod}\,2) =$$

$$= \begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}1\\0\end{vmatrix} + \begin{vmatrix}0\\1\end{vmatrix}(\mathrm{mod}\,2) = \begin{vmatrix}0\\0\end{vmatrix};$$

$$\alpha_6^T = C_3 = h^{123}\cdot\alpha_1^T + h^{23}\cdot\alpha_2^T + h^{33}\cdot\alpha_3^T\,(\mathrm{mod}\,2) = I\cdot\alpha_1^T + h^2\cdot\alpha_2^T +$$

$$+ h^1\cdot\alpha_3^T\,(\mathrm{mod}\,2) = \begin{vmatrix}1 & 0\\0 & 1\end{vmatrix}\cdot\begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}0 & 1\\1 & 1\end{vmatrix}\cdot\begin{vmatrix}0\\1\end{vmatrix} + \begin{vmatrix}1 & 1\\1 & 0\end{vmatrix}\cdot\begin{vmatrix}1\\0\end{vmatrix}(\mathrm{mod}\,2) =$$

$$= \begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}1\\1\end{vmatrix} + \begin{vmatrix}1\\1\end{vmatrix}(\mathrm{mod}\,2) = \begin{vmatrix}1\\1\end{vmatrix}.$$

Итак, передаче подлежит код

$$A = \alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6 = 11\,01\,10\,00\,00\,11.$$

Декодирование РС-кода выполняется в два этапа. На первом этапе рассчитывается так называемый синдром ошибки, т.е. вспомогательная величи-

на, которая дает возможность устанавливать в дальнейшем факт наличия ошибки, ее место и значение:

$$S_R^T = H \cdot AT.$$

Пример 7.20. Рассчитать синдром ошибки для закодированного прежде слова, считая, что искажение возникло во втором обобщенном символе и вместо символа $\alpha_2 = 01$ получен символ = 10. Тогда

$$S_R^T = \begin{vmatrix} S_3 \\ S_2 \\ S_1 \end{vmatrix} = \begin{vmatrix} I & I & I & I & 0 & 0 \\ I & h^1 & h^2 & 0 & I & 0 \\ I & h^2 & h^1 & 0 & 0 & I \end{vmatrix} \cdot \begin{vmatrix} \alpha_1^T \\ \alpha_2^T \\ \vdots \\ \alpha_6^T \end{vmatrix}.$$

Выполнив все необходимые операции, получим: $S_3 = \begin{vmatrix} 1 \\ 1 \end{vmatrix}$, $S_2 = \begin{vmatrix} 0 \\ 1 \end{vmatrix}$, $S_1 = \begin{vmatrix} 1 \\ 0 \end{vmatrix}$.

На втором этапе декодирования РС-кода вычисляется индикатор ошибки - система равенств, которая дает возможность определить место возникновения ошибки.

Если отказ (искажение) возник в разрядах *j*-го обобщенного символа, то искажение можно подать в виде 0...0*E*0...0, где *E* - обобщенный *b*-значный символ, который равняется нулю при отсутствии ошибки и приобретает значения от 0 до $(s-1)$ при ее наличии. Поскольку третья (верхняя) строка *H*-матрицы (7.48) содержит единичные подматрицы, то пакет разрядов синдрома будет S_3 содержать единицы в разрядах, которые отвечают искаженным разрядам пакета ошибки.

В самом деле:

$$S_3 = \sum_{i=1}^{k+1} \alpha_i^T = \sum_{i=1}^{j-1} \alpha_i^T + \sum_{i=1}^{k+1} \alpha_i^T + (\alpha_i^T + E^T) \pmod 2 = \sum_{i=1}^{k} \alpha_i^T + \alpha_{k+1}^T + ET \pmod 2.$$

Поскольку $\sum_{i=1}^{k} \alpha_i^T = \alpha_{k+1}^T$, это

$$\sum_{i=1}^{k} \alpha_i^T + \alpha_{k+1}^T \pmod 2 = 0 \text{ и } S_3 = ET. \qquad (7.51)$$

Итак, анализируя выражение (7.51), приходим к выводу о наличии (при $S_3 \neq 0$) и значении ошибки. Известно, что место искажения (номер искаженного символа) можно найти из системы уравнений

$$S_1 = h^{i1} S_3 = 0, S_2 = h^{i2} S_3 = 0, \qquad (7.52)$$

которое называется *индикатором искажений.* Эта система совместная только в том случае, когда в *i*-м обобщенном символе есть искажения.

К преимуществам РС-кода следует отнести сопоставимую простоту реализации матричных и модульных операций (поскольку модуль равняется 2)

при вычислении как синдрома (7.51), так и индикатора (7.52) искажений. Но учитывая большое количество этих операций суммарные аппаратурные затраты могут быть значительными.

Двоичные циклические коды. Циклические коды образовывают класс кодов, которые исправляют ошибки кодирования и декодирования, алгоритм которых основывается на полиномном представлении. Простая реализация этих кодов использует регистры сдвига и логические схемы.

Порождающие и проверочные полиномы. Обозначим C как линейный блочный (n, k) код. Пусть u - сообщение и v - соответствующее ему кодовое слово кода C. Циклические коды имеют такие свойства, благодаря которым они становятся удобными для аппаратурной реализации. Поставим в соответствие каждому кодовому слову v полином $v(x)$:

$$v = (v_0, v_1, ..., v_{n-1}) \to v(x) = v_0 + v_1 x + ... v_{n-1} x^{n-1}.$$

Переменная x является индикатором относительного положения элемента v_i в кодовом слове в виде произведения (монома) $v_i x^i$ полинома $v(x)$.

Линейный блочный код C является циклическим *тогда и только тогда, когда любой циклический сдвиг любого кодового слова представляет собой другое* (*или то же*) *кодовое слово*, *т.е.*

$$v = (v_0, v_1, ..., v_{n-1}) \to v(x) = v_0 + v_1 x + ... v_{n-1} x^{n-1}.$$

В полиномном представлении циклический сдвиг на одну позицию, обозначенный $v^{(1)}(x)$, соответствует умножению на x по модулю

$$v(x) \in C \Leftrightarrow v'(x) = xv(x) \bmod(x^n - 1) \in C.$$

Операция циклического сдвига реализуется на *регистре сдвига* (рис. 7.28).

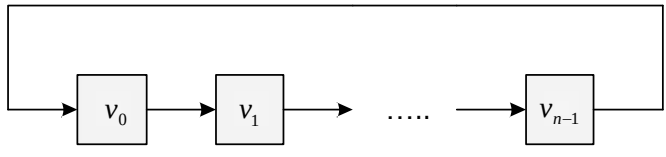


Рис. 7.28. Циклический регистр сдвига

Пример 7.21. Рассмотрим случай $n = 7$. Циклический сдвиг на одну позицию вектора равен $v^{(1)} = (1010101)$. В полиномном представлении и двоичной арифметике получаем

$$v(x)x + x^3 + x^5 + x^6,$$

$$v^{(1)}(x) = xv(x) = x^2 + x^4 + x^6 + x^7 \bmod(x^7 + 1) =$$

$$= x^2 + x^4 + x^6 + x^7 + (x^7 + 1) = 1 + x^2 + x^4 + x^6.$$

Порождающий многочлен. Важным свойством циклических кодов является то, что все кодовые слова-полиномы кратные одному фиксированному полиному, который называется *порождающим полиномом* кода. Этот полином (многочлен), как и любой другой, задается своими корнями, которые по обыкновению называют нулями кода. Легко показать, что порождающий полином $g(x)$ является делителем бинома $(x^n - 1)$. (По аналогии с целыми числами «$a(x)$ делит $b(x)$» (иначе $a(x)|b(x)$), если $b(x) = a(x)q(x)$). Итак, чтобы найти некоторый порождающий многочлен, нужно знать разложение бинома $(x^n - 1)$ на множители $\varphi_j(x), j = 1,2,..., l$.

$$(x^n - 1) = \varphi_1(x)\varphi_2(x)....\varphi_l(x). \tag{7.53}$$

Заметим, что в двоичной арифметике операции $a + b$ и $a - b$ (по модулю 2) дают одинаковый результат. Поскольку рассматриваем только двоичные коды или коды над конечными полями характеристики два, т.е. такие, которые используют двоичную арифметику, то в дальнейшем не будем различать соответствующие операции (знаки «+» и «-»).

В результате получаем

$$g(x) = \prod_{j \in J \subset \{1,2,...l\}} \varphi_j(x). \tag{7.54}$$

Пример 7.22. На множестве двоичных многочленов, т.е. полиномов с коэффициентами из множества $Z_2 = \{0,1\}$, бином $x^7 - 1$ имеет такое расписание:

$$x^7 + 1 = (x+1)(x^3 + x + 1)(x^3 + x^2 + 1).$$

Приведем примеры циклических кодов длины 7.

Двоичный циклический код Хэмминга с порождающим полиномом $g(x) = x^3 + x + 1$.

Двоичный циклический код с проверкой на парность порождается полиномом $g(x) = (x+1)$.

Дуальный код Хэмминга (код максимальной длины) имеет порождающий многочлен $g(x) = (x+1)(x^3 + x + 1)$.

Кодирование и декодирование двоичных циклических кодов. Размерность двоичного циклического (n, k) кода

$$k = n - \deg[g(x)],$$

где deg[.] - степень аргумента. Поскольку циклический код является линейным кодом, то любое множество k линейно независимых векторов (кодовых слов) можно взять как порождающую матрицу кода. В частности, двоичные векторы, которые ассоциируются с многочленами $g(x), xg(x), ..., x^{k-1}g(x)$, линейно независимы. Эти векторы можно использовать как строки порождающей матрицы кода C. В этом случае реализуется

несистематическое кодирование. Другими словами, сообщение не появляется в неизмененном виде на каких-нибудь позициях кодового слова.

Пример 7.23. Рассмотрим циклический код Хэмминга с порождающим полиномом $g(x) = x^3 + x + 1 \Leftrightarrow (1101)$. Порождающая матрица этого кода имеет вид

$$G = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

В другом варианте проверочную часть порождающей матрицы циклического кода можно построить с помощью таких полиномов:

$$x^{n-1} \bmod g(x),$$
$$\vdots$$
$$x^{n-k-1} \bmod g(x)$$
$$x^{n-k} \bmod g(x).$$

С их помощью реализуется *систематическое* кодирование, рассмотренное в приведенном дальше примере.

Пример 7.24. Пусть C - циклический код Хэмминга с порождающим многочленом $g(x) = x^3 + x + 1$. Тогда имеем:

$$x^6 \bmod(x^3 + x + 1) = x^2 + 1,$$
$$x^5 \bmod(x^3 + x + 1) = x^2 + x + 1,$$
$$x^4 \bmod(x^3 + x + 1) = x^2 + x,$$
$$x^3 \bmod(x^3 + x + 1) = x + 1.$$

Итак, систематическая порождающая матрица кода C имеет вид

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Кодирование циклического кода может быть систематическим или несистематическим в зависимости от того, что именно происходит с сообщением.

Несистематическое кодирование

$$v(x) = u(x)g(x). \tag{7.55}$$

Систематическое кодирование

$$v(x) = x^{n-k}u(x) + [x^{n-k}u(x) \bmod g(x)]. \tag{7.56}$$

Проверочный полином. Полином, который можно ассоциировать с проверочной матрицей циклического кода, называется *проверочным полиномом*. Порождающие и проверочные полиномы связаны соотношением

$$g(x)h(x) = x^n + 1. \tag{7.57}$$

Если известен порождающий полином, то проверочный полином легко вычисляется как $h(x) = (x^n + 1) / g(x) = h_0 + h_1 x + ... + h_k x^k$. Проверочную матрицу кода C легко построить, воспользовавшись как строками $n - k - 1$ циклическими сдвигами проверочного полинома:

$$h^j(x) = x^j h(x) \bmod (x^n - 1),\ j = 0, 1, ,..., n - k - 1,$$

$$H = \begin{pmatrix} h_0 & h_1 & h_2 & \cdots & h_k & 0 & 0 & \cdots & 0 \\ 0 & h_1 & h_1 & h_2 & \cdots & h_k & 0 & \cdots & 0 \\ 0 & 0 & h_0 & h_1 & \cdots & \cdots & h_k & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & h_k \end{pmatrix}. \tag{7.58}$$

Пример 7.25. Циклический код Хэмминга с порождающим многочленом $g(x) = x^3 + x + 1$ имеет проверочный многочлен $h(x) = (x^7 + 1) / (x^3 + x + 1) = = x^4 + x^2 + x + 1$. Проверочная матрица этого кода имеет, например, такой вид:

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Равно как и для линейных кодов, систематическое кодирование циклического кода можно реализовать как решение уравнения

$$v(x)h(x) = 0 \bmod (x^n - 1).$$

Рассмотрим правило систематического кодирования. Предположим, что код имеет скорость $k / n \le 0,5$. Пусть сообщение представлено многочленом, степень которого меньше k. Пусть $v(x)$ - кодовое слово кода C, который отвечает информационному многочлену $u(x)$. На первом шаге $v_l = u_l,\ l = 0, 1,..., k - 1$.

Из циклической природы этого кода вытекает, что проверочные символы кода можно вычислить рекурсивно с помощью проверочного уравнения, где $h_{(l-k),\,j}$ - j-й элемент $(l - k)$-ãî строки матрицы (7.48):

$$v_l = \sum_{j=0}^{l-1} v_j h_{(l-k),\,j},\ l = k,\ k + 1,..., n - 1. \tag{7.59}$$

В случае высокой скорости кода $k/n > 0{,}5$ кодирование с помощью деления $x^{n-k}u(x)$ на порождающий полином эффективнее. При этом всегда кодовое слово получают в систематической форме, когда k первых его символов совпадают с символами сообщения, а последние $n-k$ являются проверочными символами.

Структурная схема кодера двоичного кода с порождающим полиномом $g(x)$ изображена на рис. 7.29. Первые k тактов переключатель (правая нижняя часть схемы) находится в положении *1*, а информационные символы передаются в канал связи и одновременно вводятся в схему умножения на x^{n-k} и деления на порождающий многочлен $g(x)$. За эти k тактов в регистре сдвига вычисляется остаток от деления, после чего переключатель переводится в положение *2* и содержимое регистра передается в канал.

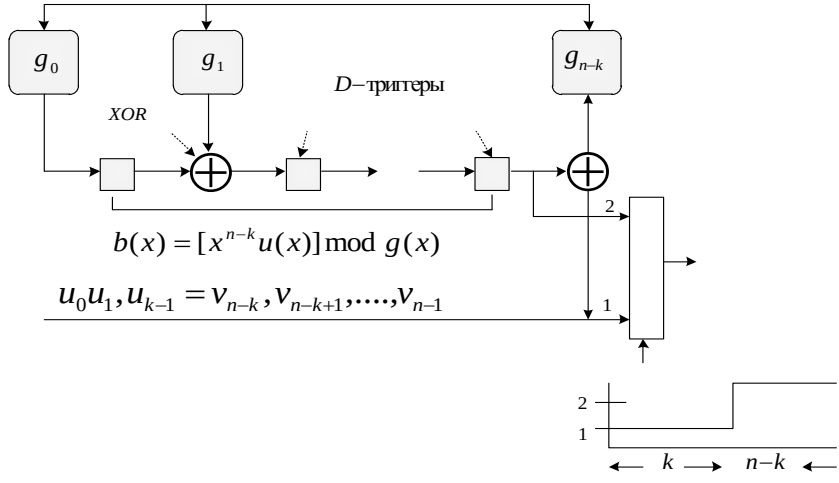


Рис. 7.29. Устройство систематического кодирования делением на $g(x)$

Дуальные циклические коды и последовательности максимальной длины. Дуальным *кодом циклического кода C*, который *порождается полиномом* $g(x)$, *является циклический код* $C^{\perp}$, *порождаемый полиномом* $h(x)$.

Важный класс представляют циклические коды, словами которых являются все сдвиги последовательности максимальной длины (MLS) и которые являются дуальными циклическому коду Хэмминга. Множество сдвигов является $(2^m - 1,\ m,\ 2^{m-1})$ циклическим кодом, который порождается полиномом $g(x) = (x^n - 1)/p(x)$, где $p(x)$ - первоначальный полином. В дальнейшем этот код будем называть *MLS кодом*.

Сокращенные циклические коды и CRC коды. Существует много практических заданий, в которых приходится использовать коды, которые исправляют ошибки согласно простым процедурам кодирования и декодиро-

вания. Тем не менее, существующие конструкции не всегда имеют нужную длину, размерность и минимальное расстояние.

Сокращение сводится к отбрасыванию информационных позиций начального кода. Пусть S - количество неупотребляемых информационных символов, которые называют *глубиной* (*длинной*) *сокращения*. Пусть C - циклический (n, k, d) код. Сокращение сообщения образовывается за счет фиксированного установления нулевых значений в некоторых (произвольных) информационных позициях. Другие позиции могут приобретать произвольные значения. Без потери для всеобщности соображений можем считать, что старшие позиции сообщения устанавливаются в нулевые состояния. Тогда $u(x) = \ = u_0 + u_1 x + ... + u_{k-1-s} x^{k-1-s}$. Данное сообщение превращается систематическим кодером в кодовое слово

$$v(x) = x^{n-k} u(x) + \left[x^{n-k} u(x) \bmod g(x) \right],$$

степень которого не превышает $n-s-1$. Таким образом, сокращенный код C_s является линейным $(n-s, k-s, d_s)$ кодом с кодовым расстоянием $d_s \geq d$. В общем случае сокращенный код не остается циклическим кодом.

Пример 7.26. Пусть C - циклический код Хэмминга с порождающим полиномом $1+x+x^3$. Новый код, образованный из C установлением в нулевое состояние двух старших информационных разрядов, имеет два информационных символа и три проверочных, вычисленные кодером кода C. Множество полученных кодовых слов представляет собой сокращенный линейный код.

Фундаментальное свойство сокращенных циклических кодов C_S заключается в том, что могут использоваться те же кодеры и декодеры, хотя эти коды и не сохраняют стойкость к циклическому сдвигу. Для компьютерного моделирования намного проще дополнять слова нулями на старших позициях и использовать те же алгоритмы кодирования и декодирования, которые здесь обсуждаются. Этот способ (дополнение нулями) широко используется в микросхемной реализации РС-кодов. Очевидно, что нули на старших позициях сообщения не должны включаться в кодовое слово. Более того, декодер модифицируется так, что выполняется умножение принятого слова $r(x)$ на x^{n-k+s} вместо умножения на x^{n-k} по модулю $g(x)$ в обычном декодере.

Еще одним возможным решениям может быть попытка построить другие классы циклических кодов с необходимыми параметрами. Интересными классами таких кодов являются *непервоначальные* коды БЧХ, евклидо-геометрические (EG) и проективно-геометрические (PG) коды. Еще одна возможность заключается в применении недвоичных циклических кодов в двоичном представлении, таких как РС-коды, которые рассматривались раньше. Двоичное отображение РС-кодов имеет дополнительную способность исправлять многократные пакеты ошибок.

CRC коды. Один из наиболее популярных стандартов помехоустойчивого кодирования обосновывается на избыточных циклических кодах для обнаружения ошибок (CRC коды). Эти циклические коды используются для обнаружения ошибок в блоках данных. CRC коды имеют длину $n \le 2^{m-1}$. Обычно CRC коды имеют порождающий полином вида $(1+x)g(x)$, где $g(x)$ - порождающий полином кода Хэмминга. Обычно значения *m* равняются 12, 16 и 32. Выбор порождающего полинома зависит от допустимой *вероятности* неопределенной ошибки, которая определяется распределением (спектром) весов кода. Вычисление вероятности неопределенной ошибки эквивалентно определению спектра веса кода. Эта задача остается чрезвычайно трудной.

Код	*m*	$g(x)$
CRC-12	12	$x^{12}+x^{11}+x^3+x^2+x+1$
CRC-16	16	$x^{16}+x^{15}+x^2+1$
CRC-CCITT	16	$x^{16}+x^{15}+x^5+1$
CRC-32	32	$x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+$ $+x^8+x^7+x^5+x^4+x^2+x+1$

Общий алгоритм декодирования циклических кодов. Пусть $r(x) = v(x)+e(x)$, где $e(x)$ - *полином ошибок*, который ассоциируется с вектором ошибок двоичного симметричного канала. Тогда синдром (синдромный полином) имеет вид

$$s(x) = r(x) \bmod g(x) = e(x) \bmod g(x). \qquad (7.60)$$

Обобщенную структуру декодера циклического кода иллюстрирует рис. 7.30. Синдром $s(x)$ используется для определения полинома ошибок $e(x)$. Поскольку циклический код является прежде всего линейным кодом, то эта структура может рассматриваться как вариант «стандартной таблицы» для циклических кодов.

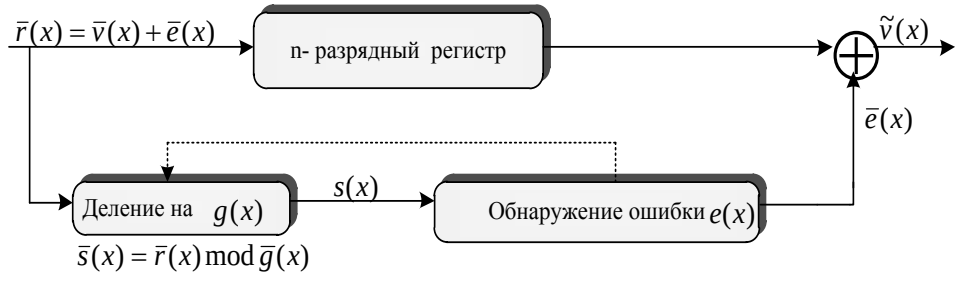


Рис. 7.30. Обобщенная структура декодера циклического кода

Проблема декодирования равноценна поиску (неизвестного) полинома ошибок $e(x)$ по известному синдрому $s(x)$. Эти полиномы связаны уравнениям, которое представляет основу синдромного декодера, известного также как *декодер Меггита* для циклического кода. Другой (но близкий) вариант декодера, который реализует алгоритм обнаружения ошибок, известный также как *декодер Касами*, проверяет совпадение синдрома с возможным вектором ошибок. Только очень немного классов кодов имеют такое сравнительно простое декодирование, как циклические коды Хэмминга и Голея.

Однако с увеличением корректирующей способности кода $t = [(d_{\min} - 1)/2]$ сложность декодера, который базируется на комбинаторном определении ошибок, становится чрезвычайно большой. Предположим, что ошибка, возникшая на первой принятой позиции, $e(x) = x^n - 1$. Соответствующий синдром $s(x) = x^{n-1} \bmod g(x)$. Если ошибка, которая искажает заданную позицию, оказывается данным циклическим кодом, то можно выявить ошибки и на других позициях за счет циклических сдвигов и соответствующей коррекции синдрома. Синдромный декодер проверяет синдром для каждой позиции принятого слова, и если оказывается полином $x^{n-1} \bmod g(x)$, то символ на этой позиции исправляется.

Пример 7.27. В этом примере рассматривается декодирование циклического кода Хэмминга с порождающим многочленом $g(x) = x^3 + x + 1$. Символы, которые принимаются, нагромождаются в регистре сдвига и одновременно вводятся в схему деления на $g(x)$. После приема седьмого бита содержимое этого регистра сдвигает на один разряд в каждом такте, а схема деления модифицирует синдром и проверяет совпадение с полиномом

$$x^6 \bmod(1 + x + x^3 = 1 + x^2 \Leftrightarrow 101) \text{ (в двоичной записи).}$$

Едва лишь на выходе схемы проверки появится 1, будет исправлено ошибку в позиции x^6. В тот самый момент исправления вводится за обратной связью в схему деления, обнуляя тем самым остаток от деления. Нулевой остаток может рассматриваться как сигнал об успешном завершении декодирование. Проверка на нулевой остаток схемы деления дает возможность обнаруживать некоторые аномалии после окончания процедуры декодирования.

7.4. Сверточное помехоустойчивое кодирование

Структурная схема информационно-коммуникационной системы передачи данных с использованием помехоустойчивого сверточного кодирования/декодирования приведена на рис. 7.31. Информационное сообщение на выходе кодера источника информации обозначается последовательностью

$m = m_1, m_2, ..., m_i, ...$, где m_i - двоичный знак (бит), а u - индекс времени (см. разд. 7.1).

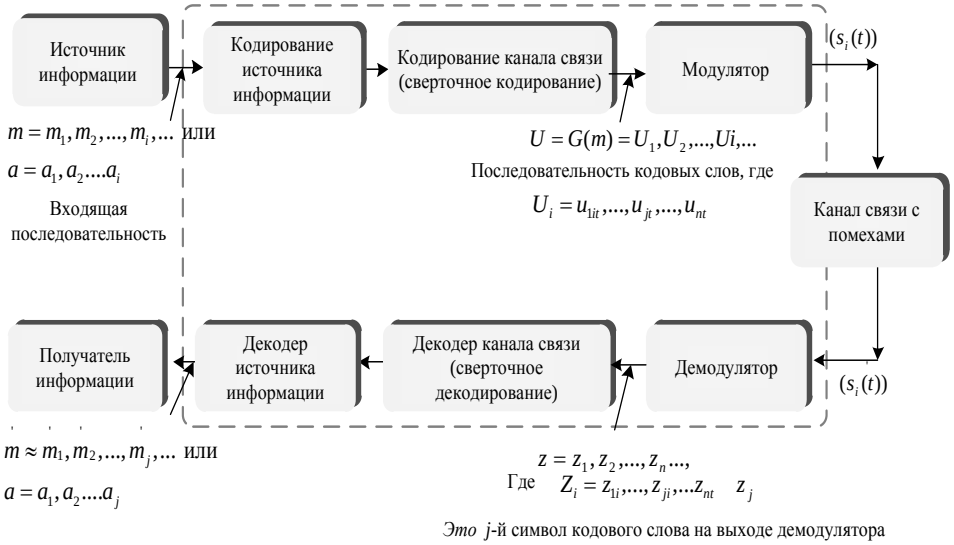


Рис. 7.31. Кодирование/декодирование и модуляция/демодуляция в каналах связи

Если быть точным, то элементы m следовало бы дополнять индексом, который соответствует типу кодирования источника согласно определенной раньше классификации и индексу времени. Однако для упрощения будет использоваться только индекс, который обозначает время (или размещение элемента внутри последовательности). Допустим, что все m_i равновероятные, и равняются единице или нулю и независимы между собой. Будучи независимой, последовательность бит имеет потребность в некоторой избыточности, т.е. знание о бите m_i не даст ни одной информации о бите m_j (при $i \neq j$). Кодер превратит каждую последовательность m в уникальную последовательность кодовых слов $U = G(m)$. И хотя последовательность m однозначно определяет последовательность U, ключевой особенностью сверточных кодов является то, что данный k-кортеж внутри m неоднозначно определяет связанные с ним n-кортежи внутри U, поскольку кодирование каждого из k-кортежей является функцией не только k-кортежей, а и предыдущих $k-1$ k-кортежей (кодирование с памятью). Последовательность U можно поделить на последовательность кодовых слов:

$$U = U_1, \ U_2, \ ..., \ U_i, ...$$

Каждое кодовое слово U_i состоит из двоичных кодовых символов, которые часто называют *канальными символами*, или *битами кода*, а соответствующую процедуру кодирования - *канальным помехоустойчивым кодированием*.

Сверточные (***рекурентные***, ***цепи***) ***коды*** *используются для кодирования непрерывной последовательности двоичных символов путем введения в эту последовательность специальных проверочных* (*избыточных*) *символов с целью обнаружения и исправление искажений в информационных сообщениях.*

Например, для каждого информационного символа исходной последовательности $A(a_i)$, которые различаются между собой на шаг кодирования (добавление, перемежевание) λ, формируется один специальный проверочный символ П.

Передаются проверочные символы с определенной задержкой относительно информационных. На стороне приемника из информационных символов a' формируются новые проверочные, которые в отличие от проверочных символов Π', называются *контрольными*. Контрольные символы сравниваются с проверочными, и в случае их расхождения формируется вывод о наличии искажения. Проверочные и контрольные символы формируются добавлением по модулю 2 информационных символов.

Передача	Прием
$a_i \oplus a_{i+\lambda} = \Pi_{i,i+\lambda}$	$a'_i \oplus a'_{i+\lambda} = K_{i,i+\lambda}$
$a_{i+1} \oplus a_{i+\lambda+1} = \Pi_{i,i+\lambda+1}$	$a'_{i+1} \oplus a'_{i+\lambda+1} = K_{i,i+\lambda+1}$
…	…
$a_{i+\lambda} \oplus a_{i+2\lambda} = \Pi_{i+\lambda,i+2\lambda}$	$a'_{i+\lambda} \oplus a'_{i+2\lambda} = K_{i+\lambda,i+2\lambda}$
$a_{i+\lambda+1} \oplus a_{i+2\lambda+1} = \Pi_{i+\lambda+1,i+2\lambda+1}$	$a'_{i+\lambda+1} \oplus a'_{i+2\lambda+1} = K_{i+\lambda+1,i+2\lambda+1}$

и т.д.

Одна проверка охватывает те информационные символы, разность номеров которых равняется шагу кодирования:

$$i+\lambda-i = i+\lambda+1-(i+1) = i+2\lambda+1-(i+\lambda) = \ldots = \lambda.$$

Каждый контрольный символ сравнивается с соответствующим проверочным:

$$S_{i,i+\lambda} = K_{i,i+\lambda} \oplus \Pi'_{i,i+\lambda};$$

$$S_{i+1,i+\lambda+1} = K_{i+1,i+\lambda} \oplus \Pi'_{i+1,i+\lambda+1} \quad \text{и т.д.}$$

Признаком отсутствия искажений есть то, что все суммы равняются нулю:

$$S_{i,i+\lambda} = S_{i+1,i+\lambda+1} = \ldots = 0.$$

Искаженным может быть как информационный, так и проверочный символ.

На искажение одного проверочного символа указывает то, что одна из сумм равняется единице. Например, при искажении $\Pi'_{i,i+\lambda}$ получаем $S_{i,i+\lambda}=1$, поскольку $K_{i,i+\lambda}$ не совпадает с $\Pi'_{i,i+\lambda}$. Если дальнейшей передачи этой последовательности не происходит (ретрансляция отсутствующая), то и исправлений нет.

Наличие двух сумм, которые равны единице и отдалены (сдвинуты) между собой на шаг добавления λ, свидетельствует об искажении двух проверочных символов.

Пример 7.28. Рассмотрим последовательный сверточный код с шагом кодирования $\lambda=3$ при условии, что проверочные символы приняты без искажений. Пусть с искажением приняты три информационных символа. Формирование проверочных и контрольных символов иллюстрирует рис. 7.32

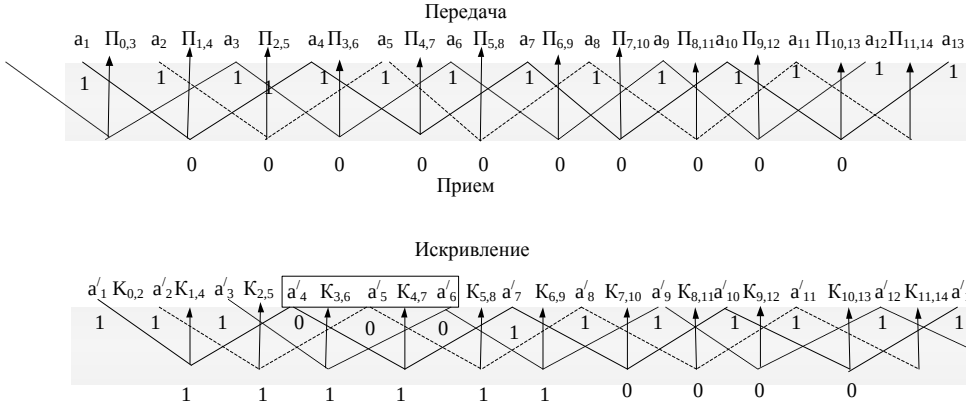


Рис. 7.32. Формирование проверочных и контрольных символов при наличии искажения в символах с номерами 4, 5, 6

Все проверочные символы, сформированные для условий примера, равны нулю, а что касается контрольных, то среди них все символы, сформированные с привлечением искаженных информационных, равняются единице, а остальные, как и проверочные, равны нулю.

Нетрудно увидеть, что сравнение контрольных элементов, которые при приеме сформированы с привлечением неискаженных информационных символов, с соответствующими проверочными дает значение, равное нулю. А сравнив контрольные элементы, которые при приеме сформированы с привлечением искаженных информационных символов, с соответствующими проверочными, получим:

$$S_{1,4}=K_{1,4}\oplus\Pi'_{1,4}=1\oplus 0=1;$$

$$S_{2,5}=K_{2,5}\oplus\Pi'_{2,5}=1\oplus 0=1;$$

$$S_{3,6} = K_{3,6} \oplus \Pi'_{3,6} = 1 \oplus 0 = 1;$$

$$S_{4,7} = K_{4,7} \oplus \Pi'_{4,7} = 1 \oplus 0 = 1;$$

$$S_{5,8} = K_{5,8} \oplus \Pi'_{5,8} = 1 \oplus 0 = 1;$$

$$S_{6,9} = K_{6,9} \oplus \Pi'_{6,9} = 1 \oplus 0 = 1;$$

Итак, образовались три пары сумм, которые равны единице и сдвинуты на шаг кодирования $\lambda = 3$:

$$S_{1,4} \text{ и } S_{4,7};\ S_{2,5} \text{ и } S_{5,8};\ S_{3,6} \text{ и } S_{6,9}.$$

Согласно последним выражениям приходим к выводу, что искажены те информационные символы, номера позиций которых являются общими в каждой паре сумм, т.е. a_4, a_5, a_6. *Значение этих символов необходимо исправить на противоположные*: *принято 0, а должно быть 1, и наоборот.*

Таким образом, исправлено три искажения, т.е. количество искажений, которые исправляются, равно шагу добавления.

Из этого примера вытекает еще и такой вывод: *значение* λ *определяет не только шаг добавления, а и количество отдельных независимых цепочек кода.*

В самом деле, при $\lambda = 3$ имеем три независимых цепочки:

$$a_1 \to \Pi_{1,4} \leftarrow a_4 \to \Pi_{4,7} \leftarrow a_7 \to \Pi_{7,10} \leftarrow a_{10} \to \Pi_{10,13} \leftarrow a_{13} \ldots$$

$$a_2 \to \Pi_{2,5} \leftarrow a_5 \to \Pi_{5,8} \leftarrow a_8 \to \Pi_{8,11} \leftarrow a_{11} \to \Pi_{11,14} \leftarrow a_{14} \ldots$$

$$a_3 \to \Pi_{3,6} \leftarrow a_6 \to \Pi_{6,9} \leftarrow a_9 \to \Pi_{9,12} \leftarrow a_{12} \to \Pi_{12,15} \leftarrow a_{15} \ldots$$

Пример 7.29. Значение информационных и проверочных символов при передаче такие же, как и в предыдущем примере (см. рис. 7.32). Проверочные символы приняты без искажений, тогда как информационные символы a'_1, a'_4, a'_5, a'_6, a'_{12} приняты с искажениями (рис. 7.33).

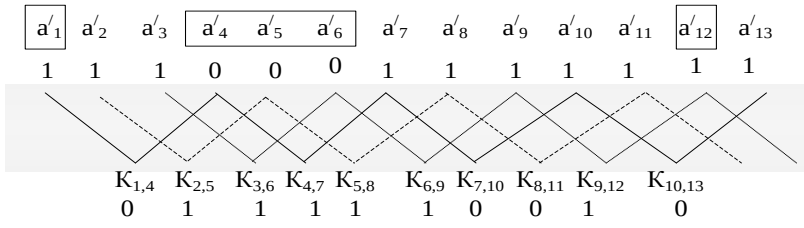


Рис. 7.33. Формирование проверочных и контрольных символов при наличии искажений в символах с номерами 1, 4, 5, 6, 12

Приведем результат сравнения контрольных и проверочных символов для каждой из цепочек:

Первая цепочка: $S_{1,4}=0;\ S_{4,7}=1;\ S_{7,10}=0;\ S_{10,13}=0;$

Вторая цепочка: $S_{2,5}=1;\ S_{5,8}=1;\ S_{8,11}=0;$

Третья цепочка: $S_{3,6}=1;\ S_{6,9}=1;\ S_{9,12}=1;$

На основании добытых результатов суммирования контрольных и проверочных символов рассмотрим возможность исправления искажений в каждой независимой цепочке кода.

Первая цепочка содержит $a'_1,\ a'_4,\ a'_7,\ a'_{10},\ a'_{13}$ и соответствующие проверочные символы $\Pi'_{1,4},\ \Pi'_{4,7},\ \Pi'_{7,10},\ \Pi'_{10,13}$. По условию искаженными являются $a'_1,\ a'_4$. В этой цепочке только одна из сумм равна единице:

$$S_{4,7}=K_{4,7}\oplus\Pi'_{4,7}=1.$$

Согласно правилам обнаружения искажений можно утверждать, что искаженным является проверочный символ $\Pi'_{4,7}$. Тем не менее, это противоречит условию примера: проверочные символы приняты без искажений. Итак, искаженных информационных символов, которые входят в первый цепочку, не обнаружено. Более того, неправильно «исправленным» будет проверочный символ $\Pi'_{4,7}$.

Вторая цепочка содержит $a'_2,\ a'_5,\ a'_8,\ a'_{11}$, и $\Pi'_{2,5},\ \Pi'_{5,8},\ \Pi'_{8,11}$. По условию искажения a'_5. Тот факт, что в этой цепочке равны единице две суммы $(S'_{2,5}=S'_{5,8}=1)$, означает, что искаженным является один информационный символ с общим для этих сумм индексом, т.е. a'_5.

Во второй цепочке искажение будет исправлено правильно.

В третью цепочку входят $a'_3,\ a'_6,\ a'_9,\ a'_{12}$, и $\Pi'_{3,6},\ \Pi'_{6,9},\ \Pi'_{9,12}$. Согласно условию примера искаженными являются a'_6, и a'_{12}. В этой цепочке три суммы равняются единице: $S'_{3,6}=S_{6,9}=S_{9,12}=1$. Согласно равенству $S'_{3,6}=S_{6,9}=1$ можно исправить a_6; равенство $S_{9,12}=1$ указывает на искажение $\Pi'_{9,12}$, что противоречит условию. Таким образом, искажение a'_{12} не выявлено и неправильно исправлено $\Pi'_{9,12}$. Если взять $S_{6,9}=S_{9,12}=1$, то можно считать искаженным a'_9, а это также неправильно.

Рассмотрев два примера с разным количеством искаженных информационных символов, можно прийти к таким выводам:

1. Сверточный код исправляет групповое искажение с λ информационных символов (см. первый пример). Чем большая длина групповой помехи, тем большим должен быть шаг добавления. Но с увеличением шага добавления возрастает и сложность кода преобразователей.

2. Правильное исправление искажений возможно, если в каждой независимой цепочке кода справа и слева от искажения есть два неискаженных символа (во втором примере - первая и третья цепочки). А поскольку информационные символы в каждой цепочке размещены на расстоянии λ от информационных символов других цепочек, то между групповыми искажениями должны быть по меньшей мере 2λ неискаженных информационных символов.

Местоположение проверочных символов определяется двумя обстоятельствами: во-первых, групповая помеха не должна одновременно охватывать информационные и соответствующие проверочные символы; во-вторых, не должно быть ошибочного исправления информационных символов.

Из этих соображений каждый проверочный символ располагается на расстоянии $2\lambda+1$ от ближайшего своего информационного символа. Например, проверочный символ $П_{i,i+\lambda}$, созданный из информационных a_i и $a_{i+\lambda}$, должен занимать $[(i+\lambda)+2\lambda+1]$ позицию.

Поскольку разность между информационными символами каждой независимой цепочки представляет λ информационных символов, то и расстояние между соответствующими проверочными символами также равняется шагу кодирования.

Важной характеристикой сверточного кода является минимально допустимое расстояние между сопредельными групповыми искажениями, когда еще возможно исправлять искажение. Это расстояние должно обеспечивать правильный прием 2λ информационных символов после искажения и проверочных, которые охватывают искаженные информационные символы. Поэтому минимально допустимое расстояние равняется $2\lambda+1$ символ.

В информационных сообщениях, которые передаются беспрерывно, проверочные и информационные символы чередуются: a_1, П, a_2, П, a_3,.... Это значит, что информационные символы, которые входят в одну проверку, отличаются один от другого на λ информационных и λ проверочных символов. Т.е. при шаге добавления λ информационных символов сверточный код исправляет $b=2\lambda$ информационных и проверочных символов, а минимально допустимое расстояние между групповыми искажениями $M=4\lambda+1=2b+1$ информационных и проверочных символов.

Диаграммы состояний и процедуры сверточного кодирования. Сверточный кодер принадлежит к классу устройств, известных как *конечный автомат*. Это общее название охватывает системы, которые имеют память о прошлых сигналах. Процедура использования конечных автоматов при помехоустойчивом сверточном кодировании непосредственно определяется алгоритмом кодирования - кодирование с памятью. Прилагательное «конечный» показывает, что существует ограниченное количество состояний, которые могут возникнуть в системе. В обобщенном смысле состояние содержит наименьшее количество информации, на основании которой вместе с теку-

щими входными данными можно определить данные на выходе системы. Состояние дает некоторое представление о прошлых событиях (сигналы) и об ограниченном наборе возможных исходных данных в будущем, т.е. будущие состояния ограничиваются прошлыми состояниями. Для сверточного кода со степенью кодирования $1/n$ состояние подается содержимым $K-1$ крайних правых разрядов. Знание состояния плюс знание следующих данных на входе являются необходимым и достаточным условием для определения данных на выходе.

Итак, пусть состояние кодера в момент времени t_i определяется как $X_i = m_i - 1, m_i - 2, ..., m_i - K + 1$. При этом u ветка кодовых слов U_i целиком определяется состоянием X_i и введенными в данное время битами m_i. Таким образом, состояние X_i описывает предысторию кодера для определения данных на его выходе. Состояние кодера считается марковским в том смысле, что вероятность $P(X_i + 1 | X_i, ..., X_0)$ пребывания в состоянии $X_i + 1$, обусловленная всеми предыдущими состояниями, зависит только от последнего состояния X_i, т.е. она равняется $P(X_i + 1 | X_i)$.

Одним из способов представления простых кодирующих устройств является *диаграмма состояния* (см. рис. 7.34). Состояния, изображенные в рамках диаграммы, представляют собой возможное содержимое $K-1$ крайних правых разрядов регистра, а пути между состояниями - кодовые слова веток на выходе, что является результатом переходов между такими состояниями. Состояние регистра взяты такие: $a = 00,\ b = 10,\ c = 01$ *и* $d = 11$.

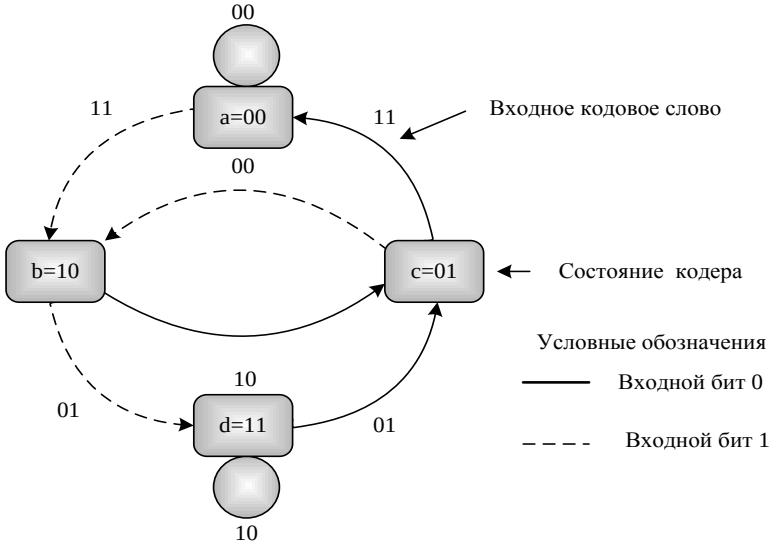


Рис. 7.34. Диаграмма состояний кодера (*степень кодирования* $1/2, K = 3$)

Существует всего два исходных перехода из каждого состояния, что соответствует двум возможным входным битам. Дальше для каждого пути между состояниями записано кодовое слово на выходе, связанное с переходами между состояниями. При изображении путей сплошной линией обозначают путь, связанный с нулевым входным битом, а пунктирной - путь, связанный с единичным входным битом.

Заметим, что за один переход *невозможно* перейти из данного состояния в любое произвольное. Поскольку за единицу времени перемещается только один бит, существует только два возможных перехода между состояниями, в которые регистр может переходить за время прохождения каждого бита. Например, если состояние кодера 00, то при следующем сдвиге может возникнуть только состояние 00 или 10.

Древовидные диаграммы кодирования. Несмотря на то, что диаграммы состояний целиком описывают кодер, их, в сущности, нельзя использовать для легкого отслеживания переходов кодера в зависимости от времени, поскольку диаграмма не отражает динамики изменений. Древовидная диаграмма прибавляет к диаграмме состояния временное измерение. Древовидную диаграмму сверточного кодера в каждый следующий момент прохождения входного бита процедуры кодирования можно описать с помощью перемещения по диаграмме слева направо, причем каждая ветка дерева описывает кодовое слово на выходе.

Правило разветвления для отыскания последовательности кодовых слов такое:

если входным битом является нуль, то он связывается со словом, которое отыскивается перемещением в следующую (по направлению вверх) правую ветку;

если входной бит - единица, то кодовое слово отыскивается перемещением в следующую (за направлением вниз) правую ветку.

Предполагается, что сначала кодер содержал одни нули. Диаграмма показывает, что когда первым входным битом был нуль, то кодовым словом ветки на выходе будет 00, а если первым входным битом была единица, то кодовым словом на выходе будет 11. Аналогично, если первым входным битом была единица, а вторым - нуль, на выходе вторым словом ветки будет 10. Если первым входным битом была единица и вторым входным битом была единица, вторым кодовым словом на выходе будет 01. Согласно этой процедуре видим, что входная последовательность 11011 подается жирной линией, изображенной на древовидной диаграмме (рис. 7.35). Этот путь соответствует исходной последовательности кодовых слов 11 01 01 00 01.

Дополнительное измерение времени в древовидной диаграмме (сравнительно с диаграммой состояний) допускается динамическое описание кодера как функции конкретной входной последовательности. Но при попытке описания с помощью древовидной диаграммы последовательности произвольной

длины возникает проблема. Количество ответвлений возрастает как 2^L, где L - это количество кодовых слов веток в последовательности.

Решетчатая диаграмма. Исследование древовидной диаграммы, приведенной на рис. 7.35, показывает, что в этом примере после третьего разветвления в момент времени t_4 структура повторяется (в общем случае древовидная структура повторяется после *K* ответвлений, где *K* - длинна кодового ограничения). Обозначим каждый узел в дереве (см. рис. 7.35), поставив в соответствие четыре возможных состояния в регистре смещения: $a = 00$, $b = 10$, $c = 01$ и $d = 11$.

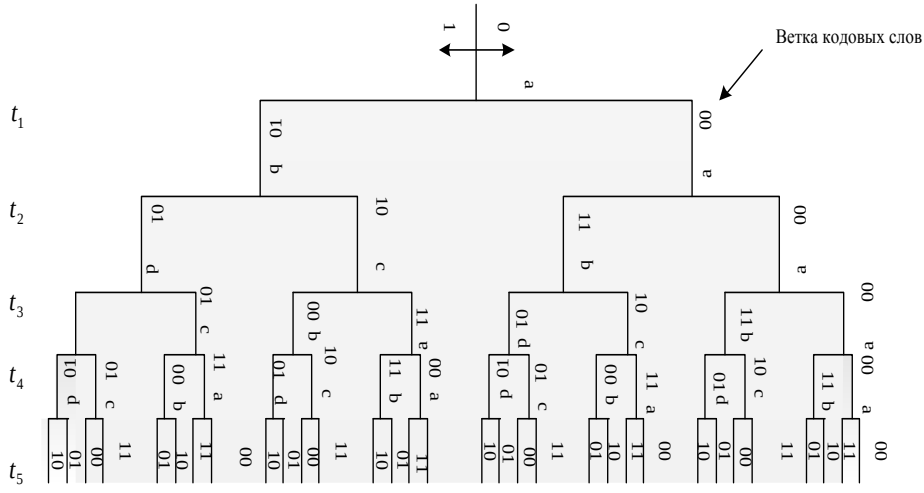


Рис. 7.35. Дерево алгоритма сверточного кодирования
(степень кодирования $1/2, K = 3$)

Первое разветвление древовидной структуры в момент времени t_1 дает пары узлов, обозначенных как *a* и *b*. При каждом следующем разветвлении количество узлов удваивается. Второе разветвление в момент времени t_2 дает в результате четыре узла, обозначенные как *a, b, c* і *d*. После *третьего* разветвления всего есть восемь узлов: два - *a*, два - *b*, два - *c* и два - *d*.

Можно увидеть, что все ветки выходят из двух узлов того же состояния, образовывая идентичные ветки последовательностей кодовых слов. В этот момент дерево делится на идентичные верхнюю и нижнюю части. Когда четвертый входной бит входит в кодер слева, первый входной бит справа отбрасывается и больше не влияет на кодовые слова на выходе.

Итак, входные последовательности $100xy...$ и $000xy...$, где крайний левый бит наиболее ранний, после *K*-го $(K = 3)$ разветвления генерируют одинаковые кодовые слова веток. Это означает, что любые состояния, которые имеют одинаковую метку в тот самый момент t_i, можно соединить, поскольк-

ку все следующие пути будут различаемыми. Если мы выполним это для древовидной структуры, изображенной на рис. 7.35, получим другую диаграмму, которая называется *решетчатой*. *Решетчатая диаграмма*, которая использует повторяемую структуру, дает более удобное описание кодера сравнительно с древовидной диаграммой. Решетчатая диаграмма для сверточного кодера изображена на рис. 7.36.

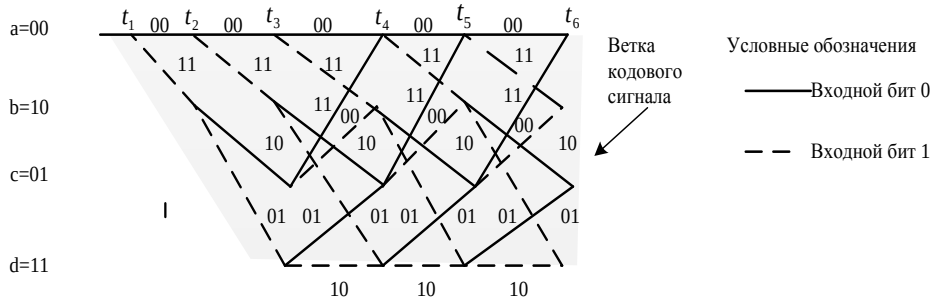


Рис. 7.36. Решетчатая диаграмма кодера (степень кодирования $1/2$, $K=3$)

При изображении решетчатой диаграммы мы воспользовались теми же условными обозначениями, что и для диаграммы состояния: сплошная линия обозначает исходные данные, которые генерируются входным нулевым битом, а пунктирная - исходные данные, которые генерируются входным единичным битом. Узлы решетки дают состояния кодера: первый ряд узлов отвечает состоянию $a=00,$ второй и следующие - состоянию $b=10,$ $c=01$ и $d=11.$ В каждый момент времени для представления 2^{K-1} возможных состояний кодера решетки нужно 2^{K-1} узлов. В нашем примере после достижения глубины решетки, которая равна трем (в момент времени t_4), делаем замечание, что решетка имела фиксированную периодическую структуру. В общем случае фиксированная структура реализуется после достижения глубины K. Итак, из этого момента в каждое состояние можно попасть из каждого из двух предыдущих состояний. Также из каждого состояния можно перейти в одно из двух состояний. Из двух исходных веток одна отвечает нулевому входному биту, а другая - единичному входному биту. На рис. 7.36 кодовые слова на выходе соответствуют переходам между состояниями, которые изображены как метки на ветках решетки.

Один столбец временного интервала решетчатой структуры кодирования, которая сформировалась, целиком определяет код. Несколько столбцов изображены лишь для визуализации последовательности кодовых символов как функции времени. Состояние сверточного кодера подается содержимым крайних правых $K-1$ разрядов в регистре кодера. Некоторые авторы описывают состояние с помощью крайних левых $K-1$ разрядов. Оба описания правильные. Каждый переход имеет начальное и конечное состояния. Край-

ние правые $K-1$ разряды описывают начальное состояние для текущих входных данных, которые содержатся в крайнем левом разряде (степень кодирования предполагается равной $1/n$). Крайние левые $K-1$ разряды являются конечным состоянием для такого перехода. Последовательность кодовых символов характеризуется N ветками (которые подают N бит данных), которые занимают N интервалов времени. Она связана с конкретным состоянием в каждый из $N+1$ интервалов времени (от начала до конца). Таким образом, мы запускаем биты в моменты времени $t_1, t_2, ..., t_N$ и интересуемся метрическим свидетельством состояния в моменты времени $t_1, t_2, ..., t_{N+1}$. Здесь использовано такое условие: текущий бит размещается в крайнем левом разряде, а крайние правые $K-1$ разряды стартуют из состояния со всеми нулями. Этот момент времени обозначим как *начальное время* t_1. Время завершения последнего перехода обозначим как *время прекращения работы* t_{N+1}.

Декодирование по методу максимального правдоподобия. Если все входные последовательности сообщений равновероятные, минимальная вероятность ошибки достигается при использовании декодера, который сравнивает условные вероятности и выбирает максимальную. Условные вероятности также называются *функциями правдоподобия декодирования* $P(Z|U^{(m)})$, где Z - принятая последовательность, а $U^{(m)}$ - одна из возможных переданных последовательностей. Декодер выбирает $U^{(m')}$, если за всеми $U^{(m)}$

$$P(Z|U^{(m')}) = \max P(Z|U^{(m)}). \qquad (7.61)$$

Принцип *максимального правдоподобия*, которое определяется уравнением (7.61), является фундаментальным достижением теории принятия решений на фоне статистических данных. При рассмотрении двоичного симметричного канала предполагалась передача только двух равновероятных сигналов $s_1(t)$ и $s_2(t)$. Итак, принятие двоичного решения на основе принципа максимального правдоподобия, которое касается данного полученного сигнала, означает, что переданный сигнал выбирается $s_1(t)$, если $p(z|s_1) > p(z|s_2)$. В противоположном случае считается, что передавался сигнал $s_2(t)$. Параметр z зависит от T, т.е. представляет собой величину $z(T)$ - значение принятого сигнала к детектированию в конце каждого периода передачи символа $t = T$.

В случае использования принципа максимального правдоподобия в задаче сверточного декодирования считается, что в сверточном коде имеется память (полученная последовательность является суперпозицией текущих и предыдущих двоичных разрядов). Таким образом, принцип максимального правдоподобия при декодировании данных, закодированных сверточным

кодом, применяется в контексте выбора *наиболее возможной последовательности,* как показано в уравнении (7.61). Обычно существует множество возможных переданных последовательностей кодовых слов. Что же касается двоичного кода, то последовательность из L кодовых слов является составляющей набора из 2^L возможных последовательностей. Итак, в контексте максимального правдоподобия можно сказать, что как переданную последовательность декодер выбирает $U^{(m')}$, если правдоподобие $P(Z|U^{(m')})$ больше правдоподобия всех других последовательностей, которые могли быть переданными.

Оптимальным декодером *является такой декодер, который использует алгоритм максимального правдоподобия и минимизирует вероятность ошибки в информационном сообщении при условии, что все переданные последовательности равновероятные.*

Функцию правдоподобия задают или вычисляют исходя из спецификации канала.

Предположим, что мы имеем дело с аддитивным белым гауссовым шумом с нулевым средним, а соответственно, с каналом без памяти, т.е. шум влияет на каждый символ кода *независимо* от других символов. Если степень кодирования сверточного кода равна $1/n$, правдоподобие можно выразить как

$$P(Z \mid U^{(m)}) = \prod_{l=1}^{\infty} P(Z_i \mid U_i^{(m)}) = \prod_{l=1}^{\infty} \prod_{j=1}^{\infty} P(z_{ji} \mid u_{ji}^{(m)}), \qquad (7.62)$$

где Z_i - *i -я* ветка принятой последовательности Z; $U_i^{(m)}$ - ветка отдельной последовательности кодовых слов $U^{(m)}$; z_{ji} - *j*-й кодовый символ Z_i; $u_{ji}^{(m)}$ - *j*-й кодовый символ $u_i^{(m)}$, а каждая ветка состоит из n кодовых символов.

При вычислениях удобнее пользоваться логарифмом функции правдоподобия, поскольку это дает возможность произведение заменить суммированием. Можем воспользоваться таким преобразованием, поскольку логарифм как монотонно возрастающая функция не внесет изменений в выбор окончательного кодового слова. Логарифмическую функцию правдоподобия можно определить как

$$\gamma_U(m) = \lg P(Z \mid U^{(m)}) = \sum_{l=1}^{\infty} \lg P(Z_i \mid U_i^{(m)}) = \sum_{l=1}^{\infty} \sum_{j=1}^{n} \lg P(z_{ji} \mid u_{ji}^{(m)}). \qquad (7.63)$$

Теперь задача декодирования состоит в выборе пути вдоль дерева на рис. 7.35 или решетке на рис. 7.36 таким образом, чтобы $\gamma_u(m)$ было максимальным. При декодировании сверточных кодов можно использовать как древовидную, так и решетчатую структуру. В случае древовидного представления кода игнорируется то, что пути снова объединяются. Для двоичного кода количество возможных последовательностей, которые состоят из L

кодовых слов, равна 2^L. Поэтому декодирование полученных последовательностей, которое обосновывается на принципе максимального правдоподобия с использованием древовидной диаграммы, нуждается в методе исчерпывающего сравнения 2^L нагроможденных логарифмических метрических свидетельств правдоподобия, которые описывают все варианты возможных последовательностей кодовых слов. Поэтому рассматривать декодирование на основе принципа максимального правдоподобия с помощью древовидной структуры практически невозможно.

В предыдущем разделе было показано, что при решетчатом представлении кода декодер можно построить так, чтобы можно было отказываться от путей, которые не могут выступать как максимально правдоподобная последовательность. Путь декодирования выбирается из какого-то сокращенного набора путей, которые остались. Такой декодер, тем не менее, является оптимальным в том смысле, что путь декодирования такой же, как и путь, полученный с помощью декодера критерия максимального правдоподобия, хотя предыдущий отказ от неудачных путей снижает сложность декодирования.

***Модели каналов*: *мягкое или твердое принятие решений*.** Перед тем как начать разговор об алгоритме, который задает схему принятия максимально правдоподобного решения, сначала рассмотрим модель канала. Последовательность кодовых слов $U^{(m)}$, определенная словами ветки, каждое из которых состоит из n кодовых символов, можно рассматривать как бесконечный поток в отличие от блочного кода, где исходные данные и их кодовые слова делятся на блоки строго определенного размера. Последовательность кодовых слов, показанная на рис. 7.31, выдается сверточным кодером и подается на модулятор, где кодовые символы превращаются в сигналы. Модуляция может быть низкочастотной (например, модуляция импульсными сигналами) или полосовой. Вообще за такт кодирования в сигнал $s_i(t)$ превратится l символов, где l - целое, причем $i = 1, 2, ...,$ а $M = 2^l$. Если $l = 1$, модулятор превратит каждый кодовый символ в двоичный сигнал. Предполагается, что канал, по которому передается сигнал, искажает сигнал гауссовым шумом. После того как искаженный сигнал принят, он сначала обрабатывается демодулятором, а потом подается на декодер.

Рассмотрим ситуацию, когда двоичный сигнал передается за отрезок времени $(0, T)$, причем двоичная единица подается сигналом $s_1(t)$, а двоичный нуль - сигналом $s_2(t)$. Принятый сигнал имеет вид $r(t) = s_i(t) + n(t)$, где $n(t)$ - взнос гауссовой помехи с нулевым средним. *Детектирование* $r(t)$ *происходит в два основных этапа.*

На первом этапе принятый сигнал переводится в число $z(T) = a_i + n_0$, где a_i - компонент сигнала $z(T)$, а n_0 - компонент шума. Компонент шума n_0 - это случайная переменная, значение которой имеют гауссовое распреде-

ление с нулевым средним. Итак, $z(T)$ также будет случайной гауссовой величиной со средним a_1 или a_2 зависимо от того, какую величину было отправлено - двоичную единицу или двоичный нуль.

На втором этапе процесса детектирование принимается решение о том, какой сигнал было передано. Это решение принимается на основе сравнения $z(T)$ с порогом. Условные вероятности $z(T)$, $p(z|s_1)$ и $p(z|s_2)$, изображенные на рис. 7.37, обозначены как правдоподобие s_1 и s_2. Демодулятор, приведенный на рис. 7.31, превращает упорядоченный по времени набор случайных переменных $\{z(T)\}$ в кодовую последовательность Z и подает ее на декодер.

Выход демодулятора можно настроить по-разному. Можно реализовать его в виде жесткой схемы принятия решений относительно того, единицу или нуль подает $z(T)$. В этом случае выход демодулятора квантуется на два уровня - нулевой и единичный, а дальше соединяется с декодером. Поскольку декодер работает в режиме жесткой схемы принятия решений, которые принимает демодулятор, такое *декодирование называется жестким.*

Аналогично демодулятор можно настроить так, чтобы он подавал на декодер значения $z(T)$, квантованное более чем на два уровня. Такая схема обеспечивает декодер большим количеством информации, чем твердая схема решений. Если выход демодулятора имеет более чем два равных квантования, то *декодирование называется мягким.*

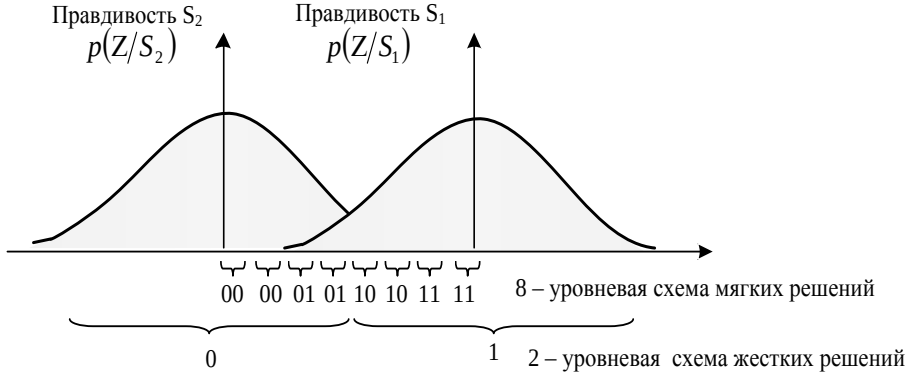


Рис. 7.37. Жесткая и мягкая схемы декодирования

На рис. 7.37 на оси абсцисс изображены восемь (3-битовых) уровней квантования. Если в демодуляторе реализована твердая схема принятия двоичных решений, он отправляет на декодер только один двоичный символ. Если в демодуляторе реализована мягкая двоичная схема принятия решений, квантованная на восьми уровнях, он отправляет на 3-декодер битовое слово, которое описывает интервал, который соответствует $z(T)$. В сущности, по-

ступление такого 3-битового слова вместо одного двоичного символа эквивалентно передаче декодеру *меры вероятности* вместе с решением относительно кодового символа. Согласно рис. 7.37, если из демодулятора поступила на декодер последовательность 111, это означает, что с очень высокой степенью вероятности кодовым символом была 1, тогда как переданная последовательность 100 означает, что с очень низкой степенью вероятности кодовым символом была 1. Вполне понятно, что, в конце концов каждое решение, принятое декодером относительно сообщения, должно быть твердым. То, что после демодулятора *не принимается твердое решение* и на декодер поступает больше данных (мягкое принятие решений), можно понимать как промежуточный этап, необходимый для того, чтобы на декодер поступило больше информации, с помощью которой он потом сможет восстановить последовательность сообщения (с высшей вероятностью передачи сообщения сравнительно с декодированием в рамках твердой схемы принятия решений).

Сейчас существуют блочные и сверхточные алгоритмы декодирования, которые функционируют на основе твердой или мягкой схемы принятия решений. Однако при блочном декодировании мягкая схема принятия решений, как правило, не используется, поскольку ее значительно сложнее реализовать, чем схему твердого принятия решений. Чаще всего мягкая схема принятия решений применяется в *алгоритме сверточного декодирование Витерби*, поскольку при декодировании Витерби мягкое принятие решений лишь немного усложняет вычисление.

Способность сверточного кода к коррекции t ошибок *характеризуется количеством ошибочных кодовых символов, которое можно исправить в каждом блоке кода путем декодирования по методу максимального правдоподобия.*

Вместе с тем при декодировании сверточных кодов способность кода к коррекции ошибок нельзя сформулировать так лаконично. При декодировании по принципу максимального правдоподобия код способен исправить *t* ошибок в пределах нескольких длин кодового ограничения (от 3 до 5). Точное значение длины зависит от распределения ошибок. Для конкретного кода и модели ошибки длину можно ограничить с использованием методов передаточной функции кода.

Систематический сверточный код *- это код, в котором входной k-кортеж тоже фигурирует как часть исходного n-кортежа кодового слова, которое отвечает этому k-кортежу.*

Двоичный систематический кодер со степенью кодирования $1/2$ и $K = 3$ изображен на рис. 7.38. Для линейных блочных кодов любой несистематический код можно превратить в систематический с такими самыми пространственными характеристиками блоков. В случае использования сверточных кодов это не так. Причина в том, что сверточные коды сильно зависят от *просвета*. При построении сверточного кода в систематической форме, когда

даны длина кодового ограничения и степень кодирования, максимально возможное значение просвета *снижается*.

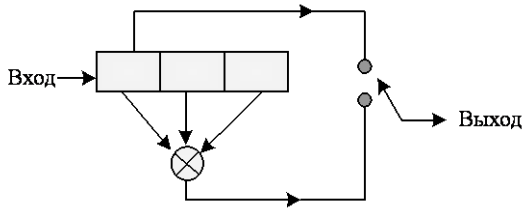


Рис. 7.38. Систематический сверточный кодер (степень кодирования $1/2$, $K=3$)

Максимальный просвет в случае, если степень кодирования равна $1/2$, для систематического и несистематического кодов при K, что изменяется от 2 до 8, приведен в табл. 7.15. При большой длине кодового ограничения результаты отличаются еще сильнее.

Катастрофическая ошибка возникает, когда конечное количество ошибок в кодовых символах вызывает бесконечное количество битовых ошибок в декодированных данных.

Месси (Massey) и Сейн (Sain) указали необходимые и достаточные условия для сверточного кода, при которых возможно распространение катастрофических ошибок. Условием распространения катастрофических ошибок для кода со степенью кодирования $1/2$, реализованного на полиномиальных генераторах, будет наличие у генераторов общего полиномиального множителя (степени, не меньшей единицы).

Таблица 7.15

Длина кодового ограничения	Просвет систематического кода	Просвет несистематического кода
2	3	3
3	4	5
4	4	6
5	5	7
6	6	8
7	6	10
8	7	10

Например, на рис. 7.39, *а* изображен кодер $(K=3$, степень кодирования $1/2)$ со старшим полиномом $g_1(X)$ и младшим $g_2(X)$:

$$g_1(X)=1+X,$$
$$g_2(X)=1+X^2. \qquad (7.64)$$

Генераторы $g_1(X)$ и $g_2(X)$ имеют общий полиномиальный множитель $1+X$, поскольку $1+X^2=(1+X)(1+X)$.

Итак, в кодере, показанном на рис. 7.39, *а*, может происходить *накопление катастрофической ошибки*.

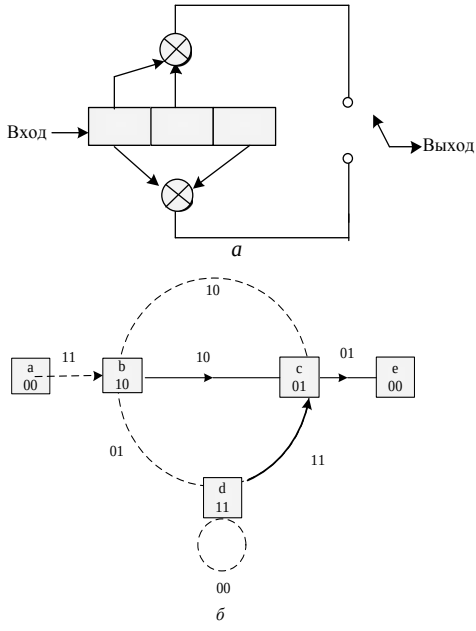


Рис. 7.39. Кодер, в котором возможно накопление катастрофической ошибки:
а - кодер; *б* - диаграмма состояния

Катастрофическая ошибка может появиться *тогда и только тогда, когда любая петля пути на диаграмме состояний кода произвольной степени кодирования имеет нулевой весовой коэффициент (нулевое расстояние к нулевому пути).*

Пример 7.30. Чтобы проиллюстрировать это, рассмотрим пример, приведенный на рис. 7.39. На диаграмме (рис. 7.39, *б)* узел состояния $a = 00$ разбит на два узла *а* и *е*, как и раньше. Предположим, что нулевой путь является правильным, тогда неправильный путь $abdd...dce$ имеет точно 6 единиц независимо от того, сколько раз мы обойдем вокруг петли в узле *d*. Поэтому, например, для канала BSC к выбору этого неправильного пути могут привести три канальные ошибки. На таком пути может появиться значительно большее количество ошибок (две плюс количество раз обхода петли). Для кодов со степенью кодирования $1/n$ можно увидеть, что когда каждый сумматор в кодере имеет парное количество соединений, петли, которые соответствуют информационным состояниям со всеми единицами, будут иметь нулевой вес, а соответственно, *код будет катастрофическим.*

Единственное преимущество описанного прежде систематического кода заключается в том, что он никогда не будет катастрофическим, поскольку каждая петля должна содержать, по крайней мере, одну ветку, порождаемую

ненулевым входным битом. Итак, каждая петля должна содержать ненулевой кодовый символ. Тем не менее, можно показать, что только небольшая часть несистематических кодов (кроме того, в котором все сумматоры имеют парное количество соединений) является катастрофической.

Граница рабочих характеристик сверточных кодов. Можно показать, что при когерентной модуляции в канале с адаптивным белым гауссовым шумом вероятность битовой ошибки

$$P_B \le Q\left(\sqrt{2d_f \frac{E_c}{N_0}}\right)\exp\left(d_f \frac{E_c}{N_0}\right)\left.\frac{dT(D,N)}{dN}\right|_{N=1,D=\exp(-E_c/N_0)}, \qquad (7.65)$$

где E_c / N_0 - отношение энергии канального символа к спектральной плотности мощности шума, $E_c / N_0 = rE_b / N_0$ ($r = k / n$ - степень кодирования; E_b / N_0 - отношение энергии информационного бита к спектральной плотности мощности шума). Итак, для кода со степенью кодирования $1/2$ и просветом $d_f = 5$ в случае использования когерентной схемы BPSK и твердой схемы принятия решений при декодировании можем записать

$$P_B \le Q\left(\sqrt{\frac{5E_b}{N_0}}\right)\exp\left(\frac{5E_b}{2N_0}\right)\frac{\exp(-5E_b/2N_0)}{\left[1-2\exp\left(-E_b/2N_0\right)\right]^2} \le \frac{Q(\sqrt{5E_b/2N_0})}{\left[1-2\exp\left(-E_b/2N_0\right)\right]^2}. \qquad (7.66)$$

Эффективность сверточного кодирования *определяется как уменьшение* (*выраженное в децибелах*) *отношения* E_b / N_0, *необходимого для достижения определенной вероятности появления ошибок в кодированной системе сравнительно с некодированной системой с той самой модуляцией и характеристиками канала.*

Верхние границы эффективности кодирования приведены в табл. 7.16.

Таблица 7.16

Коды со степенью кодирования $1/2$			Коды со степенью кодирования $1/2$		
K	d_f	Верхняя граница, дБ	*K*	d_f	Верхняя граница, дБ
3	5	3,97	3	8	4,26
4	6	4,76	4	10	5,23
5	7	5,43	5	12	6,02
6	8	6,00	6	13	6.37
7	10	6,99	7	15	6,99
8	10	6,99	8	16	7,27
9	12	7,78	9	18	7,78

Они сравниваются с некодированным сигналом с когерентной модуляцией для нескольких значений минимальных просветов сверточного кода.

Длина кодового ограничения в гауссовом канале с твердой схемой принятия решений при декодировании изменяется от 3 до 9. В таблице отображен тот факт, что даже при использовании простого сверточного кода можно достичь значительной эффективности кодирования. Реальная эффективность кодирования будет изменяться в зависимости от необходимой вероятности появления битовых ошибок.

Оценки эффективности кодов, которые сравниваются с некодированным сигналом с когерентной модуляцией, реализованной аппаратным путем или путем моделирования на компьютере, в гауссовом канале с мягкой схемой принятия решений при декодировании приведены в табл. 7.17.

Таблица 7.17

Некодированное значение E_b / N_0, дБ	Степень кодирования	$\frac{1}{3}$			$\frac{1}{2}$		$\frac{2}{3}$		$\frac{3}{4}$	
	P_b	7	8	5	6	7	6	8	6	9
6,8	10^{-3}	4,2	4,4	3,3	3,5	3,8	2,9	3,1	2,6	2,6
9,6	10^{-5}	5,7	5,9	4,3	4,6	5,1	4,2	4,6	3,6	4,2
11,3	10^{-7}	6,2	6,5	4,9	5,3	5,8	4,7	5,2	3,9	4,8

Некодированное значение E_b / N_0 представлено в крайнем левом столбце. Из табл. 7.17 можно видеть, что эффективность кодирования возрастает с уменьшением вероятности появления битовой ошибки. Однако эффективность кодирования не может возрастать бесконечно. Как вытекает из табл. 7.17, она имеет верхнюю границу. Эту границу (в децибелах) можно выразить как эффективность кодирования

$$\leq 10 \lg (rd_f). \qquad (7.67)$$

где r - степень кодирования, а d_f - просвет.

Основные выводы

Форматирование источника сообщений - это процесс аналогово-цифрового преобразования информационного сигнала источника сообщения в цифровой сигнал с помощью дискретизации и квантования сигнала и его представления в двоичной системе исчисления.

Кодирование источника сообщений проводится с целью обеспечения компактного представления данных, сокращение объема информации, кото-

рая вырабатывается источником, и для повышения скорости ее передачи или сокращения полосы частот.

Существуют два типа систем сжатия данных:

системы сжатия без потерь информации (неразрушительное сжатие);

системы сжатия с потерями информации (разрушительное сжатие).

Выбор системы неразрушительного или разрушительного сжатия зависит от типа данных, которые подлежат сжатию.

Передача данных по каналам связи и их хранение всегда происходит при наличии помех. Поэтому принятое (воспроизведенное) сообщение всегда определенной мерой отличается от переданного, т.е. на практике невозможна абсолютно точная передача при наличии помех в канале связи (в системе хранения).

Источники данных имеют ограниченный динамический диапазон и вырабатывают начальные сообщения с определенным уровнем искажений и ошибок. Этот уровень может быть большим или меньшим, но абсолютной точности воспроизведения достичь невозможно.

Критерием качества кода относительно кодирования источника сообщения является средняя длина кодовых слов.

Арифметическое кодирование нуждается в большой (в пределе бесконечной) точности вычислений, которая приводит к недопустимо высокой сложной реализации.

Простым и широко используемым для сжатия изображений и звуковых сигналов методом неразрушительного кодирования является метод дифференциального кодирования и кодирование длины повторений.

Задача помехоустойчивого корректирующего кодирования - обеспечение целостности информационных объектов с применением помехоустойчивых корректирующих кодов.

Помехоустойчивым корректирующим кодированием называется такой вид кодирования, который дает возможность реализовать программные, аппаратные или программно-аппаратные средства выявления и устранение искажений в информационных сообщениях.

Сверточным кодированием называется алгоритм кодирования, согласно которому кодер зависит не только от информационных символов в данный момент, а и от предыдущих символов на его входе или выходе.

Оптимальным есть такой декодер, который работает по принципу максимального правдоподобия и минимизирует вероятность ошибки в информационном сообщении при условии, что все переданные последовательности равновероятные.

Сверточные (рекурентные, цепи) коды используются для кодирования непрерывной последовательности двоичных символов путем введения в эту последовательность специальных проверочных (избыточных) символов с целью выявления и исправление искажений в информационных сообщениях.

Вопросы для самоконтроля

1. Раскройте содержание разрушительного и неразрушительного сжатия информации.

2. В каком случае код является однозначно декодированным?

3. Раскройте понятие «префиксности кода».

4. Выведите необходимое и достаточное условия префиксности.

5. Определите алгоритм построения кодового дерева Хаффмана.

6. В чем заключается кодирование последовательностей сообщений?

7. Раскройте алгоритм декодирования арифметического кода.

8. Раскройте понятие «код с конечной памятью».

9. На какие группы можно поделить все словарные методы кодирования?

10. В каких случаях используется метод дифференциального кодирования?

11. Объясните основные способы (механизмы) обеспечения целостности (и в определенном значении - доступности) информации в условиях естественных действий.

12. Объясните особенности способов обеспечения целостности информационных объектов с применением помехоустойчивых корректирующих кодов.

13. Объясните принципы построения помехоустойчивых кодов.

14. Что такое относительная скорость и избыточность кода?

15. Объясните принципы построения двоичных кодов с проверкой на парность или непарность (контроль по модулю 2).

16. Объясните принцип построения кодов Хэмминга. Какие их возможности относительно обнаружения и исправления ошибок?

17. Объясните алгоритмы кодирования и декодирование с применением кодов Хэмминга.

18. Построение циклических кодов. Чем определяется их корректирующая способность?

19. Объясните алгоритмы кодирования и декодирования с применением циклических кодов.

20. Объясните принципы построения сверточных кодов. Какие их возможности относительно обнаружения и исправления искажений?

21. Объясните алгоритмы кодирования и декодирования с применением сверточных кодов.

22. Коды Рида - Соломона. Принципы построения и возможности.

The main conclusions

Formatting of a source of messages is a process of analog-to-digital conversion of an information signal of a source of messages in a digital signal by means of digitization and quantization of a signal and its representation in binary system of calculation.

The coding of a source of messages is conducted with the purpose of support of compact data representation, reduction of information volume that is made by a source and for rise of its transfer rate or reduction of a band.

There are two types of systems of data compression: systems of compression without losses of the information (nondestructive compression); systems of compression with losses of the information (destructive compression).

The choice of system of nondestructive or destructive compression depends on a type of data which are subjected to compression.

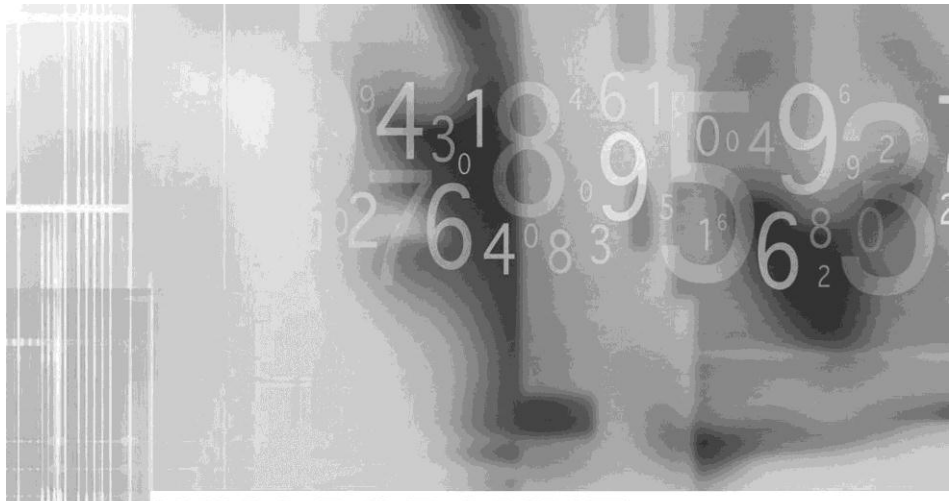
The transmission of data on communication channels and their storage are always conducted at presence of interferences of different type. Therefore the accepted (reproduced) message always differs from transferred to certain extent, that is in practice absolutely exact transmission at presence of interferences in a communication channel(in system of storage) is impossible.

The optimal decoder is a decoder that works on a principle of maximal plausibility and it minimizes probability of an error in an information message on condition that all transmitted sequences are equally possible.

Folding (recurrent, chain) codes are used for the coding of continuous sequence of binary symbols by bringing of special checking (redundant) symbols in this sequence with the purpose of revealing and correction of distortions in information messages.

Ключевые слова

Русский	**Английский**
кодирование источника сообщений	encoding of source of reports
коды с памятью	codes with memory
коды без памяти	codes without memory
помехоустойчивый корректирующий код	antigambling correcting code
сверточный двоичный код	binary displacing code
корректирующая способность кода	correcting ability of code
избыточность кода	code surplus
циклический код	cyclic redundancy check



ШИФРОВАНИЕ И ДЕШИФРОВАНИЕ ИНФОРМАЦИИ

8

8.1. Модели, задачи и системы шифрования
8.2. Шифрование в каналах связи
8.3. Алгоритмы и системы симметричного и асимметричного шифрования (криптографической защиты информации)
8.4. Электронная цифровая подпись
8.5. Стеганографические методы защиты информации
8.6. Методы криптоанализа

8.1. Модели и системы шифрования

Желание общаться конфиденциально было присущим человеку издавна. История секретного общения богата уникальными изобретениями, часть из которых мы рассмотрим дальше. Термины *шифрование* и *кодирование* означают преобразование сообщений, которые выполняются передатчиком, а термины *дешифрации* и *декодирования* — соответствующие обратные преобразования, которые осуществляются получателем информации.

Греческое слово «криптография» происходит от слов «kryptos» (тайный, спрятанный) и «graphy» (запись) и *охватывает разработку методов и средств обеспечения безопасности передачи информационных сообщений.*

Зарождение криптографии началось в глубокой давности, из которой до нас дошло немало систем шифрования, скорее всего появившихся одновременно с письменностью в IV тысячелетии до н.э. Методы секретного переписывания были независимо найдены во многих древних обществах (Египет, Шумер, Китай). По свидетельству Геродота, в Древнем Египте роль шифра как правило выполнял специально созданный жрецами язык, в котором параллельно существовали три алфавита: письменный, священный и таинственный (последний применялся пророками для скрытия содержания сообщений).

С распространением письменности криптография стала формироваться как самостоятельная наука. Первые криптосистемы появляются уже в начале нашей эры. Так, Цезарь в своей переписке использовал систематический шифр, которому было присвоено его имя.

Разработка и внедрение методов и средств безопасности передачи информационных сообщений на основе криптографических преобразований является задачей криптографов. Криптоаналитики нарушают эту безопасность, используя криптоанализ, сориентированный на взлом шифротекстов (шифров), которые представляют собой данные, поданные в зашифрованной форме (со скрытым семантическим содержанием), то есть образованные после шифрования (криптографического преобразования) открытого текста (с нескрываемым семантическим содержанием).

До появления информационных технологий криптография состояла из алгоритмов на символьной основе. Разные криптографические алгоритмы либо заменяли одни символы другими, либо переставляли символы.

Стремительное развитие криптографические системы приобрели в годы первой и второй мировых войн. Появление вычислительных средств в 1950-х годах ускорило разработку и усовершенствование методов шифрования.

Современная криптография базируется на последних достижениях математики, ряда фундаментальных физических и инженерных дисциплин в сочетании с новейшими информационными технологиями.

Основными направлениями использования криптографических методов стали передача конфиденциальной информации в каналах связи, установление подлинности переданных сообщений, хранение информации (документов, баз данных) на носителях в зашифрованном (закрытом) виде.

Появление новых мощных вычислительных систем, новейших сетевых технологий и нейронных моделей постоянно побуждает к созданию новых криптосистем и внедрению тщательного анализа и совершенствованию уже известных методов.

В основу каждой закрытой информационной системы положено использование алгоритмов шифрования как основного средства сохранения конфиденциальности информации.

Шифрование *- процесс преобразования алфавита сообщения* $(i = 1, 2, \ldots, K)$ $A\{\lambda_i\}$ $(i = 1, 2, \ldots, K)$ *в алфавит выбранных кодовых символов* $R\{x_j\}$ $(j = 1, 2, \ldots, N)$ *на основе строго определенного алгоритма или математической функции с целью обеспечения конфиденциальности, то есть предотвращение исключения полезной информации из канала передачи данных несанкционированным пользователем (рис. 8.1).*

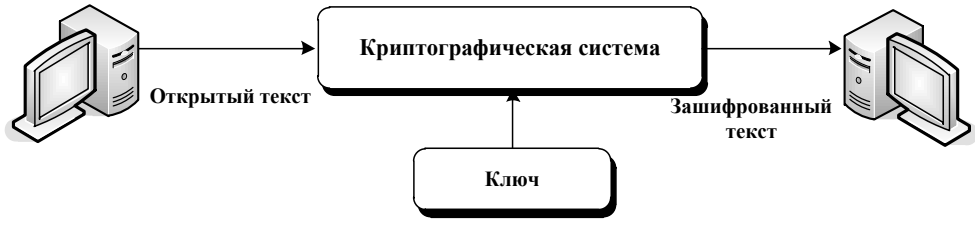


Рис. 8.1. Структурная схема процесса шифрования

Криптографическим алгоритмом называют алгоритм или математическую функцию, которая используется для шифрования и расшифровывания открытого текста из алфавита сообщения $A\{\lambda_i\}$

Алфавит сообщения $A\{\lambda_i\}$ $(i = 1, 2, \ldots, K)$ *- конечное множественное число знаков, используемых для шифрования информации.*

Текст (***открыт текст***) *- упорядоченный набор знаков* M *из элементов алфавита* $A\{\lambda_i\}$, *который содержит полезную информацию.*

Под ***шифровкой в широком смысле*** *понимают процесс преобразования исходного текста, который называют также открытым текстом, в шифрованный текст.*

Расшифровывание *- процесс, обратный шифрованию, то есть восстановление открытого текста из алфавита сообщения* $A\{\lambda_i\}$ $(i = 1, 2, \ldots, K)$ *на основе определенного ключа (алгоритма или обратной математиче*

ской функции) с целью получения полезной информации из шифрованного текста (рис. 8.2).

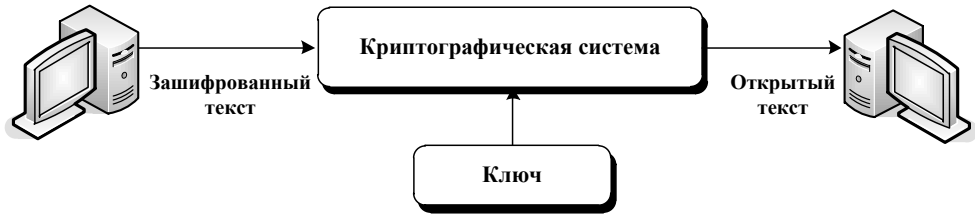


Рис. 8.2. Структурная схема процесса расшифровывания

Ключ *- информация (обратная математическая функция или алгоритм), необходимая для беспрепятственного шифрования и расшифровывания текстов.*

Пространство ключей (K) *- это набор возможных значений ключа, который может представлять собой алгоритм, обратную математическую функцию или последовательность символов алфавита.*

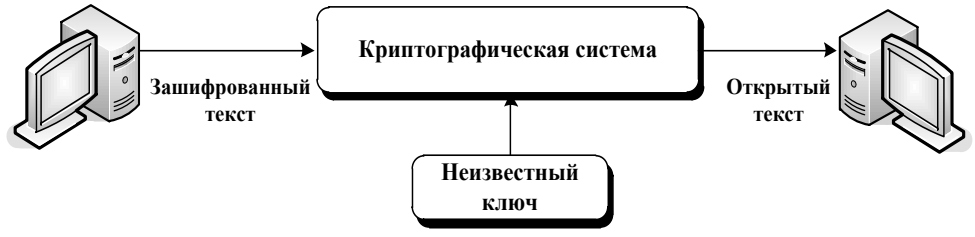


Рис. 8.3. Структурная схема процесса дешифрации

Дешифрация *- процесс, обратный к шифрованию, то есть восстановление открытого текста из алфавита сообщения* $A\{\lambda_i\}$ $(i = 1, 2, \ldots, K)$ *без знания ключа (алгоритма или обратной математической функции) с целью несанкционированного исключения необходимой информации из зашифрованного текста (рис. 8.3).*

Криптостойкость *- временная характеристика шифра, который определяет его стойкость к дешифрации без знания ключа.*

Процесс криптографического закрытия данных может осуществляться как программно, так и аппаратно. Аппаратная реализация имеет существенно большую стоимость, хотя ей присущи и важные преимущества: высокая производительность, простота, защищенность. Программная реализация наиболее практична благодаря гибкости в использовании.

В общей системе шифрования открытый текст обозначается буквой M. Это может быть поток битов, текстовый файл, битовое изображение, цифровое видео и т.п.

Обозначим шифротекст как C (ciphertext), функцию шифрования — как E. Математическая модель шифрования будет иметь вид $E(M) = C$, а функция расшифровывания $D(C) = M$.

Поскольку содержанием шифрования и расшифровывания сообщения является восстановление начального открытого текста, должно выполняться равенство:

$$D\big(E(M)\big) = M.$$

Криптографическая система шифрования (***криптосистема***) — это *совокупность методов и средств криптографической защиты информации, использование которых обеспечивает надлежащий уровень защищенности информации, которая обрабатывается, сохраняется или передается по информационно-коммуникационным каналам связи.*

Алгоритмы и системы шифрования можно классифицировать (рис. 8.4).

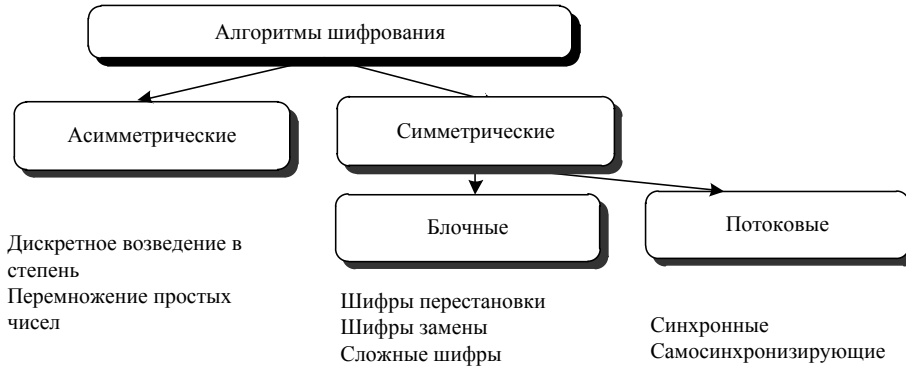


Рис. 8.4. Классификация криптографических алгоритмов

Симметричной криптосистемой *называется система, которая осуществляет процесс шифрования и расшифровывания полезного сообщения (открытого текста) на основе использования одного ключа (рис. 8.5).*

Поточными алгоритмами (***поточными шифрами***) *называются алгоритмы с последовательной (битной или байтной) обработкой открытых текстов.*

Блочными алгоритмами (***блочными шифрами***) *называются алгоритмы, которые работают с группами последовательностей (битов, байтов), которые образуют открытый текст, или с группами текстов.*

Для алгоритмов, которые используются в компьютерных модемах, типичный размер блока составляет 64 бита. (До появления компьютеров алгоритмы обычно обрабатывали открытый текст посимвольно. Такой вариант может рассматриваться как поточный алгоритм, который обрабатывает поток символов.)

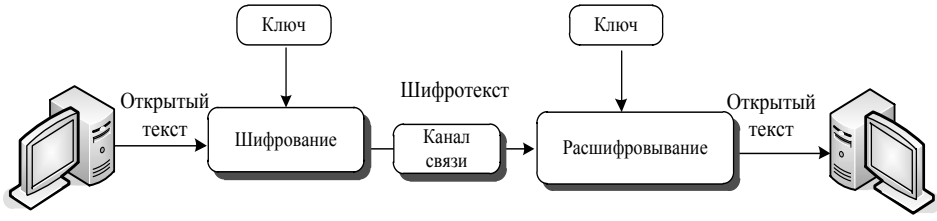


Рис. 8.5. Шифрование и дешифрование с ключом

Шифрование и расшифровывание с использованием симметричного алгоритма описывается как

$$E_K(M) = C,$$
$$D_K(C) = M.$$

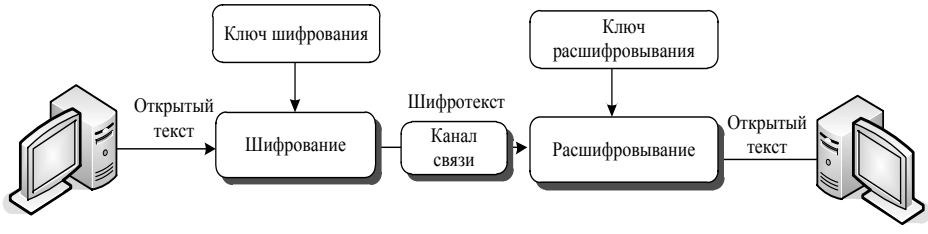


Рис. 8.6. Шифрование и дешифрование с двумя разными ключами

Асимметричной криптосистемой *называется система, которая осуществляет шифрование и расшифровывание полезного сообщения* (*открытого текста) с использованием двух ключей: открытого и закрытого (секретного*), *математически связанных друг с другом* (рис. 8.6).

Информация шифруется с помощью открытого ключа, доступного всем, а расшифровывается с помощью закрытого ключа, известного только получателю сообщения.Более того, ключ расшифровывания нельзя раскрыть ключом шифрования.

Шифрование с открытым ключом E обозначается как $E_K(M) = C.$

Хотя открытый и закрытый ключи разные, расшифровывание с соответствующим закрытым ключом обозначается $D_K(C) = M.$

Иногда сообщения шифруются закрытым ключом, а расшифровываются открытым, который используется для цифровой подписи.

Требования к современным криптографическим системам защиты информации:

зашифрованное сообщение должно поддаваться чтению только при наличии ключа;

количество операций, необходимых для определения использованного ключа шифрования за фрагментом шифрованного сообщения и соответствующего ему открытого текста, должно быть не меньше от общего количества возможных ключей;

количество операций, необходимых для дешифрации информации путем перебора любых ключей, должно выходить за пределы возможностей современных вычислительных систем (с учетом возможности использования сетевых вычислений) или нуждаться в высоких расходах на эти вычисления;

знание алгоритма шифрования не должно влиять на надежность защиты;

незначительное изменение ключа должно приводить к существенному изменению вида зашифрованного сообщения даже при шифровании того же открытого текста;

незначительное изменение открытого текста должно приводить к существенному изменению вида зашифрованного сообщения даже при использовании того же ключа;

структурные элементы алгоритма шифрования должны быть неизменными;

дополнительные биты, что вводятся в сообщение в процессе шифрования, должны быть полно и надежно спрятанные в шифрованном тексте;

длина шифрованного текста должна не превышать длину исходного текста;

не должно быть простых и легко устанавливаемых зависимостей между ключами, последовательно используемыми в процессе шифрования;

любой ключ из множественного числа возможных значений должен обеспечивать надежную защиту информации;

алгоритм должен допускать как программную, так и аппаратную реализацию, при этом изменение длины ключа не должно привести к качественному ухудшению алгоритма шифрования.

Хорст Фейстель (Horst Feistel, 1915—1990), родился в Берлине, а в 1934 году переехал в США. Закончил Гарвардский университет со степенью магистра в области физики. На время работы в компании IBM приходятся наиболее известные работы Фейстеля в области криптографии. Фейстель одним из первых ученых (которые не работали на заказ правительства) начал изучать теорию блочных шифров. Наибольшую популярность ему принес общий метод создания алгоритмов шифрования, который получил его имя. Работы Фейстеля заложили основу для шифров Luciferi Data Encryption Standards (DES).

8.2. Шифрование в каналах связи

Одной из важных характеристик любой информационной сети есть ее деление на так называемые уровни модели ISO/OSI, каждая из которых отвечает за соблюдение определенных условий и выполнение функций. Такое деление на уровни имеет фундаментальное значение для создания стандартных информационно-коммуникационных сетей (см. главу 7).

В теории шифрования данных передача по каналам связи информационно-коммуникационной сети может осуществляться на любом уровне модели OSI. На практике это обычно делается или на самых низких, или на наивысших уровнях. Если данные шифруются на нижних уровнях, шифрование называется *канальным.* Если шифрование данных выполняется на верхних уровнях, то оно называется *сквозным.* Оба этих подхода к шифрованию данных имеют свои преимущества и недостатки.

Канальное шифрование. При канальном шифровании шифруются абсолютно все данные, которые проходят через каждый канал связи, включая открытый текст сообщения, а также информацию о его маршруте трансляции и об используемом коммуникационном протоколе (рис. 8.7). Однако в этом случае любой интеллектуальный сетевой узел будет вынужден расшифровывать входной поток данных, чтобы соответствующим образом его обработать и опять зашифровать для передачи на другой узел сети.

Рис. 8.7 Канальное шифрование

Канальное шифрование является эффективным средством защиты информации в информационно-коммуникационных сетях. Поскольку шифрованию подвергаются все данные, которые передаются от одного узла сети к другому, у криптоаналитика нет никакой дополнительной информации о том, что служит источником переданных данных, их назначение, какая их структура и т.п. А если еще позаботиться и о том, чтобы, пока канал простаивает, передавать по нему случайную битовую последовательность, посторонний наблюдатель не сможет даже сказать, где начинается и где заканчивается текст переданного сообщения. Не слишком сложной является и работа с ключами. Одинаковыми ключами стоит обеспечивать только два соседних узла сети связи, которые потом могут изменять используемые ключи независимо от других пар узлов.

Наибольший недостаток канального шифрования связан с тем, что данные придется шифровать при передаче по каждому физическому каналу компьютерной сети. Отправление информации в незашифрованном виде по ка-

кому-то из каналов ставят под угрозу обеспечение безопасности всей сети в целом. В итоге стоимость реализации канального шифрования в больших сетях может оказаться излишне большой. Кроме того, при использовании канальной шифрования дополнительно нужно защищать каждый узел информационной сети, через который проходят переданные по сети данные. Если абоненты сети полностью доверяют друг другу и каждому ее узлу, размещенному в защищенном от проникновения злоумышленников месте, на этот недостаток канальной шифрования можно не обращать внимания. Однако на практике такая ситуация случается редко.

Сквозное шифрование. При сквозном шифровании криптографический алгоритм реализуется на одном из верхних уровней модели OSI. Шифрованию подлежит только содержательная часть сообщения, которое нужно передать по сети. После зашифровывания к ней добавляется служебная информация, необходимая для маршрутизации сообщения, и результат направляется на низшие уровни с целью отправления адресату. Теперь сообщения не нужно постоянно расшифровывать и зашифровывать при прохождении через каждый промежуточный узел сети связи. Сообщение остается зашифрованным на всем пути от отправителя к получателю (рис. 8.8).

Рис. 8.8. Сквозное шифрование

Основная проблема, с которой сталкиваются пользователи информационно коммуникационных сетей, где применяется сквозное шифрование, связано с тем, что служебная информация, используемая для маршрутизации сообщений, передается по сети в незашифрованном виде. Опытный криптоаналитик может добыть для себя полезную информацию, зная кто, с кем, как долго и в какие часы общается через компьютерную сеть. Для этого ему даже не нужно быть в курсе предмета общения.

Сквозное шифрование, в отличие от канального, характеризуется более сложной работой с ключами, поскольку каждая пара пользователей компьютерной сети должна быть обеспеченной одинаковыми ключами, прежде чем они смогут связаться друг с другом. К тому же, поскольку криптографический алгоритм реализуется на верхних уровнях модели OSI, приходится также иметь дело со многими существенными разногласиями в коммуникационных протоколах и интерфейсах в зависимости от типов компьютерных сетей и подсоединенных к сети компьютеров. Все это осложняет практическое применение сквозного шифрования.

Комбинированное шифрование. Комбинирование канального и сквозного шифрования данных в информационной сети обходится значительно

дороже, чем канальное или сквозное шифрование отдельно. Однако именно такой подход дает возможность как можно лучше защитить данные, переданные по сети. Шифрование в каждом канале связи не позволяет злоумышленнику анализировать служебную информацию, используемую для маршрутизации. А сквозное шифрование уменьшает вероятность доступа к незашифрованным данным в узлах сети.

При комбинированном шифровании робота с ключами происходит раздельно: сетевые администраторы отвечают за ключи, которые используются при канальном шифровании, а о ключах, которые применяются при сквозном шифровании, заботятся сами пользователи.

Аппаратное шифрование. Большинство средств криптографической защиты данных реализовано в виде специализированных аппаратных устройств. Эти устройства встраиваются в линию связи и осуществляют шифрование всей переданной по ней информации. Преимущество аппаратного шифрования над программным предопределяется несколькими причинами. Аппаратное шифрование имеет большую скорость. Криптографические алгоритмы состоят из огромного количества сложных операций, что проделываются над битами открытого текста. Современные универсальные компьютеры плохо приспособлены для эффективного выполнения этих операций. Специализированное оборудование дает возможность выполнять их намного быстрее. Аппаратуру более легко физически защитить от проникновения извне. Программа, которая выполняется на персональном компьютере, практически беззащитна. Вооружившись отладчиком, злоумышленник может тайком внести в нее изменения, чтобы снизить стойкость используемого криптографического алгоритма, и никто ничего не заметит. Что же касается аппаратуры, то она обычно размещается в специальных контейнерах, которые делают невозможным изменение схемы ее функционирования. Чип покрывается сверху специальным химическим составом, и в результате любая попытка взломать защитный слой этого чипа приводит к самоуничтожению его внутренней логической структуры. И хотя иногда электромагнитное излучение может быть хорошим источником информации о том, что происходит внутри микросхемы, от этого излучения легко избавиться, экранировав микросхему. Аналогично можно экранировать и компьютер, хотя сделать это намного сложнее, чем в случае миниатюрной микросхемы.

Программное шифрование. Любой криптографический алгоритм можно реализовать в виде соответствующей программы. Преимущества такой реализации очевидны: программные средства шифрования легко копируются, они простые в использовании, их нетрудно модифицировать в соответствии с конкретными потребностями.

Во всех распространенных операционных системах есть встроенные средства шифрования файлов. Обычно они предназначены для шифрования отдельных файлов, и работа с ключами полностью положена на пользователя. Поэтому применение этих средств нуждается в особенном внимании:

во-первых, в любом случае нельзя хранить ключи на диске вместе с зашифрованными с их помощью файлами, а во-вторых, незашифрованные копии файлов необходимо стереть сразу же после шифрования.

Обычно злоумышленник может добраться до компьютера и незаметно внести нежелательные изменения в программу шифрования. Однако основная проблема не в этом. Если злоумышленник сможет проникнуть в помещение, где установлен компьютер, он вряд ли будет возиться с программой, а просто установит скрытую камеру в стене, подслушивающее устройство в телефон или датчик для ретрансляции электромагнитного излучения в компьютер.

Сжатие и шифрование. Алгоритмы сжатие данных весьма пригодные для общего использования с криптографическими алгоритмами. На это есть две причины. При раскрытии шифра криптоаналитик больше всего полагается на избыточность, присущую любому открытому тексту. Сжатие помогает лишиться этой избыточности.

Шифрование данных является достаточно трудоемкой операцией. При сжатии уменьшается длина открытого текста, и тем самым сокращается время, которое будет потрачено на его шифровку. Нужно только не забыть сжать файл, прежде чем он будет зашифрован, а не после того.

После шифрования файла с помощью высококачественного криптографического алгоритма полученный шифротекст сжать не удастся, поскольку его характеристики будут близкими к характеристикам совсем случайного набора букв. Кстати, сжатие может служить своеобразным тестом для проверки качества криптографического алгоритма: если шифротекст поддается сжатию, этот алгоритм есть смысл заменить на лучший.

Для обеспечения надежной защиты передачи информации стоит воспользоваться автономным аппаратным шифрованием. В отличие от программного средства защиты, такое устройство обойти по большей части невозможно.

Тахер Ель-Гамаль (Taher ElGamal, 1955), криптограф, потомок египтян, которые эмигрировали к США. В 1985 году опубликовал статью "Криптосистемы с открытым ключом и схема подписи, которая основывается на дискретных логарифмах". Схема подписи Ель-Гамаля стала основанием для алгоритма цифровой подписи DSA. Он также принимал участие в разработке протоколов оплаты кредитной карточкой и интернет-схем оплаты. Ель-Гамаль был директором по разработкам компании RSA Security, прежде чем основал и возглавил в 1998 году компанию Security.

8.3. Алгоритмы и системы симметричного и асимметричного шифрования (криптографической защиты информации)

Рассмотрим методы шифрования, которые чаще всего используются на практике. Речь идет о блочных шифрах. Блочные алгоритмы шифрования являются основным средством криптографической защиты информации, которая сохраняется или передается по общедоступной сети.

Преимущества практического приложения:

возможность эффективной программной реализации на современных аппаратно-программных средствах;

высокая скорость шифрования/расшифровывания как при аппаратной, так и при программной реализации;

высокая гарантированная стойкость; при этом стойкость алгоритма блочной шифрования может быть доказана с помощью математического аппарата.

Входная последовательность блочных алгоритмов шифрования разбивается на участки определенной длины, преобразования в алгоритме блочного шифрования происходят над каждым блоком отдельно. Соответственно исходная последовательность алгоритма блочного шифрования состоит из блоков, длина которых равняется длине входных блоков. В случае, когда длина открытого текста не кратна длине входных блоков в алгоритме шифрования, применяется операция *дополнения* последнего блока открытого текста к необходимой длине. Дополнение осуществляется приписыванием необходимого количества нулей или случайного набора символов. В общем случае содержание того, чем мы дополняем блок открытого текста, не играет никакой роли с точки зрения криптографической стойкости. На приемной стороне необходимо знать, какое количество символов было прибавлено, потому рядом с отмеченными дополнениями приписывается длина этих данных.

Вариантом практического применения блочных алгоритмов является использование их для обеспечения имитозащиты переданной по каналам связи информации.

Имитозащитой сообщения *называется процесс введения дополнительного блока (имитовставки) в конец разбитого на блоки информационного сообщения. Для внедрения имитовставки применяются алгоритмы блочного шифрования в режиме гамирования с самовосстановлением или шифрования со сцеплением блоков.*

Специфика организации разных типов защищенных каналов связи обусловила появление алгоритмов блочного шифрования (рис. 8.9):

метод простои замены*, или режим электронной кодовой книги (Electronic Codebook Mode — ECB);*

метод гаммирования*, или режим по модулю, который равняется мощности алфавита;*

метод гаммирования с самовосстановлением, *или гаммирования с обратной связью (Cipher-Feedback mode — CFB);*

метод многоалфавитной подстановки;

метод шифрования со сцеплением блоков (Cipher *Block Chaining mode — CBC)*;

метод гаммирования с обратной связью *на выходе (Output-Feedback mode — OFB).*

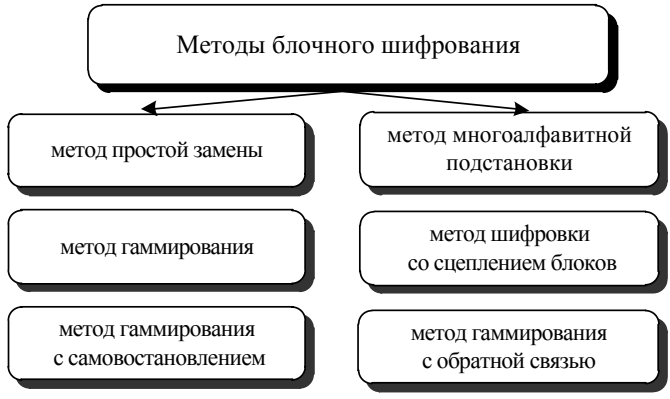


Рис.8.9. Методы алгоритмов блочного шифрования

Все разнообразие симметричных криптосистем основывается на описанных дальше базовых методах.

Метод многоалфавитной подстановки — это самый простой вид преобразований, который заключается в замене символов исходного текста на других (того же алфавита) по более-менее сложному правилу. В случае многоалфавитной подстановки каждый символ исходного текста за определенным законом превращается в символ шифрованного текста. При этом закон преобразования изменяется от символа к символу.

Для обеспечения высокой криптостойкости системы нужно использование ключей большого размера. К этому классу принадлежат криптосистемы с одноразовым ключом, которые имеют абсолютную теоретическую стойкость.

Метод перестановки — метод криптографического преобразования, которое заключается в перестановке символов исходного текста по некоторому правилу. Шифры перестановки не используются в чистом виде, поскольку их криптостойкость недостаточна.

Метод гаммирования— это преобразование исходного текста, при котором его символы добавляются (по модулю, который равняется мощности алфавита) к символам псевдослучайной последовательности, сгенерированной по некоторому правилу. Гаммирование нельзя полностью выделить в отдельный класс криптографических преобразований, поскольку отмеченная псевдослучайная последовательность может генерироваться, например, с по-

мощью блочного шифра. В случае, когда последовательность случайна (например, снятая из физического датчика) и каждый ее фрагмент используется только один раз, получаем криптосистему с одноразовым ключом.

Метод простои замены. В этом режиме блоки открытого текста шифруются независимо от других блоков на одном ключе (рис. 8.10). Этот режим назван *режимом электронной кодовой книги*, поскольку теоретически существует возможность создать книгу, в которой каждому блоку открытого текста будет отвечать блок зашифрованного текста.

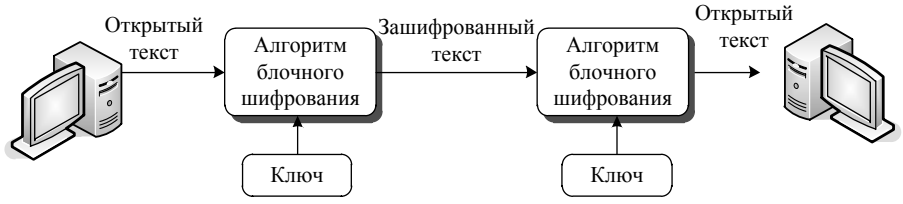


Рис. 8.10. Режим простой замены

Однако в случае, когда длина блока равняется 64 битам, книга содержит 264 записей, и каждая книга будет отвечать одному ключу.

Шифрование может быть описано зависимостью $C_i = F(P_i)$ для $i = 1, ..., N$, где C_i и P_i - блоки соответственно зашифрованного и открытого текста, а F - криптографическое преобразование, реализованное алгоритмом блочного шифрования.

Идентичные блоки открытого текста на том же ключе будут зашифрованы одинаково. С точки зрения криптоанализа этот режим является наиболее «слабым» (поскольку существует большое количество криптографических атак).

Метод усложненного гаммирования. В этом режиме алгоритм блочного шифрования используется для осложнения предыдущей гаммы, выработанной одноканальной линией задержки (рис. 8.11). Ошибка во время передачи всего сообщения приводит к искажению при расшифровывании только одного блока. Таким образом, в случае использования этой методики делается невозможным распространение ошибки за счет рассинхронизации узлов вычисления предыдущей гамы на передающей и приемной стороне.

Для предотвращения этого нежелательного явления на практике применяются устройства синхронизации работы шифраторов, если шифратор реализуется аппаратно. Начальное состояние узла создания исходной гаммы задается инициализирующим вектором (синхропосылка), который передается по открытым каналам связи в зашифрованном или открытом виде. Гамма, полученная узлом генерации предыдущей гами, обрабатывается согласно алгоритму блочного шифрования, после чего результирующая гамма подытоживается по модулю с блоком открытого текста.

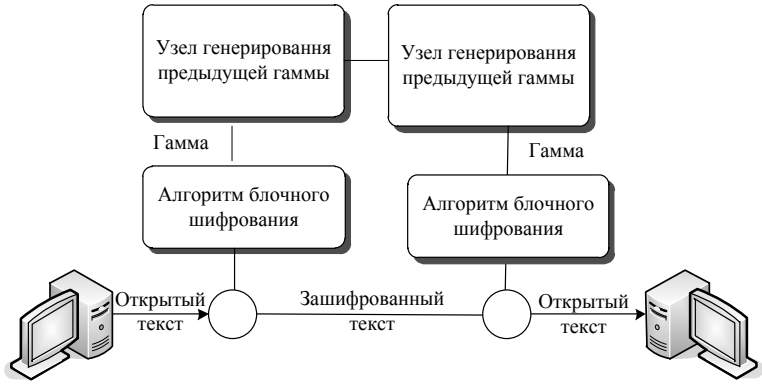


Рис. 8.11. Режим гаммирования

Шифрование можно представить как $C_i = P_i \oplus F(Y_i)$, $i = 1, \ldots, N$, где Y_j - созданная гамма; Y_t - синхропосылка.

Метод гаммирования с самовосстановлением. Этот режим характеризуется тем, что шифратор в этом случае имеет свойство самосинхронизации (рис. 8.12).

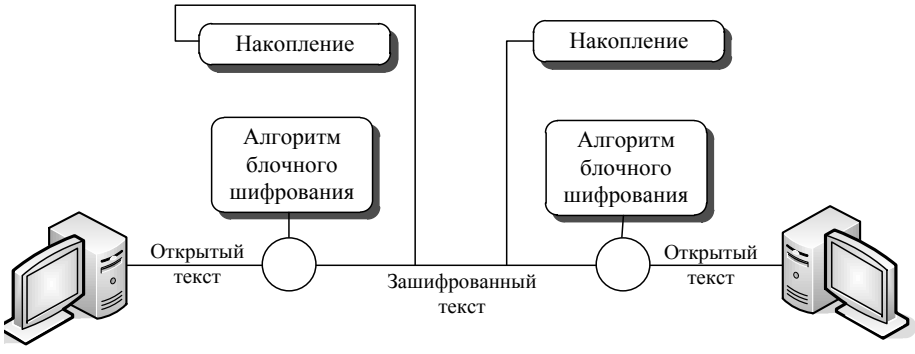


Рис. 8.12. Режим гаммирования с самовосстановлением

Начальное заполнение накопления, которое на практике обычно реализуется в виде регистра сдвига, есть синхропосылкой Y_t, которая передается по открытому каналу передачи данных. Математическое представление шифрования имеет вид $C_1 = P_1 \oplus F(Y)$, $C_i = P_i \oplus F(Y_{i-1})$. Соответственно расшифровывание имеет вид $P_1 = C_1 \oplus F(Y)$, $P_i = C_i \oplus F(C_{i-1})$.

Алгоритм блочного шифрования Файстеля (Стандарт DES). Алгоритм шифрования данных (DES — Data Encryption Standard) 1977 года был принят в США как федеральный. В стандарт входит описание блочного шифра типа

шифра Файстеля, а также разных режимов его работы как составляющих нескольких процедур криптографического преобразования данных.

Обычно под аббревиатурой DES понимают именно *блочный шифр,* который в стандарте отвечает процедуре шифрования в режиме электронной кодовой книги (ECB — Electronic Codebook Mode) (рис. 8.13).

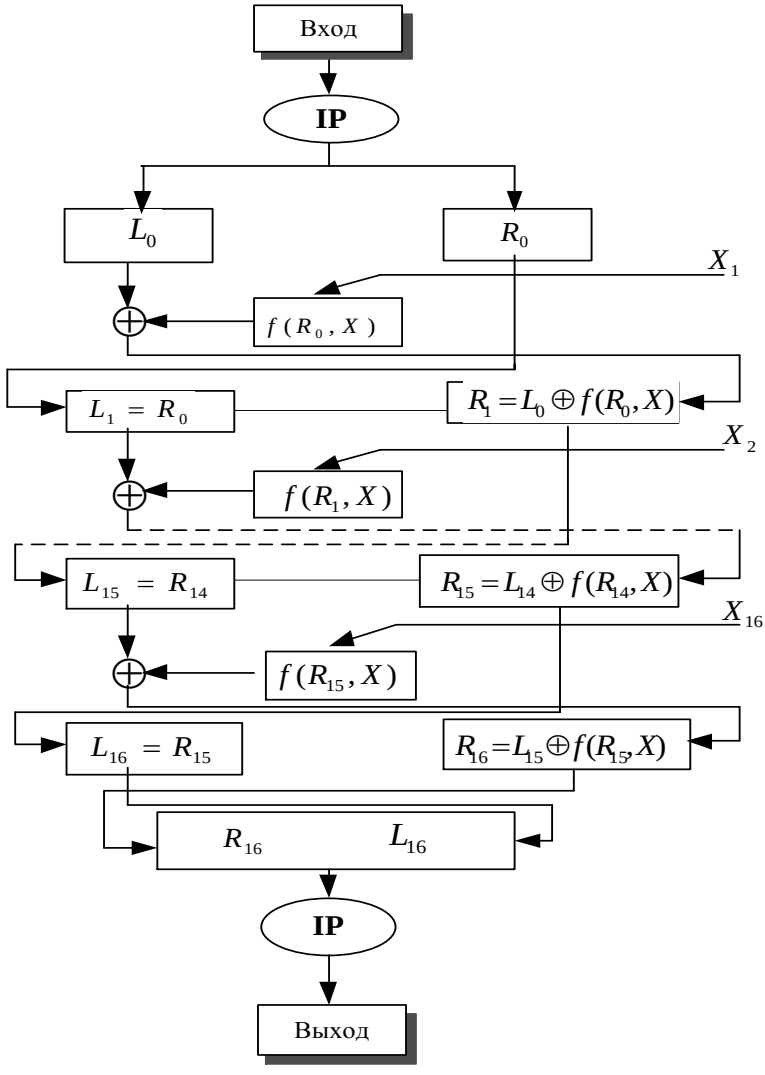


Рис. 8.13. Алгоритм DES

Название объясняется тем, что любой блочный шифр является простым шифром подстановки и подобный кодовой книге (табл. 8.1).

Таблица 8.1

58	50	42	34	26	18	10	2
60	52	44	36	28	20	12	4
62	54	46	38	30	22	14	6
64	56	48	40	32	24	16	8
57	49	41	33	25	17	9	1
59	51	43	35	27	19	11	3
61	53	45	37	29	21	13	5
63	55	47	39	31	23	15	7

Пример. *Процедура формирования подключей.* На каждом цикле (рис. 8.14) из ключа X длиной 56 бит формируется ключ X_i размером 48 бит.

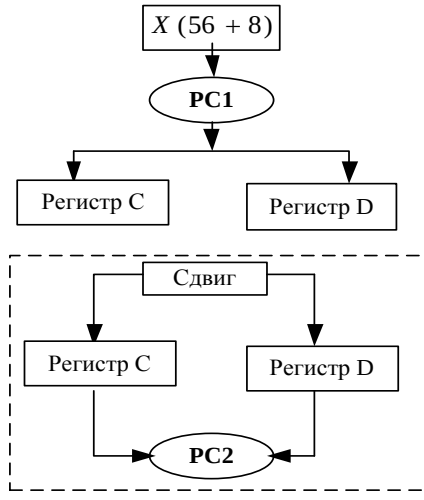


Рис.8.14. Алгоритм формирования ключей DES

Сам ключ X размещается в восьмибайтовом слове, причем восемь разрядов каждого байта являются контрольными и в ключ не входят. Перед шифрованием в соответствии с процедурой выбора $PC1$ (табл. 8.2) из X выбираются 56 бит, которыми заполняются два регистра (C и D) длиной 28 бит каждый.

В дальнейшем при входе в дежурный цикл с номером u регистры сдвигаются циклически влево. Размер сдвига зависит от номера цикла, но является фиксированным и предварительно известным.

Таблица 8.2

Заполнение *C*							**Заполнение *D***						
57	49	41	33	25	17	9	63	55	47	39	31	23	15
1	58	50	42	34	26	18	7	62	54	46	38	30	22
10	2	59	51	43	35	27	14	6	61	53	45	37	29
19	11	3	60	52	44	36	21	13	5	28	20	12	4

После сдвига оба подблока совмещаются в порядке (*C, D*). Дальше в соответствии с функцией выбора *Pc2 (*табл. 8.3) из них выбираются 48 бит подключа *Xu*.

Таблица 8.3

14	17	11	24	1	5	3	28
15	6	21	10	23	19	12	4
26	8	16	7	27	20	13	2
41	52	31	37	47	55	30	40
51	45	33	48	44	49	39	56
34	53	46	42	50	36	29	32

Шифрование и расшифровывание отличаются направлением сдвигов (табл. 8.4).

Таблица 8.4

Номер цикла	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Сдвиг влево (шифрование)	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	1
Сдвиг вправо (расшифровывание)	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	1

Выбор бит по табл. 8.2—8.4 из соответствующих блоков выполняется таким способом. Таблица рассматривается как последовательность ее строк, записанных друг за другом, начиная с первой строки. Биты блока данных нумеруются слева и справа, начиная с единицы. Каждая ячейка s таблицы рассматривается как номер бита b_s в блоке данных. Преобразование заключается в замене всех элементов s на биты.

Циклическая функция выполняет такие действия.

1. Расширение блока R_{i-1} до 48 бит за счет повторения битов блока с помощью функции расширения *EP* (табл. 8.5).

2. Поразрядное добавление результата к ключу.

3. Преобразование полученной суммы с помощью замены (с использованием так называемых *S-блоков*), в результате которого образуется блок длиной 32 бита.

Таблица 8.5

32	1	2	3	4	5
4	5	6	7	8	9
8	9	10	11	12	13
12	13	14	15	16	17
16	17	18	19	20	21
20	21	22	23	24	25
24	25	26	27	28	29
28	29	30	31	32	1

4. Применение перестановки *P* (табл. 8.6), которое дает значение функции.

Таблица 8.6

16	7	20	21	29	12	28	17
1	15	23	26	5	18	31	10
2	8	24	14	32	27	3	9
19	13	30	16	22	11	4	25

Механизм действия *S*-блоков. Преобразование, с помощью которого 48-розрядный блок превращается в 32-разрядный, сводится к выбору восьми тетрад из семи таблиц (*S*-блоков*)* размером 4×16. Из каждого *S*-блока выбирается одна тетрада. Для этого 48-разрядный блок разделяется последовательно на 8 комбинаций, по 6 бит каждая.

Первая комбинация (слева) является входом в первый *S*-блок, вторая - во второй и т. д. При этом первый и последний биты комбинации задают номер строки, а остальные 4 бита - номер столбца *S*-блока, на пересечении которых содержится соответствующая тетрада.

На практике чаще всего используется тройной алгоритм шифрования Triple DES (рис. 8.15), который имеет большую криптостойкость.

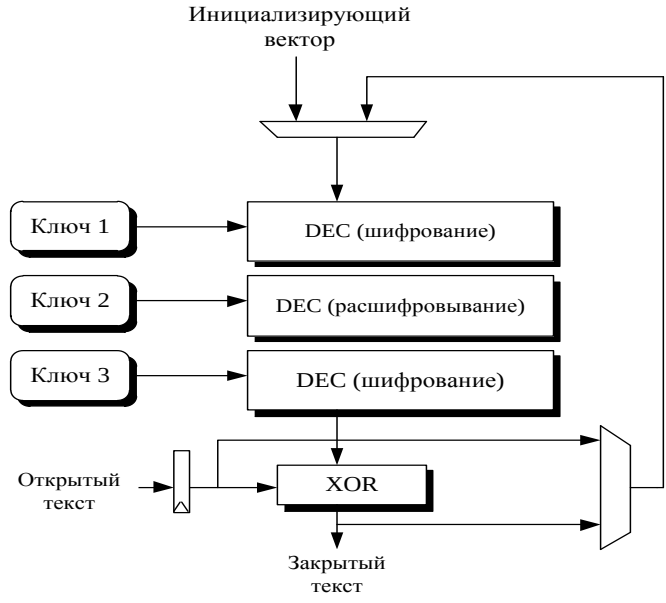


Рис. 8.15. Алгоритм шифрования Triple DES

Это, в сущности, тот же DES, только шифрование происходит трижды, следовательно, можем задавать три разных ключа и вектор инициализации. Шифрование будет происходить не за 16, а за 48 тактов (рис. 8.16). То есть нам нужно установить вектор инициализации, три ключа и данные и после 48-го такта на выходе получить результат.

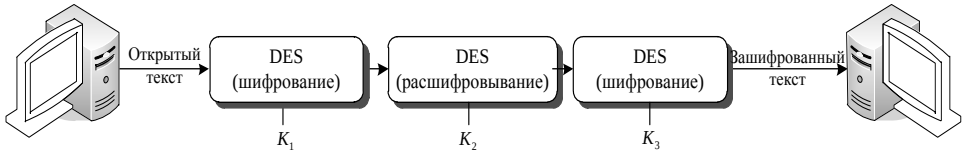


Рис. 8.16.Схема тройного DES

Многолетний опыт эксплуатации DES и его открытость (исходные тексты алгоритма и документацию на него можно найти в открытых источниках) привели к тому, что DES стал одним из более популярных алгоритмов с точки зрения проверки тех или других методов расшифровывания и криптоанализа. Среди основных недостатков DES, которые существенно снижают уровень его безопасности, можно выделить такие:

***наличие слабых ключей**, предопределенное тем, что при генерации ключевой последовательности используются два регистра сдвига, которые работают независимо друг от друга;*

***небольшая длина ключа** 56 бит (или 64 бита с контролем парности). На современном уровне развития микропроцессорных средств данная длина*

ключа не может обеспечивать надлежащей защиты для некоторых типов информации. Применение тройного DES (*Triple DES*) *не дает ощутимого результата, хотя в новой версии используются три разных ключа* (K_1, K_2, K_3). *Суть в том, что в окончательном итоге робота с тремя ключами эквивалентная зашифровыванию на другом ключе* K_4, *т. е. для любых* K_1, K_2, K_3 *найдется такой ключ* K_4, *когда*

$$EK_3\left(DK_2\left(EK_1(P)\right)\right) = EK_4(P);$$

избыточность ключа*, которая предопределена контролем четности для каждого отдельного байта ключа;*

использование статических подстановок *в S-блоках, что, несмотря на большое количество раундов, дает возможность проводить атаки на этот алгоритм.*

Асимметричные системы *характеризуются тем, что для шифрования и дешифрации используются разные ключи, связанные между собой некоторой математической зависимостью. При этом отмеченная зависимость такая, что найти один ключ, зная другой, путем вычислений очень трудно.*

Один из ключей (например, ключ шифрования) может быть общедоступным, и в этом случае проблема получения общего секретного ключа для связи отпадает. Если сделать общедоступным ключ расшифровывания, то на базе полученной системы можно построить систему аутентификации переданных сообщений. Поскольку по большей части один ключ из пары делается общедоступным, такие системы получили также название ***криптосистем с открытым ключом.***

Криптосистема с открытым ключом определяется тремя этапами реализации алгоритмов: генерации ключей, шифрования и расшифровывания.

Алгоритм *генерации ключей* открыт, любой пользователь может подать ему на вход случайную строку *r* надлежащей длины и получить пару ключей (k_1, k_2). Один из ключей (например, k_1) публикуется — он называется открытым, а второй — его называют *секретным* — сохраняется в тайне.

Алгоритмы шифрования E_{K_1} *и расшифровывания* D_{K_2} *таковы, что для любого открытого текста:*

$$m\ D_{K_2}\left(E_{K_1}(m)\right) = m.$$

Рассмотрим теперь гипотетическую атаку злоумышленника на эту систему. Злоумышленнику известен открытый ключ K_1, но неизвестен соответствующий секретный ключ K_2. Противник перехватил криптограмму *d* и пытается найти сообщение *m*, где $d = E_{K_1}(m)$. Поскольку алгоритм шифрования открыт, противник может просто последовательно перебрать все возможные сообщения длины *n*, вычислить для каждого такого сообщения m_i

криптограмму $d_i = E_{K_1}(m_i)$ и уравнять ее с *d*. То сообщение, для которого $d_i = d$, и будет открытым текстом. Если система имеет такие технические возможности, то открытый текст будет найден достаточно быстро. А в худшем случае перебор будет выполнен за время порядка $2^n(n)$, где $T(n)$ — время, необходимое для шифрования сообщения длины *n*. Если сообщения имеют длину порядка 1000 бит, то такой перебор невыполним на практике ни на одном из мощных компьютеров.

Мы рассмотрели лишь один из возможных способов атаки на криптосистему и самый простой алгоритм поиска открытого текста, названный алгоритмом *полного перебора*. Используется также другое название: *метод грубой силы*. Другой самый простой алгоритм поиска открытого текста — угадывание. Этот очевидный алгоритм требует небольших вычислений, но срабатывает с очень малой вероятностью (при больших длинах текстов). В действительности противник может пытаться атаковать криптосистему разными способами, используя разные более утонченные алгоритмы поиска открытого текста. Кроме того, злоумышленник может попробовать возобновить секретный ключ, используя знание (в общем случае несекретные) о математической зависимости между открытым и секретным ключами. Естественно считать криптосистему стойкой, если любой такой алгоритм нуждается практически в неисполнимом объеме вычислений или срабатывает с очень малой вероятностью. (При этом противник может использовать не только детерминированные, но и вероятностные алгоритмы). Это и есть теоретически сложный подход к определению стойкости.

Для его реализации относительно того или другого типа криптографических систем необходимо выполнить такие действия:

1) *дать формальное определение системы соответствующего типа;*

2) *дать формальное определение стойкости системы;*

3) *доказать стойкость конкретной конструкции системы этого типа.*

Односторонние (однонаправленные функции асимметричных криптографических систем. Центральным понятием в теории криптографических систем является понятие односторонней функции.

Односторонней функцией называется эффективно вычисляемая математическая функция, для обращения которой (т.е. для поиска хотя бы одного значения аргумента которой по заданному значению функции) не существует эффективных алгоритмов восстановления.

Формальное понятие односторонней функции описывается так.

Парная функция f называется односторонней, если:

1) существует алгоритм *A*, который для любого *x* вычисляет *f(x)*;

2) для любой полиномиальной вероятности

$$P\left\{ f\left(A\left(f\left(x\right)\right)\right) = f\left(x\right)\right\} \le 1/p(n).$$

Второе условие качественно означает такое. Любая полиномиальная вероятностная машина Тьюринга А *может* по данным *y найти* x *из* уравнения $f(x) = y$ лишь с очень малой вероятностью.

Отметим, что требование правдивости опустить нельзя. Поскольку длина входного слова $f(x)$ машины *A* равняется *m*, ей может просто не хватить полиномиальной от *m* времени на выписывание строки *x*.

Существование односторонних функций является необходимым условием стойкости многих криптосистем. Рассмотрим функцию f, такую что $f(r) = K_1$. Она вычисляется с помощью алгоритма *G*. Покажем, что когда f — не односторонняя функция, то криптосистема неустойчива.

Отметим, что существует полиномиальный вероятностный алгоритм *A*, вращающий f с вероятностью по крайней мере $1/p(n)$, для некоторого полинома *p*. Злоумышленник может подать на вход K_1 и получить с указанной вероятностью некоторое значение r' из прообраза. Дальше злоумышленник подает r' на вход алгоритма *G* и получает пару ключей (K_1, K_2'). Хотя K_2' не обязательно совпадает с K_2, по определению криптосистемы

$$D_{K_2'}(E_{K_1}(m)) = m$$

для любого открытого текста *m*.

Поскольку K_2' найдено с вероятностью $1/p(n)$, схема нестойкая.

Функцией-ловушкой *называется односторонняя функция, для которой обратную функцию вычислить просто, если есть некоторая дополнительная информация, и сложно, если такой информации нет.*

Криптосистемы с открытым ключом основываются на односторонних функциях-ловушках. При этом открытый ключ определяет конкретную реализацию функции, а секретный ключ подает информа-

Рональд Лин Ривест (Ronald Linn Rivest, 1947), криптограф. Наиболее известен изобретением алгоритма RSA вместе с Леонардом Адлеманом и Ади Шамиром. Ривест также изобретатель симметричных алгоритмов шифро-вания RC2, RC4, RC5 и соавтор RC6. Ему также принадлежит авторство криптографических хеш-функций MD2, MD4 и MD5. В 2006 г. опубликовал свое изобретение системы голосования Three Ballot - инновационной системы, способной различать, учтен ли голос конкретного избирателя при сохранении тайны голосования.

цию о ловушке.

Кто-либо, кому известна ловушка, может легко вычислять функцию в обоих направлениях, но тот, у кого такой информации нет, может выполнять вычисление только в одном направлении.

Прямое направление используется для шифрования и верификации цифровых подписей, а обратный - для расшифровывания и выработки цифровой подписи.

Во всех криптосистемах с открытым ключом чем большая длина ключа, тем более существенное разногласие между усилиями, необходимыми для вычисления функции в прямом и обратном направлениях (для того, кто не имеет информацию о ловушке).

Алгоритм шифрования Эль-Гамаля. *Криптосистема Эль-Гамаля — это криптосистема с открытым ключом, который основывается на свойствах логарифмизации. Система содержит как алгоритм шифрования, так и алгоритм цифровой подписи.*

Алгоритм Эль-Гамаля базируется на том, что свойство дискретного логарифмирования в конечном простом поле является сложной задачей с вычислительной точки зрения, которая нуждается в значительных вычислительных ресурсах. Множественное число параметров системы содержит простое число p и целое число g, степень которого по модулю p порождает множество элементов Z_p. У пользователя A есть секретный ключ a и открытый ключ y, где $y = g^a \pmod p$. Допустим, что пользователь B желает послать сообщение m пользователю A. Сначала B выбирает случайное число k, меньше p. Дальше он вычисляет $y_1 = g^k \pmod p$ и $y_2 = m \oplus (y^k \pmod p)$, где $\oplus$ означает побитовое «ИЛИ». B посылает A пары $(y_1,\ y_2)$

После получения шифрованного текста пользователь A вычисляет выражение для определения открытого текста на основе операции по модулю p: $m = (y_1^a \bmod p) \oplus y_2$. Известен вариант этой схемы, когда операция $\oplus$ заменяется умножением по модулю p. Это удобнее в том понимании, что в первом случае текст (или значение хэш-функции) необходимо разбивать на блоки той же длины, что и число $y^k \pmod p$. Во втором случае в этом нет потребности и можно обрабатывать блоки текста предварительно заданной фиксированной длины (меньше длины числа p).

Алгоритм шифрования Ривеста – Шамира - Адлемана (*RSA*). Наиболее известной асимметричной криптосистемой стала система на базе алгоритма RSA, который является первой практической реализацией на основе однонаправленной функции, предложенной Диффи и Хелманом (рис. 8.17).

Рассмотрим процесс формирования ключа шифрования и расшифровывания с помощью RSA.

Формирование ключа. Для того, чтобы сформировать ключ, получателю необходимо выполнить такие операции.

1. Выбрать два случайных простых числа p и q, что удовлетворяют условию $|p| \approx |q|$.

2. Вычислить $N = pq$.

3. Вычислить $\phi(N) = (p-1)(q-1)$.

4. Выбрать случайное целое число $e < \varphi(N)$, что удовлетворяет условию $\gcd(e, \varphi(N)) = 1$, и найти такое целое число d, что $ed \equiv 1 (\mathrm{mod}\, \varphi(N))$. (Поскольку $\gcd(e, \varphi(N)) = 1$, это уравнение имеет решение d, которое можно найти с помощью расширенного алгоритма Евклида.)

5. Использовать пару (N, e) как параметры открытого ключа, тщательным образом защитить числа p, q и $\varphi(N)$ и запомнить число d как закрытый ключ.

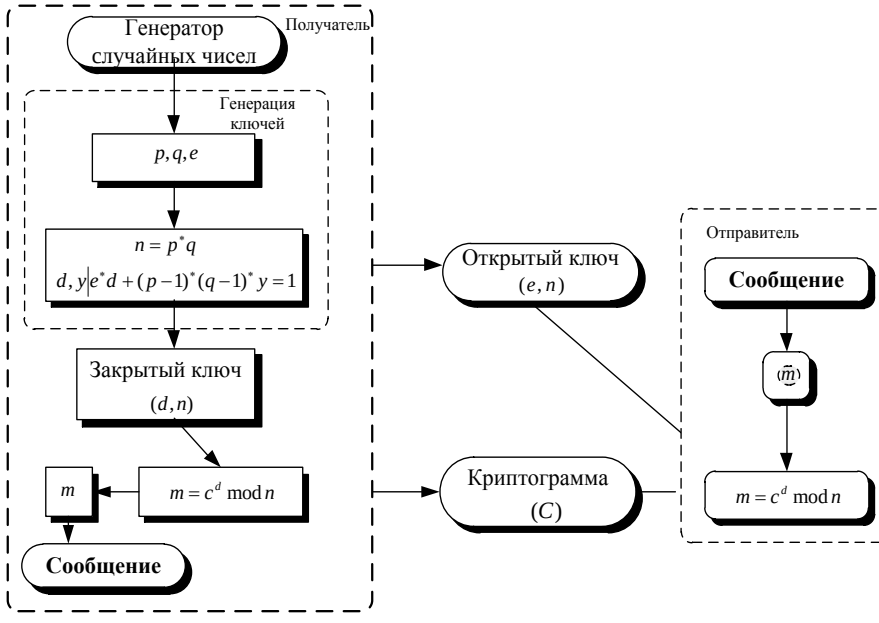


Рис. 8.17. Общая схема криптосистемы RSA

Шифрование. Для того, чтобы переслать получателю секретное сообщение, которое имеет длину $m < N$, отправитель создает шифротекст c

$$c \leftarrow m^e (\mathrm{mod}\, N).$$

С точки зрения отправителя, пространство начальных сообщений являет собой множественное число всех положительных чисел, меньше числа N.

Ади Шамир (Adi Shamir, 1952), израильский криптограф. Один из изобретателей (вместе с Роном Ривестом и Леонардом Адлеманом) алгоритма RSA. Является одним из изобретателей схемы аутентификации Feige-Fiat-Shamir. Сделал существенный вклад в развитие криптографии и информатики.

Расшифровка. Для того, чтобы расшифровать шифротекст c, получатель выполняет вычисление по формуле

$$m \leftarrow c^d \pmod N.$$

Криптосистемы, базирующиеся на алгоритме эллиптических кривых. Алгоритм Эль-Гамаля базируется на том, что операция логарифмирования в конечном простом поле является сложным техническим заданием. Однако конечные поля являются не единственными структурами алгебраизма, в которых можно поставить задание относительно вычисления дискретного логарифма. В 1985 году Коблиц и Миллер независимо друг от друга предложили использовать для построения криптосистем структуры алгебраизма, определенные на множественном числе точек на эллиптических кривых. Мы рассмотрим случаи определения эллиптических кривых над простыми конечными полями произвольной характеристики и над полями Галуа характеристики 2.

Пусть $p > 3$ — простое число; $a, b \in GF(p)$ такие, что $4a^2 + 27b^2 \neq 0$. ***Эллиптической кривой E над полем*** $GF(p)$ (эллиптической кривой в форме Веерштрасса) называется множество решений (x, y) уравнения *над полем* $GF(p)$ *вместе с дополнительной точкой* ∞*, которую называют бесконечно отдаленной точкой*

$$y^2 = x^3 + ax + b. \qquad (8.1)$$

Обозначим количество точек на эллиптической кривой E через $\#E$. Верхний и нижний пределы для $\#E$ определяются теоремой Хассе:

$$p + 1 - 2\sqrt{p} \le \#E \le p + 1 + 2\sqrt{p}.$$

Зададим бинарную операцию на E (в аддитивной записи) такими правилами:

1) $\infty + \infty = \infty$,

2) $\forall (x, y) \in E, (x, y) + \infty = (x, y)$,

3) $\forall (x, y) \in E, (x, y) + (x, -y) = \infty$,

4) $\forall (x_1, y_1) \in E, (x_2, y_2) \in E,\ x_1 \neq x_2, (x_1, y_1) + (x_2, y_2) = (x_3, y_3)$,

где $x_3 = \lambda^2 - x_1 - x_2$, $x_3 = \lambda^2 - x_1 - x_2$, $y_3 = \lambda(x_1 - x_3) - y_1$ и $\lambda = \dfrac{y_2 - y_1}{x_2 - x_1}$,

5) $\forall (x_1, y_1) \in E,\ y_1 \neq 0, (x_1, y_1) + (x_1, y_1) = (x_2, y_2)$,

где $x_2 = \lambda^2 - 2x_1$, $y_2 = \lambda(x_1 - x_3) - y_1$ и $\lambda = \dfrac{3x_1^2 + a}{2y_1}$.

Множественное число точек эллиптической кривой E из заданной таким образом операции образует абелевую группу (рис. 8.18).

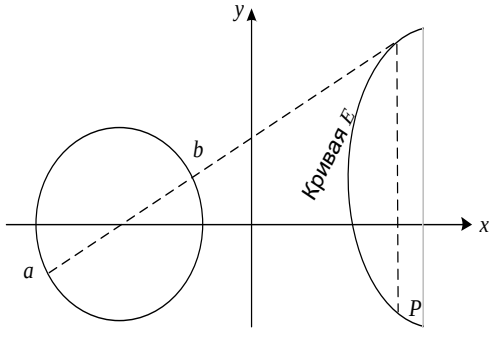


Рис. 8.18. Эллиптическая кривая

Если $\#E = p + 1$, то кривая E называется *суперсингулярной*.

Эллиптическая кривая, которая не является суперсингулярной кривой E над полем $GF(2^m)$ характеристики 2, задается таким способом.

Пусть $m > 3$ — целое число.

Пусть $a, b \in GF(2^m), b \neq 0$.

Эллиптической кривой E над полем $GF(2^m)$ называется множественное число решений *(x, в)* уравнения

$$y^2 + xy = x^3 + ax + b \tag{8.2}$$

над полем $GF(2^m)$ вместе с бесконечно отдаленной точкой.

Количество точек на кривой E также определяется теоремой Хассе:

$$q + 1 - 2\sqrt{q} \leq \#E \leq q + 1 + 2\sqrt{q},$$

где $q = 2^m$. Более того, #E парное.

Операция сложения с E в этом случае задается такими правилами:

1) $\infty + \infty = \infty$;

2) $\forall (x, y) \in E, (x, y) + \infty = (x, y)$;

3) $\forall(x, y) \in E, (x, y) + (x, x + y) = \infty$;

4) $\forall(x_1, y_1) \in E, (x_2, y_2) \in E, x_1 \neq x_2, (x_1, y_1) + (x_2, y_2) = (x_3, y_3)$,

где $x_3 = \lambda^2 + \lambda + x_1 + x_2 + a$, $y_3 = \lambda(x_1 + x_3) + x_3 + y_1$ и $\lambda = \dfrac{y_1 + y_2}{x_1 + x_2}$;

5) $\forall(x_1, y_1) \in E, x_1 \neq 0, (x_1, y_1) + (x_1, y_1) = (x_2, y_2)$,

где $x_2 = \lambda^2 + \lambda + a$, $y_2 = x_1^2 + (\lambda + 1)x_3$ и $\lambda = x_1 + \dfrac{y_1}{x_1}$.

В этом случае множество точек эллиптической кривой *E* из заданной таким способом операции также образует абелевую группу.

Пользуясь операцией добавления точек на кривой, можно естественно определить операцию умножения точки $P \in E$ на произвольное целое число *n*:

$$nP = P + P + \ldots + P,$$

где операция добавления выполняется *n* раз.

Теперь построим одностороннюю функцию, на основе которой можно будет создать криптографическую систему.

Пусть *E* — эллиптическая кривая, $P \in E$ — точка на этой кривой. Выберем целое число $n < \#E$. Тогда как прямую функцию выберем произведение nP. Для его вычисления по оптимальному алгоритму понадобится не более чем $2\log_2 n$ операций добавления. Обратную задачу сформулируем таким образом: по заданной эллиптической кривой *E*, точкой $P \in E$ и произведению nP найти *n*.

Теперь мы можем описать криптографический протокол, аналогичный известному протоколу Диффи—Хеллмана. Для установления защищенной связи двое пользователей *A* и *B* совместно выбирают эллиптическую кривую *E* и точку *P* на ней. Дальше каждый из пользователей выбирает свое секретное целое число — соответственно *a и b*. Пользователь *A* вычисляет произведение aP, а пользователь *B* – произведение bP. Дальше они обмениваются вычисленными значениями. При этом параметры самой кривой, координаты точки на ней и значения произведений являются открытыми и могут передаваться по незащищенным каналам связи. Потом пользователь *A* умножает полученное значение на *a,* а пользователь *B* - на *b*. Согласно свойствам операции умножения на число выполняется равенство $a \cdot bP = b \cdot aP$. Таким образом, оба пользователя получат общее секретное значение (координаты точки abP), которое смогут использовать для получения ключа шифрования.

Заметим, что злоумышленнику для возобновления ключа придется развязать сложную с вычислительной точки зрения задачу определения *a и b* по известным *E*, *P*, *aP* и *bP*.

8.4. Электронная цифровая подпись

Использование электронной цифровой подписи (ЭЦП) - это процесс, который обеспечивает целостность сообщений (документов), переданных незащищенными информационно коммуникационными каналами общего пользования в системах обработки информации разного назначения, с гарантированной идентификацией ее автора (лица, которое подписало документ).

Цифровая подпись (цифровая сигнатура) *- цифровая последовательность данных, которая образуется в результате асимметричного криптографического преобразования начальной информации (открытого текста) и дает возможность получателю проверить источник и целостность данных, а также осуществить защиту информации от фальсификации или подделки.*

Практическая реализация электронной цифровой подписи базируется на использовании однонаправленной функции, которая шифруется секретным ключом отправителя с целью расшифровывания лишь части сообщения, - его дайджеста (ключ - идентификатор), который защищает информацию от несанкционированного изменения. Процедуру эффективной генерации ЭЦП иллюстрирует рис. 8.19.

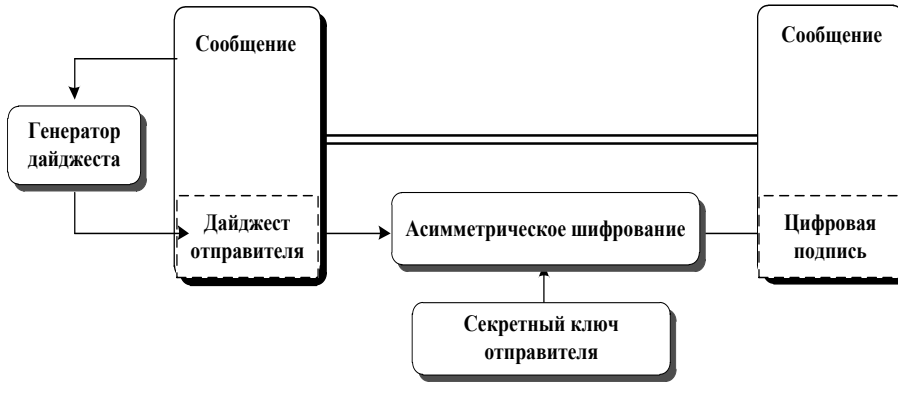


Рис. 8.19. Схема генерации ЭЦП

Проверку ЭЦП можно осуществить методом, который иллюстрирует рис. 8.20.

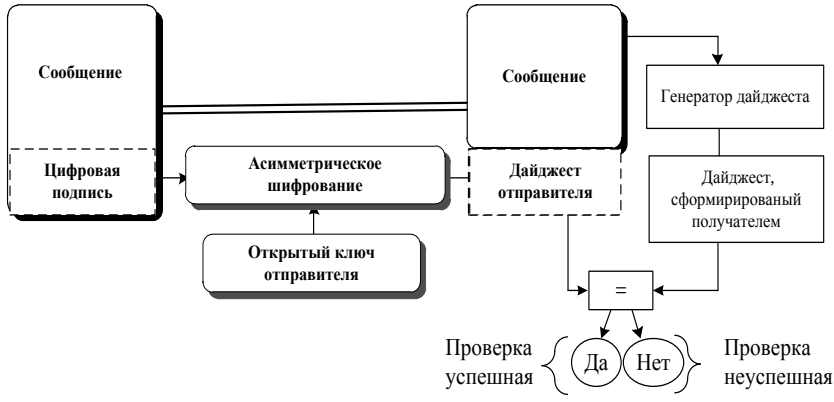


Рис. 8.20. Проверка сгенерованной ЭЦП

Основные угрозы относительно использования цифровой подписи (рис. 8.21):

отказ - отправитель впоследствии отказывается от переданного сообщения;

фальсификация - получатель подделывает сообщение;

модификация - получатель вносит изменения в сообщение;

маскирование - пользователь маскируется под другого.

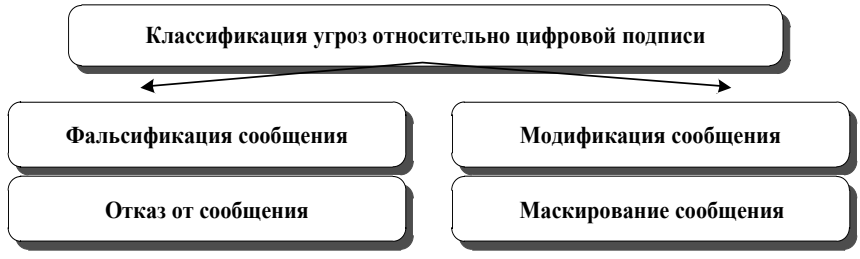


Рис. 8.21. Угрозы относительно цифровой подписи

Цифровые подписи, **основанные на асимметричных криптосистемах**. Для формирования системы ЭЦП чаще всего используют криптографическую систему Ривеста – Шамира - Эйделмана.

Пример. *Формирование системы ЭЦП на базе алгоритма Ривеста – Шамира - Эйделмана.*

Пользователь *A* производит цифровую подпись, предназначенную для пользователя *B* сообщения *M*, с помощью такого преобразования

$$SIG(M) = E_{e_b,\, n_B}\left(E_{d_A,\, n_A}(M)\right).$$

При этом он использует свое секретное преобразование $E_{e_B,\,n_B}$ и открытое преобразование пользователя *B*. Дальше он передает пользователю *B* пары <*M, SIG(M)*>.

Пользователь *B* может верифицировать это подписание сообщения сначала с помощью своего секретного преобразования $E_{d_B,\,n_B}$ с целью задержки

$$E_{d_A,\,n_A}(M) = E_{d_B,\,n_B}\left(SIG(M)\right) = E_{d_B,\,n_B}\left(E_{e_B,\,n_B}\left(E_{d_A,\,n_A}(M)\right)\right), \qquad (8.3)$$

а затем открытого преобразования $E_{e_A,\,n_A}$ пользователя *A* для получения сообщения *M*:

$$M = E_{e_A,\,n_A}\left(E_{d_A,\,n_A}(M)\right). \qquad (8.4)$$

Затем пользователь *B* сравнивает полученное сообщение *M* с тем, которое он получил в результате проверки ЭЦП, и принимает решение относительно действительности или подложности полученного сообщения.

В рассмотренном примере проверить подлинность ЭЦП может только пользователь *B*. А если нужно обеспечить возможность верификации ЭЦП произвольным пользователем (например, при циркулярной рассылке документа), то алгоритм обработки ЭЦП упрощается, и подпись выполняется по формуле

$$SIG(M) = E_{d_A,\,n_A}(M).$$

В таком случае любой пользователь осуществляет верификацию с использованием открытого преобразования отправителя (пользователя *A*):

$$M = E_{e_A,\,n_A}(SIG(M)) = E_{e_A,\,n_A}(E_{d_A,\,n_A}(M)). \qquad (8.5)$$

Вместо криптосистемы *RSA* для подписи сообщений можно использовать и любую другую асимметричную криптосистему.

Недостатком такого подхода является то, что производительность асимметричной криптосистемы может оказаться недостаточной. Возможным решением является приложение специальной эффективной вычислительной функции, названной *хеш-функцией*, или *функцией хеширования.* Входом этой функции является сообщение, а выходом — слово фиксированной длины, намного более малой, чем длина исходного сообщения.

ЕЦП производится по той же схеме, но при этом используется не именно сообщение, а значение хеш-функции от него. Это существенно убыстряет генерирование и верификацию ЭЦП.

Иногда желательно, чтобы ЭЦП была разной, даже если дважды подписывается то же сообщение. Для этого в процесс генерирования ЭЦП необходимо внести элемент «случайности». Соответствующий способ предложил Эль-Гамаль аналогично тому, как это делается в системе шифрования, которая носит его имя.

Пример. *Формирование системы ЭЦП на базе алгоритма Эль-Гамаля.*

Выбирается большое простое число p и целое число g, что является первобытным элементом в Z_p. Эти числа публикуются. Потом выбирается секретное число x и вычисляется открытый ключ для проверки подписи $y = g^x (\mathrm{mod}\, p)$.

Дальше для подписи сообщения M вычисляется его хеш-функция $m = h(M)$. Выбирается случайное целое k, где $1 < k < (p-1)$, взаимно простое из p-1 и вычисляется $r = g^k (\mathrm{mod}\, p)$. После этого с помощью расширенного алгоритма Эвклида решается относительно s уравнение $m = xr + ks (\mathrm{mod}\, p-1)$.. Подпись представляет собой пару чисел (r, s). После генерирования подписи значение k уничтожается.

Получатель подписанного сообщения вычисляет хэш-функцию сообщения $m = h(M)$ и проверяет выполнение равенства $y^r r^s (\mathrm{mod}\, p) = g^m$. Корректность этого равенства очевидна:

$$y^r r^s = g^{x \cdot r} g^{k \cdot s} = g^{x \cdot r + k \cdot s} = g^m (\mathrm{mod}\, p). \qquad (8.6)$$

Еще одну аналогичную схему предложил Шнорр. Как обычно, p - большое простое число; q - простой делитель $(p-1)$; g - элемент порядка q в Z_p; $к$ - случайное число; x и $y = g^x (\mathrm{mod}\, p)$ - соответственно секретный и открытый ключ. Уравнения генерирования подписи приобретают вид

$$r = g^k (\mathrm{mod}\, p);\ e = h(m, r);\ s = k + xe (\mathrm{mod}\, q). \qquad (8.7)$$

Подписью является пара (r, s). На приемном конце вычисляется значение хэш-функции $e = h(m, r)$ и проверяется выполнение равенства $r = g^s y^{-e} (\mathrm{mod}\, p)$, при этом действия по показателю степени генерируются по модулю q.

Стандарт цифровой подписи (***алгоритм DSS***). В США принят стандарт на генерирование и верификацию ЭЦП, который достал название DSS (Digital Signature Standard). В соответствии с этим стандартом ЭЦП производится по такой схеме:

1. Предварительный этап. Выбираются числа p, q и g, где p - простое число длины l, где l кратное 64 и $512 \le l \le 1024$; q - простой делитель числа $p-1$ длиной 160 бит; g - элемент порядка q в Z_p. Эти три числа являются открытыми данными. Выбирается секретный ключ x, $1 \le x \le q$ и вычисляется открытый ключ для проверки подписи $y = g^x (\mathrm{mod}\, p)$.

2. Генерирование ЭЦП. Вычисляется значение хеш-функции от сообщения $h(m)$. При этом используется алгоритм безопасного хеширования SHA (Secure Hashing Algorithm), на который ссылается стандарт. Значение хеш-

функции $h(m)$ имеет длину 160 бит. Дальше тот, кто ставит подпись, выбирает случайное значение $k, 1 \le \le k < q,$ вычисляет значение $k^{-1} (\mathrm{mod}\, q)$ и пары значений:

$$r = g^k (\mathrm{mod}\, p)(\mathrm{mod}\, q); \quad s = k^{-1}(h(m) + xr)(\mathrm{mod}\, q).$$

Эта пара значений (r, s) и является электронной подписью под сообщением M. После генерирования цифровой подписи значение k уничтожается.

3. Верификация ЭЦП. Пусть было принято сообщение m_1. Тогда уравнение проверки выглядит таким образом:

$$r \equiv g^{h(m_1)\cdot s^{-1}} \cdot y^{r\cdot s^{-1}} (\mathrm{mod}\, p)(\mathrm{mod}\, q). \qquad (8.8)$$

Действительно:

$$g^{h(m)s^{-1}} \cdot y^{r\cdot s^{-1}} (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) = g^{h(m)s^{-1}} \cdot g^{x\cdot r\cdot s^{-1}} (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) =$$
$$= g^{s^{-1}(h(m)+xr)} (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) = g^{(k^{-1}(h(m)+x\cdot r)^{-1}(h(m)+xr)} (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) =$$
$$= g^{(k^{-1})^{-1}(h(m)+xr)^{-1}(h(m)+xr)} (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) = g^k (\mathrm{mod}\, p)\ (\mathrm{mod}\, q) \equiv r.$$

Цифровые подписи, основанные на симметричных криптосистемах. Двухключевая криптография возникла, потому что ряд новых криптографических протоколов типа протокола цифровой подписи не удалось эффективно реализовать на базе традиционных криптографических алгоритмов. Однако это возможно. И первыми, кто обратил на это внимание, были родоначальники криптографии с открытым ключом У. Диффи и М. Хеллман, которые предложили подход, дающий возможность выполнять процедуру цифровой подписи одного бита с помощью блочного шифра. Прежде чем изложить эту идею, сделаем несколько замечаний о сущности и реализации цифровой подписи.

Практически все современные алгоритмы ЭЦП базируются на так называемых сложных математических задачах типа факторизации больших чисел или логарифмирования в дискретных полях. Однако невозможность эффективного развязывания этих задач численными методами математики не доказана, потому весьма возможно, что в ближайшем будущем эти задачи будут развязаны, а соответствующие схемы сломаны.

Пример. *Алгоритм ЭЦП на основе классического блочного шифра Диффи и Хеллмана.*

Допустим, в нашем распоряжении есть алгоритм шифрования, который оперирует блоками данных X размера n, и ключ, который используется, *размером* $nK : |X| = n, |K| = nK$. Структура ключевой информации в схеме такая: *секретный ключ подписи* k_S *выбирается как произвольная (случайная) пара ключей* k_0, k_1 *используемого блочного шифра.*

Леонард Макс Адлеман (Leonard Max Adleman, 1945), профессор информатики и молекулярной биологии в университете Южной Калифорнии. Известен изобретением (1977) вместе из Рональдом Ривестом и Ади Шамиром криптосистемы RSA (Rivest – Shamir - Adleman) и исследованием ДНК-вычислений. В 1994 г. опубликовал статью, посвященную молекулярным вычислениям решения комбинаторных задач и описания экспериментального использования ДНК как вычислительной системы. Адлеман является также одним из изобретателей оригинального теста Адлемана – Померанча -Румели (Adleman – Pomerance - Rumely) на простоту чисел.

Таким образом, размер ключа подписи равняется удвоенному размеру ключа используемого блочного шифра:

$$\left|K_S\right| = 2\left|K\right| = 2n_K. \qquad (8.9)$$

Ключ проверки является результатом шифрования двух блоков текста X_0 и X_1 с ключами k_0 и k_1 соответственно

$$k_V = (C_0, C_1) = \left(E_{k_0}(X_0),\, E_{k_1}(X_1)\right),$$

где параметром схемы является блоки данные, несекретные и известные стороне, которая проверяет подпись. Таким образом, размер ключа проверки подписи равняется удвоенному размеру блока использованного блочного шифра:

$$\left|k_V\right| = 2\left|X\right| = 2n.$$

Алгоритм Sig генерирования ЭЦП для бита $t\left(t \in \{0,1\}\right)$ заключается просто в выборе соответствующей половины из пары, которая составляет секретный ключ подписи:

Алгоритм Ver проверки подписи базируется на проверке уравнения $E_{k_t}(X_t) = C_t$, который должен выполняться для нашего t. Получателю известны все используемые при этом величины. Таким образом, функция проверки подписи:

$$\mathrm{Ver}(t,\, s,\, k_V) = \begin{cases} 1, & E_s(X_t) = C_t, \\ 0, & E_s(X_t) \neq C_t. \end{cases} \qquad (8.10)$$

Покажем, что эта схема работоспособна, для чего проверим выполнение необходимых свойств схемы цифровой подписи.

1. *Невозможность подписать бит t, если неизвестен ключ подписи.* Действительно, для выполнения этого злоумышленнику пришлось бы развязать уравнение $E_S(X_t) = C_t$ относительно s, что эквивалентно определению ключа для известных блоков шифрованного и соответствующего ему открытого текста, что вычислительно невозможно в результате использования стойкого шифра.

2. *Невозможность подписать бит* t, *если неизвестен ключ подписи.* Действительно, для выполнения этого злоумышленнику пришлось бы развязать уравнение $E_S(X_t) = C_t$ относительно s, что эквивалентно определению ключа для известных блоков шифрованного и соответствующего ему открытого текста, что вычислительно невозможно в результате использования стойкого шифра.

3. *Невозможность подобрать другое значение бита* t, *что подходило бы под заданную подпись.* Возможных значений бита всего два, а вероятности выполнения двух приведенных дальше условий одновременно очень малые, учитывая использование криптостойкого алгоритма

$$E_S(X_0) = C_0,$$

$$E_S(X_1) = C_1.$$

Предложенная Диффи и Хеллманом схема ЭЦП на основе классического блочного шифра имеет такую же стойкость, как и блочный шифр, и при этом достаточно простая. Однако она имеет два существенных недостатка.

Первый недостаток заключается в том, что эта схема дает возможность подписать лишь один бит информации. В блоке большего размера придется отдельно подписывать каждый бит, потому даже с учетом хеширования сообщения все компоненты подписи - секретный ключ, проверяющая комбинация и подпись - выходят достаточно большими по размеру и более чем на два порядка превышают размер блока, который подписывается. Предположим, что в схеме используется криптографический алгоритм E_K с размером блока и ключа соответственно n и n_K. Предполагая также, что используется функция хэширования с размером исходного блока n_H. Тогда размеры основных рабочих блоков таковы:

размер ключа подписи: $n_{kS} = 2n_H n_K$.

размер ключа проверки подписи: $n_C = 2n_H n$.

размер подписи: $n_S = n_H n_K$.

Второй недостаток этой схемы заключается в том, что пары ключей генерирования подписи и проверки подписи можно использовать только один раз. Действительно, выполнение процедуры подписи бита сообщения приводит к раскрытию половины секретного ключа, после чего он уже не является полностью секретным, а следовательно, его нельзя использовать повторно. Поэтому для каждого сообщения, которое подписывается, необходим свой комплект ключей подписи и проверки. Это практически делает невозможным использование рассмотренной схемы Диффи - Хеллмана в предложенном варианте в реальных системах ЭЦП.

Березин и Дорошкевич предложили модификацию схемы Диффи - Хеллмана, что фактически устраняет ее недостатки.

Пример. *Алгоритм ЭЦП на основе классического блочного шифра Березина и Дорошкевича.*

Центральным в этом подходе является алгоритм «одностороннего криптографического прокручивания», которое в известной мере может рассматриваться как аналог операции подъема к степени. Предположим, что в нашем распоряжении есть криптографический алгоритм E_K с размером блока данных и ключа соответственно n и n_K бит, причем $n \le n_K$.

Пусть у нас также есть некоторая функция отображения битовых блоков данных у n-бит и $Y = P_{n \to n_K}(X)$, $|X| = n$, $|Y| = n_K$. Определим рекурсивную функцию R_k «одностороннего прокручивания» блока данных T размером n бит k раз $(k \ge 0)$ с помощью такой формулы:

$$R_k(T) = \begin{cases} T, \; k = 0, \\ E_{P_{n \to n_K}(R_{k-1}(T))}(X), \; k > 0, \end{cases}$$

где X - произвольный несекретный n-битовый блок данных, являющихся параметром процедуры прокручивания.

Идея функции одностороннего прокручивания чрезвычайно проста: достаточно всего лишь нужное количество раз k выполнить такие действия: расширить n-битовый блок данных T к размеру n_K ключа использованного алгоритма шифрования; на полученном расширенном блоке как на ключе зашифровать блок данных X; результат шифрования занести на место исходного блока данных T. Операция $R_k(T)$ имеет важные свойства.

1. Аддитивность и коммутативность по количеству прокручиваний:

$$R_{k+k'}(T) = R_{k'}\left[R_k(T)\right] = R_k\left[R_{k'}(T)\right]. \qquad (8.11)$$

2. Односторонность или необратимость прокручивания: если известно только некоторое значение функции $R_k(T)$, то численными методами невозможно найти значение $R_{k'}(T)$ для любого $k' < k$. Если бы это было возможно, то в нашем распоряжении был бы способ определить ключ шифрования по известным входным и исходным блокам алгоритма E_K, что противоречит предположению о стойкости шифра.

Теперь покажем, как отмеченную операцию можно использовать для подписи блока T, состоящего из n_T битов.

Секретный ключ подписи k_S выбирается как произвольная пара блоков k_0, k_1, которые имеют размер блока данных используемого блочного шифра, т.е. размер ключа генерирования подписи равняется удвоенному размеру блока данных использованного блочного шифра $|k_S| = 2n$.

Ключ проверки подписи вычисляется как пара блоков, которые имеют размер блоков данных использованного алгоритма за такими формулами:

$$k_C = (C_0, C_1) = \left[R_{2^{nT}-1}(K_0), R_{2^{nT}-1}(K_1) \right]. \tag{8.12}$$

В этих вычислениях также используются несекретные блоки данных X_0 и X_1, которые являются параметрами функции «одностороннего прокручивания», они непременно должны быть разными. Таким образом, размер ключа проверки подписи также равняется удвоенному размеру блока данных использованного блочного шифра $|k_S| = 2n$.

Вычисление и проверку ЭЦП выполняют таким образом.

Алгоритм SIGNT получения ЭЦП для *n*-битового блока T заключается в исполнении «одностороннего прокручивания» обеих половин ключа подписи соответственно T и $2^{nT} - 1 - T$ *раз:*

$$s = \mathrm{Sig}_{nT}(T) = (s_0, s_1) = R_T(k_0), R_{2^{nT}-1-T}(k_1). \tag{8.13}$$

Алгоритм VERNT проверки подписи состоит из проверки истинности соотношений

$$R_{2^{nT}-1-T}(s_0) = C_0,\ R_T(s_1) = C_1,$$

которые должны выполняться для настоящего блока данных T

$$R_{2^{nT}-1-T}(s_0) = R_{2^{nT}-1-T}(R_T(k_0)) = R_{2^{nT}-1-T+T}(k_0) = R_{2^{nT}-1}(k_0) = C_0,$$
$$R_T(s_1) = R_T(R_{2^{nT}-1-T}(k_1)) = R_{T+2^{nT}-1-T}(k_1) = R_{2^{nT}-1}(k_1) = C_1.$$

Таким образом, функция проверки подписи приобретает вид

$$\mathrm{Ver}(T, s, k_C) = \begin{cases} 1,\ R_{2^{nT}-1-T}(s_0) = C_0 \mid R_T(s_1) = C_1, \\ 0,\ R_{2^{nT}-1-T}(s_0) \neq C_0 \mid R_T(s_1) \neq C_1. \end{cases} \tag{8.14}$$

Покажем, что для этой схемы выполняются необходимые условия работоспособности схемы подписи. Допустимо, что в распоряжении злоумышленника есть *n*-битовый блок T, его подпись $s = (s_0, s_1)$ и ключ проверки $k_C = \ = (C_0, C_1)$. Пользуясь этой информацией, злоумышленник пытается найти правильную подпись $s' = (s_0', s_1')$ для другого *n-битового* блока T'. Для этого ему нужно решить такие уравнения относительно s_0' и s_1':

$$R_{2^{nT}-1-T}(s_0') = C_0, \qquad R_T(s_1') = C_1.$$

В распоряжении злоумышленника есть блок данных T с подписью $s = (s_0, s_1)$, что дает ему возможность вычислить одно из значений s_0' и s_1', даже не имея ключа подписи

1) если $T < T'$, то $s'_0 = R_{T'}(k_0) = RT' - T(R_T(k_0)) = RT' - T(s_0)$;

2) если $T > T'$, то $s'_1 = R_{2^{nT}-1-T'}(k_1) = R_{T-T'}(R_{2^{nT}-1-T}(k_1)) = R_{T-T'}(s_1)$.

Однако для нахождения второй половины подписи s'_0 и s'_1 соответственно в случае 1 и 2 ему необходимо выполнить прокручивание в противоположном направлении, т.е. найти $R_k(X)$, имея только значение для большего k, что численными методами невозможно. Таким образом, злоумышленник не может подработать подпись под сообщением, если не имеет в своем распоряжении секретный ключ подписи.

Второе требование также выполняется: вероятность подобрать блок данных T', отличающийся от блока T, имея такую же цифровую подпись, настолько мала, что ею можно пренебречь. Действительно, пусть цифровая подпись блоков T и T' одна и та же. Тогда подписи обоих блоков таковы

$$s = S_{nT}(T) = (s_0, s_1) = \left[R_T(k_0), R_{2^{nT}-1-T}(k_1) \right],$$
$$s' = S_{nT}(T') = (s'_0, s'_1) = \left[R_{T'}(k_0), R_{2^{nT}-1-T'}(k_1) \right],$$

но $s = s'$, следовательно, имеем

$$R_T(k_0) = R_{T'}(k_0) \text{ и } R_{2^{nT}-1-T}(k_1) = R_{2^{nT}-1-T'}(k_1). \qquad (8.15)$$

Возьмем для определенности $T \le T'$, тогда выполняются такие равенства:

$$RT' - T(k_0^*) = k_0^*, \quad RT' - T(k_1^*) = k_1^*, \text{ где } k_0^* = R_T(k_0),$$

Последнее условие значит, что прокручивание двух разных блоков данных одинаковое число раз оставляет их значения неизменными. Вероятность такого события чрезвычайно мала, и ее можно не принимать во внимание.

Таким способом только что рассмотренная модификация схемы Диффи—Хеллмана делает возможной подпись не одного бита, а целой битовой группы. Это дает возможность в несколько раз уменьшить размер подписи и ключей подписи/проверки отмеченной схемы.

Однако нужно понимать, что увеличение размера подписываемых битовых групп приводит к экспоненциальному увеличению объема необходимых вычислений, а следовательно, начиная с некоторого значения робота схемы становится неэффективной. Предел «умного размера» подписываемой группы составляет около 10 бит, а блоки большего размера одинаково необходимо подписывать «вразнобой».

Размер ключей и подписи. Найдем размеры ключей и подписи, а также объем необходимых для реализации схемы вычислений.

Пусть размер хеш-блока и блока используемого шифра одинаковые и равняются *n*, а размер подписываемых битовых групп *равняется nT*. Предположим также, что когда последняя группа содержит меньшее количество битов, то обрабатывается она так же, как полная *n-битовая* группа. Тогда размеры ключей подписи/проверки и самой подписи совпадают:

$$|K_S| = |K_C| = |s| = 2n\left\lceil \frac{n}{n_T} \right\rceil \approx 2\frac{n^2}{n_T} \text{ бит},$$

где $\lceil x \rceil$ - округление числа *x* к ближайшему целому в направлении роста.

Количество операций шифрования $E_K(X)$, необходимое для реализации процедур схемы, определяется приведенными дальше соотношениями.

При генерировании ключевой информации

$$W_K = 2(2^{nT} - 1)\left\lceil \frac{n}{n_T} \right\rceil \approx \frac{2^{nT+1} n}{n_T};$$

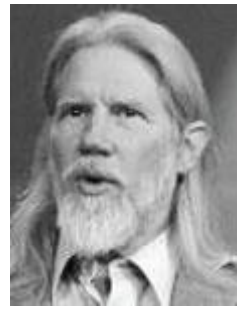
при генерировании и проверке подписи

$$W_S = W_C = (2^{nT} - 1)\left\lceil \frac{n}{n_T} \right\rceil \approx \frac{2^{nT} n}{n_T}.$$

Размер ключа подписи и проверки подписи можно дополнительно уменьшить такими приемами.

1. Нет необходимости сохранять ключи подписи отдельных битовых групп, их можно динамически производить в нужный момент времени с помощью генератора криптостойкой гаммы. Ключом подписи в этом случае будет обычный ключ, использованный в схеме подписи блочного шифра.

2. Аналогично нет необходимости хранить массив ключей проверки подписи отдельных битовых



Витфилд Диффи (Whitfield Diffie, 1944), получил степень бакалавра по математики в Массачусетском технологическом институте (1965). В 1976 году вместе с еще одним известным специалистом в области криптографии Мартином Хеллманом опубликовал статью "Новые направления в криптографии", в которой предлагался радикально новый метод работы с распределенными ключами, который решал одну из фундаментальных проблем криптографии - так называемый метод обмена ключей Диффи - Хеллмана.

групп блока, достаточно хранить значение хэш-функции этого массива. При этом алгоритм генерирования ключа подписи и алгоритм проверки подписи будут дополнены еще одним шагом - вычислением проверяющих комбинаций отдельных битовых групп.

Таким образом, проблема размера ключей и подписи решена, однако второго недостатка схемы - одноразовости ключей - не устранено, поскольку это невозможно в рамках подхода Диффи - Хеллмана.

Для практического использования такой схемы, рассчитанной на подпись N сообщений, отправителю необходимо хранить N ключей подписи, а получателю - N ключей проверки, что достаточно неудобно. Эту проблему можно решить так же, как была решена проблема ключей для множественных битовых групп - генерацией ключей подписи для всех N сообщений из одного мастерского ключа и свертыванием всех проверяющих комбинаций в одну контрольную комбинацию с помощью алгоритма вычисления хэш-функции.

Такой подход решил бы проблему размера сохраненных ключей, но привел бы к необходимости вместе с подписью каждого сообщения высылать отсутствующие $N-1$ проверяющих комбинаций, необходимых для вычисления хеш-функции массива всех контрольных комбинаций отдельных сообщений. Понятно, что такой вариант не имеет преимущества по сравнению с выходным.

Упоминавшиеся уже авторы предложили механизм, который дает возможность значительно снизить остроту проблемы. Его основная идея — вычислять контрольную комбинацию (ключ проверки подписи) не как хеш-функцию от линейного массива проверяющих комбинаций всех сообщений, а попарно — с помощью бинарного дерева. На каждом уровне проверяющая комбинация вычисляется как хеш-функция от конкатенации двух проверяющих комбинаций младшего уровня. Чем высший уровень комбинации, тем больше отдельных ключей проверки «учитывается» в ней.

Допустим, что наша схема рассчитана на 2^L сообщений. Обозначим через $C_i^{(l)}$ i-ю комбинацию l-го уровня. Если нумерацию комбинаций и уровней начинать с нуля, то выполняется такое условие: $0 \le i < 2^{L-l}$, а i-я проверяющая комбинация l-го уровня рассчитана на 2^l сообщений с номерами от $i \cdot 2^l$ к $(i+1) \cdot 2^l - 1$ включительно. Комбинаций нижнего, нулевого, уровня будет 2^L, а последнего, верхнего, *L-го* уровня — одна, которая и является контрольной комбинацией всех сообщений, на которые рассчитана схема.

На каждом уровне, начиная с первого, проверяющие комбинации рассчитываются по формуле

$$C_i^{(l+1)} = H(C_{2i}^{(l)} \| C_{2i+1}^{(l)}),$$

где через $A \| B$ *обозначен результат конкатенации двух блоков данных* A и B, а через $H(X)$ - процедура вычисления хэш-функции блока данных X.

При использовании отмеченного подхода вместе с подписью сообщения необходимо передать не как в исходном варианте, а только $\log_2 N$ контрольных комбинаций. Передаваться должны комбинации, которые отвечают смежным отраслям дерева на пути от конечной вершины, которая отвечает номеру использованной подписи, к корню.

Пример. *Организация проверяющих комбинаций.*

Организацию проверяющих комбинаций в виде двоичного дерева в схеме на восемь сообщений иллюстрирует рис. 8.22.

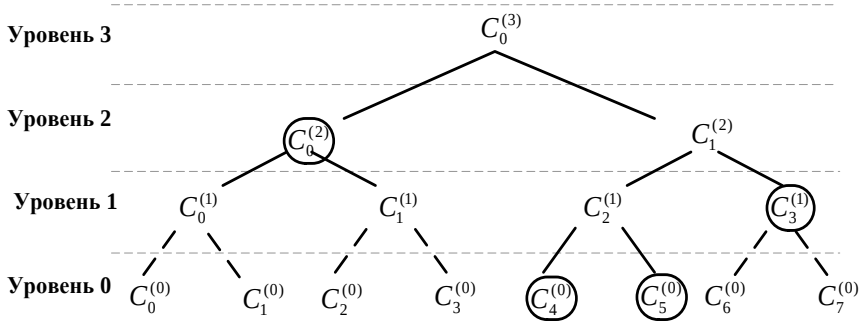


Рис. 8.22. Организация проверочных комбинаций в виде двоичного дерева

Так, при передаче сообщения № 5 (контрольная комбинация выделена рамкой) вместе с его подписью должна быть передана контрольная комбинация сообщения № 4 $C_4^{(0)}$, общая для сообщений № 6-7 $C_3^{(1)}$ и общая для сообщений № 0-3 $C_0^{(2)}$, все они выделены на рисунке другим фоном. При проверке подписи значение $C_5^{(0)}$ будет вычислено из сообщения и его подписи, а итоговую контрольную комбинацию, которая подлежит сравнению с эталонной, по формуле

$$C = C_0^{(3)} = H(C_0^{(2)} \| H\left(H\left(C_4^{(0)} \| C_5^{(0)} \| C_3^{(1)}\right)\right). \qquad (8.16)$$

Необходимость отправлять вместе с подписью сообщения дополнительную информацию, нужную для проверки подписи, в действительности не очень обременительная. Действительно, в системе на 1024 = 210 подписей вместе с сообщением и его подписью необходимо дополнительно передавать 10 контрольных комбинаций, а в системе на 1048576 = 220 подписей - всего 20 комбинаций. Однако в случае большого количества подписей, на которые рассчитана система, появляется другая проблема - хранение дополнительных ком-

бинаций, если они рассчитаны предварительно, или их генерирование в момент формирования подписи.

Дополнительные контрольные комбинации, которые передаются вместе с подписью и используются при его проверке, производятся при формировании ключа проверки за ключом подписи и могут сохраняться в системе и использоваться в момент формирования подписи или вычисляться заново в этот момент.

Первый подход допускает расходы дисковой памяти, поскольку необходимо хранить значения хэш-функции всех уровней, а второй нуждается в большом объеме вычислений в момент формирования подписи. Можно использовать и компромиссный подход - хранить все хэш-комбинации начиная с некоторого уровня l^*, а комбинации меньшего уровня вычислять при формировании подписи.

В рассмотренной схеме подписи на 8 сообщений можно хранить все 14 контрольных комбинаций, используемых при проверке (всего их 15, но последняя верхняя не используется), тогда при проверке подписи их не придется вычислять заново. Можно хранить 6 комбинаций, начиная с уровня 1 $\left(C_0^{(1)}, C_1^{(1)}, C_2^{(1)}, C_3^{(1)}, C_0^{(2)}, C_1^{(2)}\right)$, тогда при проверке подписи сообщения 5 необходимо будет заново вычислить комбинацию $C_4^{(0)}$, а другие $\left(C_0^{(2)}, C_3^{(1)}\right)$ взять из таблицы и т.д. Отмеченный подход дает возможность достичь компромисс между быстродействием и требованиями к занятому количеству дискового пространства.

Отметим, что отказ от хранения комбинаций одного уровня дает возможность экономить память, но вычислительные расходы при этом растут почти вдвое, т.е. зависимость имеет экспоненциальный характер.

Функции хэширования. Функция хеширования может использоваться для выявления модификации сообщения. Т.е. она может служить криптографической контрольной суммой (ее называют также *кодом выявления изменений*, или *кодом аутентификации сообщения)*.

Хэш-функция H (функция хэширования; функция расстановки) — функция, которая используется для математического преобразования сообщения переменной длины M в блок данных фиксированной длины $H(M)$ (значение функции) и однозначно отображает произвольно выбранный открытый текст (аргумент функции).

Значение хэш-функции - множество значений целых чисел, которые принадлежат заданному диапазону и образованы в результате вычисления хеш-функции (рис. 8.23).

Теоретически возможно, что два разных сообщения могут быть сжаты в ту же свертку (так называемая ситуация «столкновения»). Поэтому для обеспечения стойкости функции хеширования необходимо предусмотреть способ избежать столкновения. Полностью столкновений избежать нельзя, поскольку в

общем случае количество возможных сообщений превышает количество возможных исходных значений функции хеширования. Однако вероятность столкновения должна быть низкой.

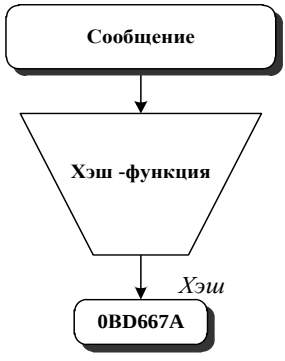


Рис. 8.23. Схематическое изображение хэш-функции

Для того чтобы функция хэширования могла быть должным образом использована в процессе аутентификации, она должна иметь такие свойства:

1) хэш-функцию H можно применить к аргументу любого размера;

2) исходное значение H является фиксированным;

3) значение $H(x)$ достаточно просто вычислить для любого x. Скорость вычисления хэш-функции должна быть такой, чтобы скорость генерирования и проверки ЭЦП при использовании хэш-функции была значительно больше, чем при использовании самого сообщения;

4) для любого y с вычислительной точки зрения невозможно найти x, такое что $H(x) = y$;

5) для любого фиксированного x с вычислительной точки зрения невозможно найти $x' \neq x$, где $H(x') = H(x)$.

Свойство 5 гарантирует, что нельзя найти другое сообщение, которое дает ту же свертку. Это предотвращает подделку и также позволяет использовать H в качестве криптографической контрольной суммы для проверки целостности.

Свойство 4 эквивалентно потому, что H является односторонней функцией. Стойкость систем с открытыми ключами зависит от того, что открытое криптопреобразование является односторонней функцией-ловушкой. Напротив, функции хэширования являются односторонними функциями, которые не имеют ловушки.

Пример. *Функция безопасного хеширования* (*SHA*).

Алгоритм безопасного хеширования SHA (Secure Hash Algorithm), принятый как стандарт США 1992 г. предназначен для использования вместе с алгоритмом цифровой подписи, определенным в стандарте DSS. При введении сообщения M алгоритм производит 160-битовое исходное сообщение,

названное сверткой (Message Digest), которая используется при генерировании ЭЦП.

Рассмотрим работу алгоритма подробнее. В первую очередь исходное сообщение дополняется так, чтобы его длина стала кратной 512 бит. При этом сообщение дополняется даже тогда, когда его длина уже кратна отмеченной. Процесс происходит таким образом: добавляется единица, потом столько нулей, сколько необходимо для получения сообщения, длина которого на 64 бит меньше, чем та, что кратная 512, и потом добавляется 64-битовое представление длины исходного сообщения.

Дальше инициирующие пять 32-битовых переменных такими 16-ричными константами:

$$A = 67452301;$$
$$B = EFCDAB89;$$
$$C = 98BADCFE;$$
$$D = 10325476;$$
$$E = C3D2E1F0.$$

Дальше эти пять переменных копируются соответственно в новые переменные b, c, a, b, c, d и главный цикл можно достаточно просто описать на псевдокоде как

$$\text{for}(t = 0; t < 80; t++)$$
$$\{\text{temp} = (a <<< 5) + f_t(b,c,d) + e + W_t + K_t;$$
$$e = d;\ d = c;\ c = b <<< 30;\ b = a;\ a = \text{temp};\},$$

где $<<<$ - операция циклического сдвига влево; K_t - 16-ричные константы, которые определяются такими формулами

$$K_t = \begin{cases} 5A827999, & t = 0,\ ...,19; \\ 6ED9EBA1, & t = 20,\ ...,39; \\ 8F1BBCDC, & t = 40,\ ...,59; \\ CA62C1D6, & t = 60,\ ...,79; \end{cases}$$

функции $f_t(x, y, z)$ задаются выражениями

$$f(x, y, z)_t = \begin{cases} X \wedge Y \vee \neg X \wedge Z, & t = 0,\ ...,19; \\ X \oplus Y \oplus Z, & t = 20,...,39,\ 0,\ ...,79; \\ X \wedge Y \vee X \wedge Z \vee Y \wedge Z, & t = 40,\ ...,59, \end{cases}$$

значения W_t образуются из 32-битовых подблоков 512-битового блока расширенного сообщения по правилу

$$W_t = \begin{cases} M_t, & t = 0,\ ...,19; \\ (W_{t-3} \oplus W_{t-8} \oplus W_{t-14} \oplus W_{t-16}) <<< 1, & t = 16,\ ...,79. \end{cases}$$

После окончания главного цикла значения *a*, *b*, *c*, *d* и прибавляются соответственно к содержимому *A*, *B*, *C*, *D* и *E*, и происходит переход к обработке следующего 512- битового блока расширенного сообщения. Исходное значение хэш-функции является конкатенацией значений *A*, *B*, *C*, *D* и *E*.

Инфраструктура открытых ключей (*Public Key Infrastructure — PKI*) *- это интегрированный комплекс методов и средств* (*набор служб*) *для обеспечения внедрения и эксплуатации криптографических систем с открытыми ключами (рис. 8.24).*

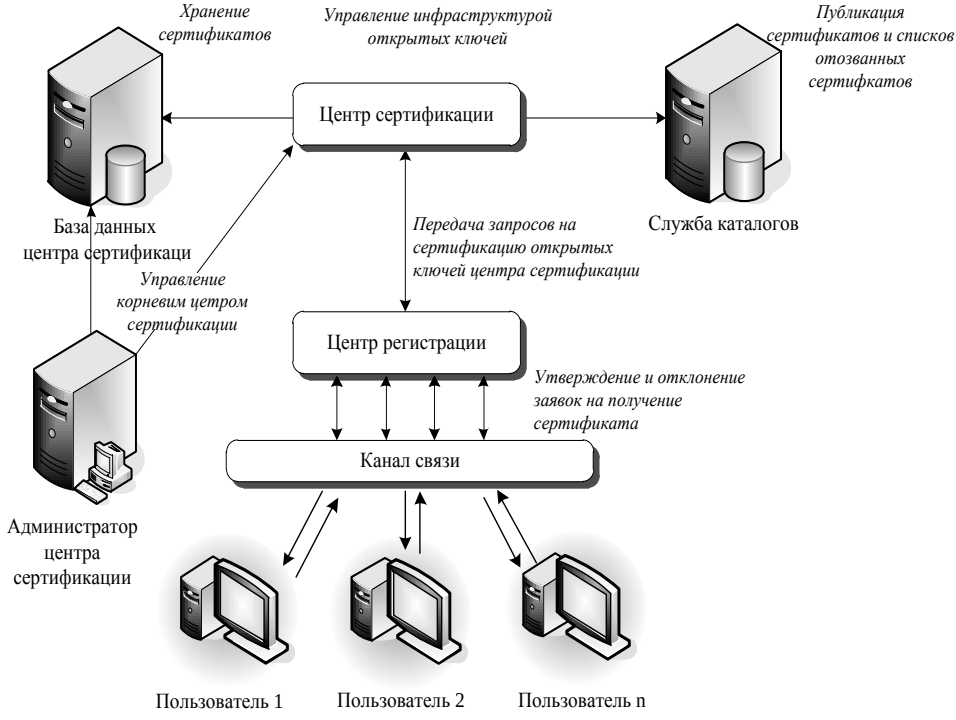


Рис. 8.24. Компоненты инфраструктуры открытых ключей

Технология PKI заключается в использовании двух математически связанных цифровых ключей, которые имеют такие свойства:

один ключ может использоваться для шифрования сообщения, которое может быть расшифровано только с помощью второго ключа;

даже если известен один ключ, с помощью вычислений невозможно определить второй. Один из ключей открыт для всех, а второй имеет частный характер и сохраняется в защищенном месте. Эти ключи могут использоваться для аутентификации или шифровки цифровой подписи электронных данных.

PKI служит не только для создания цифровых сертификатов, но и для хранения огромного количества сертификатов и ключей, обеспечения резервирования и возобновления ключей, взаимной сертификации, ведения списков аннулированных сертификатов и автоматического возобновления ключей и сертификатов по окончании срока их действия.

Основными атрибутами сертификата является имя и идентификатор субъекта, информация об открытом ключе субъекта, имя, идентификатор и цифровая подпись уполномоченного из выдачи сертификатов, серийный номер, версия и срок действия сертификата, информация об алгоритме подписи и тому подобное. Важно, что цифровой сертификат содержит цифровую подпись на основе секретного ключа доверительного центра.

Центр сертификации (Certificate Authority - CA), или ***доверительный центр*** - объект, уполномоченный создавать, подписывать и публиковать сертификаты. Центр имеет также полномочия идентифицировать пользователей. Основными операциями, которые выполняет доверительный центр, являются издание, возобновление и аннулирование сертификата.

Действия Центра сертификации ограничены политикой сертификации, которая диктует ему, какую информацию он должен вмещать в сертификат. Центр сертификации публикует свою политику сертификации так, чтобы пользователи могли проверить соответствие сертификатов этой политике.

Список аннулированных сертификатов (Certificate Revocation List — Crl) ***CRL*** - список сертификатов, признанных недействительными в период их действия в случае компрометации секретного ключа или изменения атрибутов сертификата с момента его выпуска.

Хранилище сертификатов - специальный объект PKI, где сохраняются выпущенные сертификаты и списки отозванных сертификатов. Оно не является обязательным компонентом PKI, но значительно упрощает доступ к ресурсам и управлению системой.

К хранилищу предъявляются такие требования: простота доступа; доступ должен быть стандартным; возобновление информации; встроенная защищенность; простое управление; совместимость с другими хранилищами (необязательное требование).

Хранилище упрощает систему распространения сертификатов.

Фактически действующим стандартом доступа к хранилищу является LDAP (Light-weight Directory Access Protocol), упрощен протокол доступа к каталогу. Он наиболее адекватен как стандарт для хранения и вытягивания сертификатов после их генерации, поддерживается большинством серверных операционных систем и баз данных и достаточно открытый для того, чтобы его могли поддерживать практически любые инфраструктуры с открытыми ключами.

Центр регистрации (Registration Authority — RA) является дополнительным компонентом системы PKI, которая дает возможность авторизированному CA аутентифицировать пользователей и проверять информацию,

которая заносится в сертификат. К его функциям могут принадлежать генерирование и архивирование ключей, сообщения об аннулировании сертификатов и тому подобное. В некоторых системах CA выполняет функции RA. CA выдает сертификат RA (если он присутствует в системе), причем RA выступает как объект, подчиненный CA. Но RA не может выпускать сертификаты.

Конечный пользователь (End Entity — EE) — пользователь сертификата PKI и владелец сертификата. То есть конечный пользователь — это объект, который использует некоторые услуги и функции системы PKI. Конечный пользователь может быть владельцем сертификата или объектом, который спрашивает сертификат.

В настоящее время не существует общепризнанного аналога срока, который берет начало в отрасли шифрования с открытыми ключами, — Certificate Authority. Это понятие получило много разных названий: *служба сертификации*, *уполномоченный из выпуска сертификатов, распорядитель сертификатов, орган выдачи сертификатов, доверительный центр, центр сертификации и т.п.*

Взаимодействие между разными компонентами инфраструктуры открытых ключей иллюстрирует рис. 8.25.

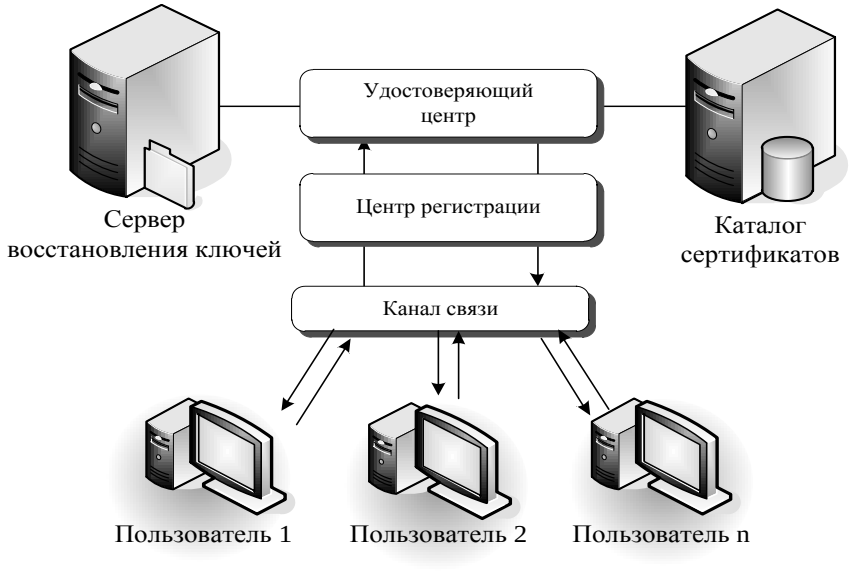


Рис. 8.25. Взаимодействие между компонентами инфраструктуры открытых ключей

8.5. Стеганографические методы защиты информации

В переводе с греческого *стеганография* означает тайнопись (steganos - тайна, секрет; graphy - запись), а ее история насчитывает тысячелетие. Стеганографическую защиту информации можно реализовать с помощью разных технических, химических, физических и психологических методов.

История стенографии достигает времен создания клинописных табличек (знаки, которые выжимались на сырой глине, — Шумер, в 3000 г. до н. е.), где первая надпись (секретный текст) замазывалась слоем глины, на котором выполнялась другая клинопись. Сегодня считается, что родиной стеганографии является Египет, но в литературе первое упоминание о ней было в Древней Греции, где текст наносился на деревянные дощечки с восковым покрытием. Еще во времена историка Геродота известен случай с греком Демератом, который был в изгнании в Персии и передал в Спарту сообщение о плане царя персов Ксеркссеса захватить Грецию. Демерат использовал обычную дощечку со слоем воска, под которым и было написано соответствующее сообщение. Когда охранники обнаружили дощечку, на которой было послание, то они восприняли ее как обычную заготовку для записей, которая не вызывала в них никакого подозрения. Известно, что в Древнем Китае письма писались на полосках шелка. Для укрывательства записанных сообщений такие полоски свертывали в шарики, надежно покрывали слоем воска, после чего глотались курьером и доставлялись им соответствующему адресату.

Значительное количество трудов в направлении стеганографической и криптографической защиты информации приходится на XVI—XVII ст. Первой из известных трудов была книга немецкого аббата Плохая Тритемия (1462—1516) «Стеганография» (1499), которая описывала системы гадания и пророчества, но включала сложную систему защиты данных от несанкционированного доступа. В труде Гаспара Скотта (1608—1666) под названием «Стеганография» (1665) раскрывается возможность укрывательства текста в музыкальной партитуре, где каждой ноте отвечает определенная буква. Одной из известных фигур того времени был епископ Джон Вилкинс (1614—1672), который разработал много стеганографических схем, — от невидимых чернил (1641) к укрывательству сообщений в музыкальных произведениях и изображением геометрических фигур — и описал принципы частотного анализа.

Стеганография не заменяет, а дополняет криптографию. Сокрытие сообщения методами стеганографии значительно снижает вероятность выявления самого факта передачи сообщения. А если это сообщение к тому же зашифровано, то оно имеет еще один, дополнительный, уровень защиты.

Стеганография *— это совокупность методов утайки факта укрывательства исходного информационного сообщения* (*открытого текста*) *в другом сообщении, которое имеет аналоговую природу* (*оцифрованный непрерывный сигнал*) *с целью обеспечения конфиденциальности, т.е. предотвра-*

щение исключения полезной информации из информационного потока несанкционированным пользователем.

Модели и задачи систем стеганографического шифрования. Классификация известных стеганографических методов приведена на рис. 8.26.

Рис. 8.26. Классификация стеганографических методов

Материальные методы *— используют для укрывательства информации на базе физических или химических свойств стеганографического контейнера* (*объекта, в какой внедряется секретная информация*) *или просто контейнера, а также средств внедрения в него информации.*

Такими свойствами, например, может быть прозрачность, габаритные размеры, цвет контейнера или способность внедренной информации проявляться в результате определенного влияния. Разработка и исследование таких стеганографических методов связана с изучением свойств различных материальных носителей информации и соответствующих способов (не общепринятых) ее внедрения.

К материальным стеганографическим методам можно отнести невидимые чернила, микроточки и т.п. В современном контексте стандартными носителями информации является аудио-, видео- и вычислительная техника.

Информационные методы *- используют для укрывательства данных на базе свойств информационного наполнения контейнера*. Методы этого класса разделяют на лингвистические и цифровые.

Лингвистические методы используют избыточность языка или другой среды, которая не содержит ни букв, ни цифр (рисунки, взаимное расположение объектов и т.п.). К этому классу можно также отнести метод генерации текста, необходимого для укрывательства секретного сообщения, а также методы, которые базируются на изменении положения строк на странице или слов в предложении и т.п.

Цифровые методы базируются, с одной стороны, по большей части на том, что файлы, которые не нуждаются в абсолютной точности, могут быть несколько видоизменены без потери функциональности, а с другой - на отсутствии специального инструментария или неспособности органов чувств человека различать незначительные изменения в таких файлах (рис. 8.27).

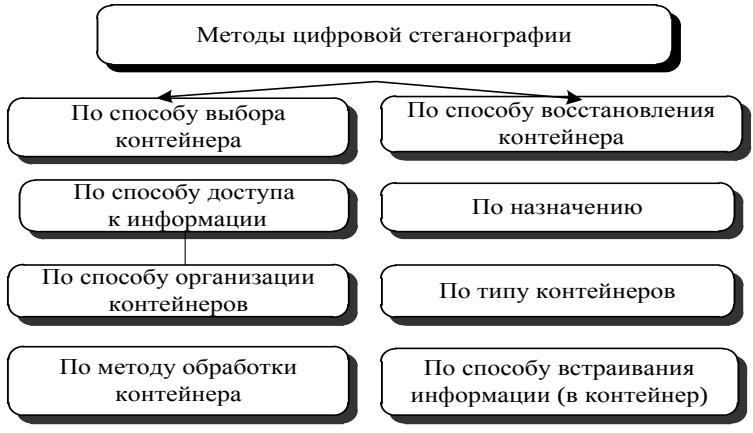


Рис. 8.27 Классификация методов цифровой стеганографии

1. ***По способу выбора контейнера*** различают методы стеганографии безальтернативной, суррогатной, стеганографии, что выбирает, и стеганографии, что конструирует. Безальтернативная стеганография использует определенный заранее известный контейнер и не предусматривается возможность его выбора. Суррогатная стенография — может быть применена к определенному контейнеру, случайно выбранному из набора допустимых. Стеганография, что выбирает, базируется на генерации большого количества альтернативных контейнеров, из которых дальше за отдельными характеристиками выбирается наиболее эффективный для внедрения информации. Стеганография, что конструирует, предусматривает в зависимости от сообщения саму генерацию контейнера, при этом моделируются его необходимые статистические свойства.

2. ***По способу доступа к информации*** различают поточные и фиксированы методы. Поточные характеризуются включением в контейнер в режиме реального времени непрерывных битовых потоков данных сообщения, а со-

ответствующие контейнеры имеют большие размеры, что делает их привлекательными для передачи значительных объемов секретной информации.

Фиксированные методы используют контейнеры определенной длины, например текстовые, кодовые, графические, звуковые или другие файлы. В отличие от поточных здесь заранее известные длина контейнера и его наполнения, а следовательно, могут быть априорно оценены с точки зрения их эффективности относительно заданного сообщения и стеганографического преобразования. Чаще на практике используют именно контейнеры фиксированной длины как более удобные и доступные.

3. ***По способу организации контейнеров*** различают систематические и несистематические методы стеганографии. В систематических информационные биты контейнера можно отделить от шумовых, в которые и будет внедрена секретная информация. В несистематических методах такого деления нет, а потому для выделения сообщения необходимо прорабатывать все биты контейнера.

4. ***По способу возобновления сообщения*** стеганографические методы бывают эталонные и безэталонные. При возобновлении секретной информации в некоторых стеганографических методах необходимо иметь эталон контейнера, чтобы обеспечить его надежное хранение и защита от несанкционированного использования. Большинство современных методов не нуждаются в наличии такого эталона контейнера, а для его формирования осуществляется специальная обработка стеганограммы.

5. ***По методу обработки контейнера*** цифровые методы разделяются на непосредственные и спектральные. При использовании непосредственных методов обработке подлежат биты самого контейнера, как например, в методе наименьшего значимого бита. Спектральные методы базируются на использовании дискретных унитарных преобразований, например Фурье, Уолша, Вейвлета, высокой и низкой корреляции и др. При использовании методов этой группы обработке подлежит соответствующий спектральный контейнер.

6. ***По назначению*** различают методы, направленные на скрытую передачу информации (информацию внедрено в контейнер), защиту прав на цифровые объекты (защите подлежит контейнер), аутентификации данных и скрытую аннотацию документов.

В пределах стеганографии защита прав на цифровую интеллектуальную собственность осуществляется с помощью внедрения в объект защиты цифровых водяных знаков, что активно применяется для защиты от копирования и несанкционированного использования цифровых фотографий, аудио- и видеозаписей и других данных.

7. ***По типу контейнеров*** различают методы, которые используют текстовые, графические, аудио- и видеосреды. Каждый выделенный класс сориентирован на максимальное использование особенностей соответствующей среды. Например, графические методы используют особенности человече-

ского зрения, такие как чуствительность к контрасту, размеру, форме, цвету, местоположению. Аудиометоды используют модель человеческого слуха и основные психоакустические принципы.

8. ***По способу встраивания информации*** **(*в контейнер*)** методы разделяются на форматные и неформатные. Форматные базируются на особенностях формата хранения данных, которые представляют собой файл-контейнер. В рамках таких методов формат хранения пустого контейнера анализируется с целью отыскания тех служебных полей в заглавии файла, изменение которых в конкретных условиях не повлияет на функциональность контейнера. Это могут быть служебные поля, которые не используются современными программами, не полностью заполнены поля комментариев и т.п. Неформатные основываются на внедрении информации не в заглавие, а непосредственно в данные пустого файла-контейнера. Они базируются на двух принципах. Во-первых, некоторые виды файлов-контейнеров не нуждаются в абсолютной точности представления своих внутренних данных, например файлы, которые содержат оцифрованные изображение или звук, могут быть в некоторой степени видоизменены без потери их функциональности. Во-вторых, сигнальная система человека неспособна отличить незначительные градации, например в цвете изображения или в качестве звука. Неформатные методы всегда ведут к появлению в контейнере дополнительного шума, который инициирует несущественное ухудшение изображения или звука, но они являются более перспективными (сравнительно с форматными) благодаря лучшей стойкости и пропускной способности создаваемого стеганографического канала.

Стеганографические системы и их модели. В настоящее время стеганография фактически является одним из путей поддержки информационной безопасности и дает возможность организовать связь, которая скрывает сам факт наличия секретных данных. Стенографические методы активно используются для защиты информации от несанкционированного доступа, противодействия системам мониторинга и управления ресурсами сетей, маскировки программного обеспечения от незарегистрированных пользователей, защите авторского права на некоторые виды интеллектуальной собственности, а также для аутентификации цифровых объектов.

Стеганографическая система (***стеганосистема***) *является совокупностью методов и средств относительно формирования секретного информационного потока данных с целью укрывательства факта передачи полезной информации* (*открытого текста*).

Стеганографическим контейнером *называется сообщение, в которое будет размещена* (*скрыто*) *полезная информация или секретные данные.*

Любой файл или поток данных может быть *цифровым контейнером*, если контейнер не содержит секретного сообщения, то его называют пустым, а тот, который содержит полезные данные, — *заполненным*, или стеганоконтейнером (*стеганограммою*).

Стеганографическим *каналом* (***стеганоканал***) *называется информационно коммуникационный канал связи, по которому передается стеганоконтейнер.*

Секретный ключ, необходимый для внедрения информации в контейнер, называется *стеганоключом*, или просто *ключом*. В зависимости от количества уровней защиты в стеганосистеме может использоваться один и больше ключей.

Факт отправления контейнера к получателю не должен быть подозрительным и не должно наблюдаться заметных отклонений контейнера от нормы.

Обозначим через C (container) множество всех возможных контейнеров, через P (plaintext) множество всех возможных открытых текстов, а $|C|$ и $|P|$ — количество контейнеров и количество открытых текстов соответственно в C и P, так как $|P| \le |C|$. Для упрощения предположим, что любой открытый текст из P есть бинарным $P = \{0,1\}^p$. Модель стеганографической системы можно подать схематично (рис. 8.28).

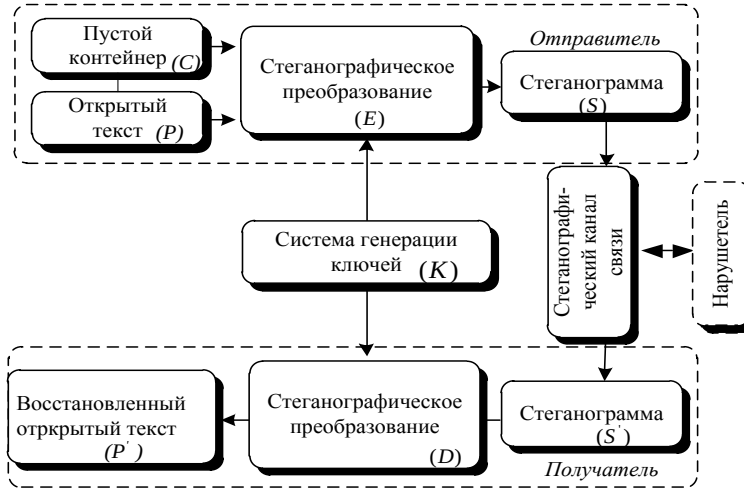


Рис. 8.28. Обобщенная модель стеганографической системы

Стенографические системы разделяются на такие виды:

бесключевые, ***симметричные*** (*одноключевые, системы с секретным ключом*);

асимметричные (*двуключевые, системы с открытым ключом*).

Безключевой стеганосистемой называется совокупность (C, P, E, D), базирующаяся на стеганографических преобразованиях ($E: C \times P \to C$ и $D: C \to P$), которые применяются для внедрения и восстановление открыто-

го текста из контейнера, причем $D\left[E(c, p)\right] = p$ для $\forall p \in C$, где C - множество возможных контейнеров; P - множество возможных открытых текстов.

Из определения вытекает, что функционирование безключевых стеганосистем происходит без стеганографического ключа, и потому их безопасность базируется на укрывательстве функций E и, что противоречит принципу Керкгофа, в соответствии с которым стойкость системы должна определяться только секретностью ключа.

Симметричной стеганосистемой *называется совокупность* (C, P, K, E_k, D_k), *где* C *- множество возможных контейнеров;* P *- множество возможных открытых текстов;* K *- алгоритм генерации ключей;* E_k *- преобразования, которые моделирует процесс вживления (если в качестве начальных данных взяты* $c \in C$, $p \in C$ *и ключ* (k), *порожденный* K) *и создает стеганоконтейнер* $s \in C$; D_k *- преобразования, которые моделирует процесс восстановления, если* $s' = s$ *и* k, *и формирует бинарное* p. *Если объект* s *содержал скрытый открытый текст* p, *то* $p' = p$.

Такой тип стеганосистем нуждается в наличии закрытого информационного потока данных, недоступного никому, кроме отправителя и получателя и предназначенного для обмена стеганографическими ключами. Отметим, что для некоторых алгоритмов при возобновлении скрытой информации необходим начальный контейнер или другие данные, которых нет в стеганограме. Такие алгоритмы будем считать случаем части симметричных стеганосистем, для которых $k = c$ или $k = c \times k'$, где k' — дополнительный набор секретных ключей.

Асимметричной стеганосистемой *называется совокупность* $(C, P, K, E_{K1}, D_{K2})$, *где* C *- множество возможных контейнеров;* P *- множество возможных открытых текстов;* K *- алгоритм генерации ключей, который порождает пары ключей* (k_1, k_2), *где* k_1 *и* k_2 *соответственно открытый и секретный ключе, которые используются для внедрения и восстановление открытого текста;* $E_{K1} : C \times P \times k_1 \to C$ *- преобразования, которые моделирует процесс внедрения открытого текста;* $D_{K2} : C \times k_2 \to P$ *- преобразования, которые моделируют процесс восстановления открытого текста, причем*

$$D_{K2}\left[E_{K1}(c, p, k_1, k_2)\right] = p \text{ при } \forall p \in P \text{ и } \forall c \in C.$$

При построении стеганосистем необходимо учитывать следующие положения.

Стойкость системы полностью определяется секретностью ключа, с помощью которого можно установить факт наличия внедренного открытого текста и его содержание, а структура стеганосистемы, детали ее реализации, характеристики множественных чисел открытых текстов и контейнеров известны. Единственная неизвестная информация — это ключ.

Только при наличии соответствующего стеганоключа можно найти и возобновить скрытый открытый текст сообщения или доказать его существование.

Если становится известным факт существования скрытого открытого текста, то это не должно давать возможность его возобновить из контейнера, пока ключ сохраняется в секрете.

Отсутствует возможность статистически доказать существование скрытого открытого текста, а его отыскание без знания ключа является сложной вычислительной задачей.

Обеспечивается необходимая пропускная способность создаваемого системой стеганографического канала связи.

Стеганосистема должа быть приемлема по сложности вычислительной реализации.

Нарушитель не должен иметь никаких технических или других преимуществ перед пользователем в распознавании или раскрытии содержания скрытых открытых текстов.

Основные характеристики стеганографических систем: *стойкость*; *вычислительная сложность*; *пропускная способность стеганографичних каналов связи*.

Рассмотрим подробнее стойкость стеганосистемы, с помощью которой оценивается безопасность ее использования.

Стойкостью стеганосистемы *называется способность системы скрывать от несанкционированного пользователя факт передачи внедрения открытых текстов и противостоять попыткам разрушить, обезобразить, возобновить или заменить их, а также способность подтвердить или опровергнуть подлинность информации.*

В стеганографии от нарушителя скрывается сам факт существования секретной информации.

Стойкой стеганосистемой *является система, для которой нарушители, наблюдая за информационным обменом между отправителем и получателем, не смогут обнаруживать, а тем более считывать внедренные в контейнер скрытые данные.* По уровню обеспечения секретности стеганографические системы разделяются на теоретически стойкие, практически стойкие и неустойчивые системы (рис. 8.29).

Рис. 8.29. Классификация стеганосистем по уровню обеспечения секретности

Принципы защиты информации в стеганографических системах. Нарушитель стеганографических систем (рис. 8.30) может быть пассивным, активным, злонамеренным и в зависимости от этого он может создавать разные угрозы.

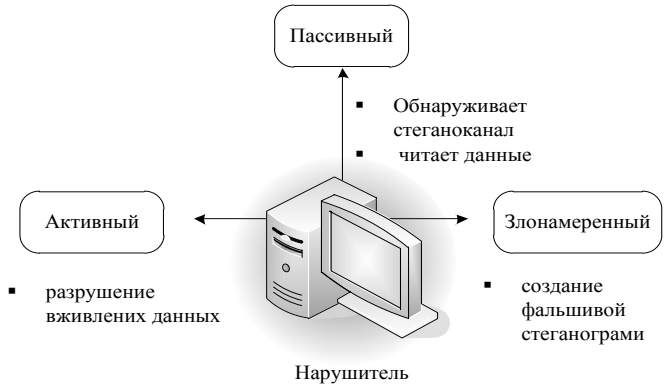


Рис. 8.30. Возможные типы нарушителей в стеганографических системах

Пассивный нарушитель может лишь обнаружить факт наличия стеганоканала и (возможно) читать внедренные данные. Активный нарушитель может влиять на информацию, которая проходит каналом связи (не только обнаруживать и читать скрытые данные, но и полностью или частично разру-

шать их). Такой нарушитель может изменять контейнер независимо от того, пустой он или заполненный, тем более, что разрушить внедренные данные иногда легче, чем прочитать. Злонамеренный нарушитель наиболее опасен, он имеет возможность не только разрушать, но и создавать фальшивую стеганограму.

Для осуществления той или другой угрозы нарушитель (аналитик) в основном применяет типы атак, приведенные на рис. 8.31.

Рис. 8.31. Типы атак на стеганографическую систему

Атака на основе известного заполненного контейнера. Нарушитель имеет одну или несколько стеганограмм и выполняет задание обнаруживать факт существования стеганоканала (основное), а также возобновлять открытый текст или определять ключ для возможности анализа других стеганограмм.

Атака на основе известного внедренного открытого текста. На основе нескольких примеров скрытых открытых текстов и соответствующих стеганограмм нарушитель осуществляет соответствующий анализ для получения ключа. Такие атаки в значительной мере присущи системам защиты интеллектуальной собственности, когда как водяной знак используется известный логотип фирмы.

Атака на основе выбранного скрытого открытого текста. В этом случае аналитик (стеганоаналитик) имеет возможность предлагать для передачи собственные открытые тексты и анализировать стеганограммы.

Адаптивная атака на основе выбранного скрытого открытого текста. Эта атака является случаем части предыдущей и характеризуется возможностью аналитика адаптивный выбирать сообщение для навязывания в зависимости от результатов анализа предыдущих стеганограмм.

Атака на основе выбранного заполненного контейнера. Стеганоаналитик имеет возможность создавать стеганограмы для выбранного им открытого текста с целью определения образцов стеганограмм, что могут идентифицировать использование анализируемой стеганосистемы.

Атака на основе известного пустого контейнера. В этом случае стеганоаналитик сравнением известного пустого контейнера с предусмотренной стеганограммой может всегда определить факт наличия стеганоканала. Контейнер может быть известен приближенно, с некоторой погрешностью. В этом случае есть возможность построения стойкой стеганосистемы.

Атака на основе выбранного пустого контейнера. В этом случае стеганоаналитик должен возможность заставить отправителя пользоваться предложенным контейнером.

Атака на основе известной математической модели контейнера или его части. При этом атакующий пытается определить разницу между подозрительным внедренным открытым текстом и известной ему моделью. Отправитель и атакующий могут иметь разные модели, тогда выигрывает тот, кто имеет лучшую модель.

В общем случае сигналы-контейнеры могут поддаваться возмущениям двух типов:

1. *Стандартные процедуры «легальной» обработки контейнеров всегда связаны с внесением в них дополнительного шума.*

2. *Контейнер может поддаваться определенному влиянию со стороны активного или злонамеренного нарушителя.*

Например, возмущения, применяемые к контейнерам изображений, можно классифицировать так:

аддитивний и мультипликативный шум (гауссовский, равномерный, дискретный и другой);

линейная фильтрация (низкочастотная, высокочастотная, узкополосная) и приглаживание;

нелинейная фильтрация (фильтрация медианы и др.);

сжатие с потерями (например, JPEG-компресия);

исключение строк или столбцов пикселей изображения;

локальные или глобальные преобразования (линейные преобразования со сдвигом (в частности, вращение), изменение масштаба в обе стороны, перенесения);

редукция данных (например, путем вырезки или добавления части данных);

композиция данных (например, путем встраивания дополнительного логотипа и т.д.);

изменение формата данных (например, путем перехода от формата GIF к JPEG).

Рассмотренные возмущения по большей части не преднамеренные, а являются следствием естественных возмущений, присущих каналам передачи

данных и разным вариантам обработки информации типа печати, сканирования и т.п. Известен список возмущений, которые могут преднамеренно вноситься в контейнер (большинство из них были перечислены раньше). Дополнительно могут применяться такие возмущения:

геометрические искажения: вращение на малый угол, пространственное масштабирование, внесение нелинейных искажений в отдельные области изображения, сдвиг изображения, вырезки текстуры и встраивания ее как текстура в другие области;

добавление постоянного сдвига в значение пикселей;

локальная перестановка отдельных пикселей;

квантование изображения и изменение варианта квантования;

последовательное приложение преобразований типа аналог/код и код/аналог;

встраивание сообщения в изображение, которое уже содержит сообщение;

редукция цветов;

печатание изображения и замена исходного изображения на отсканированное;

изменение размера изображения (в пикселях).

Количественная оценка стойкости стеганографической системы защиты от внешних влияний является сложным заданием, которое на практике обычно выполняется методами системного анализа, систематического моделирования или экспериментального исследования. Как правило, стеганосистема должна обеспечивать следующую модель защиты информации (рис. 8.32).

Рис. 8.32. Трехуровневая стеганографическая модель защиты информации

Рассмотрим модель нарушителя, который пытается противодействовать укрывательству информации. Вслед за К. Шенноном и согласно принципу Д. А. Керкхофса (Dutchman A. Kerckhoffs) назовем эту модель теоретико-информационной, считая, что нарушителю известно полное описание стеганосистемы и вероятностные характеристики сообщения (какие скрываются), контейнеров, ключей, стеганограмм, а также то, что он имеет неограниченные вычислительные ресурсы, устройства запоминания достаточно большой вместимости, имеет в своем распоряжении бесконечно большое время для стеганоанализа и ему известно достаточно большое число перехваченных стеганограм. Единственное, что неизвестно нарушителю, — ключ стеганосистемы.

Если в соответствии с данной моделью нарушитель не в состоянии обнаружить стеганограмму, то назовем такую стеганосистему теоретически информационно стойкой к атакам пассивного нарушителя, или совершенной.

Нужно различать стойкость разных систем, например, относительно факта передачи (существование), возобновления и уничтожения скрываемой информации; навязывание фальшивых сообщений в канале скрытой связи (имитоскойкости); возобновление секретного ключа стеганосистемы.

Методы компьютерной стеганографии. ***Метод наименьшего значимого бита*** (НЗБ) или LSB-метод (Least Significant Bits) является одним из наиболее распространенных не форматных методов. Его сущность заключается в замене нескольких младших битов в байтах данных файла-контейнера битами скрываемого сообщения. Существуют модификации LSB-методу, которые отличаются выбором подмножества данных файла-контейнера, в которые внедряется сообщение, стратегией изменения значений данных и другими деталями. Абсолютное большинство из модификаций метода наименьшего значащего бита сориентировано на графические растровые форматы файлов-контейнеров.

Спектральные методы цифровой стеганографии. Рассмотренный метод наименьшего значимого бита имеет ряд преимуществ, связанных с простотой реализации (и, как следствие, высоким быстродействием программных продуктов на его основе) и сравнительно высокой пропускной способностью созданного стеганоканалу. Однако его можно использовать только для решения заданий, которые не требуют высокой стеганостойкости, поскольку LSB-метод не является стойким практически относительно всех активных атак противника, например относительно дополнительного зашумления, фильтрации, конвертации цветов, сжатия с потерями, геометрических преобразований и т.п.

Более стойкими к разнообразным искажениям являются методы, которые скрывают сообщение в спектральной области *файла-контейнера*. В отличие от LSB-метода, здесь возможно внедрение информации в те зоны контейнера-изображения, которые несут наиболее существенную информацию, - в низкие частоты. Попытка возобновить или отобразить сообщение в таком

случае может привести к заметным искажениям самого изображения, что сделает его непригодным для эксплуатации. Спектральные методы цифровой стеганографии достаточно разнообразны и некоторые из них используются в комбинации с LSB-методом и широкополосной модуляцией.

Для частотного представления данных файла-контейнера используют дискретные ортогональные преобразования, такие как дискретное преобразование (ДКП) косинуса, дискретное преобразование Фурье, вейвлет-преобразования, преобразования Карунена - Лоева, Адамара, Хаара и др.

Приложение к контейнерам-изображений дискретных ортогональных преобразований выполняет три основных задания (рис. 8.33).

В общем случае дискретное изображение является матрицей отсчетов функции, которая описывает распределение яркости на условно непрерывном изображении . Использование дискретного ортогонального преобразования может значительно снизить межэлементную корреляцию, при этом достигается концентрация максимально возможной части энергии выходного дискретного сигнала в минимально возможном количестве спектральных коэффициентов.

Рис. 8.33. Основные задачи применения дискретных ортогональных преобразований к контейнерам-изображениям

8.6. Методы криптоанализа

Криптоанализ - это совокупность методов относительно получения и восстановление открытого текста (полезного сообщения) из зашифрованного текста без знания секретного ключа (алгоритма или математической функции восстановления).

Успешно проведенный криптоанализ может раскрыть открытый текст или ключ. Он может также обнаружить слабые места в криптосистемах, что в конечном итоге приведет к предыдущему результату. (Раскрытие ключа не криптологическими способами называется компрометацией.)

Попытка реализации криптоанализа называется *раскрытием***.** Основное предположение криптоанализа, в первый раз сформулированное в XИХ столетии Датчманом А. Керкхофсом, заключается в том, что безопасность полностью определяется ключом. В реальном мире криптоаналитики не всегда имеют доклад-ну информацию, такое предположение является производительной рабочей гипотезой. Существует несколько типов криптоаналитического раскрытия открытого текста (рис. 8.34).

Рис. 8.34. Типы криптоаналитических раскрытий открытого текста

Относительно каждого из них, как правило, предусматривается, что криптоаналитик имеет всю полноту знания об используемом алгоритме шифрования:

1. *Раскрытие с использованием только шифротекста*. У криптоаналитика есть шифротексты нескольких сообщений, зашифрованных тем самым алгоритмом шифрования. Задание криптоаналитика заключается в раскрытии открытого текста как можно большего количества сообщений или, что лучше, в получении ключа (ключей), использованного для шифрования сообщений, для дешифрации других сообщений, зашифрованных тем самым ключом.

Дано: $C_1 = E_k(P_1),\ C_2 = E_k(P_2),\ ...,\ C_i = E_k(P_i).$

Получить: или $P_1, P_2, \ldots, P_i$; k; или алгоритм получения P_{i+1}, если известно $C_{i+1} = E_k(P_{i+1})$.

2. *Раскрытие с использованием открытого текста.* У криптоаналитика есть доступ не только к шифротекстов нескольких сообщений, но и к открытому тексту этих сообщений. Его задание заключается в получении ключа (или ключей), использованного для шифрования сообщений, для дешифрации других сообщений, зашифрованных тем самым ключом (ключами).

Дано: $P_1,\ C_1 = E_k(P_1),\ P_2,\ C_2 = E_k(P_2),\ ...,\ P_i,\ C_i = E_k(P_i).$

Получить: или k; или алгоритм получения P_{i+1}, если известно $C_{i+1} = E_k(P_{i+1})$.

3. *Раскрытие с использованием выбранного открытого текста*. У криптоаналитика не только есть доступ к шифротекстов и открытых текстов нескольких сообщений, но и возможность выбирать открытый текст для шифрования. Это предоставляет больше вариантов, чем раскрытие с использованием открытого текста, поскольку криптоаналитик может выбирать шифрованные блоки открытого текста, который может дать больше информации о ключе. Его задание заключается в получении ключа (или ключей), использованного для шифрования сообщений, или алгоритма, который дает возможность дешифровать новые сообщения, зашифрованные тем самым ключом (или ключами).

Дано: $P_1, C_1 = E_k(P_1), P_2, C_2 = E_k(P_2), \ldots, P_i, C_i = E_k(P_i),$ где криптоаналитик может выбирать $P_1, P_2, \ldots P_i$.

Получить: или k; или алгоритм получения P_{i+1}, если известно $C_{i+1} = E_k(P_{i+1}).$

4. *Адаптивное раскрытие с использованием открытого текста*. Это частный случай раскрытия с использованием выбранного открытого текста. Криптоаналитик не только может выбирать шифрованный текст, но также может строить свой следующий выбор на базе полученных результатов шифрования. При раскрытии с использованием Выбранного открытого текста криптоаналитик мог выбрать для шифрования только один большой блок открытого текста, при адаптивном раскрытии с использованием выбранного открытого текста он может выбрать меньший блок открытого текста, потом выбрать следующий блок, используя результаты первого выбора, и так далее

Существует по крайней мере еще три типа криптоаналитического раскрытия.

5. *Раскрытие с использованием выбранного шифротекста*. Криптоаналитик может выбрать разные шифротексти для дешифрации и имеет доступ к дешифрованным открытым текстам. Например, у криптоаналитика есть до-

ступ к «черному ящику», который выполняет автоматическую дешифрацию. Его задание заключается в получении ключа.

Дано: $C_1, P_1 = D_k(C_1), C_2, P_2 = D_k(C_2), \ldots C_i, P_i = D_k(C_i)$

Получить: k.

Раскрытие из использованного выбранного шифротекста также эффективно для симметричных алгоритмов. (Иногда раскрытие с использованием выбранного открытого текста и раскрытия с использованием выбранного шифротекста вместе называют раскрытием с использованием выбранного текста.)

6. *Раскрытие с использованием выбранного ключа*. Такой тип раскрытия значит не то, что криптоаналитик может выбирать ключ, а что у него есть некоторая информация о связи между разными ключами.

7. *Преступный криптоанализ*. Криптоаналитик угрожает, шантажирует или истязает кого-нибудь, пока не получит ключ. Взяточничество иногда называется раскрытием с покупкой ключа. Это мощные способы раскрытия, которое часто является наилучшим путем сломать алгоритм.

Безопасность алгоритмов. Разные алгоритмы предоставляют разные уровни безопасности в зависимости от того, насколько трудно сломать алгоритм. Если стоимость излома алгоритма выше, чем стоимость зашифрованных данных, вы, скорее всего, в безопасности. Если время излома алгоритма больше, чем время, в течение которого зашифрованы данные должны сохраняться в секрете, то вы также, скорее всего, в безопасности. Если объем данных, зашифрованных одним ключем, меньше, чем объем данных, необходимый для излома алгоритма, и тогда вы, скорее всего, в безопасности.

Важно, чтобы значимость данных всегда оставалась меньше, чем стоимость излома системы безопасности, которая защищает данные.

Ларс Кнудсен (Lars Knudsen) разбил раскрытие алгоритмов за такими категориями (приведенными в порядке уменьшения значимости):

1. *Полное раскрытие*. Криптоаналитик получил ключ *K*, такой что $D_k(C) = P$.

2. *Глобальная дедукция*. Криптоаналитик получил альтернативный алгоритм *A*, эквивалентный $D_k(C)$ без знания *K*.

3. *Местная* (*или локальная*) *дедукция*. Криптоаналитик получил открытый текст для перехваченного шифротекста.

4. *Информационная дедукция*. Криптоаналитик получил некоторую информацию о ключе или открыт текст. Такой информацией могут быть биты ключа, ведомости о форме открытого текста и тому подобное.

***Алгоритм является безусловно безопасным**, если независимо от объема шифротекстов у криптоаналитика информации для получения открытого текста недостаточно. В сущности, только шифрование одноразовыми блокнотами невозможно раскрыть в случае бесконечных ресурсов.*

Все другие криптосистемы поддаются раскрытию с использованием сведений из полученного шифротекста простым перебором возможных ключей. Это называется раскрытием *грубой силой*.

Криптография больше интересуется криптосистемами, которые трудно сломать вычислительным способом. Алгоритм считается *вычислительный безопасным* (или, как иногда его называют, сильным), если он не может быть сломан с использованием доступных ресурсов теперь или в будущем. Срок «доступные ресурсы» достаточно расплывчатым.

Сложность раскрытия можно измерять разными способами:

1. *Сложность данных*. Объем данных, используемых на входе операции раскрытия.

2. *Сложность обработки*. Время, нужное для проведения раскрытия. Часто называется коэффициентом работы.

3. *Требования к памяти*. Вместимость памяти, необходимая для раскрытия.

Как эмпирический метод сложность раскрытия определяется по максимальному из этих три коэффициентов. Некоторые операции раскрытия допускают взаимосвязь коэффициентов: более быстрое раскрытие возможно за счет увеличения требований к памяти.

Сложность выражается порядком величины. Если сложность обработки для данного алгоритма составляет 2128, то 2128 операций нужно для раскрытия алгоритма. (Эти операции могут быть сложными и длительными.) **Например,** если предусматривается, что ваши вычислительные мощности способны выполнять миллион операций за секунду и вы используете для развязывания задачи миллион параллельных процессоров, и на получение ключа вам понадобится свыше 1019 лет, что в миллиард раз превышает время существования Вселенной.

Тогда как сложность раскрытия остается постоянной (пока какой-либо криптоаналитик не придумает лучший способ раскрытия), мощность компьютеров растет. За последние 50 лет вычислительные мощности феноменально выросли, и нет никаких причин подозревать, что эта тенденция не продлится. Много криптографических приемов пригодны для параллельных компьютеров: задача разбивается на миллиарды маленьких фрагментов, для решения которых не нужно межпроцессорное взаимодействии. Объявление алгоритма безопасным просто потому, что его нелегко взломать, используя современную технику, по крайней мере ненадежно. Хорошие криптосистемы проектируют стойкими к взлому с учетом развития вычислительных средств на многие годы вперед.

Основные выводы

Криптоалгоритм — алгоритм, предназначенный для реализации любого метода шифрования данные.

Криптографический ключ — последовательность символов, которая обеспечивает возможность шифрования и дешифрации.

Криптографическая система с открытым ключом основывается на криптографии с открытым ключом. Самыми известными практическими реализациями этого типу есть системы Диффи—Хеллмана, RSA и Ель-Гамаля.

Хеш-функция (хеш-функция; функция хеширования; функция расстановки) — функция, которая используется для выработки блока данных фиксированной длины (значение функции), которая однозначно отображает произвольно выбранный открытый текст (аргумент функции).

Значение хеш-функции — множественное число значений целых чисел, которые принадлежат заданному диапазону и образованные в результате вычисления хеш-функции.

Цифровая подпись (цифровая сигнатура) — цифровая последовательность данных, которая образуется в результате асимметричного криптографического преобразования начальной информации и позволяет получателю проверить источник и целостность данных, а также осуществить защиту от фальсификации или подделки.

Алгоритм побайтовой блочной шифрования — криптографический алгоритм, в котором во время зашифрования и расшифровывания используются только операции над байтами.

Стеганографическая защита — обеспечение укрывательства самого факта существования конфиденциальных сведений при их передаче, хранении или обработке.

Стеганографическая система — это совокупность средств и методов, которые используются с целью формирования скрытого (незаметного) канала передачи информации.

Направления стеганографии: внедрение информации с целью ее скрытой передачи; внедрение цифровых водяных знаков; внедрение идентификационных номеров; внедрение заглавий.

Процесс проведения стеганоанализа — оценка перехваченного контейнера на предмет наличия в нем скрытого сообщения.

Сообщение и контейнер — основные стеганографические понятия.

Конфиденциальное сообщение — секретная информация, наличие которой необходимо скрыть.

Контейнер — несекретная информация, которую можно использовать для укрывательства сообщения.

Пустой контейнер (или так называемый контейнер-оригинал) — это контейнер, который не содержит скрытую информацию.

Заполненный контейнер (контейнер-результат) — контейнер, который содержит скрытое сообщение.

Поточный контейнер — последовательность битов, что непрерывно изменяется.

Фиксирован контейнер — размеры и его характеристики являются предварительно известными.

Теоретически стойкая (абсолютно надежная) стеганосистема - это система, которая осуществляет укрытие информации лишь в тех фрагментах контейнера, значения элементов которых не превышают уровень шумов или ошибок квантования, и при этом теоретически доказано, что невозможно создать стегано аналитический метод выявления скрытой информации.

Практически стойкой стеганосистемой называется система, которая проводит такую модификацию фрагментов контейнера, изменения которых могут быть обнаружены, но известно, что на данный момент необходимы стеганоаналитические методы у нарушителя отсутствуют или пока еще не разработаны.

Неустойчивой стеганосистемой называется система, которая скрывает информацию таким образом, что существующие стегано аналитические средства позволяют ее обнаружить.

Криптография объединяет принципы, методы и средства преобразования данные с целью маскировки (шифровка) содержания информации для гарантирования ее конфиденциальности и целостности.

Криптоанализ (криптографический анализ) — изучение системы защиты сообщений и (или) исследования ее входных и исходных сообщений с целью выделения скрытых переменных или истинных данных, включая начальный текст.

Вопросы для самоконтроля

1. *Какие известны два базовых направления теории тайнописи*?
2. *В чем назначение стеганографии*?
3. *Чем отличается криптография от стеганографии*?
4. *Как классифицируются стеганографичні методы*?
5. *Объясните, как происходит в пределах стеганографии защита прав на цифровую интеллектуальную собственность.*
6. *В чем заключается сущность метода наименьшего значимого бита*?
7. *Как классифицируются спектральные методы цифровой стеганографии*?
8. *Какие существуют основные виды атак на стеганографическую систему*?
9. *Какие существуют основные типы криптоаналитического раскрытия информации*?
10. *Какой принцип положен в основу алгоритмов из постановочным и перестановочным шифром*?
11. *Где нашли приложение программные постановочные шифры*?
12. *Какие два основных типа алгоритмов основываются на ключах*?
13. *На какие категории разделяются симметричные алгоритмы*?
14. *Почему асимметричные алгоритмы называются алгоритмами «с открытым ключом»*?
15. *Какая разница между симметричной и асимметричной криптосистемами?*
16. *Какие базовые криптографические преобразования используются в алгоритме шифрования RSA*?
17. *Приведите примеры симметричных алгоритмов шифрования данные.*

The main conclusions

Cryptoalgorithm is the algorithm intended for realization of any method of an enciphering of data.

The cryptographic key is a sequence of symbols, which provides possibility of an enciphering and deciphering.

Cryptographic system with open key is based on cryptography with the open key. The most known practical realizations of its type are the systems of Diffie-Hellman, RSA and Ell-Gamal.

Hash function (hash function; function of hashing; function of arrangement) is a function that is used for making a block of the fixed length (value of function) that displays arbitrarily selected opened text unambiguously (argument of function).

Value of hash function is multitude of values of integers that belong to the set range and are formed as a result of calculation of hash function.

The digital signature is a digital sequence of data that is formed as a result of asymmetric cryptographic transformation of the initial information and allows the receiver to check up a source and data integrity and also to make the protection from falsification or a fake.

Algorithm of byte block enciphering is a cryptographic algorithm where only the operations on bytes are used during an enciphering and deciphering.

Steganographic system is a collection of means and methods that are used with the purpose of creation of the hidden (imperceptible) channel of transmission of the information.

The directions of steganography are the following: implantation of the information with the purpose of its hidden transmission; implantation of digital watermarks; implantation of identification numbers; implantation of titles.

The process of conducting of steganalysis is an estimation of the tapped container for presence of the hidden message in it.

The message and container are the main steganographic concepts.

Confidential message is the classified information the presence of which is necessary to hide.

Container is the unclassified information that can be used for hiding of the message.

The empty container (or the so-called container-original) is a container that does not contain the hidden information.

The filled container (container-result) is the container that contains the hidden message.

The streaming container is the sequence of bits that continuously varies.

The fixed container has preliminary known sizes and characteristics.

Theoretically firm (absolutely reliable) steganosystem is a system that carries out the hiding of the information only in the fragments of the container, the value of units of which do not exceed a level of noise or errors of quantization and thus it is theoretically proved that it is impossible to create steganalytical method of revealing of the hidden information.

Practically firm steganosystem is the system that conducts such modification of fragments of the container, changes of which can be found out, but it is known, that at present the intruder has not got the necessary steganalytical methods or they have not been developed yet.

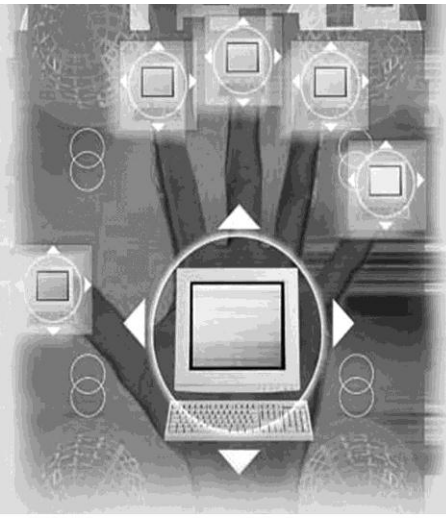
Unstable steganosystem is the system that hides the information in such a way that existing steganalytical facilities allow to find it out.

Cryptography unites principles, methods and facilities of transformation of data with the purpose of masking (enciphering) of a content of the information for guaranteeing its confidentiality and integrity.

Cryptoanalysis (the cryptographic analysis) is learning of system of protection of messages and (or) research of its incoming and outgoing messages with the purpose of singling out the latent variables or true data, including the initial text.

Ключевые слова

Русский	**Английский**
криптография	cryptography
стеганография	steganography
шифр	cipher
цифровая подпись	digital signature
контейнер	container



ЗАЩИТА ИНФОРМАЦИИ ОТ НЕСАНКЦИОНИРОВАННОГО ДОСТУПА

9

9.1. Методы несанкционированного доступа к ресурсам информационных систем
9.2. Средства защиты от несанкционированного доступа
9.3. Моделирование систем и процессов защиты информации
9.4. Противодействие сетевому несанкционированному доступу

9.1. Методы несанкционированного доступа к ресурсам информационных систем

Проблема несанкционированного доступа (НСД) к ресурсам информационных систем обострялась с развитием информационных технологий и тотального использования компьютерных сетей во всех сферах деятельности общества.

Решение задач разработки и выбора соответствующих эффективных методов и средств защиты в значительной мере зависит от ряда факторов, связанных с самым процессом несанкционированного доступа. Поскольку спектр несанкционированных действий достаточно разнообразный, то основой классификации могут быть базовые признаки (рис. 9.1).

Рис. 9.1. Классификация несанкционированного доступа по базовым признакам

Мануальный НСД (подсматривание, собирание мусора, изъятие информационных носителей, подмена положений включателей режимов и т.п.) реализуется за прямого участия человека без использования любых специальных средств.

Пример реализации мануального НСД - сбор промышленного мусора и изъятие информационных носителей - приведен на рис. 9.2.

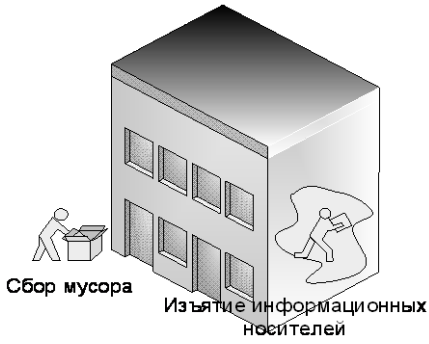


Рис. 9.2. Пример мануального НСД

Автоматизированный НСД (рис.9.3) осуществляется при постоянном участии оператора с использованием широкого спектра программных и аппаратных средств и связан с суперзапингом, снупингом, снифингом, подключением дополнительных терминалов, использованием сетевых анализаторов и т.п. Пример автоматизированного НСД, когда неавторизованная сторона с помощью аппаратных и программных средств (например, путем использования утилит отдаленного управления) реализовывает НСД к рабочей станции при прямом подключении к сети, приведен на рис. 9.3.

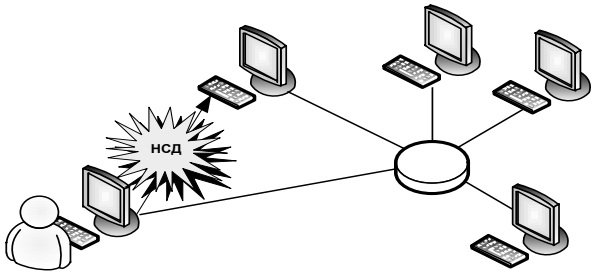


Рис. 9.3. Пример автоматизированного НСД

Автоматический НСД реализуется без участия человека, как правило, с использованием специализированных программных средств, функционирования которых базируется на вирусных технологиях. Примером автоматического НСД

может быть заражение компьютера вирусом во время подключения к глобальной сети Интернет без участия человека (рис. 9.4).

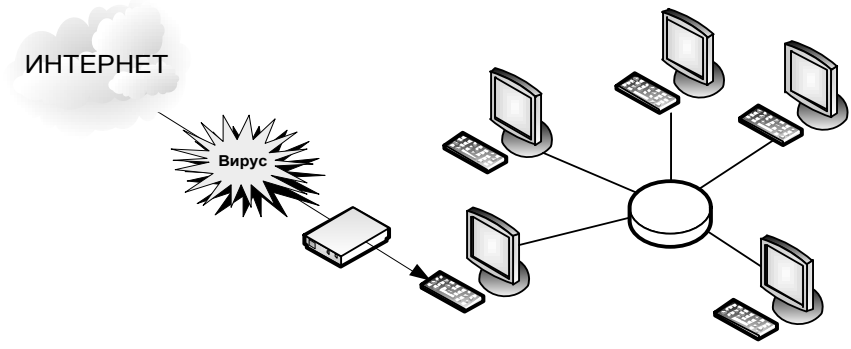


Рис. 9.4. Пример автоматического НСД

Постполитизационный НСД базируется на использовании недостатков в уже реализованной политике безопасности. Такими недостатками могут быть неправильно построенные правила разделения доступа, использование программных и аппаратных средств с недостаточным уровнем защищенности, счета при блокировании каналов утечки информации с ограниченным доступом и т.п. Например, если пользователю и администратору сети предоставлены одинаковые права доступа к серверу, то неавторизованная сторона, получив права доступа пользователя 1, сможет реализовать НСД к информационным ресурсам на уровне прав администратора (рис. 9.5, *а*).

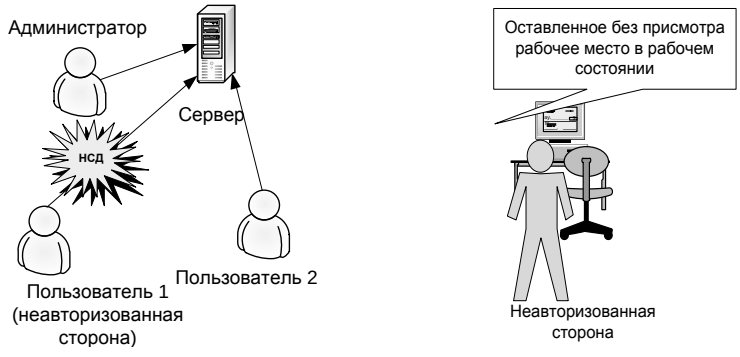
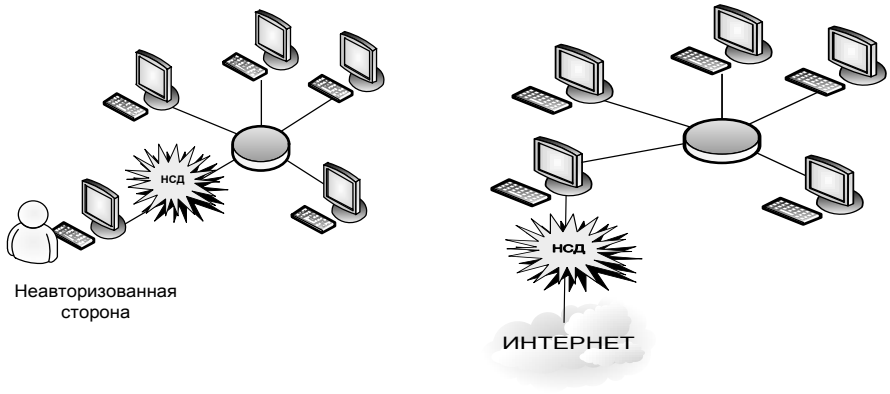


Рис. 9.5. Пример постполитизационого и деполитизационного НСД

Деполитизационный НСД связан с ошибками и небрежностью, которые случаются при реализации мероприятий с обеспечением уже имеющейся политики безопасности. Это прежде всего связано с человеческим фактором

(недостаточная административная поддержка, некорректное выполнение функций защиты, несвоевременное реагирование на нештатные ситуации и т.п.). Например, если пользователь оставляет свое рабочее место, не придерживаясь политики безопасности (не заканчивает сеанса работы или не блокирует компьютер), то неавторизованная сторона за время его отсутствия имеет доступ к информационным ресурсам через некорректность выполнения пользователем функций защиты (см. рис. 9.5, *б)*.

Если НСД к ресурсу осуществляется в локализованной области его расположения (локальная вычислительная сеть, рабочая станция, принтер, носитель информации, операционная система, приложения и т.п.), то он называется *локальным*, а в противоположном случае - *отдаленным.* Например, локальный НСД может быть реализованный внутри сегмента (физического объединения станций с помощью коммутационных устройств не выше от канального уровня). При этом источник НСД и ресурс, который подвергнулся несанкционированному действию, будут находиться в пределах одного сегмента. Примеры реализаций локального и отдаленного НСД изображено соответственно на рис. 9.6, *а* и 9.6, *б*. Локальный НСД реализовывается неавторизованной стороной в одном сегменте сети, а отделенный несанкционированный доступ происходит через сеть Интернет.



а *б*

Рис. 9.6. Реализация локального и отдаленного НСД

Межсегментный НСД может быть примером отдаленного НСД, когда источник и ресурс находятся в разных сегментах компьютерной сети, например кампусной. В первом случае инициатором НСД бывает легальный пользователь, который, например, из файла-сервера через рабочую станцию осуществляет копирование конфиденциальных данных на внештатный носитель информации.

В результате своего действия (например, визуального просмотра данных из терминала) *пассивный НСД* не осуществляет непосредственного влияния на ре-

сурсы и может не нарушить их характеристик безопасности, например при перехвате зашифрованных данных.

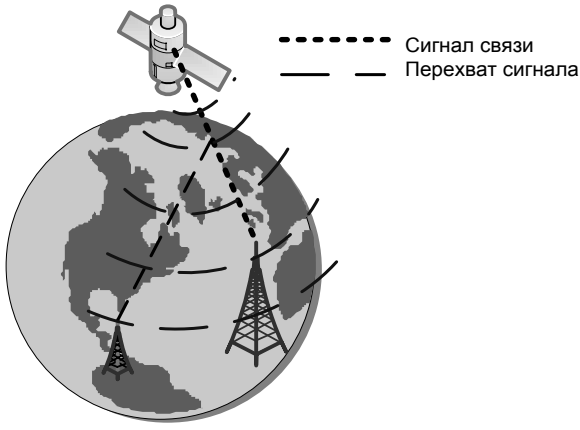


Рис. 9.7. Реализация пассивного НСД

В результате *активного НСД* на ресурсы осуществляется непосредственное влияние (изменение конфигурации, нарушение функциональности и т.д.) и поднимаются их характеристики безопасности. Практически все типы отдаленного НСД являются активными, т.е. такими, что делают принципиально возможным его выявление, поскольку в результате непосредственного влияния происходят определенные изменения. А пассивный НСД, в отличие от активного, не оставляет следов вмешательства. Например, перехват спутникового сигнала без его модификации является пассивным НСД (рис. 9.7).

Перехват сообщения неавторизованной стороной с дальнейшей его модификацией является активным НСД (рис. 9.8).

Рис. 9.8. Реализация активного НСД

Условный НСД инициализируется в случае возникновения определенного события (механизм логической бомбы) и, в свою очередь, может быть как пассивным, так и активным. Примером инициализации пассивного условного НСД может быть передача от потенциальной цели запроса определенного типа, который и будет условием начала атаки. Например, таким условием могут быть DNS- и ARP-запросы в стеке протоколов TCP/IP.

Активный условный НСД осуществляет постоянный мониторинг состояния отдельных ресурсов, и в случае определенного изменения указанного состояния формируется сигнал инициализации. Примером такой ситуации может быть событие, связанное с прерыванием сеанса работы пользователя с сервером без стандартной команды, например LOGOFF. Момент инициализации безусловного НСД не сопровождается определенным изменением состояния ресурсов и определяется источником атаки.

Программный НСД базируется на специальных микро- или макрокодированных средствах (например, суперзапинговых утилитах, внутренних командах, сценариях автоматизации и т.п.), которые функционируют в пределах информационных систем для реализации своих функций.

Например, отдаленный компьютер (рис. 9.9) отправляет сообщение на рабочую станцию, к которой несанкционированно подключается неавторизованная сторона с помощью портативного компьютера и реализует заражение этой станции вирусом, который модифицирует указанное сообщение.

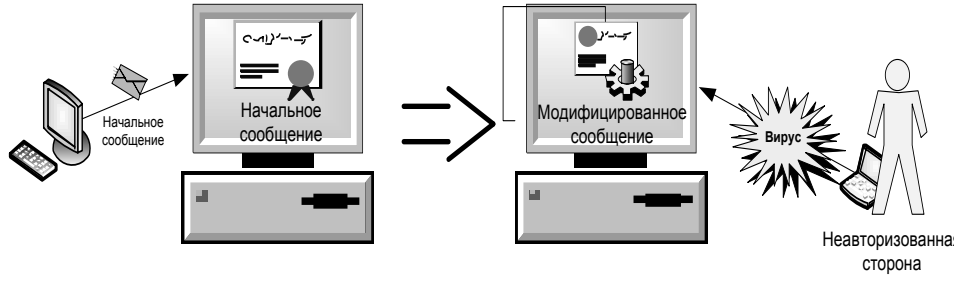


Рис. 9.9. Реализация программного НСД

Аппаратный НСД базируется на разнообразных механических, электрических, электромеханических, электронных, электронно-механических и других устройствах, которые используются автономно или в объединении с другой аппаратурой для выполнения соответствующих функций.

Например, в комнате для совещаний (рис. 9.10) может быть установлен «жучок» для прослушивания конфиденциальных разговоров. Этот пример характерный для аппаратного НСД.

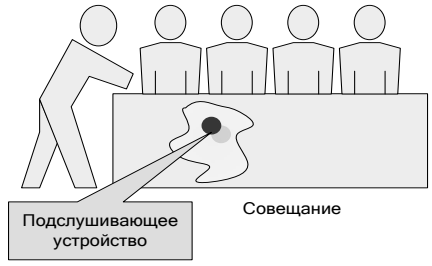


Рис. 9.10. Реализация аппаратного НСД

Нетипичный НСД реализуется на базе средств, которые не принадлежат к аппаратным или программным, таким например, как взрывчатка, радиоактивные материалы, кислоты, щелочи, насекомые, грызуны и т.п.

В процессе реализации *НСД с обратной связью* нарушитель получает от ресурса, который подвергся несанкционированным действиям, ответ на эти действия, чтобы в дальнейшем осуществлять НСД на более эффективном уровне благодаря анализу реакций объекта НСД на те или другие изменения.

Примером НСД с обратной связью является сканирование портов специальными программным обеспечением. Во время сканирования на порты сервера отправляется пакет синхронизации SYN. Если на это сообщение (рис. 9.11) приходит ответ (обратная связь) в виде пакета SYN/ACK, то это означает, что сканированный порт находится в состоянии ожидания и можно выполнять следующее действие.

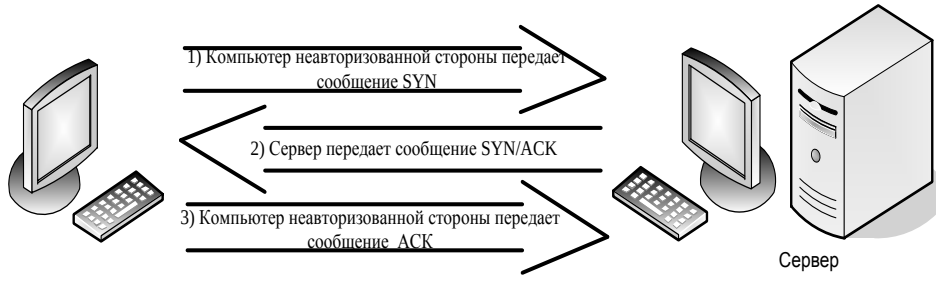


Рис. 9.11. Реализация НСД с обратном связью

НСД без обратной связи реализует свои действия независимо от реакции ресурса, который подвергнулся несанкционированным действиям. Примером таких несанкционированных действий есть отказ в обслуживании. Например, блокирование маршрутизатора (рис. 9.12), что приводит к отказу в обслуживании.

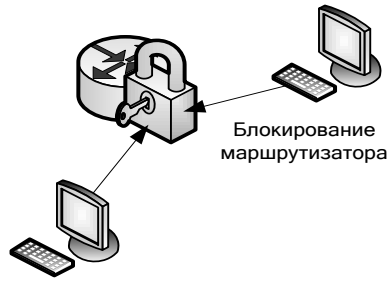


Рис. 9.12. Реализация НСД без обратной связи

При реализации НСД происходит нарушение основных характеристик безопасности ресурсов информационных систем:

конфиденциальность - характеристика безопасности ресурсов, отражающая их свойство доступности без соответствующих полномочий. Фактически ресурсы не могут быть доступными или раскрытыми неавторизованной стороне, т.е. для нее их якобы нет. В свою очередь, авторская сторона (например, обслуживающий персонал, пользователь, программы и т.д.), которой предоставлены соответствующие полномочия, имеет полный доступ к ресурсам;

целостность - характеристика безопасности ресурсов, отражающая их свойство противостоять несанкционированному изменению. Например, пользователь, нагромождающий информацию, имеет право ожидать, что содержимое его файлов останется неизменным, несмотря на целенаправленные влияния, отказы программных или аппаратных средств. По этой характеристике ресурсы не испытывают изменений со стороны неавторизованной стороны;

доступность - характеристика безопасности ресурсов, отражающая возможности их использования в заданный момент времени соответственно предоставленным полномочиям. Фактически авторская сторона в любой момент времени получает неограниченный доступ к необходимому ресурсу.

В этом контексте по типу нарушений указанных характеристик НСД бывает:

К - действенный (нарушение конфиденциальности ресурсов);

Ц - действенный (нарушение целостности ресурсов);

Д - действенный (нарушение доступности ресурсов).

Если в процессе НСД нарушаются разные характеристики безопасности, то результирующий тип будет комбинированным на базе основных, например КЦД - действенный - НСД, нарушающий конфиденциальность, целостность и доступность ресурсов.

По природе взаимодействия с ресурсами информационной системы НСД бывает *физическим* и *логическим*. Для первого характерна физическая форма взаимодействия в виде разного рода прямых блокирований, повреждений, проникновений, краж и т.п., например размыкания электрических соединений, по-

вреждение носителей информации, разукомплектование, преодоление физической границы защиты, подслушивание, перехват побочных электромагнитных излучений и наведений и т.п.

Примером физической формы НСД могут быть перехваты побочных электромагнитных излучений из монитора компьютера специализированной аппаратурой (рис. 9.13).

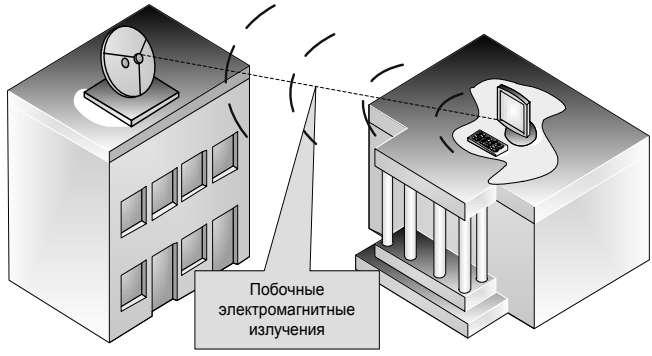


Рис. 9.13. Реализация физического НСД

Логическому НСД не присуще прямое физическое взаимодействие с ресурсами. Речь идет о вмешательстве в логику событий, например, анализ протоколов, перегрузка, определение паролей, перехваты сеансов и т.п.

Примером *логической* формы НСД может быть перехваты сеанса связи (рис. 9.14).

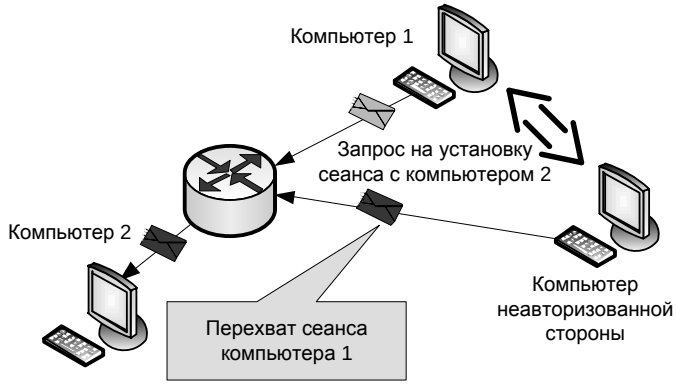


Рис. 9.14. Пример логического НСД

Мономономный НСД реализуется с одного источника на другой конкретный ресурс. Такой НСД называют также *нераспределенным.* Например, мономо-

номный НСД можно реализовать с помощью сканирования портов компьютера с определенным IP-адресом (рис. 9.15).

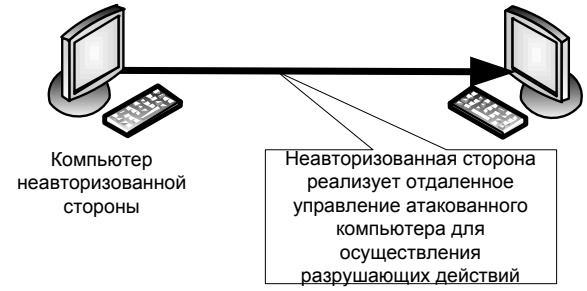


Рис. 9.15. Пример реализации мономономного НСД

Полимономный НСД осуществляется одновременно с нескольких (двух и больше) источников на один ресурс и направленный на достижение одной конкретной цели. Такой НСД называют также *распределенным* (рис. 9.16).

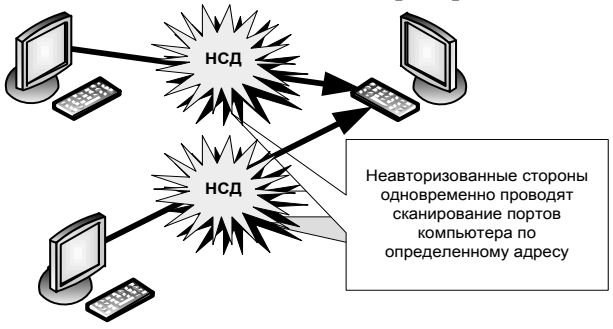


Рис. 9.16. Пример реализации полимономного НСД

Монополичный НСД реализуется с одного источника одновременно на несколько (два и больше) ресурсов и направляется на достижение конкретной цели (рис. 9.17).

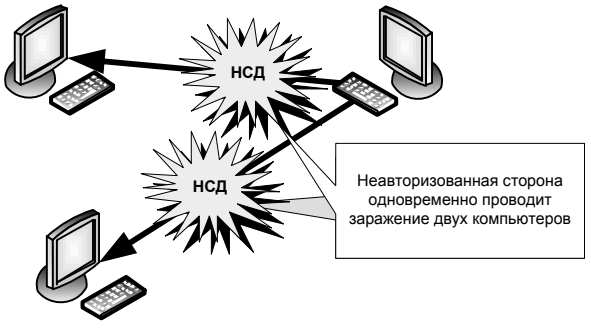


Рис. 9.17. Пример монополичного НСД

Например, такого типа НСД может базироваться на широковещательной передаче сообщения от источника на все компьютеры сегмента, адреса которых находятся под одной маской подсети.

Полиполичный НСД объединяет в себе *полимономную* и *монополичную* технологии, согласно которым множество источников осуществляет НСД на множество ресурсов, чтобы достичь одной конкретной цели (рис. 9.18).

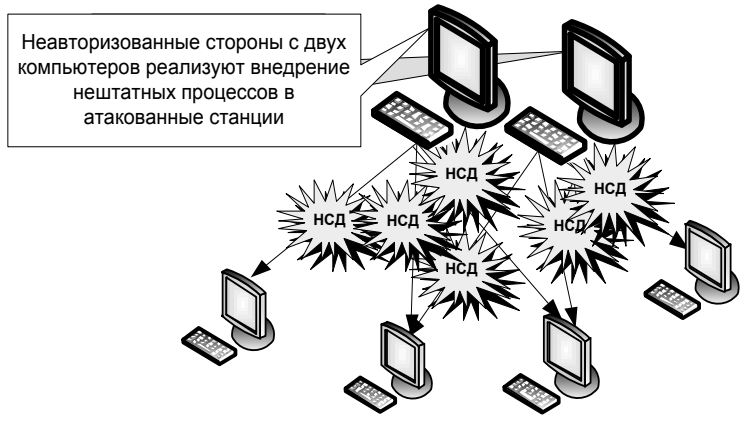


Рис. 9.18. Пример полиполичного НСД

Фрагментированый НСД базируется на принципе декомпозиции и поэтапной реализации, например на использовании механизма разбивки IP-пакетов (на множество мельчайших) и дальнейшей их передаче (рис. 9.19).

Такое фрагментирование дает возможность обходить системы выявления атак, которые не рассчитаны на противодействие декомпозиционным технологиям.

Рис. 9.19. Пример фрагментированого НСД

НСД, реализующийся без использования значений по умолчанию (например, программная закладка BackOrifice по умолчанию использует порт 31337, но это значение можно изменить, например, на 31336), сориентировн на преодоление систем выявления атак, базирующихся на сигнатурных (шаблонных) технологиях по аналогии с антивирусами или программами, предназначенными для защиты от сигнатурных вирусов.

Скрытые атаки используют разнообразные мероприятия (подмена контрольных сумм, перехват разнообразных данных, модификация ядра операционной системы, использование стандартных или похожих на стандартные имен и т.п.), которые дают возможность оставаться невыявленными в локализованной области атакованного ресурса. Технология *скрытого НСД* по своей идеологии подобна технологии *стелс-вирусов.*

Пигибекинговый НСД базируется на НСД к временно неконтролируемому ресурсу, например путем проникновения в информационную систему в результате временного отсутствия или после некорректного завершения сеанса работы легального пользователя.

Маскарадный НСД базируется на формировании такого поведения нарушителя, которое дает ему возможность выдать себя легальным источником, например посредством обмана (spoofing) атаковать вычислительную сеть (с протоколом TC/IP), присваивая IP-адрес, с помощью которого удается обойти систему защиты.

Косвенный НСД базируется на том, что нападение осуществляется через третье лицо (посредника), а истинный источник нападения остается неизвестным.

При этом часто используются маскарадные технологии (рис. 9.20). Например, воспользовавшись вариантом нападения с перенаправлением НСД, чтобы сделать невозможным выявление реального источника, нарушитель проводит или перенаправляет свой трафик через чужой компьютер, который для ресурса (подвергнувшегося НСД) и будет исходным источником.

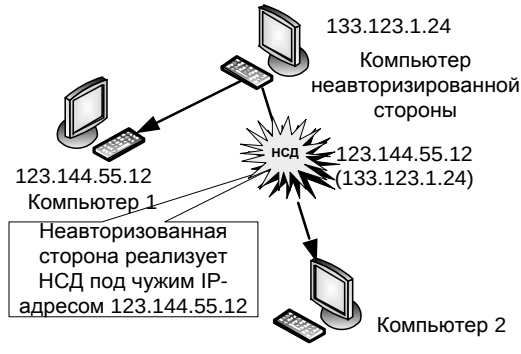


Рис. 9.20. Пример реализации косвенного НСД

Социотехнический (социоинжиниринговый) НСД связан с получением данных (например, имен пользователей, паролей, телефонных номеров отдаленного доступа и т.п.) от атакованных объектов в процессе информационного обмена (рис. 9.21).

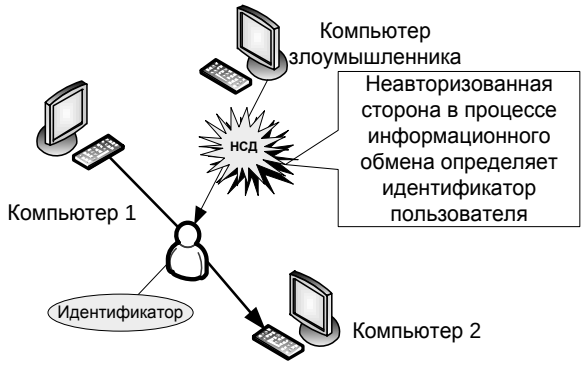


Рис. 9.21. Пример социотехнического НСД

Криптоаналитический НСД базируется на использовании широкого спектра криптоаналитических методов и средств для взлома ресурсов, защищенных разными криптографическими средствами.

К неспецифичным категориям НСД принадлежат те, которые не имеют указанных особенностей реализации. При этом следует учитывать, что технологии НСД постоянно развиваются, т.е. соответствующие возможности будут расширяться.

Расширяющий НСД (рис. 9.22) сориентирован на получение больших полномочий на права доступа к ресурсу, например, на вход в локальную вычислительную сеть с правами администратора, получение доступа на запись к полям баз данных, изменение атрибутов файлов и т.п.

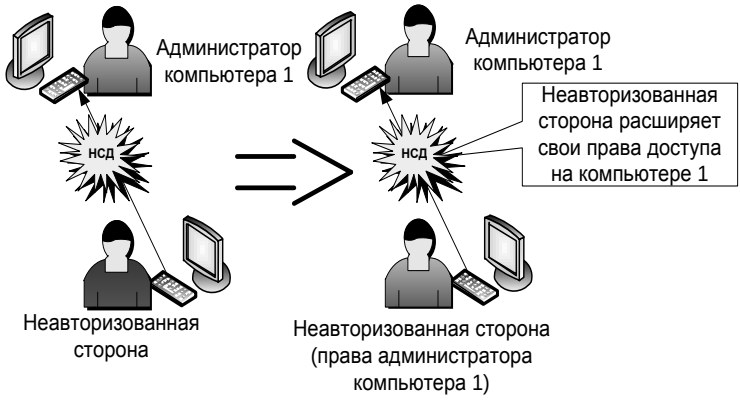


Рис. 9.22. Пример реализации расширяющего НСД

Искажаемый НСД связан с осуществлением любых прямых изменений в целевом ресурсе (например, подделка полей баз данных, подмена информационных носителей, изменение времени и дат и т.п.).

Перегрузочный НСД направлен на загрузку ресурса до такого уровня, что он становится непригодным для использования. Результатом таких несанкционированных действий может быть невозможность использования, перегрузка, препятствование использованию (отказ в обслуживании) и т.п. Примером такого НСД (рис. 9.23) может быть перегрузка маршрутизатора.

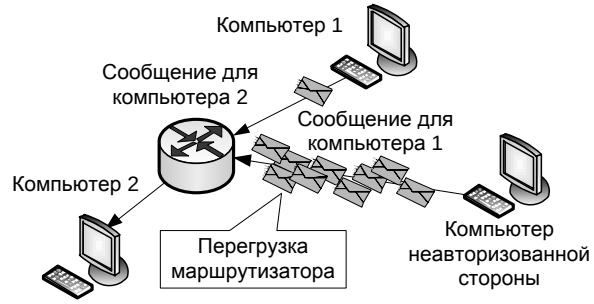


Рис. 9.23. Реализация перегрузочного НСД

Информационный НСД связан со сбором необходимых данных (как правило, для реализации дальнейших действий) и не предусматривает осуществления прямого НСД к ресурсу. Например, получение информации в результате анализа публикаций, использование системных утилит для выявления активных рабочих станции, сервисов и т.п. (рис. 9.24).

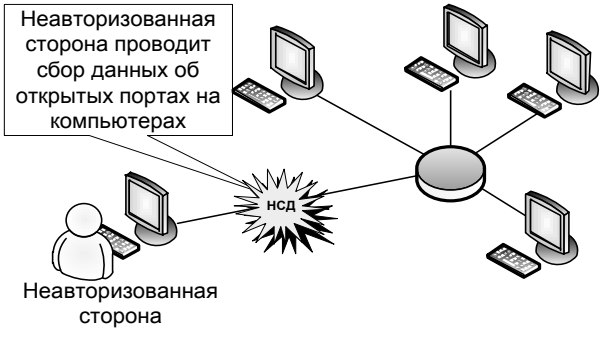


Рис. 9.24. Пример информационного НСД

Распространительный НСД направлен на получение доступа к ресурсу и его раскрытие без соответствующих полномочий, например несанкционированное получение файла данных с паролями и их публикация на хакерских сайтах или рассылка среди абонентов вычислительной сети.

Разворовывающий НСД состоит в использовании ресурса без нанесения прямого убытка, например без снижения качества обслуживания пользователей осуществляется временное изъятие части памяти (для расширения возможностей другой системы), загрузка телекоммуникационных каналов, использование рабочей станции или сетевого сервиса и т.п.

Задерживающий НСД предназначен для временной задержки ресурса с целью снижения его актуальности. Например, задержка шифрограммы на промежуточном узле при ее передаче телекоммуникационными каналами общего пользования (рис. 9.25).

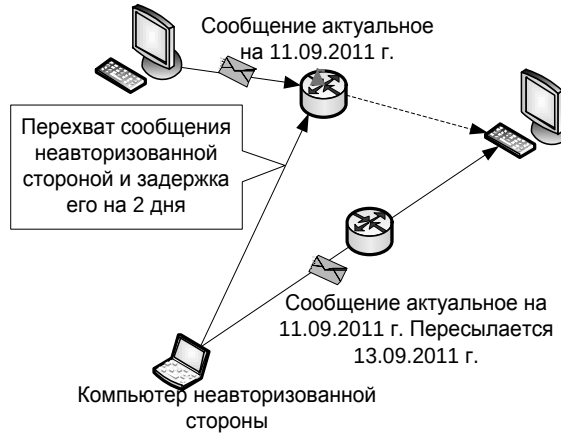


Рис. 9.25. Пример задерживающего НСД

Уничтожающий НСД сориентирован на безвозвратную ликвидацию ресурса, например, изъятие файла, измельчение информационного носителя, низкоуровневое форматирование жесткого диска с целью уничтожения данных и т.п. (рис. 9.26).

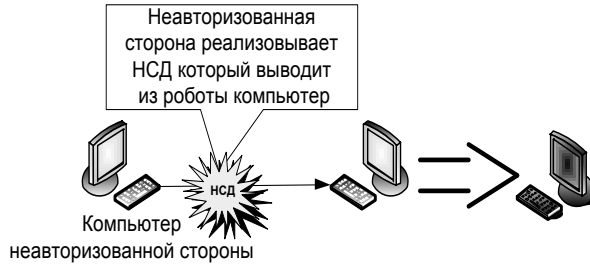


Рис. 9.26. Пример уничтожающего НСД

Простой НСД (рис. 9.27) означает несложные в реализации действия, направленные на выполнение отдельных процедур, например сканирование пор-

тов, анализ трафика, поиск активных рабочих станций, отдаленное управление и т.п.

Рис. 9.27. Пример простого НСД

Сложный НСД является комбинацией простого НСД, предназначенного для реализации ряда необходимых функций, например выявление активной рабочей станции, и действий относительно отдаленного управления ею.

Системный НСД строится на основе сформированного системного подхода с многошаговой комбинацией действий в сочетании с простым НСД для эффективной реализации специально направленного комплекса функций. Например, это может быть поиск активных рабочих станций, мониторинг трафика, сканирование, взлом ОС и рабочих приложений, несанкционированное копирование данных и заметание следов.

Международная организация стандартизации (*ISO*) предложила семиуровневую *эталонную модель* с целью разделения функций разных протоколов в процессе передачи информации от одного абонента другому. Таких классов функций выделено семь. Они получили название уровней, каждый из которых выполняет определенные задачи в процессе передачи блока информации, причем соответствующий уровень со стороны приемника тем, которые выполнены на этом уровне на передающей стороне (источнике). В этой связи *НСД по семиуровневой эталонной модели* можно определить как:

физический НСД, кодирующийся параметром $K0(2) = 20$;
канальный НСД, кодирующийся параметром $K1(2) = 21$;
сетевой НСД, кодирующийся параметром $K2(2) = 2^2$;
транспортный НСД, кодирующийся параметром $K3(2) = 23$;
сеансовый НСД, кодирующийся параметром $K4(2) = 24$;
представление данных, кодирующееся параметром $K5(2) = 25$;
прикладной НСД, кодирующийся параметром $K6(2) = 2^6$.

Симеон-Дени Пуассон (Siméon-Denis Poisson, 1781—1840), французский ученый, математик, механик и физик. Работы касаются теоретической и небесной механики, математики и математической физики. Впервые записал уравнение аналитической механики в составляющих импульса. В области небесной механики исследовал устойчивость движения планет Солнечной системы, занимался решением задач о возмущении планетных орбит и о движении Земли вокруг ее центра тяжести. В теории потенциала ввел уравнение, названное в его честь, и применил его к решению задач по гравитации и электростатике.

На физическом уровне обеспечиваются необходимые механические, электрические, функциональные и процедурные характеристики для установления, поддержание и размыкание физического соединения. На *канальном уровне* обеспечиваются функциональное и процедурное средства для установления, поддержание и высвобождение линий передачи данных между абонентами сети (например, терминалами и узлами сети). На *сетевом уровне* обеспечиваются функциональное и процедурное средства для обмена служебной информацией между двумя объектами транспортного уровня сети (т.е. устройствами, которые поддерживают протоколы на транспортном уровне с помощью сетевого соединения). Гарантируется независимость поведения объектов транспортного уровня от схемы маршрутизации и коммутации. На *транспортном уровне* обеспечивается оптимизация коммутационного обслуживания (поддерживаемого реализацией более низких уровней связи) с помощью прозрачной передачи данных между абонентами в рамках сеанса.

На уровне *представления данных* обеспечивается совокупность служебных операций, которые можно выбрать на прикладном уровне для интерпретации переданных и получаемых данных. Эти служебные операции содержат управление информационным обменом, отображение данных и управление структурированными данными. Служебные операции этого уровня представляют собой основу всей семиуровневой модели и дают возможность связывать в единое целое терминалы и средства вычислительной техники любых типов.

На *прикладном уровне* обеспечивается непосредственная поддержка прикладных процессов и программ конечного пользователя и управление взаимодействием этих программ с разными объектами сети передачи данных.

Определение класса несанкционированного действия по этим признаком такое:

ISO-*K*, где $K = K_{(2)}\,0 + K_{(2)}\,1 + K_{(2)}\,2 + K_{(2)}\,3 + K_{(2)}\,4 + K_{(2)}\,5 + K_{(2)}\,6$ - двоичный код.

Для удобства будем подавать *K* в 16-ричном коде ($K_{(16)}$).

Например, если по этой классификации НСД записано как ISO-3*A* (где $3A_{(16)} = 0111010 = 2^5 + 2^4 + 2^3 + 2^1 = K5 + K4 + K3 + K1$), то она интерпретирует реализацию на канальном, транспортном и сеансовом уровнях и на уровне представления данных, а например, запись ISO-00 означает, что класс НСД не поддерживается семиуровневой эталонной моделью.

Нетрудно заметить, что НСД, которые классифицируются по принципу признака, могут в каждом конкретном случае при определении общего класса содержать не только один, а и больше компонентов каждого из признаков. С появлением новых методов и средств реализации НСД признаки предложенной классификации могут быть расширены.

При практическом использовании классификации такой, например, НСД, как сканирование портов, можно определить как автоматизированный, постполитизированый, отдаленный, пассивный, безусловный, программный, с обратной связью, К-действенный, логический, мономономный, неспецифический, информационный, простой, ISO-36.

С помощью этой классификации можно осуществлять соответствующую формализацию возможностей систем противодействия для повышения эффективности их выбора и формирования требований при их разработке.

С НСД связаны такие понятия, как *доступ* и *перехват*. Под *доступом* (access) понимают взаимодействие между ресурсами информационных систем, которое обеспечивает передачу информации между такими ресурсами, а в процессе доступа к информации (access to information) реализуются, в частности, ее копирование, модификация, уничтожение, инициализация и т.п. Различают несанкционированный и *санкционированный доступы*. Если доступ к ресурсам системы осуществляется, например, с нарушением или вне правил размежевания доступа, то такой доступ - несанкционированный. Одним из базовых действий, которое порождает НСД, являются перехваты (intercept), под которыми понимают несанкционированное получение информации незаконным подключением к каналам связи (например, прямой перехват) визуально (например, подсматривание) или с помощью радиотехнических средств (например, косвенный перехват).

По действию на информацию различают *активный* и *пассивный перехваты*, а по типу подключения - *прямое* и *косвенное*.

Активный перехват (active eavesdropping) - это такой перехват, во время которого у неприятеля есть возможность не только перехватывать сообщение, а и влиять на него, например, задерживать или изымать сигналы, которые передаются по каналам связи.

Пассивный перехват (passive tapping) - получение информации с возможностью только наблюдать за обменом сообщениями (например, с целью выявления разной системной информации в вычислительной сети (ВС)), не влияя на него.

Прямой перехват (direct eavesdropping) - перехват информации при непосредственном подключении (например, дополнительного терминала) к линии связи. Прямой перехват можно обнаруживать проверкой линии связи.

Косвенный перехват (indirect eavesdropping) - перехват информации (например, индуктивных волн) без использования непосредственного подключения к линии связи (threat). Такой перехват тяжело определить, поскольку нет непосредственного присоединения терминального оборудования к линии связи.

Разрушительные программные влияния. ***Разрушительное программное влияние*** - это программный код или его части, с помощью которых осуществляется угроза хотя бы одной характеристике безопасности определенных ресурсов информационных систем. Разрушительные влияния можно поделить на такие группы: компьютерные вирусы (вирусы), логические бомбы, тайные хода и лазейки; программы раскрытия паролей, репликаторы, сетевые программные анализаторы, суперзапинговые утилиты, троянские кони.

Вирус. Программа, способная к многоразовому самовольному созданию своего тела, которая по обыкновению модифицирует (заражает) другие программы, записанные в файлах или системных областях, для дальнейшего воспроизведения нового тела и получения управления с целью модификации записей, уничтожения файлов, загрузка ресурсов и выполнения других разрушительных влияний в информационной системе.

Логические бомбы. Программа, которая инициируется с возникновением разных событий, например открытие определенного файла, обработка заданных записей и другие действия с целью нарушения безопасности ресурсов информационных систем. Используются, например, для разворовывания с помощью изменения определенным образом (в свою пользу) кода программы, которая реализует финансовые операции.

Тайный ход. Уязвимость в системе, которую специально создал разработчик или которая возникла случайно и фактически является дополнительным способом проникновения в систему.

Программы раскрытия паролей. Программы по обыкновению предназначены для угадывания паролей (например, архивированных файлов) подбором вариантов, возможных для использования символов или проникновение в систему с помощью словаря. Программы, которые основываются на последнем методе, осуществляют взлом системы парольной защиты подбором элементов одного или нескольких словарных файлов, сложенных специально или взятых из серверов или жестких дисков локальных станций.

Репликаторы. Программы, которые при выполнении создают несколько своих копий в информационной системе. Например, когда репликатор создает

только одну копию и после этого выполняет ее, то память системы быстро переполняется, чем ограничивается доступ к определенным компонентам системы.

Сетевые анализаторы. Программно-аппаратные средства (в отдельных случаях программы, которые запускаются из рабочей станции, подключенной к сети), предназначенные для считывания любых параметров потока данных в информационной системе.

Суперзапинг. Разрушительное влияние, связанное с несанкционированным использованием утилит для модификации, уничтожения, копирования, раскрытия, вставки, применения или запрета применения данных информационной системы.

Троянские кони. Специализированная программа, которая, как правило, выступает от лица других программ и разрешает действия, отличные от определенных в спецификации, которые используются программным обеспечением.

Со временем этот перечень может дополняться новыми составляющими, поскольку уровень роста программного и аппаратного обеспечения настолько интенсивный, что его даже тяжело спрогнозировать.

Некоторые из разрушительных программных влияний, например логические бомбы или троянские кони, могут быть реализованные в виде программных закладок («жучков»), которые умышленно внедряются в тело определенных программ с целью реализации разрушительных действий на характеристики безопасности ресурсов информационных систем. Например, в командный процессор операционной системы можно установить программную закладку, которая является резидентной, выполняет функции логической бомбы (активизируется во время запуска DISKREET.EXE с нортоновских утилит) и назначается для записи в определенное место на диске (по обыкновению сектор, помеченный как сбойный) паролей, которые вводятся из клавиатуры. В этом случае такая закладка нарушает не только целостность командного процессора, а и конфиденциальность данных, которые передаются к нужной утилите с клавиатуры.

Классификация компьютерных вирусов. Одним из наиболее распространенных разрушительных программных влияний являются компьютерные вирусы. В мире созданы тысячи вирусов, а количество их разновидностей постоянно увеличивается. С учетом анализа указанных разрушительных программных влияний их удобнее классифицировать по таким признакам (рис. 9.28): *среде распространения*; *способу маскировки в среде*; *инфицированным объектам*; *способу инфицирования объекта*; *способу размещения в инфицированном объекте*; *разрушительным влияниям*; *способности к изменению*; *стилю написания*; *типу кода*.

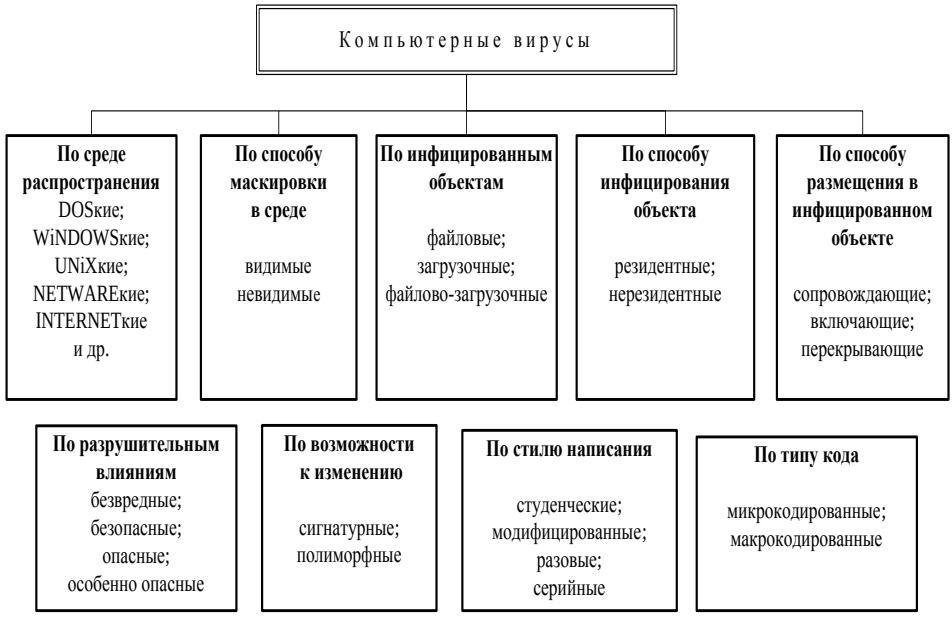


Рис. 9.28. Классификация компьютерных вирусов

По *среде распространения* вирусы можно поделить на такие, которые функционируют соответственно в среде DOS, WINDOWS, UNIX, NETWARE, INTERNET и других системах, функции и программные приложения которых используют вирусы. Название вирусам дают по имени среды, в которой они распространяются.

По *способу маскировки* в среде вирусы разделяют на видимые и невидимые. *Видимые вирусы* довольно легко выявить, используя простейшие средства. Например, с помощью команды DOS dir можно увидеть изменение размера инфицированного файла или, запустив простой шестнадцатеричный редактор, найти, например, сигнатуру вируса и т.д. *Невидимые вирусы* (stealth; стелс-вирусы; вирусы-невидимки) содержат алгоритмы, которые позволяют им маскироваться в среде распространения. Такие вирусы, находясь в системе, перехватывают ее обращения к инфицированным объектам и заменяют их инфицированные участки на оригинальные. После таких действий, например по команде dir, нельзя идентифицировать изменение размера файла, а 16-ричный редактор, который использует в своей работе функции ОС, не даст возможности найти, например, сигнатуру вируса и т.д.

По *типу инфицированных объектов* среди вирусов различают файловые, загрузочные, файлово-загрузочные. *Файловые вирусы* размещаются в файлах разных форматов (COM, EXE, SYS, DOC и т.д.) и, как правило, инициализируются первыми во время их обработки. *Загрузочные вирусы*

размещаются в Boot-секторах дисков или в секторе винчестера с системным загрузчиком и активизируются во время начальной загрузки ОС. *Файло-загрузочные вирусы* (соединение файловых и загрузочных) инфицируют как файлы, так и указанные секторы, а алгоритм их работы значительно усложняется с учетом бинарного действия.

По *способу инфицирования* объекта среди вирусов выделяют резидентные и нерезидентные. *Резидентные вирусы* после запуска оставляют в оперативной памяти компьютера свою резидентную часть, которая перехватывает обращение системы к объектам, которые подлежат инфицированию, и заражает их. Такой вирус остается в памяти вплоть до отключения компьютера или его перезагрузки. Иногда для того, чтобы инфицировать все файлы, например все выполняемые EXE-файлы текущего каталога диска A:, довольно вставить незащищенную дискету с файлами в дисковод A и выполнить команду dir A:. *Нерезидентные вирусы* инициализируются в период обработки системой инфицированного объекта и не являются активными (не заражают память) после этого периода. Поэтому нерезидентный вирус не может инфицировать другие файлы, если не будет инициирован в момент инфицирования носитель-файл-носитель.

По *способу размещения в инфицированном объекте* вирусы разделяют на сопроводительные, включающие и перекрывающие. *Сопроводительные вирусы* случаются очень редко; они инфицируют EXE-файли косвенно, т.е. для указанного файла создается новый с таким самым именем, но с COM-расширением, куда и вмещается тело вируса. В момент запуска файла с именем первым активизируется COM-файл (т.е. вирус), который выполняет свои функции и дальше запускает одноименный EXE-файл. *Включающие вирусы* встраиваются в начало, конец или во внутреннюю часть файла, при этом границы последнего расширяются и размеры файлов соответственно увеличиваются на размер введенного тела вируса, а сам инфицированный объект остается неповрежденным и сохраняет свою работоспособность. *Перекручивающие вирусы* бесповоротно повреждают инфицированный объект, поскольку в случае его заражения тело вируса накладывается на код объекта. Файлы, пораженные перекручивающим вирусом, не вылечиваются и потому по обыкновению изымаются.

По *разрушительному влиянию* среди вирусов различают безвредные, безопасные, опасные и особенно опасные. *Безвредные вирусы* соответственно их алгоритму не наносят прямой ущерб информационной системе, за исключением занятого дискового пространства и части оперативной памяти (особенно, если вирус резидентный). Но любой вирус, даже безвредный, может проявить себя иначе в новых условиях (например, в случае изменения версии операционной системы или среды, характерной для распространения вирусов, форматов дисков и т.д.) и привести к непредусмотренным следствиям. *Безопасные вирусы* характеризуются проявлениями разных звуковых и видеоэффектов и аналогично безвредным вирусам уменьшают размер свободной памяти. *Опасные*

вирусы не только уменьшают размер свободного дискового пространства и оперативной памяти, а и приводят к серьезным сбоям в работе системы. *Особенно опасные вирусы* наносят наибольший ущерб пользователю и системе. Они уничтожают программы, данные, разные записи на дисках и другую информацию, которая может привести к безвозвратным потерям и серьезным отказам в работе системы.

По *способности к изменению* вирусы разделяют на сигнатурные и полиморфные. *Сигнатурные вирусы* всегда имеют постоянный код, или в теле вируса можно по крайней мере выделить неизменный код (сигнатуру), за которым его можно распознать. *Полиморфные вирусы* очень сложно проявить, поскольку они не имеют сигнатур. Другими словами, сложно найти два одинаковых тела вируса. Такой эффект достигается через трансформацию его тела , например с помощью несложного шифрования (по обыкновению с использованием операции XOR), а также сменой программы шифрования. Указанные вирусы называют *призраками* и довольно часто при их построении используют стелс-технологию, которая делает их невидимыми в соответствующей среде.

По *стилю написания* среди вирусов выделяют студенческие, модифицированные, разовые и серийные. *Студенческие вирусы* содержат по обыкновению много ошибок, построенные по несложному алгоритму и являются первыми попытками авторов в этой области программирования. *Модифицированные вирусы* создают путем изменения кода вируса, который раньше создали другие авторы. Разовые вирусы создают большей частью опытные программисты после кропотливой работы по обыкновению с целью проверки своих профессиональных способностей. *Серийные вирусы* создают авторы или группа авторов и выполняют их в едином стиле. Исследуя их, можно увидеть, как возрастает мастерство производителя.

По *типу кода* вирусы разделяют на микрокодированные и макрокодированные. *Микрокодированные вирусы* образовываются, как правило, вследствие работы компилятора и являются бинарным кодом, который присоединяется к инфицированному объекту. *Макрокодированные вирусы* по обыкновению обрабатываются интерпретатором макрокода и являются последовательностью макрокодов, которые сохраняются преимущественно в файлах (например, с расширением DOC), созданных прикладными программами, которые имеют возможность создания макросов (например, MS WORD for WINDOWS).

Такие развернутые классификации делают возможным создание более эффективных комплексных систем защиты информации.

9.2. Средства защиты от НСД

Индустрия современных средств защиты информации от НСД определяется широким номенклатурным арсеналом. Такие средства с практической точки зрения можно поделить на такие классы: аппаратные, программные, программно-аппаратные, криптографические, стенографические, организационные, законодательные и морально-этические.

Аппаратные средства — разнообразные механические, электрические, электромеханические, электронные, электронно-механические и другие устройства и системы (например, источники бесперебойного питания, криптографические вычислители, электронные идентификаторы и ключи, устройства для выявления «жучков», генераторы шума и т.п.), функционирующие автономно или встраивающиеся или соединяющиеся с другой аппаратурой с целью блокирования действий дестабилизирующих факторов и решения других задач защиты информации.

Программные средства — специальные программы (например, антивирусы, шифрование данных, реализация алгоритмов цифровой подписи, разделение доступа, оценки рисков, определение уровня безопасности, организация экспертиз и т.п.), которые функционируют в пределах информационных систем для решения задач защиты информации.

Программно-аппаратные средства — взаимосвязанные аппаратные и программные средства (например, банковские системы электронных платежей, информационные системы конфиденциальной связи, автоматизированные системы контроля доступа персонала и транспортных средств в режимные зоны и т.п.), функционирующие автономно или в составе других систем с целью решения задач защиты информации.

Криптографические средства — средства, предназначенные для защиты информации путем криптографического преобразования информации (шифрование, дешифровка), которое реализуется с помощью асимметричных или симметричных криптографических систем. Асимметричные криптографические системы базируются на криптографии с открытым ключом. Например, известнейшими практическими реализациями этого типа являются системы Диффи - Хелмана, RSA и Эль-Гамаля. Симметричные криптографические системы базируются на криптографии с секретным ключом, наиболее известными из которых являются, например, DES, ГОСТ и т.п.

Практическое использование современных криптографических средств тесно связано с фундаментальными исследованиями в этой области и осуществляется через соответствующие аппаратные, программные и аппаратно-программные средства, например системы Тессера, Клиппера, Криптона и т.п. Следует отметить, что с этим классом средств тесно связан криптоанализ, эффективно используемый для испытания надежности криптографических систем.

Стеганографические средства сориентированы на утаивание информации в такой форме, когда сам факт ее наличия не очевидный, например утаивание данных в звуковых или графических файлах, которые входят в состав ОС Windows.

Организационные средства защиты информации — это множество процессов и действий (например, контроль за утилизацией носителей информации с ограниченным доступом, планирование мероприятий по восстановлению утраченной информации, аудит систем защиты, реализация экспертиз и т.п.), осуществляемых на всех технологических этапах (проектирование, изготовление, модификация, эксплуатация, утилизация и т.п.), и ведут к созданию, усовершенствованию, упорядочению и согласованности взаимосвязей и взаимодействия их компонент с целью решения задач защиты информации. Разрабатывая организационные средства, необходимо учитывать, чтобы в общем множестве механизмов защиты они могли самостоятельно или в комплексе с другими средствами решать задачи защиты, обеспечивать эффективное использование средств других классов, а также рационально объединять все средства в единую целостную систему защиты. Следует отметить, что множество всех нужных и потенциально возможных организационных средства не определены и не существует формальных методов формирования их перечня и содержания. Учитывая это, основными методами формирования организационных средств можно считать лишь неформально-эвристические.

Законодательные средства защиты информации являются множеством нормативно-правовых актов (конвенции, законы, указы, постановления, нормативные документы и т.п.), действующих в определенном государстве и обеспечивающих юридическую поддержку для решения задач защиты информации. Вообще с помощью законодательных средств определяются права, обязанности и ответственность относительно правил взаимодействия с информацией, нарушение которых может повлиять на состояние ее защищенности. В мировой практике основу указанных средств составляют патентное и авторское право, национальные законы о государственной тайне и обработке информации в информационных системах, лицензирование, страхование, сертификация, классификационные нормативные документы и т.п. (глава 10).

Морально-этические средства — моральные нормы и этические правила, которые сложились в обществе, коллективе и объекте информационной деятельности, нарушения которых отождествляется с несоблюдением общепринятых дисциплинарных правил и профессиональных идеалов. Примером таких средств может быть кодекс чести, этикет, этика хакера и т.п.

Базовые требования к любому средству защиты информационных систем можно разбить на пять категорий (рис. 9.29).

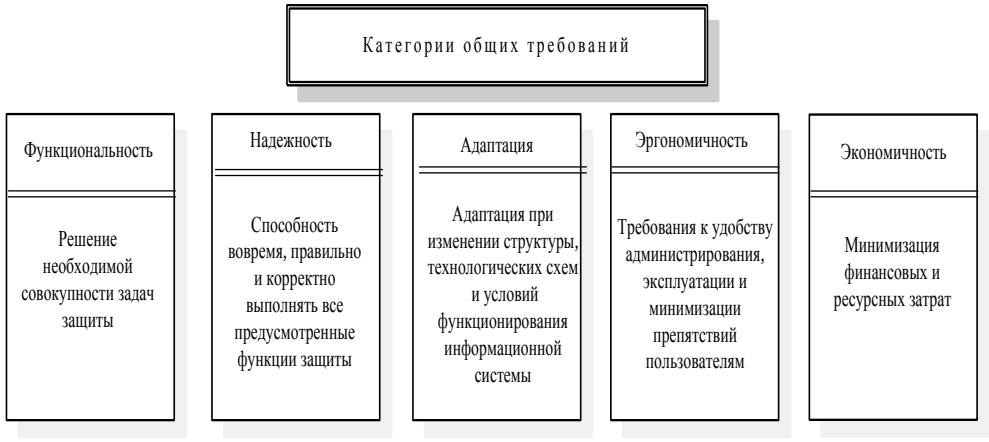


Рис. 9.29 Базовые требования к средствам защиты информационных систем

Категории средств защиты программного обеспечения. Как показывает практика, наиболее дорогой составляющей информационной системы есть ее программное обеспечение. Поэтому разработка программных средств защиты информации для защиты программного обеспечения с практической точки зрения - наиболее привлекательные задачи. Вопрос защиты программного обеспечения от хакеров (относительно его копирования и динамических и статических исследовательский приемов) острее возникает перед его разработчиками и владельцами. Из системных позиций защита программного обеспечения осуществляется целым комплексом средств, который начинается законодательными актами и заканчивается конкретными аппаратурными разработками. Известны такие категории средств защиты программного обеспечения (рис. 9.30): собственные, в составе информационной системы, с запросом информации, активные и пассивные.

Средства собственной защиты определяют элементы защиты, которые содержатся в самом программном обеспечении или сопровождают его. К ним относятся документация, распространение продукции в виде исполнительных модулей, сопровождение программ разработчиком, ограничение применения, проектирование на заказ, встроенные идентификационные метки владельца и авторское право.

К *средствам защиты в составе информационных систем* относят защиту магнитных дисков, защитные механизмы самых устройств информационных систем, замки защиты доступа и изменение функции в системе.

Среди механизмов защиты магнитных дисков нужно выделить две группы:

1. Препятствие непосредственному копированию программ из диска на диск.
2. Защита программ от реассемблирования и настройщиков.

Реассемблеры и настройщики дают возможность подать программу в форме, более доступной для восприятия, а также помогают выучить логику защиты и оперативно осуществить модификацию программ.

Рис. 9.30. Категории средств защиты программного обеспечения

Первая группа защищает программу от несанкционированного воспроизведения, вторая - от несанкционированной ревизии. Эти группы не взаимосвязанные, поскольку программы, предназначенные для свободного распространения, могут быть защищенные от настройщиков и реассемблеров, и наоборот, разные программные продукты могут быть защищены не только от копирования. Наиболее надежная защита обеспечивается одновременным применением двух групп, поскольку ее преодоление в программе, скопированной из защищенного диска, будет связано с преодолением защиты второй группы. Итак, программы, защищенные от настройщиков и реассемблеров, которые содержатся на стандартном диске, могут быть легко скопированы, а те, что расположены на защищенном диске и не содержат второй группы, могут быть легко изучены и модифицированы с целью преодоления механизмов защиты.

Программы, которые используют *средства защиты с запросом информации* для дальнейшего выполнения предназначенных функций, осуществляют запрос на введение дополнительной информации, представленной, например, в виде ключевых слов. В структуру указанных средств входят пароли, криптографические шифры, сигнатуры на основе уникальных характеристик информационных систем и аппаратуры защиты.

Средства активной защиты разделяют на внешние и внутренние; они активизируются в случае возникновения определенных (нештатных) обстоятельств, подпадающих под контроль, например неправильно введенные пароли, исчерпанное время пользования или количество запусков,

неправильная контрольная сумма участков программы и т.п.

Средства пассивной защиты охватывают методы идентификации программ, устройства контроля событий, водяные знаки, которые препятствуют созданию копии, и психологические мероприятия. С помощью этих средств осуществляется подтверждение подлинности, контроль доступа, поиск доказательств несанкционированного копирования и т.п.

Программные средства защиты являются важнейшей и необходимой частью механизма защиты современных информационных систем и могут быть как автономными, так и входить в состав разного программного обеспечения. Это прежде всего связано с такими их свойствами, как универсальность, гибкость, надежность, простота реализации, возможность модификации и усовершенствование. С универсальностью связана решаемость широкого круга задач защиты программными средствами. Гибкость ассоциируется с тем, что они могут адаптироваться к конкретным условиям функционирования информационных систем, а также к структуре информационных систем, например, быть составной частью ОС, функционировать как самостоятельные пакеты программ защиты, распределяться между отдельными компонентами системы и т.д. С надежностью связана высокая программная стойкость в случае продолжительной беспрерывной работы и удовлетворение высоких требований относительно достоверности влияний управления при наличии разных дестабилизирующих факторов. Простота реализации программных средств защиты очевидна сравнительно с возможностью реализации любых других средств. Возможности их модификации и усовершенствования определяются самой природой. Основными недостатками их использования является то, что они дополнительно загружают процессор, что приводит к увеличению времени реагирования на запросы и, соответственно, к уменьшению эффективности работы, а также к уменьшению емкости доступной и внешней оперативной памяти. Следует отметить, что при случайной

Вильям Сили Госсет «Стьюдент» (William Sealy Gosset «Student», 1876—1937), химик и специалист по математической статистике, больше известный под своим псевдонимом Стьюдент (Студент). В 1899 г. окончил колледж в Оксфорде и начал работать в компании Arthur Guinness & Son, которая занималась прогрессивным агрохимическим бизнесом. Там Госсет применил свои знания по математической статистике, разработав критерий для оценивания качества пива. Учитывая коммерческую тайну, статья Госсета об *t*-распределении Стьюдента увидела мир в журнале "Биометрика", где автор выступал под указанным псевдонимом.

или злонамеренной модификации программы она может потерять способность выполнять функции защиты, а также стать дополнительным каналом несанкционированного получения информации. Кроме этого, жесткая ориентация на архитектуру определенных типов ПК, а также зависимость от ОС дополняет указанные недостатки.

Перечень программных средств защиты может быть любым, а выполняемые ими функции могут быть такие: проверка прав доступа (простым паролем, сложным паролем, разовыми паролями); распознавание пользователей по разным идентификаторам, компонентам программного обеспечения и элементам баз данных; размежевание доступа к защищенным данным по матрицам полномочий, уровню секретности и другим признакам; управление доступом к задачам, программам и элементам баз данных по специальным мандатам; регистрация обращений к системе, задачам, программам и элементам защищенных данных; подготовка к выдаче конфиденциальных документов (формирование и нумерация страниц, определение и присвоение грифа ограничения доступа, регистрация выданных документов и т.п.); проверка адресата перед выдачей защищенных данных в каналы связи; управление выдачей данных в каналы связи; криптографическое преобразование данных; контроль процессов обработки и выдачи защищенных данных; уничтожение остаточной информации в постоянном запоминающем устройстве после выполнения запросов пользователей; сигнализация попыток несанкционированных действий; блокирование работы пользователей, нарушающих правила защиты информации; организация псевдоработы с нарушителем с целью отвлечения его внимания; обеспечение комплексных средств и систем защиты; комплексная защита от вирусов и других программных разрушительных влияний; обеспечение проведения экспертиз и сертификации; управление риском и организация действий в кризисных ситуациях и т.п.

Следует отметить, что индустрии этого направления присуща стихийность развития программ защиты, которая, с одной стороны, не дает гарантий полноты имеющихся средств, а с другой — не исключает дублирования тех самых задач защиты. Учитывая сказанное, можно выделить три принципиально важных *требования к формированию программных средств защиты*: *функциональная полнота, гибкость и унифицированность использования.*

Что касается первого требования, то нетрудно убедиться, что по приведенному перечню можно создать комплексный продукт, который охватывает все классы защиты.

Удовлетворение других двух требований зависит от форм и способов представления программ защиты. Анализ показал, что полнее требованиям гибкости и унифицированности удовлетворяет такая совокупность принципов: сквозное модульное построение, полная структуризация, представление машинонезависимым языком.

Принцип сквозного модульного построения заключается в том, что каждую из программ любого уровня (объема) подают в виде системы модулей, каждый

из которых должны быть целиком автономным и иметь стандартные вход и выход, которые обеспечивают комплексирование с любыми другими модулями.

Представление машинонезависимым языком означает, что программные модули должны быть такими, чтобы их с минимальными усилиями можно было включать в состав программного обеспечения любой информационной системы. Полную универсализацию представления модулей можно обеспечивать представлением их в виде блок-схемы, детализированной настолько, чтобы каждый блок можно было реализовывать малым количеством операторов наиболее распространенных языков высокого или низкого уровня.

9.3. Моделирование систем и процессов защиты информации

В процессе развития теории и практики информационной безопасности сформировались *эмпирический, теоретический и теоретико-эмпирический методологические подходы* к оценке уязвимости информации.

Суть *эмпирического* подхода заключается в том, что на основе продолжительного сбора и обработки данных реальных проявлений угроз информации и размеров причиненного убытка исключительно эмпирическим путем устанавливаются зависимости между потенциально возможным убытком и коэффициентами, которые характеризуют частоту проявления соответствующей угрозы и значение размера убытка.

Исходной предпосылкой при разработке моделей являются предположения, что при нарушении защищенности информации наносится некоторый убыток, а обеспечение защиты информации связано с затратами. Ожидаемую стоимость защиты можно определить суммой затрат на защиту и убытками от ее нарушения (рис. 9.31).

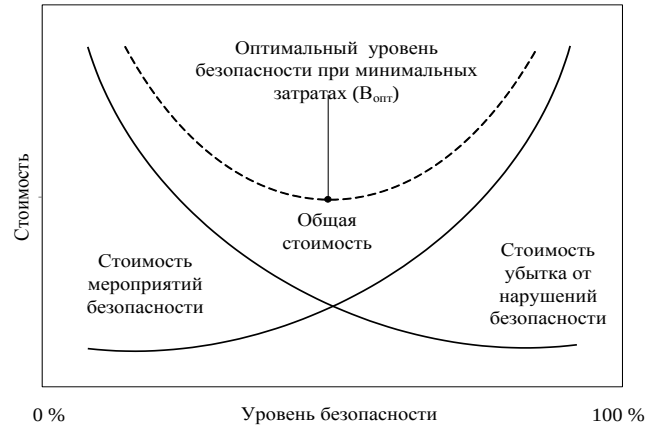


Рис. 9.31. Стоимость защиты информации

Очевидно, что оптимальным решением является выделение средств на защиту информации в размере $B_{опт}$, поскольку именно при этом обеспечивается минимизация общей стоимости защиты информации.

Для того чтобы воспользоваться этим подходом к решению проблемы, необходимо определить ожидаемые потери от нарушения защищенности информации и зависимость между уровнем защищенности и затратами на ее защиту. Затраты, которые обеспечивают необходимый уровень защищенности, зависят от полного множества угроз информации, потенциальной опасности от реализации каждой из них и размеров затрат, необходимых для их нейтрализации.

Поскольку оптимальный уровень затрат на защиту (см. рис. 9.31) отвечает уровню ожидаемых убытков от нарушений безопасности, то достаточно определить только уровень убытка. Ожидаемые убытки от *i*-ой угрозы информации можно определить за известной эмпирической зависимостью:

$$R_i = 10^{(S_i + V_i - 4)},$$

где S_i и V_i — коэффициенты, которые характеризуют возможную частоту возникновения соответствующей угрозы и значение возможного убытка при ее возникновении. Значение этих коэффициентов было определено на основе заключений экспертов.

Ожидаемая частота появления угрозы	S_i
Почти никогда	0
1 раз на 1000 лет	1
1 раз на 100 лет	2
1 раз на 10 лет	3
2 раза на неделю	6
3 раза на неделю	7

Значение возможного убытка при возникновении угрозы (USD)	V_i
1	0
10	1
100	2
1000	3
10000	4
100000	5
1000000	6
10000000	7

Суммарный убыток определяется выражением:

$$R_i = \sum_{\forall_i} \alpha_i V_i,$$

где α_i — весовой коэффициент.

Рассмотренная модель является достаточно приблизительной, а увеличение ее адекватности приводит к сложным аналитическим выкладкам, которые базируются на методах теории вероятностей и принятия решений.

В связи с этим проиллюстрируем известную динамическую модель оценки потенциальных угроз.

Пусть λ — средний коэффициент возможного появления угрозы определенного типа, тогда в общем случае этот коэффициент будет рассматриваться как случайная переменная $\bar{\lambda}$ с распределением вероятностей $f(\lambda)$. Функция распределения определится на основе количества проявлений угроз в процессе реального функционирования информационной системы.

Количеству проявлений данной угрозы $r_{\bar{t}}$ за фиксированное время (например, год) соответствует распределение вероятности $f(r/\lambda)$, а если количество проявлений угроз зависит только от продолжительности периода наблюдений и среднего коэффициента проявления, то исполняется функция распределения Пуассона

$$P(\bar{r} = r / \lambda) = \frac{(\lambda t) r e^{-\lambda t}}{r!},\ r = 0, 1, 2,\ \ ...,$$

где t — число периодов времени, за которые определено r.

По ряду значений r_i ($i = \overline{1, n}$) функция $f(\lambda)$ может быть отображена функцией гамма-распределения

$$f(\lambda / a, b) = \frac{b^a \lambda^{a-1} e^{b\lambda}}{(a-1)!},$$

где a и b — параметры распределения, которые определяются рекуррентными зависимостями:

$$a'' = a' + \sum_{i=1}^{n} r_i;\ \ b'' = b' + \sum_{i=1}^{n} t_i,$$

где r_i (i = 1, n) — количество проявлений рассматриваемой угрозы за периоды наблюдения t_i ($i = \overline{1, n}$).

Безусловное распределение вероятностей количества проявлений угроз за период времени t определяется формулой

$$f(r / a,\ b,\ t) = \int_0^{\infty} f(r / \lambda t)\, f(\lambda / a, b) d\lambda,$$

а результирующее распределение имеет вид

$$f_{nb}(r / P,\ a) = \frac{(r + a - 1)!}{r!(a-1)!} P^r (1 - P).$$

Эффективность защиты за определенные периоды функционирования системы характеризуется параметрами *a* и *b*. Ожидаемое количество проявлений угроз в последующие *t* периодов характеризуется математическим ожиданием

$$E(\overline{r} / t) = at/b$$

и дисперсией

$$V = (\overline{r} / t) = \frac{at}{b(t+b)}.$$

Подходы к оценке стоимости проявления угроз состоят вот в чем. Прежде всего рассматриваются средние стоимости проявления угроз, а потому используется нормальная функция распределения с параметрами *m* и *V*

$$f(\overline{m}, V / m', V', n', k') = f_n(m / m', n', V) f_g(V / V', \chi'),$$

где $f_n(\cdot)$, $f_g(\cdot)$ - нормальный и гамма-распределение вероятностей соответственно, а n', χ' - параметры гамма-распределения.

Если за следующие *t* периодов времени происходит *r* проявлений рассмотренной угрозы, которая приводит к убытку в размере соответственно x_i (i = 1, *n)*, то параметры распределения вероятностей ожидаемых потерь корректируются так

$$m'' = \frac{n'm' + r\overline{x}}{n''};$$

$$V'' = \frac{\chi'V'n'(m')^2 + (r-1)S^2 + r\overline{x}^2 - n''(m'')^2}{k'+2},$$

где $n'' = n' + r;\ \chi'' = \chi' + r;\ \overline{x} = \sum x_i / r;\ S^2 = \sum (x_i - \overline{x})^2 / (r-1).$

Прогнозируемое распределение убытков от возможного проявления рассмотренной угрозы формируется путем выделения неопределенных параметров *m* и *V* из функции распределения вероятностей для стоимости проявления угрозы данного типа, тогда

$$f_S(x / m'', V'', K'') = \int_{-\infty}^{+\infty} \int_{0}^{\infty} f(x / m, V) f(m / m'', V'', V) f_S(V / V'', \chi'')\, dmdV,$$

где $f(\cdot)$ — член семьи распределения Стьюдента.

Ожидаемое изменение значения $\overline{X}$ определяется параметрами

$$E(\overline{x}) = m; \qquad V(\overline{x}) = \frac{\chi''}{K'' - 2}.$$

Необходимо указать, что возможное количество проявлений *i*-й угрозы данного типа за данный период времени r_i и стоимость каждого проявления этой угрозы x_{ij} ($j = i = 1, r$) являются случайными переменными, поэтому ожидаемая полная стоимость угроз за t периодов времени $\overline{C}_t$ определяется по формуле

$$\overline{C}_t = \sum_{i=1}^{t} \sum_{j=1}^{\overline{r}_i} \overline{x}_{ij},$$

а ожидаемая стоимость проявления угрозы одного *i*-го типа - по формуле

$$\overline{C}_{i,t} = \sum_{j=1}^{\overline{r}} \overline{x}_{ij}.$$

Следует отметить, что стоимости проявления угроз являются случайными величинами, а для полной их оценки необходимо определить функции распределения их вероятностей.

Если бы удалось собрать достаточное количество фактических данных о проявлениях и следствиях угроз, то рассмотренную модель можно было бы использовать для решения достаточно широкого круга задач защиты информации.

Рассмотренная модель может иметь *игровую интерпретацию*, т.е. ее можно свести к постановке в *сроках теории игр*. Предположим, что неавторизованная сторона применяет x средств с целью преодоления механизма защиты, на создание которого израсходовано y средств. Тогда ожидаемое количество информации, получаемое неавторизованной стороной, является некоторой функцией $I(x, y)$. Если $f(n)$ для неавторизованной стороны является ценностью n единиц информации, а $g(n)$ является суммарными затратами на создание и сохранение этой самой информации, то чистая прибыль неавторизованной стороны определяется по формуле

$$V(x, y) = f[I(x, y)] - x;$$

а потери –

$$u(x, y) = g[I(x, y)] + y.$$

В соответствии с известными правилами теории игр оптимальные стратегии обеих сторон можно определить по формулам:

$$f'[I(x, y)]\frac{dI(x, y)}{dx} = 1;$$

$$g'[I(x, y)]\frac{dI(x, y)}{dy} = -1.$$

Эта модель с *теоретической точки зрения* достаточно строгая, но для практического использования необходимо знать стоимость информации, а также функции I, f и g для общего случая, который до сих пор является нерешенной проблемой.

Модель с полным перекрытием. Развитием моделей оценки угроз информационных систем являются модели их нейтрализации, т.е. модели защиты, наиболее общей из которых есть модель *системы с полным перекрытием.*

Построение этой модели базируется на том, что в механизме защиты должны содержаться по крайней мере одно средство для перекрытия любого потенциально возможного канала утечки информации. Рассмотрим методику формального описания моделирующей системы.

1. Составляется полный перечень объектов O системы, которые подлежат защите.

2. Составляется полный перечень потенциально возможных угроз T информации.

3. Составленные таким образом множества объединяются в двудольный граф с соблюдением условия: ребро $<t_i, o_j>$ существует тогда и только тогда, когда угроза t_i есть реальной для объекта o_j.

4. Для каждого ребра в графе определяется количественная мера соответствующей угрозы для соответствующего объекта.

5. Формируется множество M средств защиты информации в вычислительной системе.

6. Определяется количественная мера возможности противодействия каждого средства защиты каждой из угроз. Если возможность противодействия превышает уровень угрозы, то соответствующее ребро графа исключается.

Если множество M такое, что устраняются все ребра графа, то такая система является ***системой с полным перекрытием***.

Одной из разновидностей *теоретически* строгих моделей являются модели систем разделения доступа к ресурсам информационных систем.

В общем случае сущность этих моделей можно описать так. Информационная система *является системой множественного доступа, т.е. к тем самым ее ресурсам имеет права доступа некоторое количество пользователей* (*процессов*). Если любые из указанных ресурсов защищаются, то доступ к ним осуществляется лишь при наличии соответствующих полномочий. При этом система разделения доступа является механизмом, который регулирует такой доступ. По этому механизму не должен быть разрешен доступ пользователям (процессам), которые не имеют на это полномочий, и не должно быть отказано в доступе пользователям (процессам), которые имеют соответствующие полномочия.

Рассмотрим пример модели дискреционного доступа (АДЕПТ–50), построенной на названных механизмах. Основными структурными элементами этой модели есть объекты таких типов: пользователь u, задача j, терминал t и файл f. Объект каждого типа целиком описывается с помощью четырех характеристик: A - уровень компетенции, выраженный наибольшим грифом секретности данных (для данного объекта); C - категория доступа к данным,

по которой разрешен доступ для объекта; F - полномочие пользователей, которые имеют доступ к объекту; M - режим, выраженный перечнем процедур, разрешенных для соответствующего объекта.

На базе такого формального описания системы можно сформулировать систему формальных правил регулирования доступа. Например:

1) пользователь u получает доступ к задаче j тогда и только тогда, когда $u \in U$, где U — множество всех пользователей, зарегистрированных в системе;

2) пользователь u получает доступ к терминалу t тогда и только тогда, когда $u \in F(t)$;

3) пользователь u получает доступ к файла f тогда и только тогда, когда $u \in F(f)$; $A(u) \geq A(f)$; $C(u) \geq C(f)$; $M(u) \geq M(f)$;

4) из терминала t может быть осуществлен доступ к файла f тогда и только тогда, когда $F(t) \geq F(f)$; $A(t) \geq A(f)$; $C(t) \geq C(f)$; $M(t) \geq M(f)$.

Набор таких правил может расширяться и модифицироваться, а на их основе легко построить алгоритм управления доступом к данным.

К такому типу принадлежит известная пятимерная *модель безопасности* Хартсона, согласно которой для формального описания процесса доступа к данным в условиях защиты введено пять таких множеств: U - список зарегистрированных пользователей; R - набор имеющихся в системе ресурсов; S - множество возможных состояний ресурсов; E - набор операций над ресурсами; A - перечень возможных полномочий пользователей.

Далее вводится понятие *области безопасности* как декартового произведения указанных множеств

$$D = U \times A \times R \times S \times E.$$

В области безопасности можно выделить четыремерные подобласти, которые отвечают отдельным пользователям, группам пользователей, отдельным ресурсам и т.п.

Любой запрос на доступ можно описать кортежем

$$g = (u, r, s, e)\ (u \in U; r \in R; s \in S; e \in E).$$

Запрос получает право на доступ только в том случае, если он попадает в соответствующую подобласть области безопасности.

Описанная структуризация дает возможность построить алгоритмическую процедуру управления доступом, а обобщением моделей этого типа является модель, в которой сформулирована и строго решена задача разделения доступа.

Постановка задачи. Дана система (A, B, g), в которой:

$A = (A_1, A_2, ..., A_u)$ - конечный набор субъектов (активных элементов системы);

$B = (B_1, B_2, ..., B_v)$ - конечный набор объектов (пассивных элементов) системы;

a_i - код доступа субъекта i;

b_j - код доступа объекта j(i, j - некоторые двоичные числа);

$g = g(a_i, b_j)$ - механизм доступа, причем если $g(a_i, b_j) = 1$, то субъекту i разрешается доступ к объекту j, а если $g(a_i,b_j) = 0$, то указанный доступ не разрешается.

Задача заключается в том, чтобы при заданных наборах субъектов и объектов с их ограничениями в иерархических структурах и заданном механизме доступа выбрать такое значение разрядности кода доступа n и определить такое распределение значений кодов доступа субъектов и объектов, чтобы обеспечивались все разрешенные и минимизировались неразрешенные доступы.

Механизмом доступа может быть любая булевая функция

$$g(a_i, b_j) = \sum_{k=1}^{n} f(a_{ik}, b_{ik}),$$

которая удовлетворяет условию

$$g(a_i, b_j) = \begin{cases} 1, \text{ если} \sum_{k=1}^{n} f(a_{ij}, b_{jk}) \geq m \\ 0, \text{ в противном случае} \end{cases}.$$

Здесь $f(a_{ik}, b_{jk})$ — произвольная булевая функция двух бит; a_{ik} — значение k-го бита в коде доступа j-го объекта; m — порог доступа, причем $0 \leq m \leq n$.

С целью нахождения выражений для количественных оценок вводятся такие определения.

1. Пусть x_{ij} и y_{ij} - булевые переменные, где $x_{ij} = 1$ ($y_{ij} = 1$) означает, что A_i имеет разрешенный (неразрешенный) доступ к B_j, иначе $x_{ij} = 0$ ($y_{ij} = 0$). При этом $x_{ij} \wedge y_{ij} = 0$ для всех i ($1 \leq i \leq |A|$) и всех j ($1 \leq j \leq B_i$), где $|A|$ означает кардинальное число множеств A.

2. Пусть x_j - количество субъектов $A_i \in A$, имеющих разрешенный доступ к B_j, а y_j - количество субъектов, не имеющих такого доступа, причем

$$\begin{cases} x_j = \sum\limits_{A_i \in A} x_{ij}; \\ y_j = \sum\limits_{A_i \in A} y_{ij}. \end{cases}$$

3. Пусть $x(y)$ - среднее (относительно B) количество субъектов, которые имеют разрешенный (неразрешенный) доступ к любому $B_j \in B$, т.е.

$$\begin{cases} \overline{x} = \dfrac{\sum\limits_{B_j \in B} x_j}{|B|}; \\ \overline{y} = \dfrac{\sum\limits_{B_j \in B} y_j}{|B|}. \end{cases}$$

4. Пусть $x(y)$ - минимальное (относительно B) количество субъектов, которые имеют разрешенный (неразрешенный) доступ к любому $B_i \in B$, т.е.

$$\begin{cases} \breve{x} = \min\limits_{B_j \in B}(x_j); \\ \breve{y} = \min\limits_{B_j \in B}(y_j). \end{cases}$$

5. Пусть $x(y)$ - максимальное (относительно B) количество субъектов, которые имеют разрешенный (неразрешенный) доступ к любому $B_i \in B$, т.е.

$$\begin{cases} \hat{x} = \max\limits_{B_j \in B}(x_j); \\ \hat{y} = \max\limits_{B_j \in B}(y_j). \end{cases}$$

Тогда качество распределения кодов доступа можно оценить приведенными дальше показателями.

1. Абсолютная степень защиты

$$\delta_{\text{абс}} = (1 + \overline{y})^{-1}.$$

При этом $\delta_{\text{абс}} = 1$ будет только тогда, когда найдено такое распределение, при котором неразрешенных доступов нет вообще; в противоположном случае $\delta_{\text{абс}} < 1$. Значения этого показателя не зависит от количества субъектов и объектов, а изменяется только в зависимости от среднего количества неразрешенных доступов.

2. Относительная степень защиты

$$\delta_{\text{отн}} = \frac{|A| - \overline{x} - \overline{y}}{|A| - \overline{x}},$$

$0 \le \delta_{\text{отн}} \le 1$, причем $\delta_{\text{отн}} = 1$, если система не допускает неразрешенных доступов, а $\delta_{\text{отн}} = 0$, если система разрешает максимально возможное количество неразрешенных доступов.

3. Минимальная степень защиты

$$\breve{\delta} = (1 + \hat{y})^{-1}.$$

4. Максимальная степень защиты

$$\hat{\delta} = (1 + \breve{y})^{-1}.$$

Доказано, что оптимальное распределение кода будет при минимальном количестве наборов, которые разрешают доступ. Это достигается при равномерном распределении субъектов по всем классам доступа.

Рассмотрим дальше ***теоретико-эмпирический*** подход к оценке информации.

По определению, базовым есть показатель уязвимости информации в одном структурном ресурсе информационной системы относительно одного дестабилизирующего фактора и относительно одного нарушителя одной категории (для факторов, связанных с преступными действиями людей).

Для определения базовых показателей уязвимости разных видов применяются аналитические модели, которые дают возможность определять искомые величины (в данном случае - показатели впечатлительности информации) путем проведения вычислений по заранее установленным (выведенным) зависимостям.

Приведем некоторые аналитические модели.

1. **Нарушение физической целостности информации.** Введем такие обозначения:

$P_{i\,\text{вх}}^{(\text{ц})}$ - вероятность того, что на вход *i*-го структурного компонента поступает информация с затронутой целостностью;

$P_{ijk}^{(\text{ц})}$ - вероятность того, что целостность информации (обрабатываемой, сохраненной, переданной), которая содержится в *i*-му структурном компоненте, будет затронуто под влиянием *j*-го дестабилизирующего фактора *u* (в случае преступных действий людей) относительно одного нарушителя *k*-ой категории. Для тех дестабилизирующих факторов, которые не связанные с преступными действиями людей, индекс *k* игнорируется, т.е. значение $P_{ijk}^{(\text{ц})}$ для всех *k* одинаковые и зависят только от *i* и *j*;

$P_{ijk\,\text{вых}}^{(\text{ц})}$ - вероятность того, что целостность исходной из *i*-го компонента информации нарушена под влиянием *j*-го дестабилизирующего фактора *u* (в случае преступных действий людей) относительно одного нарушителя *k*-й категории (относительно индекса *k* исполняется высказанное только что замечание).

Соответственно теореме умножения вероятностей случайных событий величину $P_{ijk\,\text{вых}}^{(\text{ц})}$ (нарушение целостности исходной информации) можно подать такой зависимостью

$$P_{ijk\,\text{вых}}^{(\text{ц})} = 1 - \left(1 - P_{i\,\text{вх}}^{(\text{ц})}\right)\left(1 - P_{ijk}^{(\text{ц})}\right).$$

Раскрыв скобки и выполнив тождественные преобразования, окончательно получим:

$$P_{ijk\ \text{вых}}^{(\text{ц})} = P_{i\ \text{вх}}^{(\text{ц})} + \left[1 - P_{i\ \text{вх}}^{(\text{ц})}\right] P_{ijk}^{(\text{ц})} p. \qquad (9.1)$$

Графически эта зависимость представлена семейством кривых, приведенных на рис. 9.32.

Как видим, на эту модель есть определенные ограничения. Здесь предполагается такое:

1) разные дестабилизирующие факторы влияют на информацию независимо друг от друга;

2) процесс нарушения целостности ресурсов информационной системы не зависит от того, затронута ли целостность информации, которая поступает на его вход.

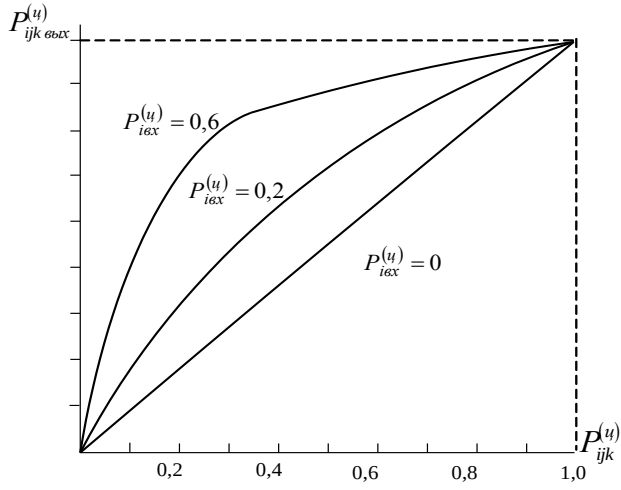


Рис. 9.32. График значений базовой вероятности нарушения физической целостности информации

Поэтому правомерность использования такой модели необходимо обосновать. События $P_{ijk}^{(\text{ц})}$ являются сложными с той точки зрения, что для их осуществления необходимо одновременное проявление соответствующего дестабилизирующего фактора и собственно нарушение целостности информации вследствие его действия. Если вероятность первого события обозначить через $P_{ijk}^{(\text{ц, п})}$, а второй — через $P_{ijk}^{(\text{ц, н})}$, то $P_{ijk}^{(\text{ц})} = P_{ijk}^{(\text{ц, п})} \cdot P_{ijk}^{(\text{ц, н})}$.

Подставляя это значение в формулу (9.1), получаем

$$P_{ijk\,\text{вых}}^{(\text{ц})} = P_{i\,\text{вх}}^{(\text{ц})} + \left[1 - P_{i\,\text{вх}}^{(\text{ц})}\right] \cdot P_{ijk}^{(\text{ц, п})} \cdot P_{ijk}^{(\text{ц, н})}.$$

Значение $P_{ijk\,\text{вых}}^{(\text{ц})}$ и является базовым показателем уязвимости с точки зрения нарушения целостности исходной информации. Тем не менее для его использования необходимы значения $P_{ijk}^{(\text{ц})}$ для всех структурных компонентов и всех дестабилизирующих факторов информационной системы. Следует отметить, что формирование всего множества значений этих величин связано с определенными трудностями.

2. **Несанкционированное получение информации**. С точки зрения несанкционированного получения информации главную опасность представляют преступные действия людей. Введем такие обозначения:

$P_{ikl}^{(\text{п, д})}$ - вероятность доступа нарушителя *k*-й категории в *l*-ю зону *i*-го ресурса информационной системы;

$P_{ijl}^{(\text{п, к})}$ - вероятность наличия (проявления) *j*-го канала несанкционированного получения информации в *l*-й зоне *i*-го ресурса информационной системы;

$P_{ijkl}^{(\text{п, н})}$ - вероятность доступа нарушителя *k*-й категории к *j*-го канала несанкционированного получения информации в *l*-й зоне *i*-го компонента при условии доступа нарушителя в зону;

$P_{ijl}^{(\text{п, і})}$ - вероятность наличия информации, которая защищается, в *j*-м канале несанкционированного получения информации в *l*-й зоне *i*-го компонента в момент доступа туда нарушителя. Укажем, что зоны являются составными, например, если $l = \overline{1,5}$ (рис. 9.33), то первая зона является внешней (первый рубеж), а пятая - внутренней (последний рубеж).

Но поскольку под базовым показателем уязвимости информации (с точки зрения несанкционированного получения) понимается вероятность несанкционированного ее получения в одном ресурсе информационной системы одной неавторизованной стороной одной категории по одному каналу несанкционированного получения информации, то выражение для базового показателя для пяти зон запишется так:

$$P_{ijk}^{(\text{п, б})} = 1 - \prod_{l=1}^{5}\left(1 - P_{ijkl}^{(\text{п, б})}\right) = 1 - \prod_{l=1}^{5}\left(1 - P_{ilk}^{(\text{п, д})} P_{ijl}^{(\text{п, к})} P_{ijkl}^{(\text{п, н})} P_{ijl}^{(\text{п, і})}\right).$$

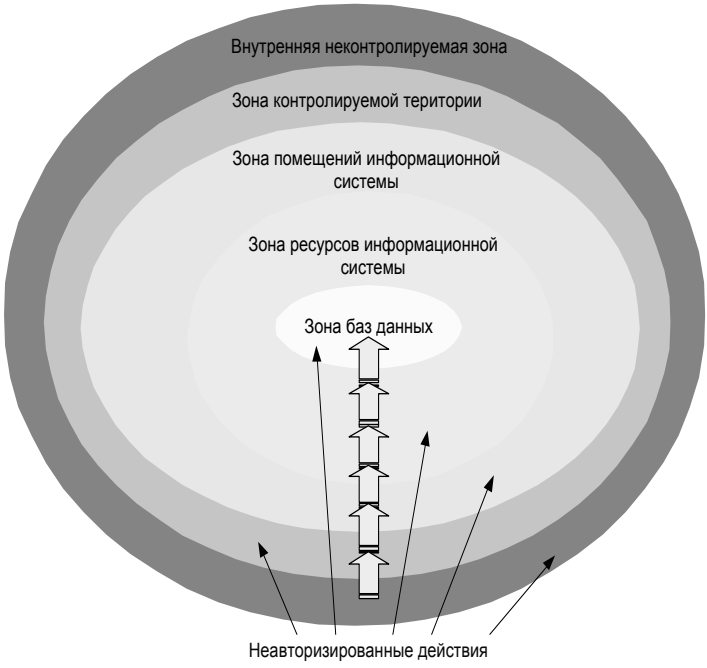


Рис. 9.33. Пример схемы возможных несанкционированных действий в информационной системе

Классические мандатные модели целостности и размежевания доступа. Модель мандатного размежевания доступа, которая получила название модели Белла - Лападула (БЛМ), до сих пор влияет на исследования и разработки в области компьютерной безопасности. Идеи, заложенные в БЛМ, могут быть использованы при построении разных политик безопасности (например, в основе «Оранжевой книги» [51]).

Принципы, которые положено в основу БЛМ, происходят от бумажного документирования информации с ограниченным доступом, а Белл и Лападула перенесли известные подходы обеспечения безопасности в информационные технологии. Здесь основным тезисом является то, что все субъекты и объекты ассоциируются с уровнями безопасности, которые варьируются от низких (неклассифицированных) уровней к высоким (совершенно секретных). Кроме того, для предотвращения утечки информации к неуполномоченным субъектам с низкими уровнями безопасности им не разрешается читать информацию из объектов с высокими уровнями безопасности. Этот тезис ведет к первому правилу БЛМ.

Простое свойство безопасности, известное также как правило «запрет чтения вверх» (NRU), говорит, что субъект с уровнем безопасности x_s может читать информацию из объекта с уровнем безопасности x_o, только если x_s преобладает над x_o. Это значит, что когда в системе, которая удовлетворяет

правилам модели БЛМ, субъект с уровнем доступа *секретный* попробует прочитать информацию из объекта, классифицированного как целиком секретный, то такой доступ не будет разрешен.

Белл и Лападула также использовали тезис, по которому субъектам не разрешается размещать информацию или записывать ее в объекты, которые имеют низший уровень безопасности. Например, когда целиком секретный документ перемещается в неклассифицированную мусорную корзину, то может состояться утечка информации, и это привело ко второму правилу БЛМ.

Свойство, известное как правило «запрет записи вниз» (NWD), говорит, что субъект безопасности x_s может записать информацию в объект с уровнем безопасности x_o только если x_o имеет преимущество над x_s. Это означает, что когда в системе, которая удовлетворяет правилам модели БЛМ, субъект с уровнем доступа *целиком секретный* попробует записать информацию в неклассифицированный объект, то такой доступ не будет разрешен. Такое свойство, например, решает проблему троянских коней, поскольку запись информации на низший уровень безопасности (типичное действие троянских коней) запрещен.

Правило запрета записи является упрощением некоторых реализаций БЛМ. Так, некоторые описания содержат более подробное понятие типа доступа (например, такие как добавление и выполнение).

Правила запрета записи и чтения БЛМ отвечают интуитивным понятиям того, как предотвратить утечку информации к неуполномоченным (неавторизованным) источникам.

Рассмотрим формализацию БЛМ. Введем обозначения: S - множество субъектов; $ОБ$ - множество объектов; L - решетка уровней безопасности; $F : S \bigcup O \to L$ - функция, которая применяется к субъектам и объектам и определяет уровни безопасности своих аргументов в данном стане; V - множество состояний (множество упорядоченных пар (F, M), где M - матрица доступа субъектов системы к объектам).

Система описывается начальным состоянием v_0 , определенным множеством запросов к системе R и функцией переходов T: $(V \times R) \to V$ таким, что система переходит из состояния в состояние после выполнения запроса. В связи с этим сформулируем определения, необходимые для доказательства основной теоремы безопасности, доказанной для БЛМ.

Определение 1. Состояние (F, M) безопасное к чтению (NRU) тогда и только тогда, когда для $\forall s \in S$ и для $\forall в \in O$ чтение $\in M$ $[s,o] \to F(s) \geq F(o)$.

Определение 2. Состояние (F, M) безопасное к записи (NWD) тогда и только тогда, когда для $\forall s \in S$ и для $\forall в \in O$ запись $\in M$ $[s, o] \to F(o) \geq F(s)$.

Определение 3. Состояние безопасное тогда и только тогда, когда оно безопасно к чтению и записи.

Теорема. Система (v_0, R, T) безопасна тогда и только тогда, когда состояние v_0 безопасно и T такое, что для любого состояния v, достигнутого с v_0 после выполнения конечной последовательности запросов из R, $T(v, c) = v^*$,

где $v = (F, M)$ и $v^* = (F^*, M^*)$, переходы системы (T) из состояния в состояние подлежат таким ограничениям для $\forall s \in S$ и для $\forall в \in ОБ$:

если чтение $\in M^*[s, o]$ и чтение $\notin M[s, o]$, то $F^*(s) \geq F^*(o)$;

если чтение $\in M[s, o]$ и $F^*(s) < F^*(o)$, то чтение $\notin M^*[s, o]$;

если запись $\in M^*[s, o]$ и запись $\notin M[s, o]$, то $F^*(o) \geq F^*(s)$;

если запись $\in M[s, o]$ и $F^*(o) < F^*(s)$, то запись $\notin M^*[s, o]$.

Доказательство. 1. *Необходимость*. Предположим, что система безопасна, а состояние v_0 безопасно по определению. Если есть некоторое состояние v, достигнутое из состояния v_0 после выполнения конечной последовательности запросов из R, таких что $T(v, c) = v^*$, хотя v^* не удовлетворяет одно из двух первых ограничений для T, то v^* будет достижимым состоянием, но таким, что противоречит ограничению безопасности по чтению. Если v^* не удовлетворяет одному из двух последних ограничений для T, то v^* будет достижимым состоянием, но противоречащим ограничению безопасности по записи. В любом случае система опасна.

2. *Достаточность*. Предположим, что система опасна. В этом случае или состояние v_0 должны быть опасным, или должно быть опасным состояние v, достижимое из состояния v_0 после выполнения конечной последовательности запросов из R . Если v_0 опасно, то все доказано, а если v_0 безопасно, то предположим, что v^* - первое в последовательности запросов опасное состояние. Это означает, что есть безопасное состояние v, такое что $T(v, c) = v^*$, где v^* - опасное. Но это противоречит четырем ограничениям безопасности на T.

Несмотря на все преимущества, выяснилось, что при использовании БЛМ в контексте практического проектирования и разработки реальных информационных систем возникает ряд технических вопросов, которые являются логическим следствием преимуществ БЛМ - ее простоты. Проблемы возникают при рассмотрении вопросов построения политик безопасности для конкретных типов систем, т.е. на менее абстрактном уровне рассмотрения. При данном рассмотрении системный компонент модели усложняется, что может привести к неадекватности БЛМ в ее классической форме, что инициировало широкую полемику по поводу применимости БЛМ для построения безопасных систем.

Мандатная модель контроля целостности (Модель Биба). Известно, что в БЛМ важность или чувствительность субъектов и объектов возрастает с повышением в иерархии уровней безопасности. При рассмотрении моделей контроля целостности запись вверх может быть угрозой в том случае, когда субъект с низким уровнем безопасности модифицирует или уничтожает данные в объекте, который лежит на более высоком уровне. Поэтому, исходя из задач целостности, существует требование запрета такой записи. Придерживаясь подобных аргументов, можно рассматривать чтения снизу как поток информации, которая идет от объекта нижнего уровня и поднимает целост-

ность субъекта высокого уровня. Поэтому достаточно вероятно, что и такое чтение необходимо запретить.

Таких два наблюдения сделал в середине 1970-х годов Кен Биба. Они были последовательно внесены в модель безопасности, которая с того времени называется *моделью целостности Биба* (или просто *моделью Биба*), который подал ее в таком же виде, что и БЛМ, хотя правила его модели являются полной противоположностью правилам БЛМ. По обыкновению рассматриваются три вариации модели Биба: мандатная модель целостности; модель снижения уровня субъекта; модель снижения уровня объекта, а общий термин «модель Биба» используется для обозначения любой или сразу всех трех моделей. Для мандатной модели контроля целостности известно формальное описание, и ее часто называют *инверсией БЛМ*. Это довольно точное название, поскольку основные правила этой модели противоположны правилам БЛМ и отображаются как «запрет чтения снизу» (NRD) и «запрет записи вверх» (NWU). Определим их в сроках субъектов, объектов и нового типа уровней безопасности — *уровней целостности*, к которым введено отношение преимущества.

Правило NRD мандатной модели целостности Биба определяется как запрет субъектам на чтение информации из объекта с более низким уровнем целостности. Правило NRD является полной противоположностью NRU БЛМ, за исключением того, что здесь используются уровни целостности (а не безопасности, как в БЛМ), а правило NWU (запрет субъектам на запись информации в объект с более высоким уровнем целостности) является полной противоположностью правилу NWD БЛМ для того самого случая уровней целостности.

Одним из преимуществ этой модели является то, что она унаследовала много важных характеристик БЛМ, включая ее простоту и интуитивность. Это означает, что проектировщики реальных систем могут легко понять сущность этих правил и использовать их для принятия решений при проектировании. Кроме того, поскольку мандатная модель целостности Биба, как и БЛМ, базируется на простой иерархии, ее легко объяснить и отобразить пользователям системы. Тем не менее, эта модель является очевидным разногласием с правилами NRU и NWD. Это означает, что когда необходимо построить систему, которая предотвращает угрозы секретности и целостности, то одновременное использование правил моделей БЛМ и Биба может привести к ситуации, в которой уровни безопасности и целостности будут достигаться противоположными способами.

Рассмотрим формальное описание модели Биба. Для этого приведем простые математические конструкции, которые помогут описать разные правила, которые образовывают эту мандатную модель.

Пусть существуют множества субъектов и объектов, где уровни целостности субъекта или объекта x обозначаются как уровень (x), и для них введено отношение *преимущества*. Используя эти определения, сформулируем

правила NRD и NWU мандатной модели целостности Биба в терминах булевой функции *РАЗРЕШИТЬ*:

NRD: ∀ *s* ∈ *субъекты*, *в* ∈ *объекты*:

РАЗРЕШИТЬ (*s*, *о*, чтение) тогда и только тогда, когда уровень (*о*) *преобладает* над уровнем (*s*).

Этот тип определения предусматривает условия, при которых функция *РАЗРЕШИТЬ* приобретает значение *истинной*. Определение утверждает, что для всех определенных субъектов и объектов операция чтения разрешена только в том случае, когда выполняется условие преимущества. Правило NWU является обратимым к использованию отношения преимущества, как это показано в таком определении:

NWU: ∀ *s* ∈ *субъекты*, *в* ∈ *объекты*:

РАЗРЕШИТЬ (*s*, *о*, запись) ⇔ clerence (s) ≥ classification (*о*).

Это определение утверждает, что для всех субъектов и объектов операция записи разрешается только в том случае, когда выполняется условие преимущества. Сходство определения этих двух правил правилам модели БЛМ может предоставить удобный способ для разработчиков системы предусмотреть возможность переконфигурирования правил БЛМ таким образом, чтобы поддерживать мандатную модель целостности Биба.

Модель снижения уровня субъекта (вторая модель Биба) связана с небольшим ослаблением правила чтения снизу. Эта модель не разрешает субъектам с высокой целостностью читать информацию из объектов с низкой целостностью. Такое правило гарантирует, что информация из объекта с низкой целостностью не нарушит целостности субъекта. Но в этой модели субъекта разрешается осуществлять чтение снизу, в результате которого уровень целостности субъекта снижается к уровню целостности объекта.

Модель снижения уровня объекта (третья модель Биба) связанная с ослаблением правил записи вверх, т.е. вместо полного запрета записи вверх модель разрешает такую запись, но снижает уровень целостности объекта к уровню целостности субъекта, который осуществляет запись.

9.4. Противодействие сетевому несанкционированному доступу

Основой современных информационных систем являются компьютерные сети, на которые также распространяются действия прежде описанных методов НСД. В общем случае сеть организации фактически является неоднородным рядом компьютеров разной конфигурации и назначения, управляемых разными операционными системами и связанных между собой с помощью сетевого оборудования. В таких условиях осуществить надежную защиту от несанкционированных действий при взаимодействии с внешними сетями можно лишь с помощью специализированных программно-аппаратных комплексов, которые обеспечивают соответствующую защиту. Такие комплексы называют межсетевыми экранами, межсетевыми фильтрами, экранированными

фильтрами, брандмауэрами или системами FireWall, которые устанавливают на стыке между внутренней и внешней сетями и возлагают на них функции противодействия.

Функции межсетевого экранирования. Как уже отмечалось, для защиты от несанкционированного межсетевого доступа брандмауэр располагается между сетью, которая подлежит защите (внутренней), и сетью, из которой возможен НСД (внешней) (рис. 9.34), при этом все внутрисетевые взаимодействия осуществляются только через экранированный фильтр, который организационно входит в состав внутренней сети.

Межсетевой экран должен учитывать протоколы информационного обмена, которые положены в основу функционирования внутренней и внешней сетей, а если протоколы отличаются, то брандмауэр должен поддерживать многопротокольный режим работы, обеспечивая протокольное преобразование уровней соответственно модели OSI для взаимодействия открытых сетей.

Для межсетевого экрана отдельно задаются правила, которые ограничивают доступ из внутренней сети во внешнюю, и наоборот, и в общем случае его работа базируется на динамическом выполнении двух групп функций:

фильтрации информационных потоков;

посредничества в межсетевом взаимодействии.

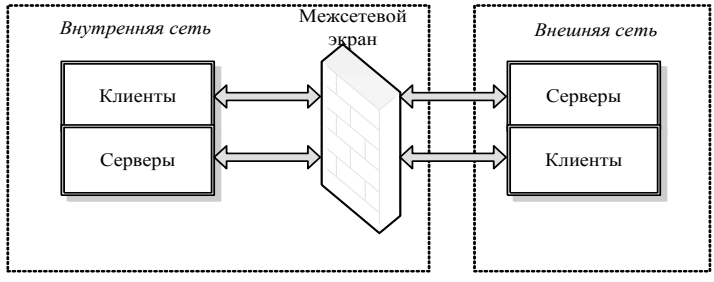


Рис. 9.34. Схема подключения брандмауэра

При этом простые межсетевые экраны сориентированы на выполнение только одной из указанных функций, а комплексные - обеспечивают совместное их выполнение. Для принятия управленческих решений относительно используемых сервисов брандмауэр должен получать, запоминать, выбирать и обрабатывать информацию, полученную от всех коммуникационных уровней и прикладных программ. Полнота и правильность управления требуют, чтобы комплексный брандмауэр имел возможность анализа и использования ряда факторов:

информации о соединении - информации от всех семи уровней в пакете;

истории соединений - информации, полученной от предыдущих соединений;

состояния уровня прикладной программы - информации о состоянии,

полученной из других прикладных программ;

агрегативного элемента - вычисление разнообразных зависимостей, которые базируются на всех перечисленных факторах.

При экранировании отдельного компьютера поддерживается доступность сетевых сервисов и уменьшается нагрузка, инициированная внешней активностью, которая снижает уязвимость защищенных внутренних сервисов компьютера, поскольку сначала неавторизованная сторона должна преодолеть механизм защиты экранированнго фильтра.

Базовая задача брандмауэра - фильтрация трафика, связанная с выборочным пропуском данных через экранированный фильтр (иногда с выполнением некоторых преобразований).

Фильтрация осуществляется на основе ряда правил, которые загружаются в экран, и сетевых аспектов, которые выражают принятую политику безопасности. С этих позиций брандмауэр удобно представить как последовательность фильтров (рис. 9.35), которые обрабатывают информационные потоки на основе интерпретации отдельных правил фильтрации путем:

анализа информации по заданным в интерпретированных правилах критериям, например, по адресам получателя и (или) отправителя, по типу прикладной программы, для которой предназначена информация, и т.п.

принятия на основе интерпретированных правил одного из таких решений:

1) не пропустить данные;

2) обработать данные от лица получателя и возвратить результат отправителю;

3) передать данные на следующий фильтр для продолжения анализа;

4) пропустить данные, игнорируя следующие фильтры.

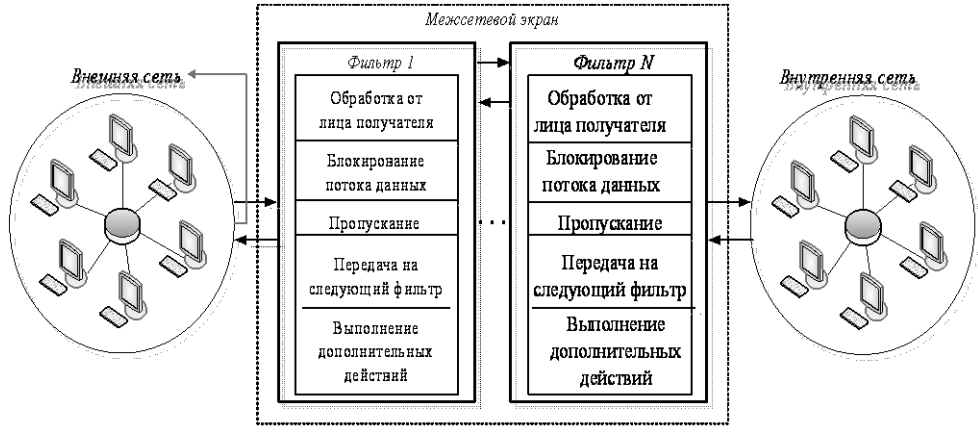


Рис. 9.35. Логическая структура межсетевого экрана

При фильтрации могут задаваться и дополнительные действия (посредничества), например преобразование данных, регистрация событий и т.п., и тогда соответствующие правила фильтрации определяют перечень условий, по которым с использованием указанных критериев анализа осуществляются: разрешение или запрет дальнейшей передачи данных; выполнение дополнительных защитных функций.

Критериями анализа информационного потока могут использоваться:

служебные поля пакетов сообщений, которые содержат сетевые адреса, идентификаторы, адреса интерфейсов, номера портов и другие важные данные;

непосредственное содержание пакетов сообщений (например, на наличие компьютерных вирусов), что подлежат проверке;

внешние характеристики информационного потока, например, временные и частотные характеристики, объем данных и т.п.

Используемые критерии анализа зависят от уровней модели OSI, на которых осуществляется фильтрация, а в общем случае действует правило: чем выше уровень модели OSI, на котором брандмауэр фильтрует пакеты, тем выше и обеспечиваемый им уровень защиты.

Еще одной базовой функцией брандмауэра является выполнение посреднических функций, которые он осуществляет с помощью специальных экранированных программ (агентов)-посредников, которые являются резидентными и контролируют (запрещают, разрешают) передачу пакетов между внешней и внутренней сетью.

В случае доступа из одной сети в другую сначала устанавливается логическое соединение с программой-посредником, которая проверяет допустимость запроса межсетевого взаимодействия и в случае соответствующего разрешения устанавливает соединение с необходимым компьютером. В дальнейшем обмен между компьютерами сетей осуществляется через программного посредника, который может выполнять фильтрацию трафика и осуществлять другие защитные функции.

Функции фильтрации межсетевой экран может выполнять без программ-посредников, обеспечивая прозрачное взаимодействие между внутренней и внешней сетями. Программные посредники, в свою очередь, могут и не осуществлять соответствующей фильтрации.

В общем случае экранированные агенты, блокируя прозрачную передачу, могут выполнять еще ряд функций (рис. 9.36).

Идентификация и аутентификация пользователей необходимы для обеспечения высокой степени безопасности при их доступе к сети, при этом для предотвращения перехвата пароль не передается в открытом виде через общие коммуникации. Наиболее эффективным способом аутентификации является использование одноразовых паролей, а также применение цифровых сертификатов, выданных доверительными органами (например, центром распределения ключей).

Рис. 9.36. Базовые функции экранирующих агентов

Проверка действительности получаемых и переданных данных является особенно актуальной функцией для аутентификации не только электронных сообщений, а и мигрирующих программ (например, Java), относительно которых может быть выполнена подделка. Эта функция связана с контролем цифровых подписей, при этом также могут применяться цифровые сертификаты.

Разделение доступа к ресурсам внутренней или внешней сети при обращении к межсетевому экрану реализуется с помощью функций идентификации и аутентификации пользователей способами, идентичными поддерживаемым на уровне операционных систем. Функция разделения доступа базируется на одном из следующих подходов:

разрешение доступа только по определенным адресам во внешней сети;

фильтрация запросов по обновленным спискам недопустимых адресов и блокирование поиска с нежелательными ключевыми словами;

накопление и возобновление санкционированных информационных ресурсов внешней сети в массиве памяти брандмауэра и полный запрет доступа во внешнюю сеть.

Фильтрация и преобразование потока сообщений выполняется посредником на основе заданного набора правил, при этом различают:

экранированные агенты, сориентированные на анализ потока сообщений для определенных видов сервиса (например, FTP, HTTP, Telnet и т.п.);

универсальные экранированные агенты, которые обрабатывают все сообщения, например, с целью поиска и обезвреживания компьютерных вирусов, прозрачного шифрования данных и т.п.

Трансляция внутренних сетевых адресов реализуется для всех пакетов, которые идут из внутренней сети во внешнюю. Для этих пакетов осуществляется

автоматическое преобразование IP-адресов компьютеров-отправителей в один IP-адрес, ассоциированный с брандмауэром, из которого передаются все пакеты, исключая прямой контакт между внутренней и внешней сетью, а IP-адрес брандмауэра становится единственным активным IP-адресом, который попадает во внешнюю сеть.

Это позволяет спрятать топологию внутренней сети и иметь внутри сети собственную систему адресации, не согласованную с внешней (например, Интернет), что эффективно решает проблему расширения адресного пространства.

Анализ и регистрация событий, реагирование на определенные события, а также анализ зарегистрированной информации и составление отчетов дают возможность выявить попытки несанкционированных действий. Эффективность брандмауэра в значительной мере зависит от эффективности выполнения этой функции, которая разрешает формировать предупредительные сигналы при выявлении нападения.

Система регистрации, сбора и анализа статистики обрабатывает адреса клиентов и сервера, идентификаторы пользователей, время сеансов и соединений, количество переданных и принятых данных, действия администратора и пользователей и т.п. и предоставляет администраторам подробные отчеты и оповещает об определенных событиях в режиме реального времени.

Хэширование данных, запрашиваемых при доступе пользователей внутренней сети к информационным ресурсам внешней, осуществляется с помощью сети накопления информации на пространстве жесткого диска брандмауэра (технология proxy-сервера). Поэтому если при запросе нужна хэшированная информация, то она предоставляется без обращения к внешней сети, что существенно ускоряет доступ.

По такой технологии все санкционированные информационные ресурсы внешней сети нагромождаются и обновляются администратором на proxy-сервере, а пользователям внутренней сети разрешается доступ только к хэшированным информационным ресурсам.

Экранированные агенты надежнее обычных фильтров и обеспечивают высокую степень защиты, но снижают производительность обмена данными между сетями и не имеют достаточной прозрачности для прикладных программ и конечных пользователей.

Особенности межсетевого экранирования. Брандмауэры поддерживают безопасность межсетевого взаимодействия на разных уровнях эталонной модели OSI, где функции защиты существенным образом отличаются. В этой связи комплексный брандмауэр подают в виде совокупности неделимых экранов, сориентированных на отдельный уровень модели OSI. Обычно комплексный экран функционирует на сетевом, сеансовом и прикладном уровнях эталонной модели, соответственно различают экранирующий маршрутизатор и шлюзы сеансового и прикладного уровня (рис. 9.37).

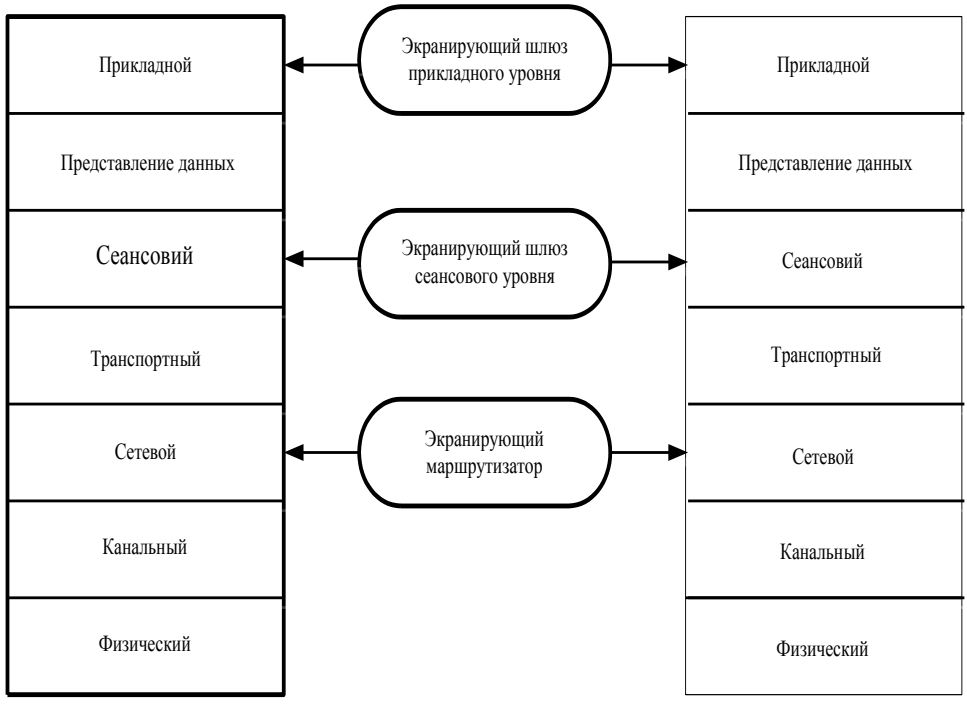


Рис. 9.37. Типы межсетевых экранов, функционирующих на отдельных уровнях модели OSI

Используемые в сетях протоколы не всегда однозначно отвечают модели OSI, и потому указанные экраны могут отображать и соседние уровни эталонной модели, например шлюз прикладного уровня может зашифровывать сообщение при их передаче и расшифровывать принятые данные. Тогда он функционирует на уровнях представления данных и прикладных.

Межсетевые экраны каждого из типов имеют свои преимущества и недостатки, но надежную защиту обеспечивают только комплексные системы, которые объединяют все виды экранирования.

Экранирующий шлюз прикладного уровня (пакетный фильтр) предназначен для фильтрации пакетов сообщений. Он обеспечивает прозрачное взаимодействие между внутренней и внешней сетями, функционируя на сетевом уровне модели OSI, а также может охватывать и транспортный. Решение о пропуске пакета принимается независимо по заданным правилам фильтрации на основе анализа пакетов сетевого и транспортного уровней (рис. 9.38).

Адреса отправителя и получателя могут быть IP-адресами, которые определяются при формировании пакета и остаются неизменными при передаче его через сеть.

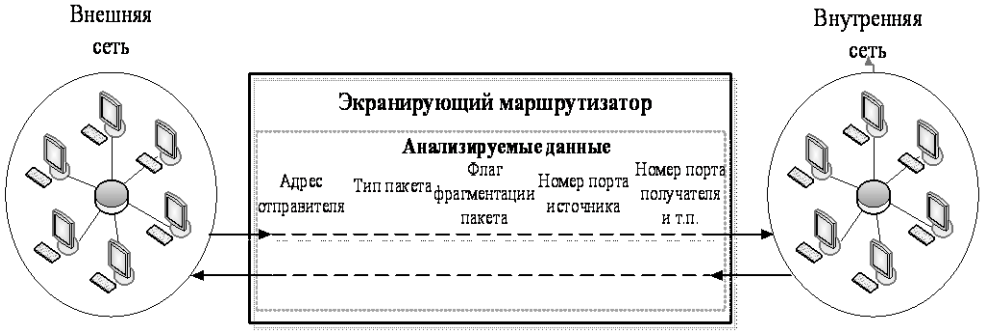


Рис. 9.38. Схема функционирования пакетного фильтра

Тип пакета содержит код протокола, который соответствует сетевому или транспортному уровню. Флаг фрагментации пакета определяет наличие или отсутствие фрагментации. Номера портов источника и получателя однозначно идентифицируют прикладную программу отправителя, а также прикладную программу, для которой предназначен этот пакет. Для возможности фильтрации пакетов по адресам портов необходимо знание принятых в сети соглашений относительно выделения номеров портов протоколам высокого уровня.

Экранирующие маршрутизаторы не обеспечивают высокой степени безопасности, так как проверяют только заголовки пакетов и не поддерживают таких функций, как аутентификация конечных узлов, шифрование пакетов, проверка их целостности и т.п. Эти маршрутизаторы чувствительны к таким атакам, как подделка исходных адресов, несанкционированная модификация содержимого пакетов сообщений, а обход таких экранов реализуется на основе формирования заголовков пакетов, которые удовлетворяют правилам фильтрации.

Экранирующий шлюз сеансового уровня предназначен для контроля виртуальных соединений и трансляции адресов (например, IP-адрес) при взаимодействии с внешней сетью. Он функционирует на сеансовом уровне модели OSI, охватывает транспортный и сетевой уровни, а защитные механизмы принадлежат к функциям посредничества.

Контроль виртуальных соединений заключается в контроле квитирования связи и передачи информации из установленных виртуальных каналов.

При контроле квитирования осуществляется контроль установки виртуального соединения между узлами внутренней и внешней сетей путем определения допустимой связи, на основе информации, которая содержится в заголовках пакетов сеансового уровня, протокола TCP. Однако если пакетный фильтр при анализе заголовков проверяет только номера портов источника и получателя, то экранирующий шлюз сеансового уровня анализирует другие поля, которые относятся к процессу квитирования связи. Определение шлюзом допустимости запроса на сеанс связи осуществляет-

ся по алгоритму (рис. 9.39).

После установления соединения с компьютером внешней сети шлюз действует от лица клиента, следит за выполнением квитирования связи, например по протоколу TCP, а примером базового критерия фильтрации может быть возможность DNS-сервера определить IP-адрес клиента и ассоциированное с ним имя.

Рис. 9.39. Алгоритм определения допустимости запроса на сеанс связи

Рассмотрим процедуру квитирования связи на примере обмена TCP-пакетами, где флагом SYN обозначается синхронизация, а ACK - подтверждение (рис. 9.40).

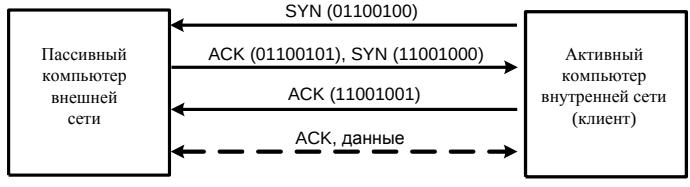


Рис. 9.40. Схема квитирования связи по протоколу TCP

Первый пакет сеанса TCP, обозначенный флагом SYN, содержит произвольное число, например 01100100, и является запросом клиента на открытие сеанса. Компьютер внешней сети, который получил этот пакет, посылает ответ, обозначенный флагом АСК, и содержит число на единицу больше, чем в принятом пакете (в нашем случае 01100101), подтверждая таким образом прием пакета SYN от клиента. Кроме того, осуществляя обратную процедуру, компьютер внешней сети посылает также клиенту пакет SYN, но уже с порядковым номером первого байта передачей данных (например, 11001000), а клиент подтверждает его получение передачей пакета АСК, содержащего число 11001001. На этом процесс квитирования связи завершается.

Для экранированного шлюза сеансового уровня (рис. 9.41) сеанс считается допустимым тогда, когда при квитировании связи флаги SYN и АСК и числа заголовков TCP-пакетов оказываются логически связанными между собой. После того как шлюз определил, что компьютеры внутренней и внешней сетей являются авторизованными участниками сеанса TCP, и проверил допустимости данного сеанса, он устанавливает соединение.

Рис. 9.41. Схема функционирования шлюза сеансового уровня

Для контроля виртуальных соединений используются специальные программы (канальные посредники), устанавливающие виртуальные каналы между внутренней и внешней сетями.

В основном экранирующие шлюзы сеансового уровня работают в комплексе с экранирующими шлюзами прикладного уровня.

С точки зрения практической реализации шлюз сеансового уровня - достаточно простая и надежная программа и дополняет экранирующий маршрутизатор функциями контроля виртуальных соединений и трансляции внутренних адресов (например, IP-адреса).

Шлюз сеансового уровня имеет почти такие же недостатки, что и экранный маршрутизатор, поэтому он применяется как дополнение к экранирующему шлюзу прикладного уровня.

Экранирующий шлюз прикладного уровня функционирует на прикладном уровне модели OSI, охватывая также уровень представления данных, и обеспечивает наиболее надежную защиту межсетевых взаимодействий. Его защитные функции принадлежат к функциям посредничества, но, в отличие от

экранированного шлюза сеансового уровня, он выполняет большее количество функций, к которым принадлежат те, что реализуются экранирующими агентами (рис. 9.42) и используются по одному для каждого обслуженного прикладного протокола.

Прикладной шлюз, равно как и шлюз сеансового уровня, перехватывает с помощью соответствующих экранирующих агентов входные и исходные пакеты, копирует и переправляет информацию через шлюз, функционируя как сервер-посредник, кроме прямых соединений между внутренней и внешней сетями. Посредники, пользующиеся прикладным шлюзом, связаны с конкретными прикладными программами и могут фильтровать поток сообщений на прикладном уровне модели OSI.

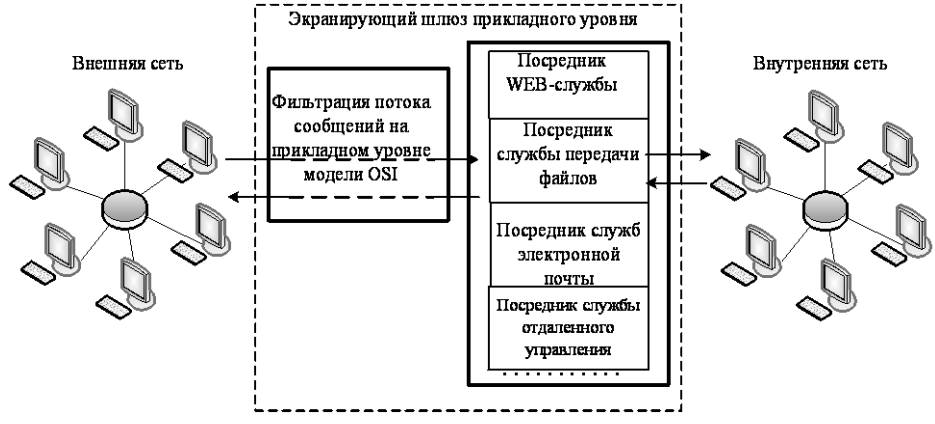


Рис. 9.42. Схема функционирования экранирующего шлюза прикладного уровня

Прикладные шлюзы в качестве посредников используют программные серверы конкретных служб (передача файлов, электронной почты, отдаленного управления и т.п.), функционирующие на брандмауэре в резидентном режиме и реализуют функции защиты.

Как и в случае экранирующего шлюза сеансового уровня, для связи между компьютерами внутренней и внешней сетей соответствующий посредник прикладного шлюза образовывает два соединения: от компьютеров (внутренней сети) к брандмауэру и от брандмауэра к месту назначения. Но, в отличие от канальных посредников, посредники прикладного шлюза пропускают только пакеты, генерированные теми прикладными программами, которые им поручено обслуживать. Например, программа-посредник Web-службы может обрабатывать лишь трафик, генерированный этой службой. Если в сети работает прикладной шлюз, то входные и исходные пакеты могут передаваться лишь для тех служб, для которых есть соответствующие посредники.

Во время налаживания прикладного шлюза и описания правил фильтрации сообщений используются такие параметры, как название сервиса, допу-

стимый временной диапазон его использования, ограничение на содержание сообщений, связанных с этим сервисом, компьютеры, из которых можно пользоваться сервисом, идентификаторы пользователей, схемы аутентификации и т.п. Основные преимущества экранированного шлюза прикладного уровня:

за счет возможности выполнения большого количества функций посредничества обеспечивает наиболее высокий уровень защиты локальной сети;

защита на уровне приложений разрешает осуществлять большое количество дополнительных проверок, уменьшая тем самым вероятность проведения успешных атак, основанных на недостатках программного обеспечения;

при нарушении трудоспособности прикладного шлюза блокируется сквозное прохождение пакетов между разделяемыми сетями, которое не снижает безопасность защищенной сети в случае отказов.

При этом экранированный шлюз прикладного уровня имеет ряд недостатков:

довольно большая сложность самого брандмауэра, а также процедур его установки и конфигурирование;

высокие требования к производительности и ресурсоемкости компьютерной платформы;

отсутствие "прозрачности" для пользователей и снижение пропускной способности при реализации межсетевых взаимодействий.

Для эффективной защиты межсетевые взаимодействия брандмауэр должен быть правильно установлен и сконфигурирован. Для этого необходимо разработать политику межсетевого взаимодействия, определить схемы подключения межсетевого экрана и наладить параметры функционирования брандмауэра.

Разработка политики межсетевого взаимодействия. Политика межсетевого взаимодействия являются той частью политики безопасности в организации, которые определяет требования к безопасности информационного обмена с внешним окружением. Эти требования непременно должны отражать два аспекта: политику доступа к сетевым серверам; политику работы межсетевого экрана.

Политика доступа к сетевым сервисам определяет правила предоставления и использования всех возможных сервисов защищаемой компьютерной сети. Должны быть заданы все сервисы, предоставляемые межсетевым экраном, определены допустимые адреса клиентов для каждого сервиса, указаны правила относительно того, когда и какие пользователи каким именно сервисом и на каком компьютере могут воспользоваться, в частности правила аутентификации компьютеров и пользователей, а также условия работы последних вне локальной сети.

Политика работы межсетевого экрана задает базовый принцип управления межсетевым взаимодействием, положенный в основу функционирования

брандмауэра. Может быть выбран один из двух таких принципов: запрещено все, что не разрешено; разрешено все, что не запрещено.

В первом случае межсетевой экран конфигурируется так, чтобы блокировать любые неразрешенные межсетевые взаимодействия. Такой подход дает возможность адекватно реализовать принцип минимизации привилегий, а значит, с точки зрения безопасности он наилучший.

Во втором случае межсетевой экран налаживается таким образом, чтобы блокировать только запрещенные межсетевые взаимодействия, благодаря чему повышается удобство использования сетевых сервисов пользователями. Тем не менее вместе с тем снижается безопасность межсетевого взаимодействия, когда администратор может учесть не все действия, которые запрещены пользователям.

Определение схемы подключения межсетевого экрана. Для подключения межсетевых экранов используются разные схемы, зависящие от условий функционирования и количества сетевых интерфейсов брандмауэра.

Брандмауэры с одним сетевым интерфейсом (рис. 9.43) недостаточно эффективны с точки зрения безопасности и из позиций удобства конфигурирования. Физически не разграничиваются внутренняя и внешняя сети, а потому не обеспечивается надежная защита межсетевых взаимодействий. Отладка таких экранов и соединение с ними маршрутизаторов - довольно сложная задача, стоимость решения которой превышает стоимость брандмауэра с двумя или тремя сетевыми интерфейсами. Поэтому будем рассматривать лишь две последних схемы подключения. При этом защищаемую локальную сеть подадим как совокупность закрытой и открытой подсетей, где открытой является та подсеть, доступ к которой со стороны внешней сети может быть целиком или частично открытым.

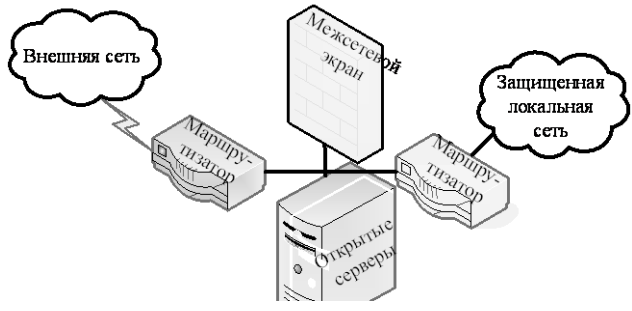


Рис. 9.43. Защита локальной сети брандмауэром с одним сетевым интерфейсом

Среди многочисленных возможных схем подключения брандмауэров типичными являются:

схема единой защиты локальной сети;

схема из защищенного закрытой и незащищенной открытой подсетей;
схема с распределенной защитой закрытой и открытой подсетей.

Схема единой защиты локальной сети является наиболее простым решением (рис. 9.44), согласно которому брандмауэр целиком экранирует локальную сеть от внешней, а между маршрутизатором и брандмауэром существует только один путь для трафика. При этом маршрутизатор настраивается так, что брандмауэр является единственным видимым извне узлом. Серверы, входящие в локальную сеть, защищены межсетевым экраном, но объединение серверов, доступных из внешней сети, вместе с другими ресурсами защищаемой локальной сети существенным образом снизит безопасность межсетевых взаимодействий. Поэтому эту схему подключения брандмауэра можно использовать лишь при отсутствии в локальной сети открытых серверов или когда они являются доступными из внешней сети только для ограниченного числа доверенных пользователей.

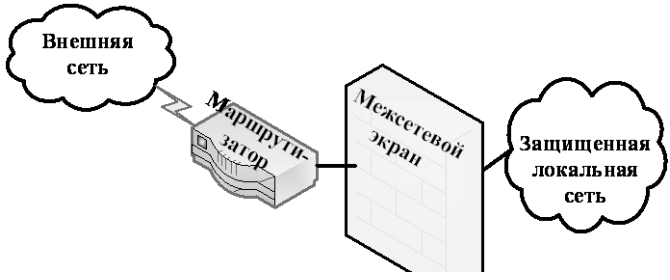


Рис. 9.44. Схема единой защиты локальной сети

При наличии в локальной сети общедоступных открытых серверов их целесообразно вынести в открытую подсеть к межсетевому экрану (рис. 9.45).

Рис. 9.45. Схема с защищенной и незащищенной открытой подсети

Такой способ обеспечивает высшую защищенность закрытой части локальной сети, но вместе с тем снижается безопасность открытых серверов.

Некоторые брандмауэры дают возможность разместить эти серверы на себе, но такое решение не является наилучшим с точки зрения загрузки компьютера и безопасности самого брандмауэра, поэтому такую схему целесообразно использовать лишь при невысоких требованиях к безопасности открытой подсети.

Если к безопасности открытых серверов выдвигаются повышенные требования, то необходимо использовать схему с распределенной защитой закрытой и открытой подсети. Такую схему можно построить на основе одного брандмауэра с тремя сетевыми интерфейсами (рис. 9.46) или двух брандмауэров с двумя сетевыми интерфейсами (рис. 9.47). В обоих случаях доступ к открытой и закрытой подсети локальной сети возможен только через межсетевой экран, при этом доступ к открытой подсети не дает возможности осуществить доступ к закрытой.

Рис. 9.46. Схема с распределенной защитой закрытой и открытой подсети на основе одного брандмауэра с тремя сетевыми интерфейсами

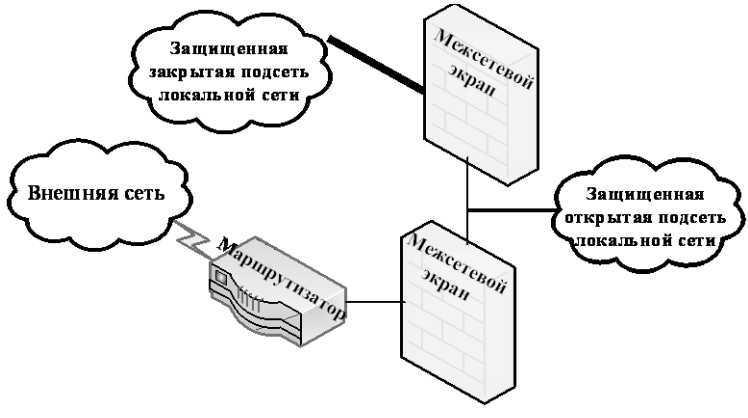


Рис. 9.47. Схема с раздельной защитой закрытой и открытой подсетей на основе двух брандмауэров с двумя сетевыми интерфейсами

Из последних двух схем большую степень безопасности межсетевых взаимодействий обеспечивает схема с двумя брандмауэрами, каждый из которых образовывает отдельный эшелон защиты закрытой подсети, а защищенная открытая подсеть здесь выступает как экранированная подсеть, конфигурируемая таким образом, чтобы обеспечить доступ к компьютерам подсети как из внешней сети, так и из закрытой локальной. Но прямой обмен информационными пакетами между внешней сетью и закрытой подсетью невозможный. При атаке системы с экранированной подсетью необходимо преодолеть по крайней мере две независимые линии защиты. Средства мониторинга состояния межсетевых экранов, как правило, проявят такую попытку атаки, и администратор системы сможет своевременно начать необходимые действия относительно предотвращения НСД.

Отладка параметров функционирования брандмауэра. Межсетевой экран является программно-аппаратным средством защиты, которая состоит из компьютера, а также операционной системы и специального программного обеспечения (которое часто называют брандмауэром), которые функционируют на нем.

Компьютер брандмауэра должны быть довольно мощным и физически защищенным (например, содержаться в специально отведенном помещении, которое охраняется) и иметь средства защиты от загрузки операционной системы из несанкционированного носителя.

Операционная система брандмауэра также должны удовлетворять ряду требований:

иметь средства размежевания доступа к ресурсам системы;

блокировать доступ к компьютерным ресурсам в обход существующего программного интерфейса;

запрещать привилегированный доступ к своим ресурсам из локальной сети;

содержать средства мониторинга (аудита) любых административных действий.

После установления брандмауэра осуществляется отладка параметров, которые состоит из таких этапов:

1) формирование правил работы межсетевого экрана соответственно разработанной политике межсетевого взаимодействия и описание правил в интерфейсе брандмауэра;

2) проверка заданных правил на непротиворечивость;

3) проверка соответствия параметров отладки брандмауэра разработанной политике межсетевого взаимодействия.

Сформированная на первом этапе база правил работы межсетевого экрана является формализованным отображением разработанной политики межсетевого взаимодействия, а компонентами правил являются защищаемые объекты, пользователи и сервисы.

К таким объектам могут принадлежать компьютеры с одним сетевым интерфейсом, шлюзы (компьютеры с несколькими сетевыми интерфейсами), маршрутизаторы, сети и т.п. Объекты, каждый из которых имеет ряд атрибутов (сетевой адрес, маска подсети и т.п.), могут объединяться в группы. Важно обращать внимание на необходимость полного описания объектов для проверки корректности заданных правил экранирования. Такое описание возможное только тогда, когда определены все сетевые интерфейсы шлюзов и маршрутизаторов.

При описании правил работы межсетевого экрана пользователям предоставляются входные имена, которые объединяются в группы. Для пользователей отмечаются допустимые исходные и целевые сетевые адреса, диапазон дат и времени работы, а также схемы и порядок аутентификации.

Определение набора и свойств используемых сервисов осуществляется на основе встроенной в дистрибутив брандмауэра базы данных, а для нестандартных сервисов параметры могут задаваться вручную с помощью специальных атрибутов.

После формирования базы правил осуществляется ее проверка на непротиворечивость. Это очень важный момент, особенно для развитых многокомпонентных сетевых конфигураций со сложной политикой межсетевого взаимодействия.

Проверка сформированных правил на непротиворечивость выполняется автоматически, а выявленные неоднозначности устраняют путем редактирования противоречивых правил. Большинство брандмауэров после формирования базы правил выполняют процесс окончательной отладки автоматически.

Проверка соответствия параметров отладки брандмауэра разработанной политике межсетевого взаимодействия может выполняться на основе анализа протоколов работы межсетевого экрана. Однако наибольшая результативность такой проверки будет достигнута в случае использования специализированных программных средств анализа защищенности сети (сетевые сканеры).

После поиска слабых мест в конфигурации межсетевого экрана подаются рекомендации относительно их коррекции. Поиск слабых мест осуществляется на основе проверки реакции межсетевых экранов на разные типы попыток нарушения безопасности. При этом выполняется сканирование всех сетевых сервисов, доступ к которым осуществляется через межсетевой экран, а для постоянной поддержки высокой степени безопасности сети сканеры монтируются в межсетевые экраны.

При отладке межсетевого экрана следует учитывать, что он не может защитить от некомпетентности администраторов и пользователей, например, через выбор пароля, который легко угадывается, неконтролируемый канал связи и т.п.

Примеры уязвимостей, возникающих при организации защищенных информационных сетей и используемые для реализации НСД, приведены на рис. 9.48.

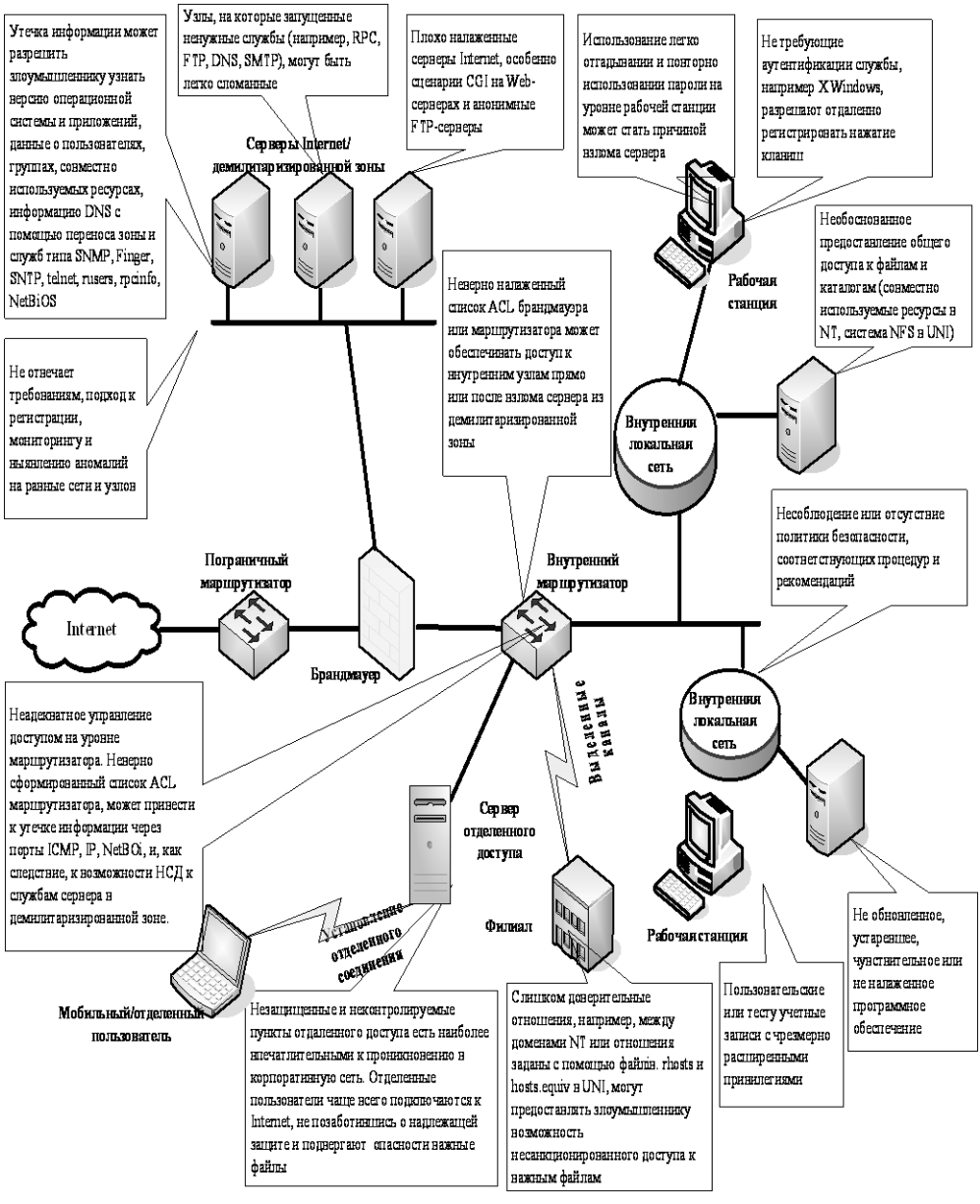


Рис. 9.48. Примеры уязвимости в информационных сетях

Общие требования к межсетевым экранам.

По целевым качествам - обеспечение безопасности внутренней защищенной сети и полный контроль над внешними подключениями и сеансами связи, а также наличие средств авторизации доступа пользователей через внешние подключения. Например, типична ситуация, когда часть персонала организации должна выезжать в командировку, и в процессе работы они нуждаются в доступе к некоторым ресурсам внутренней компьютерной сети организации. Брандмауэр должен надежно распознавать таких пользователей и предоставлять им необходимые виды доступа.

По управляемости и гибкости - наличие мощных и гибких средств управления для полного внедрения политики безопасности организации. Обеспечение простой реконфигурации системы при изменении структуры сети. Если в организации есть несколько внешних подключений, в том числе и в отдаленных филиалах, то система управления экранами должна иметь возможность централизованно обеспечивать для них проведение единой политики межсетевого взаимодействий.

По производительности и прозрачности - достаточно эффективное функционирование и обеспечение в реальном времени обработки всего входного и выходного трафика при максимальной нагрузке. Это необходимо для того, чтобы брандмауэр не перегружался большим количеством вызовов, что привело бы к нарушению его работы. Межсетевой экран должен работать незаметно для пользователей локальной сети и не усложнять выполнения ими легальных действий. В противном случае пользователи будут стараться любым способом обойти установленные уровни защиты.

По самозащищенности - наличие механизмов самозащиты от любых несанкционированных воздействий. Поскольку межсетевой экран является и ключом, и дверью к конфиденциальной информации в организации, то он должен блокировать любые попытки несанкционированного изменения его параметров отладки, а также иметь развитые средства контроля своего состояния и сигнализации, которые должными обеспечивать своевременное оповещение службы безопасности при выявлении любых несанкционированных действий и нарушения работоспособности межсетевого экрана.

Основные выводы

Конфиденциальность - характеристика безопасности ресурсов, которая отражает их свойство недоступности без соответствующих полномочий. Фактически ресурсы не могут быть доступными или раскрытыми неавторизованной стороне, т.е. для нее их якобы нет. В свою очередь, авторская сторона (например, обслуживающий персонал, пользователь, программы и т.д.), которой предоставлены соответствующие полномочия, имеет полный доступ к ресурсам.

Целостность - характеристика безопасности ресурсов, которая отражает их свойство противостоять несанкционированному изменению.

Доступность - характеристика безопасности ресурсов, которая отражает их свойство, которое заключается в возможности их использования в заданный момент времени соответственно предоставленным полномочиям. Фактически авторская сторона в случае потребности сразу в любой момент времени получает неограниченный доступ к необходимому ресурсу.

Активный перехват (active eavesdropping) - перехват, во время которого у неприятеля есть возможность не только перехватывать сообщение, а и влиять на него, например задерживать или изымать сигналы, которые передаются каналами связи.

Пассивный перехват (passive tapping**)** — получение информации с возможностью только наблюдать за обменом подключения с целью выявления разной системной информации в вычислительной сети (ОМ), не взыскивая на него никакого влияния. Пассивный перехват тяжело выявить, поскольку нет непосредственного присоединения терминального оборудования к линии связи.

Прямой перехват (direct eavesdropping) - перехват информации непосредственным подключением дополнительного терминала к линии связи. Прямой перехват можно обнаруживать проверкой линии связи.

Косвенный перехват (indirect eavesdropping) - перехват информации (например, индуктивных волн) без использования непосредственного подключения к линии связи (threat).

Аппаратные средства — разнообразные механические, электрические, электромеханические, электронные, электронно-механические и другие устройства и системы (например, источники бесперебойного питания, криптографические вычислители, электронные идентификаторы и ключи, устройства для выявления «жучков», генераторы шума и т.п.), функционирующие автономно (встраиваемые или соединенные с другой аппаратурой) с целью блокирования действий дестабилизирующих факторов и решения других задач защиты информации.

Программные средства - специальные программы (например, антивирусы, шифрования данных, реализации алгоритмов цифровой подписи, размежевание доступа, оценки рисков, определение уровня безопасности, органи-

зации экспертиз и т.п.), которые функционируют в пределах информационных систем для решения задач защиты информации.

Программно-аппаратные средства - взаимосвязанные аппаратное и программное средства (например, банковские системы электронных платежей, комплексные информационные системы конфиденциальной связи, автоматизированные системы контроля доступа персонала и транспортных средств в режимных зонах и т.п.), которые функционируют автономно или в составе других систем с целью решения задач защиты информации.

Криптографические средства - средства, предназначенные для защиты информации путем криптографического преобразования информации (шифрование, дешифрование), которое реализуется с помощью асимметрических или симметричных криптографических систем. Асимметрические криптографические системы базируются на криптографии с открытым ключом. Например, известнейшими практическими реализациями этого типа являются системы Диффи - Хеллмана, RSA и Эль-Гамаля. Симметричные криптографические системы грунтуются на криптографии с секретным ключом, наиболее известными практическими реализациями которых являются, например, DES, ГОСТ и т.п.

Стеганографические средства ориентированы на утаивание информации в такой форме, когда сам факт ее наличия не очевидный, например, утаивание данных в звуковых или графических файлах, которые входят в состав ОС Windows.

Организационные средства защиты информации - это множество процессов и действий (например, контроль за утилизацией носителей информации с ограниченным доступом, планирование мероприятий по восстановлению утраченной информации, аудит систем защиты, реализация экспертиз и т.п.), осуществляемые на протяжении всех технологических этапах (проектирование, изготовление, модификация, эксплуатация, утилизация и т.п.) существования соответствующих ресурсов информационных систем и ведущие к созданию, усовершенствованию, упорядочению и согласованности взаимосвязей и взаимодействия их компонент с целью решения задач защиты информации.

Законодательные средства защиты информации - это множество нормативно-правовых актов (конвенции, законы, указы, постановления, нормативные документы и т.п.), которые действуют в определенном государстве и обеспечивают юридическую поддержку для решения задач защиты информации. Вообще с помощью законодательных средств определяются права, обязанности и ответственность относительно правил взаимодействия с информацией, нарушение которых может повлиять на состояние ее защищенности.

Морально-этические средства - моральные нормы и этические правила, которые сложились в обществе, коллективе и на объекте информационной деятельности, нарушения которых отождествляется с несоблюдением об-

щепринятых дисциплинарных правил и профессиональных идеалов. Примером таких средств может быть кодекс чести, этикет, этика хакера и т.п.

Разрушительное программное влияние — это программный код или его части, с помощью которых осуществляется угроза хотя бы одной характеристике безопасности определенных ресурсов информационных систем. Разрушительные влияния можно поделить на такие группы: компьютерные вирусы (вирусы), логические бомбы, тайные хода и лазейки; программы раскрытия паролей, репликаторы, сетевые программные анализаторы, суперзапинговые утилиты, троянские кони.

Вирус — программа, способная к многоразовому самовольному созданию своего тела, которая по обыкновению модифицирует (заражает) другие программы, записанные в файлах или системных областях, для дальнейшего воспроизведения нового тела и получения управления с целью модификации записей, уничтожения файлов, загрузки ресурсов и выполнения других разрушительных влияний в информационной системе.

Логические бомбы — программа, которая инициируется с возникновением разных событий, например открытие определенного файла, обработка заданных записей и другие действия с целью нарушения характеристик безопасности ресурсов информационных систем. Используются, например, для разворовывания с помощью изменения определенным образом (в свою пользу) кода программы, которая реализует финансовые операции.

Тайный ход - уязвимость в системе, которую нарочно создает ее разработчик, или возникшая случайно и фактически являющаяся дополнительным способом проникновения в систему.

Программы раскрытия паролей - программы, по обыкновению предназначенные для угадывания паролей (например, архивированных файлов) через перебор вариантов, возможных для использования символов или проникновение в систему с помощью словарей.

Репликаторы — программы, которые при выполнении создают несколько своих копий в информационной системе. Например, когда репликатор создает только одну и после этого выполняет ее, то память системы быстро переполняется, что ограничивает доступ к определенным компонентам системы.

Сетевые анализаторы — программно-аппаратные средства (в отдельных случаях программы, которые запускаются из рабочей станции, подключенной к сети), предназначенные для считывания любых параметров потока данных в информационной системе.

Суперзапинг — разрушительное влияние, связанное с несанкционированным использованием утилит для модификации, уничтожения, копирования, раскрытия, вставки, применения или запрета применения данных информационной системы.

Троянские кони — специализированная программа, которая, как правило, выступает от лица других программ и разрешает действия, от-

личные от определенных в спецификации, которые используются программным обеспечением.

Экранирующий шлюз прикладного уровня (пакетный фильтр) - устройство, предназначенное для фильтрации пакетов сообщений, обеспечивающее прозрачное взаимодействие между внутренней и внешней сетями.

Экранирующий шлюз сеансового уровня - это устройство, предназначенное для контроля виртуальных соединений и трансляции адресов (например, IP-адрес) при взаимодействии с внешней сетью.

Политика работы межсетевого экрана — это политика, которая задает базовый принцип управления межсетевым взаимодействием, положенный в основу функционирования брандмауэра.

Вопросы для самоконтроля

1. *Как классифицируются НСД по автоматизации*?
2. *Как классифицируются НСД по инициализированным условиям*?
3. *Как реализуется НСД с обратной связью?*
4. *Возможно ли реализовать мономономный НСД с нескольких источников*?
5. *На чем базируется пигибекинговый НСД*?
6. *Что можно отнести к неспецифичным категориям НСД?*
7. *Как классифицируются средства защиты от НСД*?
8. *Приведите классификацию компьютерных вирусов.*
9. *В чем заключается сущность эмпирического подхода*?
10. *Раскройте содержание модели системы с полным перекрытием.*
11. *Какими показателями может быть оценено качество распределения кодов доступа*?
12. *В чем заключается сущность теоретико-эмпирического подхода при построении моделей защиты информации*?
13. *Раскройте основные функции межсетевого экранирования.*
14. *Какие есть типы межсетевых экранов*?
15. *Что нужно для эффективной защиты межсетевого взаимодействия?*
16. *Раскройте сущность моделей Белла - Лападула и Биба.*

The main conclusions

Confidentiality is the characteristic of safety of resources that displays their quality of undetection and availability without appropriate authorities. Actually, resources cannot be accessible or opened to unauthorized side that is they are not present for it supposedly. In turn, the authoring side (for example, serving staff, users, programs and others), that is given the appropriate authorities, has complete access tothe resources.

Integrityis the characteristic of safety of resources that displays their quality to resist to unauthorized change.

Availability is the characteristic of safety of resources which displays their quality of possibility of their use in the set moment of time according to presented authorities. Actually authoring side, if necessary, gets unlimited access to the necessary resource at once, at any moment of time.

Active eavesdropping is eavesdropping, during which the opponent has a possibility not only to eavesdrop the message, but also to influence it, for example, to delay or except signals that are transmitted by communication channels.

Passive tapping is getting of the information with possibility only to observe of the exchanging of the messages (for example, with the purpose of revealing of different system information in computer network (CN), not having any influence on it.

Direct eavesdropping is eavesdropping of the information by direct connection (for example, the additional terminal) to the communication line. Direct eavesdropping can be found out by checking of the communication line.

Indirect eavesdropping is eavesdropping of the information (for example, inductive waves) without use of direct connection to the communication line (threat). It is difficult to find out passive eavesdropping as there is no direct connection of terminal equipment to the communication line.

Hardware are various mechanical, electrical, electromechanical, electronic, electronically-mechanical and other devices and systems (for example, uninterrupted power supplies, cryptography calculators and VLSI-processors, electronic identifiers and keys, devices for revealing of bugs, generators of noise and so on) that function autonomously or are built in or connect to other equipment with the purpose of blocking of operations of destabilizing factors and solution of other problems of information protection.

Software are special programs (for example, antiviruses, enciphering of data, realizations of algorithms of the digital signature, differentiation of access, estimations of risks, definitions of a level of safety, organizations of reviews and so on) that function within the information systems for solving of problems of information protection.

Firmware are interconnected hardware and software (for example, bank systems of electronic payments, complex information systems of confidential communication, computer-based systems of control of access of the staff and vehicles in regime areas and so on) that function autonomously or in structure of other systems with the purpose of solving of problems of information protection.

Cryptographic facilities are the facilities intended for information protection by cryptographic transformationof the information (enciphering, deciphering) that is realized by means of asymmetric or symmetric cryptographic systems. Asymmetric cryptographic systems are based on public key cryptography. For example, the most known practical realizations of its type are the systems of Diffie-Hellman, RSA and Ell-Gamal. Symmetric cryptographic systems are based on hidden-key cryptography, the most known practical realizations of which are, for example, DES, GOST (State All-Union standard) and so on.

Steganographic facilities are oriented to concealment of the information in form when the fact of its presence is not obvious, for example, concealment of data in sound or graphics files that are the part of Windows OS.

Organizational facilities of information protection are the plural of processes and operations (for example, the control of utilization of media with the limited access, planning of measures on restoration of the lost information, audit of protection systems, realization of reviews and so on) that are carried out during all technological stages (designing, production, modification, maintenance, utilization and so on) of existence of appropriate resources of information systems and lead to creation, improvement, ordering and coordination of interrelations and interaction of their components with the purpose of solving of problems of information protection.

Legislative facilities of information protection are the plural of normative-legal acts (conventions, laws, decrees, decisions, normative documents and so on) that operate in the certain state and provide legal support for solving of the problems of information protection. In general, the rights, duties and the responsibility of rules of interaction with the information, violation of which can influence on a state of its protectability are determined by means of legislative facilities.

Moral and ethical facilities are moral standards and ethical rules that have formed in society, collective and the object of information activity, violation of which is identified with non-observance of the standard disciplinary rules and professional ideals. Code of honour, netiquette, hacker ethics and so on can be the example of such facilities.

Destructive program influence is a program code or its parts, by means of which threat even to one characteristic of safety of the certain resources of information systems is carried out. Destructive influences can be divided into such groups: computer viruses (viruses), logic bombs, back doors and trap doors; programs of disclosure of passwords, replicators, network program analyzers, superzapping utilities, Trojan horses.

The virus is the program that is capable to multiple unauthorized creation of its body and usually modifies (infects) other programs written in files or system areas for the subsequent reconstruction of new body and obtaining of control with the purpose of modification of records, destructions of files, loading of resources and executions of other destructive influences in information system.

Back door is vulnerability in system that is purposely created by its developer or that has appeared accidentally and actually is additional way of penetration into the system.

Programs of disclosure of passwords are the programs usually intended for guessing passwords (for example, archived files) by means of sorting out of the variants, possible for use of the characters or penetration into the system with the help of dictionaries.

Replicators are the programs that while executing create a few their own copies in information system. For example, when the replicator creates only one and after that execute it, then the memory of system is quickly overfilled, that limits access to the certain components of the system.

Network analyzers are firmware (in individual cases the program that is started from the workstation connected to the networks) intended for reading of any parameters of the dataflow in information system.

Superzapping is destructive influence that is linked with unauthorized use of the utilities for modification, destruction, copying, disclosure, insertions, applications or prohibitions of application of data of information system.

Trojan horse is a specialized program that, as a rule, comes out on behalf of other programs and allows the operations different from certain in specification that are used by the software.

The shielding gateway of an application layer (packet filter) is intended for filtering of packets of messages and provides transparent interaction between internal and external networks.

The policy of operation of the firewall sets a base principle of control of the internetworking, put in a basis of functioning of a firewall.

Ключевые слова

Русский	**Английский**
несанкционированный доступ	unauthorized access
компьютерная сеть	computer network
неавторизованная сторона	unauthorized side
атака	attack
сканирование портов	port scanning
средства защиты	security facilities
модель Белла-Лападула	Bell-LaPadula model

ЗАКОНОДАТЕЛЬНОЕ ОБЕСПЕЧЕНИЕ ЗАЩИТЫ ИНФОРМАЦИИ

10

10.1. Правовые основы защиты информации
10.2. Организационные меры защиты информации
10.3. Научно-методическое обеспечение защиты информации
10.4. Международные стандарты информационной

10.1. Правовые основы защиты информации

Правовую основу защиты информации составляют законодательные средства защиты информации - множество нормативно-правовых актов (конвенции, законы, указы, постановления, нормативные документы и т.п.), действующие в определенном государстве и обеспечивающие юридическую поддержку для решения задач защиты информации. Например, одним из эффективных средств защиты от несанкционированного копирования программного обеспечения может стать соответствующий закон о защите авторских прав. С помощью законодательных средств определяются права, обязанности и ответственность относительно правил взаимодействия с информацией, нарушение которых может повлиять на состояние ее защищенности. В мировой практике основу указанных средств составляют патентное и авторское право, национальные законы о государственной тайне и обработке информации в компьютерных системах, акты относительно лицензирования, страхования и сертификации, классификационные нормативные документы и т.п.

Основные законодательные основы информационной безопасности отображаются в информационном праве, базовым для которого является понятие информации. Сам термин «информация» используется для раскрытия ее как правовой категории и употребляется в различных нормативно-правовых актах.

В правовом поле существует немало определений информации, характеризующих ее различные формы и свойства. Например: «информацией является документированные или публично объявленные сведения о событиях и явлениях, которые имели или имеют место в обществе, государстве и окружающей среде» или «информация - сведения, представленные в виде сигналов, знаков, звуков, подвижных или не подвижных изображений и т.п.» (см. также главы 1 и 2).

Анализ предметной сферы определений этого термина позволяет сформулировать такое обобщающее определение: ***информация*** *— это данные, представленные в любой организационной форме и в произвольном виде, на любых носителях, о любых событиях и явлениях независимо от места и времени.*

Тогда к термину «информация» можно отнести сведения:

о любых событиях и явлениях, которые имели или имеют место в обществе, государстве и окружающей среде;

любой формы и вида, представленные на любых носителях;

документированные или публично объявленные;

недокументированные или публично не объявленные.

Обобщающее определение информации дает возможность отнести к содержимому этого понятия сведения о любых событиях и явлениях, которые имеют место не только внутри определенного общества, государства и

в окружающей естественной среде, а и в других странах. При этом представление сведений не связывается с конкретными формами, видами, типами носителей, которые разрешает учесть новые возможности представления сведений. Также это определение вводит в сферу правового регулирования общественное отношение, связанные с информацией, которая циркулирует в информационных системах, т.е. которая публично не объявлена и не документируется, что важно в условиях быстрого распространения современных информационных технологий (телекоммуникационных, компьютерных, телевизионных и т.п.).

В законодательстве рядом с понятием информации используется термин «информационная система», имеющийт ряд разнообразных определений. Обобщающим для них будет такое: ***информационная система*** *- взаимосвязанное организованное множество предприятий, подразделов, специалистов нормативно-правового обеспечения, комплекса организационных и технических мероприятий, информационных технологий и ресурсов, предназначенных для обеспечения информационных процессов, в* частности *создание, распространение, использование, хранение и утилизация информации.*

Такая интерпретация термина имеет важное методологическое значение для определения объекта правоотношений в сфере информационного права благодаря тому, что высокий уровень абстракции дает возможность охватить любые информационные системы, которые обеспечивают создание, распространение, использование, хранение и утилизацию информации. Термин «***информационный ресурс***» определяется как *любая совокупность информации независимо от содержания, времени и места создания,* а если такой ресурс находится под юрисдикцией государства и доступен для использования лицом, обществом и государством, то является национальным информационным ресурсом. Такое определение дает возможность отнести к информационным ресурсам любое организованное или неорганизованное множество информации и продуктов и ввести в сферу правового регулирования информационные ресурсы разнообразнейших предметных сфер и назначений, например образования, производства, культуры, медицины, права, социологии и т.п., которые не зависят от времени создания. Информационные ресурсы, являющиеся объектом общественных отношений в государстве, могут быть созданы на любой территории, и потому возникает возможность урегулировать нормами информационного законодательства общественные отношения, связанные с информационными ресурсами, созданными в любом месте. Общественные отношения относительно национальных информационных ресурсов регулируются национальным информационным законодательством, а их особым признаком является принципиальная возможность доступа неограниченного круга лиц, субъектов общества и государства. Для обеспечения такого доступа для

национальных информационных ресурсов устанавливается особый правовой режим.

Информационные процессы. Реализация в обществе информационных отношений сопровождается обращением информации, в основу которого положены процессы, связанные с этапами ее жизненного цикла, базовыми из которых, как отмечалось, есть создание, распространение, использование, хранение и утилизация.

Создание информации связано с такими процессами:

формированием сведений, представленных в любой форме и в произвольном виде, на любых носителях, о событиях и явлениях, которые имели или имеют место в обществе, государстве и окружающей среде;

индивидуальной (группой лиц) деятельностью человека или как результат деятельности юридических лиц.

Различают первично созданную информацию (например, методы защиты информации) и вторичную, которая является результатом обработки совокупности первичной (например, сравнительный анализ методов защиты) информации. Характерным для последней является создание новой информации и получение новых знаний, которые могут использоваться для генерирования новой информации и знаний. Поэтому процесс простого соединения информации (определенной совокупности сведений) в какую-то новую форму без изменения содержания не считается ее созданием. Это важно в процессе определения правовых оснований для установления права собственности на информацию, авторского права и права интеллектуальной собственности.

Процесс распространения информации это передача сведений от ее первоисточника (реального или правомерного собственника информации) к потребителю. На рис. 10.1, *а* приведена типичная схема передачи информации (например, писем по электронной почте) непосредственно от владельца к ее потребителю, на рис. 10.1, *б* — схема с правомерным собственником. Например, для ситуации в сети Интернет, когда владелец информации передает ее организации, которая осуществляет хостинг - комплекс организационных и технических мер по обеспечению приема и хранения в пользу заказчика (владельца информационных ресурсов) его электронной информации, для размещения и хранения на веб-сервере и дальнейшей доставки этой информации к потребителю.

Для процесса распространения существует общее требование неизменности информации. Различают такие правовые режимы использования распространенной информации:

владелец после передачи информации лишается права на нее;

владелец и получатель имеют одинаковые права на дальнейшее использование информации;

получатель информации не ограничен, ограничен или частично ограничен в праве ее использования;

владелец после распространения информации сохраняет полностью или частично права на нее.

Процессы хранения информации приобрели значительное распространение в обществе благодаря тому, что она имеет уникальное свойство - может быть использована многими людьми без потери своих характеристик.

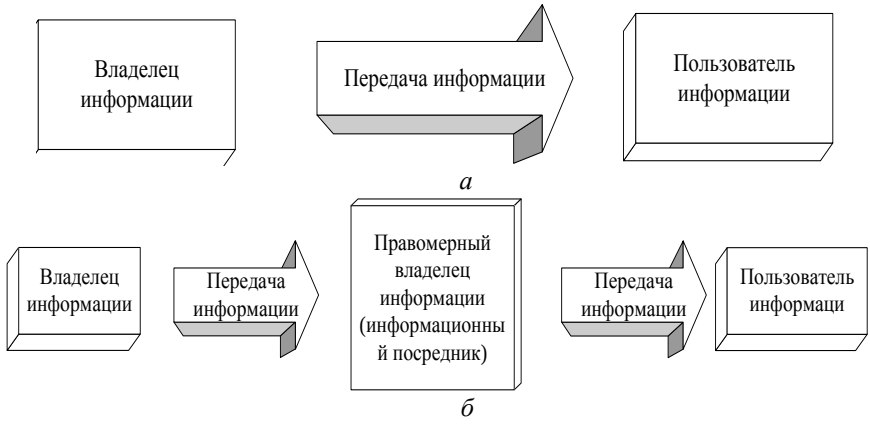


Рис. 10.1. Типичная схема передачи информации
а - от владельца к пользователю; *б* - с правомерным владельцем

Системы хранения информации должны обеспечивать ее целостность и доступность соответственно ее правовому режиму.

В системах хранения информации, даже если она открыта, существует определенный режим доступа, связанный прежде всего с большим количеством пользователей, для которых необходимо обеспечить ее хранение и возможность продолжительного использования.

Процесс использования информации разный (рис. 10.2) и имеет определенные правовые особенности. Простое потребление информации предназначено для удовлетворения индивидуальных информационных нужд личности и, как правило, не используется в профессиональной деятельности потребителя. Сюда принадлежит массовая информация, использование которой не накладывает на потребителя никаких прав или обязательств.

Если же информация потребляется с целью удовлетворения физическим или юридическим лицом профессиональных нужд, связанных с его деятельностью, то потребитель должен действовать с учетом прав и обязанностей соответственно закону или договору. Например, в научной деятельности есть нужная ссылка на источник информации, которая цитируется или используется.

Поиск и сбор информации являются необходимыми условиями реализации процесса ее использования, и он не должен быть противоправным (подменять законные права и интересы). Анализ информации как разновидность ее использования является предпосылкой для создания новой информации, новых знаний.

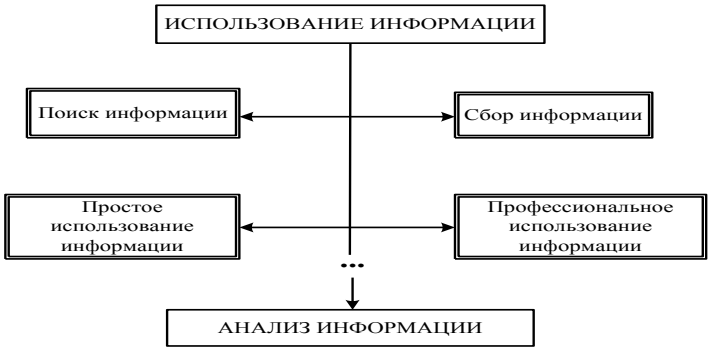


Рис. 10.2. Разновидности процесса использования информации

Жизненный цикл информации заканчивается утилизацией (прежде всего это уничтожение ее носителей), а необходимость правовой регламентации этого процесса связана с наличием права ее собственности, авторского права, определенных требований к срокам хранения и исключение несанкционированного уничтожения.

В качестве примера можно привести одну из распространенных проблем утилизации носителей информации, изготовленных с нарушением авторского права, права интеллектуальной собственности. Здесь необходимо учитывать особенности физических свойств некоторых носителей, которые предопределяют то, что после уничтожения информации на них возможно ее восстановление, и потому правовая регламентация процесса утилизации должна оказывать содействие невозможности использования информации после уничтожения.

Весомую роль в процессе правового регулирования информационных отношений сыграет определение правового режима носителей информации. В определении термина «информация» отмечается, что это сведения, представленные на любых носителях, основными из которых являются:

физические поля (электромагнитные, магнитные, электрические, акустические);

вещевые носители (бумага, фотобумага, ткани и т.п.);

фото-, кино-, магнитные пленки и т.п.;

лазерные и магнитные диски, электронные устройства памяти и т.п.

В этой связи предметом гражданских правоотношений является информация, представленная на определенных носителях, которые выступают как имущество, неразрывно связанное с какой-то конкретной информацией, и это необходимо учитывать при определении правового статуса и режима информации как объекта собственности.

Классификация информации по режиму доступа к ней. По режиму доступа информация делится на открытую (доступ к которой неограничен в правовом смысле) и с ограниченным доступом. К этому классу принадлежит информация, относительно которой установлены определенные степени ограничения доступа и предоставлены соответствующие грифы и которая делится на тайную и конфиденциальную (рис. 10.3). К тайной информации принадлежат сведения, которые представляют государственную или другую предусмотренную законом тайну, а критерием отнесения ее к тайной есть угроза нанесения вреда лицу, обществу и государству в случае разглашения этих сведений.

Существует целый ряд разных тайн (рис. 10.4), которые используются в государственном законодательстве.

Раскроем содержание некоторых из них.

Адвокатская тайна. Предметом этой тайны являются вопросы, по которых физическое или юридическое лицо обращалось к адвокату, суть консультаций , советов, разъяснений и других сведений, полученных адвокатом при осуществлении своих профессиональных обязанностей.

Банковская тайна. Информация относительно деятельности и финансового состоянмя клиента, которая стала известной банку в процессе его обслуживания и взаимоотношений с клиентом или третьими лицами при предоставлении услуг банка.

Военная тайна. Как правило, в государственном законодательстве секретную информацию в сфере обороны относят к государственной тайне, хотя понятие военной тайны используется в разных нормативно-правовых актах.

Коммерческая тайна. Информация является секретной, если она в целом или в совокупности ее составляющих неизвестна и не легкодоступна для лиц, которые по обыкновению имеют дело с видом информации, к которому она принадлежит. Коммерческой тайной могут быть сведения технического, организационного, коммерческого, производственного и другого характера.

Врачебная тайна. Сведения медицинского характера (относительно лиц), которые получены должностными лицами и медицинскими работниками учреждений здравоохранения в связи с выполнением ими профессиональных обязанностей.

Профессиональная тайна. Материалы, документы и другие сведения, которыми пользуются в процессе выполнения своих должностных

обязанностей и другие лица, которые привлекаются к работе с информацией профессионального характера.

Рис. 10.3. Классификация информации с ограниченным доступом

Служебная тайна. Сведения, которые есть в распоряжении конкретных государственных органов, служб или соответствующих должностных лиц относительно подконтрольных объектов, контролирующих, правоохранительных и других государственных органов, их работников, способов достижения определенных законодательством задач и которые по этой причине на определенный период не подлежат внешнему или внутреннему разглашению.

Тайна голосования. Выборы в органы государственной власти и органов местного самоуправления происходят на основе общего, равного и прямого избирательного права путем тайного голосования. Организация процедуры волеизъявления граждан должна обеспечивать тайну их голосования.

Тайна корреспонденции. Обеспечение тайны корреспонденции при предоставлении услуг почтовой связи и при передаче информации телекоммуникационными каналами.

Тайна переписки. Государственное законодательство гарантирует каждому лицу тайну переписки.

Тайна совещания судей. Соображения судей, которые высказываются ими во время постановления приговора в комнате для совещаний, ход обсуждения и принятие решения не подлежат разглашению.

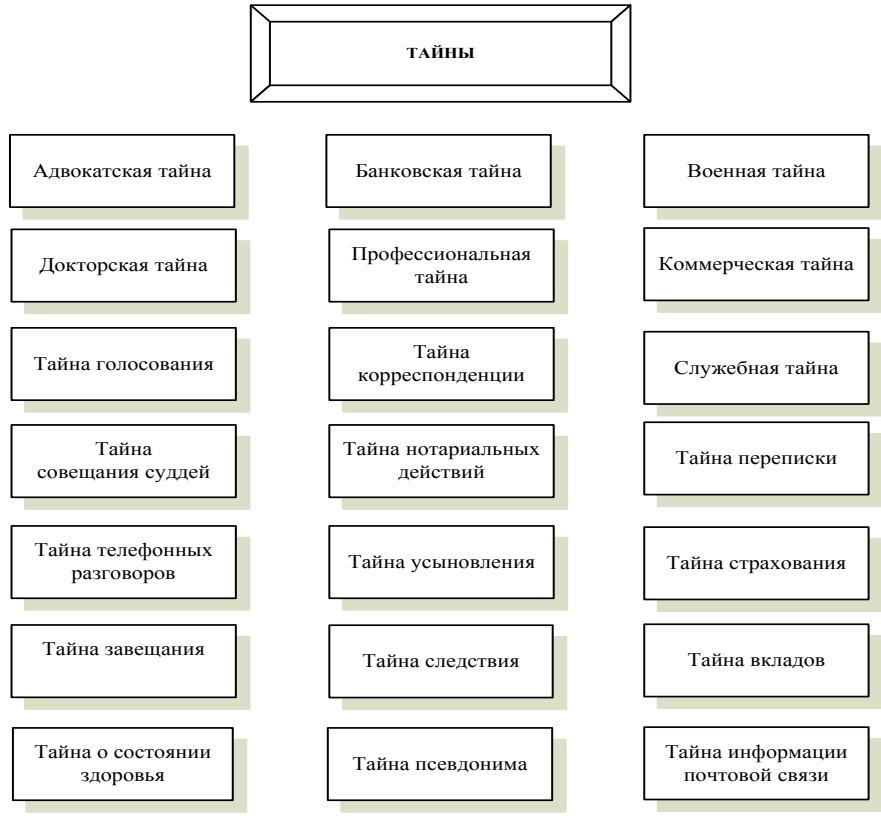


Рис. 10.4. Основные разновидности тайн государственного законодательства

Тайна нотариальных действий. Нотариусы и другие должностные лица, которые совершают нотариальные действия, обязаны соблюдать тайны этих действий и сведений, полученных ими в связи с их совершением.

Тайна страхования. Информация относительно деятельности и финансового состояния страхователя, которая стала известной ему во время взаимоотношений с клиентом или с третьими лицами.

Тайна телефонных разговоров. Каждому лицу гарантируется тайна телефонных разговоров, которые не подлежат разглашению и снятию информации из телекоммуникационных сетей.

Тайна усыновления. Лицо имеет право на тайну пребывания на учете тех, кто желает усыновить ребенка, поиска ребенка для усыновления, пода-

чи заявления об усыновлении и его рассмотрения, решение суда об усыновлении. Усыновленный ребенок имеет право на тайну, в том числе и от него самого, факта его усыновления.

Тайна вкладов (***счетов***). Кредитные союзы и их должностные лица обязаны сохранять тайну относительно счетов, вкладов и других финансовых операций, которые осуществлены членами кредитного союза.

Тайна завещания. Сведения относительно факта составления, содержания, отмены или изменения завещания нотариусом, другим должностным лицом, которое заверяет завещание, свидетелями, а также физическим лицом, которое подписывает завещание вместо завещателя, к открытию наследства.

Тайна следствия. Сведения, которые сохраняются в процессе следствия адвокатами, сотрудниками правоохранительных органов и журналистами и не разглашаются в публичных выступлениях и при подготовке материалов для средств массовой информации.

Тайна псевдонима. При необходимости любое лицо имеет право использовать псевдоним, например работники спецслужб, которые выполняют задания, или лица, которые взяты под защиту для обеспечения их безопасности и т.п.

Информационная инфраструктура. По содержанию элементов информационной инфраструктуры ее можно разбить на группы (рис. 10.5).

1. Предприятия и организации (организационные структуры), связанные с осуществлением информационных услуг и работ на всех базовых этапах жизненного цикла информации. К таким организационным структурам можно отнести те, деятельность которых относительно информационных продуктов связана:

с их производством (телерадиоорганизации, информационные и рекламные агентства, редакции, научные и государственные учреждения и т.п.);

распространением (издательства, телекоммуникационные и почтовые организации и компании, системы распространения книжных и периодических печатных изданий и т.п.);

сохранением (библиотеки, архивы, музеи и т.п.);

использованием (в основном все учреждения, предприятия и организации, физические лица);

утилизацией (в основном все учреждения, предприятия и организации, физические лица).

2. Предприятия и организации, связанные с производством и использованием информационных технологий и ресурсов. Деятельность таких организационных структур может быть связана:

с производством средств обеспечения информационной деятельности (производители теле-, кино- и аудиоаппаратуры, компьютерной, телекоммуникационной и полиграфической техники и т.п.);

производством информационных технологий (предприятия и организации, которые вырабатывают телекоммуникационные, полиграфические, телевизионные и компьютерные системы, программное обеспечение и т.п.);

использованием информационных технологий (все известные учреждения, предприятия и организации, физические лица);

использованием ресурсов (телерадиоорганизации и телекоммуникационные компании, которые используют частотный и номерной ресурсы и т.п.).

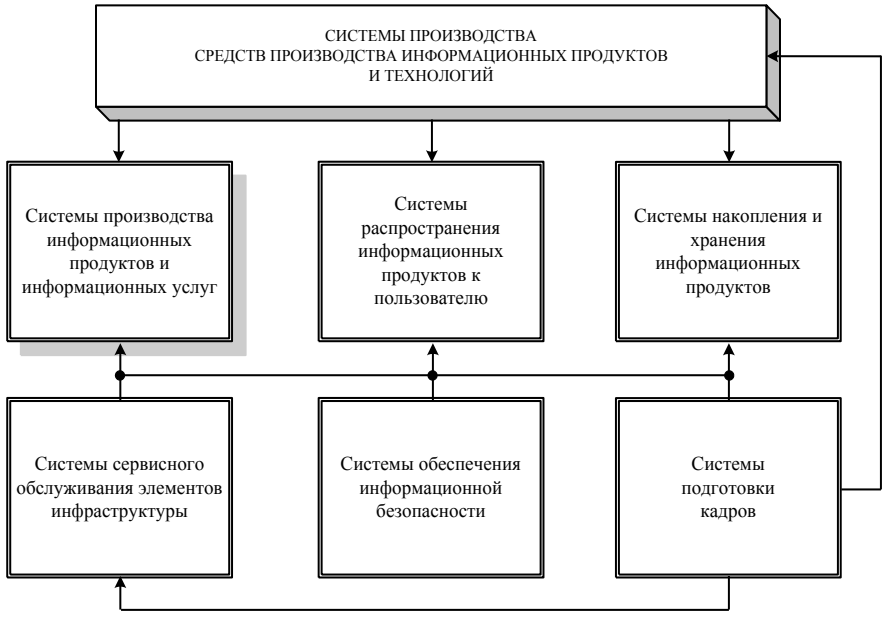


Рис. 10.5. Информационная инфраструктура

3. Предприятия и организации обеспечения информационной безопасности. Деятельность таких организационных структур связана:

с производством систем и средств, которые обеспечивают информационную безопасность (производители систем и средств технического, криптографического, стенографического и других видов защиты информации);

обеспечением информационной безопасности (предприятия и организации, которые реализовывают проведение комплексных мероприятий по обеспечению информационной безопасности).

Определение информационной безопасности. Термин «безопасность» буквально означает существование без опасности или без угрозы. Рассмотрим отдельные характерные особенности безопасности и определим понятие системы. Под термином «система» понимают объекты живой природы, социальные образования (общество, общественные организации, социальные группы, учреждения, предприятия, военные части, отдельно взятое лицо и т.п.), объекты и системы разного назначения, созданные человеком в процессе его деятельности.

Безопасность является атрибутом состояния любой системы. Для всех без исключения систем имеет место влияние внешних и внутренних действий, среди всего спектра которых есть такие, что снижают качество функционирования систем, а в некоторых случаях приводят к ее гибели. Действия, которые могут нанести ущерб, являются внешними или внутренними угрозами.

В соответствии с общей теорией систем, ее атрибутивным свойством является самосохранение, которое может быть обеспечено только тогда, когда система может нейтрализовать внешние угрозы, ликвидировать внутренние, или минимизировать возможные потери от их реализации. Такое состояние системы является безопасным, важнейшим для самосохранения и атрибутивным для любой системы.

Безопасность имеет системный характер. Кроме составных элементов системы, предназначенных только для обеспечения безопасности, в процессе нейтрализации угроз принимают участие и те элементы, для которых это не является основной функцией, тем не менее их влияние на эффективность мероприятий безопасности может быть значительным. Поэтому учет всех факторов безопасности осуществляется только при условии рассмотрения системы в целом во всей совокупности составляющих, а также внутренних и внешних связей и действий (рис. 10.6). Рассмотрение безопасности (как одной из функций системы) возможно лишь при условии учета ее взаимосвязи с другими функциями этой системы.

Безопасность может быть формально оценена. Очень важно, когда есть возможность количественно или качественно осуществить оценку состояния безопасности.

Правильно выбранная оценка позволяет органически объединить требования безопасности системы с ее основными функциональными требованиями, при этом появляется возможность применять современные математические методы и модели анализа уровня угроз и состояния безопасности для синтеза систем защиты и определения эффективности функционирования.

В основном известные подходы оценки безопасности делятся на две группы, которые соответственно базируются на оценке угроз или вреда, причиненного системе в процессе реализации угроз. При этом имеют ме-

сто качественные и количественные показатели отрицательных последствий, испытанных системой, а исследованию подлежат изменения ее основных функциональных характеристик, что дает возможность проводить формальную оценку состояния безопасности.

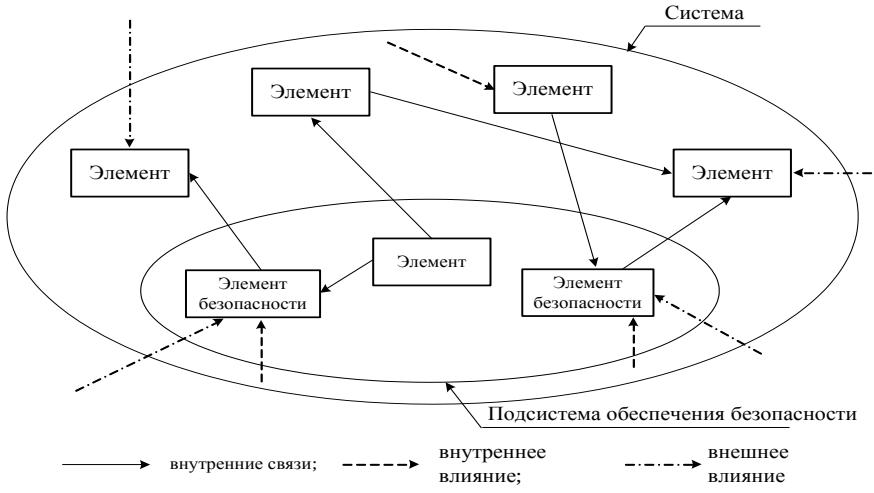


Рис. 10.6. Модель системы безопасности с учетом внутренних и внешних факторов

Подходы к такой оценке могут базироваться на определении степени уменьшения вероятности реализации угроз в результате функционирования механизмов противодействия; уменьшения вреда от реализации угроз в результате частичной нейтрализации их влияния; снижения показателей функциональных характеристик (рис. 10.7).

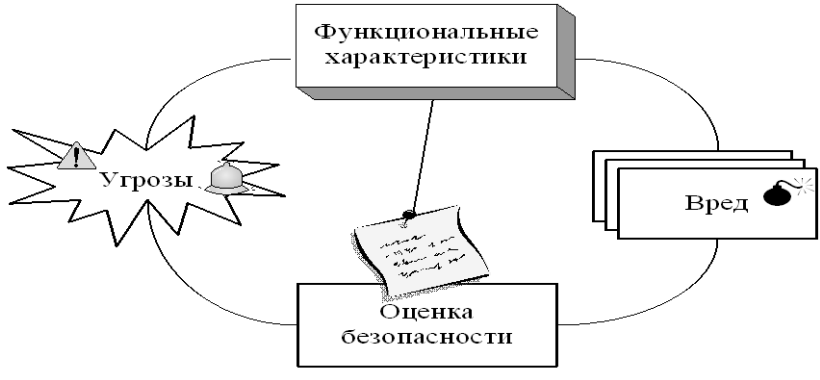


Рис. 10.7. Модель оценки безопасности “угроза - ущерб”

Оценка безопасности - величина конечная. Это означает, что количественные показатели оценки безопасности не могут быть бесконечными или в нормированном виде приближаются к 1, т.е. не может быть достигнута абсолютная безопасность с нулевым убытком от действия любых угроз.

Концептуальная причина невозможности достижения бесконечного значения оценки безопасности связана с недосягаемостью абсолютной безопасности, которая вытекает из причинно-следственной связи угрозы и реакции (адаптации) системы в процессе обеспечения безопасности. Здесь всегда имеет место временная задержка между возникновением новой (или модифицированной) угрозы и окончанием адаптации системы, во время чего и наносится вред. Процесс возникновения новых угроз базируется на диалектическом противодействии средств нападения и защиты, а развитие систем само по себе приводит к перманентному появлению угроз, которые являются новыми для текущего этапа и неизвестными для разработанной или модифицированной системы.

Безопасность является непрерывным во времени состоянием. По обыкновению рассматриваются системы, которые существуют в течение какого-то определенного непрерывного отрезка времени, а поскольку процесс реализации угроз в общем случае является случайным, то даже для простых технических систем задачи определения всего перечня возможных угроз в каждый момент времени являются довольно сложными. В связи с этим система должна постоянно находиться в состоянии обеспечения безопасности, т.е. действия, направленные на нейтрализацию угроз, должны быть превентивными относительно начального момента реализации угроз, что возможно лишь тогда, когда состояние безопасности будет непрерывным во времени.

Безопасность является понятием конкретным. Исследование безопасности осуществляется лишь в конкретных системах. Можно давать оценку безопасности лишь для конкретных систем и лишь относительно конкретного вида или типа угроз. Конкретных значений безопасности в целом или, иначе говоря, безотносительной безопасности не может быть.

Безопасность является понятием интегральным. Внешние и внутренние угрозы могут быть одновременно реализованы относительно разных элементов и связей системы и при этом могут принципиально отличаться по своей природе, механизмом действия, а вред от их реализации может быть качественным и количественным. Таким образом, безопасность системы должна иметь интегральный характер, учитывающий все множества и способы реализации угроз.

Интегральность безопасности не означает простую сумму отдельных безопасностей (отдельная безопасность — возможность нейтрализации конкретной группы родственных угроз), а является функцией их суммы, которая

может быть и простой и достаточно сложной. Например, составляющими национальной безопасности есть политическая, экономическая, экологическая, военная, информационная и т.п., при этом информационная безопасность имеет самостоятельное значение (рис. 10.8).

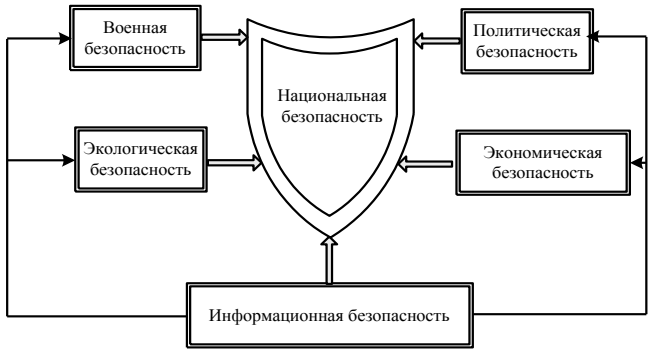


Рис. 10.8. Составляющие информационной безопасности

С учетом издоженного для систем обобщенное определение термина безопасности такое: ***безопасность*** *- такое состояние системы, при котором минимизируется ущерб в результате реализации угроз.*

Это определение имеет высокую степень универсальности и широчайшую сферу применения за счет того, что отмечаются, но не конкретизируются:

система, относительно которой могут быть реализованы угрозы;

состояние системы, для которой определяется безопасность;

виды и типы угроз, в результате реализации которых может быть причинен ущерб;

виды и размеры возможного ущерба.

Как уже отмечалось, под системой понимается объект любой природы и назначения или организованная определенным образом совокупность таких объектов, а состоянием системы в этом случае является полное множество ее количественных и качественных параметров, которые могут изменяться в процессе функционирования системы. Множество значений параметров, при которых обеспечивается допустимый уровень ущерба в случае реализации угроз, и определяет состояние, которое отвечает безопасности системы.

Для определения понятия информационной безопасности укажем, что в общем случае существуют такие ***основные требования к информационным процессам:***

информация, которая используется в системе, должны быть полной, достоверной, своевременной, а также доступной, за исключениями, определенными правилами функционирования системы;

информационные действия не должны приводить к нарушению функционирования системы;

любое изменение режима работы информационных технологий при реализации процессов создания, распространения, использования, хранения и уничтожения информации не должна повлечь за собой нарушение функционирования системы;

информация не может быть уничтожена или модифицирована, распространена или использована без соответствующей санкции.

С учетом этого сформулируем определение информационной безопасности.

Информационная безопасность *— это такое состояние защищенности жизненно важных интересов лица, общества и государства, при котором сводится к минимуму нанесение вреда из-за:*

неполноты, несвоевременности и недостоверности информации, которая используется;

отрицательного информационного влияния;

отрицательных последствий функционирования информационных технологий;

несанкционированного распространения, использования и нарушения целостности информации.

Таким образом, можно предположить, что все процессы в обществе и государстве тесно связаны с информационной безопасностью.

Общесистемные проблемы информационной безопасности. Обеспечение информационной безопасности связано с решением целого комплекса взаимосвязанных проблем, которые можно условно объединить в такие группы: концептуальные, нормативные и организационные.

В основу концепции информационной безопасности положены следующие позиции:

информационный ресурс государства должны быть отнесен к стратегическим ресурсам исходя из того, что в информационной инфраструктуре государства той или другой мерой задействованы значительные человеческие, финансовые и материальные ресурсы, которые по своим объемам приближаются к некоторым областям экономики. Следует отметить, что в современном обществе (с точки зрения обеспечения его жизнедеятельности) весомо возросла значимость информации и его впечатлительность от ее потери или неполучения, а практическое большинство проблем информационной безопасности находит свое отражение в сфере информатизации, и именно в этой сфере происходит наибольшая их концентрация;

деятельность в сфере информационной безопасности должна быть регламентирована системой законодательных и подзаконных актов, нормативных и нормативно-технических документов и осуществляться в суровом соответствии с их требованиями;

управление в сфере информационной безопасности государства должно базироваться как на методах административного управления, так и на новых методах регуляторного характера, в частности таких, что базируются на лицензировании и сертификации;

информационная безопасность государства может быть обеспечена лишь при условии создания соответствующей иерархической организационной структуры и ориентирования на отечественную научную и производственную инфраструктуру;

среди проблем в сфере информационной безопасности особое место занимает правовое регулирование. Несовершенство правового регулирования разнообразия информационных отношений тормозит не только развитие и усовершенствование политических, экономических, материальных и других отношений в обществе, а и сам процесс обеспечения информационной безопасности.

Создание общей системы обеспечения информационной безопасности должно предусматривать наличие в государстве некоторой иерархической организационной структуры, определение которой из системных позиций осуществляется на основе декомпозиции общей функции на ряд отдельных локальных функций (рис. 10.9).

Рис. 10.9. Декомпозиция общей функции информационной безопасности государства на локальные

Необходимо создать такую организационную структуру, обеспечивающую информационную безопасность в государстве, которая была бы функционально полной, рационально объединяла усилия отдельных ее элементов и не допускала неоправданного дублирования.

Таким образом, информационная безопасность имеет самостоятельное и важное значение в контексте национальной безопасности, и потому задачи ее обеспечения являются приоритетными для государства. Выполнению этой задачи содействует создание и развитие современной нормативно-правовой базы в сфере обеспечения информационной безопасности.

Правовой режим информации. Определение правового статуса информации относительно ее использования и доступа базируется на общем постулате открытости информации и ее свободного использования, связанного с принципом информационного права - принципом свободы получения, использования и распространения информации. Вместе с тем существуют объективные условия, которые требуют ограничения доступа к информации, несанкционированные действия с которой могут привести к нарушению прав и интересов лица, общества и государства.

Информация по режиму доступа делится на открытую и информацию с ограниченным доступом, поэтому основная функция органов государственной власти - обеспечение защиты рушенных прав и интересов субъектов информационных отношений в части обеспечения санкционированного доступа, использования и распространения информации. В соответствии с этим формулируются соответствующие базовые принципы правового режима информации в государственном законодательстве (рис. 10.10).

Для информации, которая находится в общественном обращении, ограничение доступа определяется государственным законодательством. Прежде всего это тайна переписки, телефонных разговоров, телеграфной и другой корреспонденции, а также запрет собирания, хранения, использования и распространения конфиденциальной информации о лице без его согласия.

К информации с ограниченным доступом относится конфиденциальная и тайная. Конфиденциальная информация – сведения, которые находятся во владении, пользовании или распоряжении отдельных физических или юридических лиц и распространяются по их желанию соответственно предусмотренным ими условий. Кроме того, к конфиденциальной информации можно отнести информацию, которая принадлежит государству.

Базовый подход к определению оснований для отнесения сведений к категории конфиденциальной информации следующий:

1. Для физических лиц:

их персональные данные;

авторское право на созданную информацию и право на интеллектуальную собственность.

Рис. 10.10. Базовые принципы правового режима информации в государственном законодательстве

2. Для юридических лиц:

право собственности на информацию, созданную за счет собственных ресурсов и собственные средства;

требования законодательства относительно отнесения информации к категории конфиденциальной для определенных видов деятельности при выполнении работ или предоставлении услуг за счет государственного бюджета;

наличие договорных отношений относительно режима доступа к конфиденциальной информации, полученной от физических или юридических лиц.

3. Для органов государственной власти и местного самоуправления основанием для отнесения сведений к категории конфиденциальной информации являются требования соответствующего законодательства.

Массив открытой информации и информационных продуктов может иметь **ограниченное использование**, например, это касается творческой информации, использование которой регламентируется законодательством об авторском праве и праве на интеллектуальную собственность.

Для авторов творческой информации устанавливается особый правовой режим охраны прав, к которым принадлежат личные неимущественные и имущественные права субъектов авторского права (авторов произведений), а как объекты этого права могут быть определены:

литературные письменные произведения разного характера (книги, брошюры, статьи и т.п.);

выступления, лекции, речи, проповеди и др. устные произведения;

компьютерные программы;

базы данных;

текстовые и музыкальные произведения;

драматические, музыкально-драматические, пантомимные, хореографические и другие произведения, созданные для сценического показа, и их постановки;

аудиовизуальные произведения;

произведения изобразительного искусства;

произведения архитектуры, градостроительства и садово-паркового искусства;

фотографические и подобные фотографии произведения;

произведения декоративного ткачества, керамики, резьба, из художественного стекла, ювелирные изделия и другие произведения искусства;

иллюстрации, карты, планы, чертежи, эскизы, относящиеся к географии, геологии, топографии, технике, архитектуре и т.п.;

энциклопедии и антологии, сборники произведений, обработки фольклора и обычных данных, другие составленные произведения при условии, что они являются результатом творческой работы по отбору, координации или приведении в порядок содержания без нарушения авторских прав на произведения, которые входят в них как составные части;

тексты переводов для дублирования, озвучивания, субтитрования аудиовизуальных произведений;

другие произведения.

При этом для правовой охраны не имеет значения, были эти произведения оглашены, завершены или не завершены, какой они целевой направленности и какого объема, назначения и жанра.

К имущественным правам автора прежде всего принадлежит исключительное право на использование произведения и его разрешение или запрет использования другими лицами. Имущественные права дают возможность автору использовать произведение в любой форме и любым способом, а также разрешать или запрещать его:

воспроизведение;

публичное выполнение, извещение, демонстрацию и показ;

любое повторное оповещение, если оно осуществляется другой организацией, чем та, что осуществила первое оповещение;

перевод;

переработку, адаптацию, аранжировки и другие подобные изменения;

включение в состав сборников, антологий, энциклопедий и т.п.;

распространение путем первой продажи, отчуждение другим способом или путем сдачи в имущественный наем или в прокат и путем другой передачи к первой продаже экземпляров произведения;

представление публике так, что ее представители могут осуществить доступ к произведениям из любого места и в любое время по их собственному выбору;

сдача в имущественный наем и коммерческий прокат после первой продажи, отчуждение другим способом оригинала или экземпляров аудиовизуальных произведений, компьютерных программ, баз данных, музыкальных произведений в нотной форме, а также произведений, зафиксированных в фонограмме, видеопрограмме или в форме, которую считывает компьютер;

импорт экземпляров произведений и др.

С одной стороны, такие имущественные права автора является стимулирующим фактором для активной творческой деятельности и использование произведений, но с другой - право автора на ограничение использования вступает в принципиальное разногласие с демократическими положениями о праве каждого свободно собирать, сохранять, использовать и распространять информацию устно, письменно или другим способом.

Частично такое разногласие решается путем ограничения имущественных прав автора при условии, что они не будут наносить ущерб использованию произведения и безосновательно не будут ограничивать законные интересы автора. Поэтому, как правило, без согласия автора, но с обязательным указанием его имени и источника заимствования разрешается:

цитирование творческой информации или воспроизведение ее части при информировании о ней;

копирование одного экземпляра творческой информации библиотеками и архивами, деятельность которых не направлена на получение прибыли при условии, если такое копирование является единичным случаем и не имеет систематического характера;

копирование или использование творческой информации учебными заведениями для аудиторных занятий, а также отрывков из творческой информации (с иллюстрациями или без них) при условии, если такое копирование является единичным случаем и не имеет систематического характера.

Без разрешения автора (или другого лица, которое имеет авторское право) и без выплаты авторского вознаграждения допускается использование (в частности, копирование) в личных целях правомерно оглашенной или правомерно приобретенной творческой информации.

Право собственности на информацию. В современном информационном обществе каждый имеет право владеть, пользоваться и распоряжаться своей собственностью, результатами своей интеллектуальной (творческой) деятельности, и потому все результаты такой деятельности могут быть объектом права владения, пользования и распоряжения, т.е. объектом права собственности. К таким результатам можно отнести и информацию в любой организационной форме и виде, на любых носителях, созданную в результате интеллектуальной (творческой) деятельности. Исходя из этого, правом собственности на информацию является урегулированное законом общественное отношение относительно владения, пользования и распоряжения информацией, т.е. сведениями, представленными в любой организационной форме и виде, на любых носителях, которые могут быть объектом права собственности граждан, юридических лиц и государства как в полном объеме, так и объектом лишь во владении, пользовании или распоряжении.

Информационные продукты (документы, книги, базы данных, иллюстрации, фотографии, голограммы, кино-, видеофильмы и т.п.) являются предметом материального мира и принадлежат к движимому имуществу, и правовые отношения в государстве относительно прав собственности на информационные продукты регулируются соответствующим гражданским законодательством.

Рассмотренные базовые вопросы нормативно-правового обеспечения информационной безопасности находят отображение в ряде международных юридических актов, например конвенциях и соглашениях, принятых Генеральной конференцией или межправительственными конференциями, созванными ЮНЕСКО самостоятельно или вместе с другими международными организациями.

К ним принадлежат:

Конвенция о международном обмене изданиями;

Конвенция об обмене официальными изданиями и правительственными документами между государствами;

Международная конвенция об охране интересов артистов-исполнителей, производителей фонограмм телерадиовещательных организаций;

Конвенция о мероприятиях, направленных на запрет и предупреждение незаконного ввоза, вывоза и передачи права собственности на культурные ценности;

Конвенция об охране интересов производителей фонограмм от незаконного воспроизведения их программ;

Конвенция о распространении программ, которые несут сигналы, которые передаются через спутники;

Всемирная конвенция об авторском праве;

Многосторонняя конвенция о предотвращении двойного налогообложения выплат авторского вознаграждения;

Соглашение о сотрудничестве в области охраны авторского права и сопредельных прав;

Соглашение о содействии распространению в международном плане наглядно-звуковых материалов образовательного, научного и культурного характера;

Соглашение о ввозе материалов образовательного, научного и культурного характера и др.

10.2. Организационные меры защиты информации

Организационные мероприятия защиты информации - это множество процессов и действий (например, контроль за утилизацией носителей информации с ограниченным доступом, планирование мероприятий по восстановлению утраченной информации, аудит систем защиты, реализация экспертизы и т.п.), осуществляемых на всех технологических этапах (проектирование, изготовление, модификация, эксплуатация, утилизация и т.п.) существования соответствующих ресурсов, ведущих к созданию, усовершенствованию, упорядочению и согласованию взаимосвязей и взаимодействия их компонент с целью решения задач. Разрабатывая организационные средства, необходимо учитывать, чтобы в общей совокупности механизмов защиты они могли самостоятельно или в комплексе с другими средствами решать задачи защиты, обеспечивать эффективное использование средств других классов, а также рационально объединять все средства в единую целостную систему защиты. Важно знать, что множество всех нужных и потенциально возможных организационных средств не определено и не существует формальных методов формирования их перечня и содержания. Исходя из этого, основными методами формирования организационных средств можно считать лишь неформально-

эвристические.

Организация защиты информации в современных информационных системах основывается на обеспечении ее базовых характеристик: конфиденциальности, целостности, доступности. При этом надо учитывать факторы комплексности, системности, унификации и непрерывности (рис. 10.11).

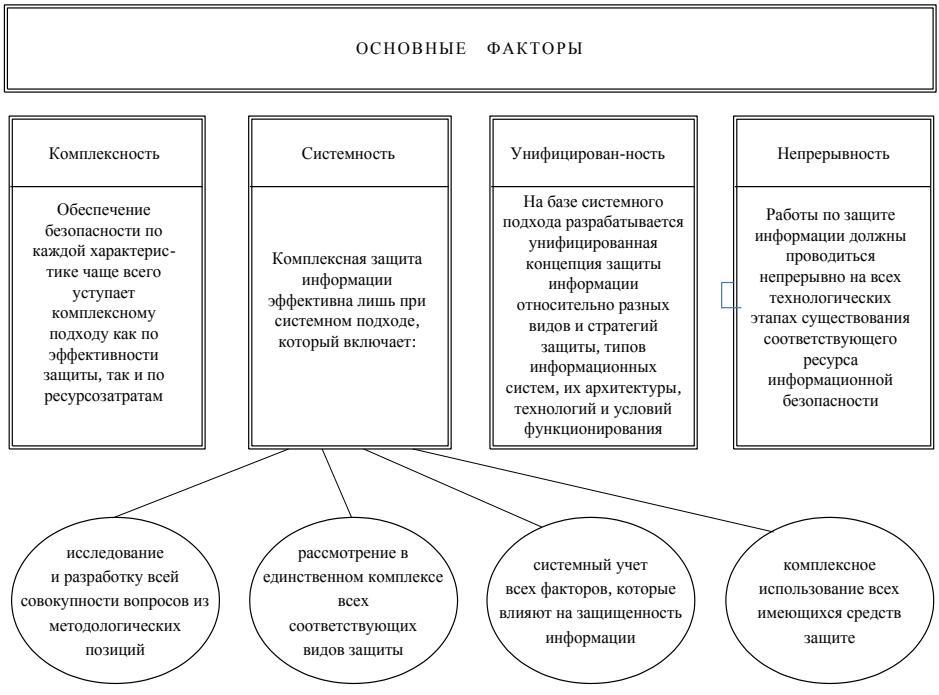


Рис. 10.11.Основные факторы, которые необходимо учитывать при организации защиты информации

Организация защиты в широкой интерпретации этого понятия является процессом создания механизмов защиты информации:

установление необходимой степени защиты информации;

назначение лица, ответственного за выполнение мероприятий по защите информации;

определение возможных причин (каналов) нарушение защищенности информации;

выделение необходимых средств на защиту информации;

выделение лиц (подразделов), которым поручается разработка механизмов защиты;

установление мероприятий контроля и ответственности за соблюдением всех правил защиты информации.

Для повышения эффективности функционирования механизмов защиты существует целый ряд дополнительных мер организационного характера, направленных в основном на обеспечение целостности и конфиденциальности информационных ресурсов (рис. 10.12).

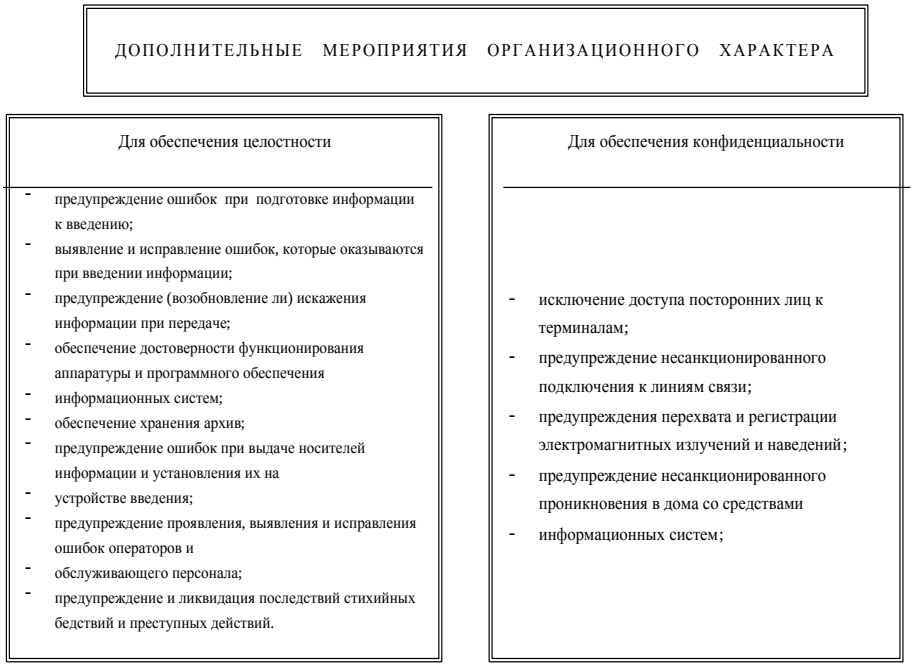


Рис. 10.12. Перечень мероприятий направленных на обеспечение характеристик безопасности

С ростом информационных преступлений совершенствовалось и расширялось множество мероприятий, которые повышают эффективность защиты. В этой связи создания эффективных механизмов защиты, поддержка и обеспечения их надежного функционирования связанные с решением специфичных задач и потому может осуществляться лишь подготовленными высококвалифицированными специалистами-профессионалами соответствующего профиля.

Работы по созданию систем защиты информации выполняются в три этапа (подготовительный, основной и завершающий), а их назначение и общее содержание являются общепринятыми (рис. 10.13).

Часто в процессе функционирования информационных систем появляются непредвиденные факторы, которые предопределяются, с одной стороны, действиями факторов, не учтенных на этапе создания системы защиты, а со второго — изменениями, которые происходят в процессе

функционирования информационных систем, и потому возникает необходимость усовершенствования системы защиты. В этой связи с практической стороны технологический процесс организации работ по защите информации является циклическим.

ОСНОВНЫЕ ЭТАПЫ		
Подготовительный	Основной	Заключительный
Изучаются и оцениваются все факторы, влияющие на защиту информации	Принимается решение относительно необходимого уровня защиты и проектируется система защиты	Оценивание защиты по критериям эффективности и технико-экономическим показателям

Рис. 10.13. Основные этапы работ по созданию систем защиты информации

Структуру и общее содержание такая технология приведена на рис. 10.14.

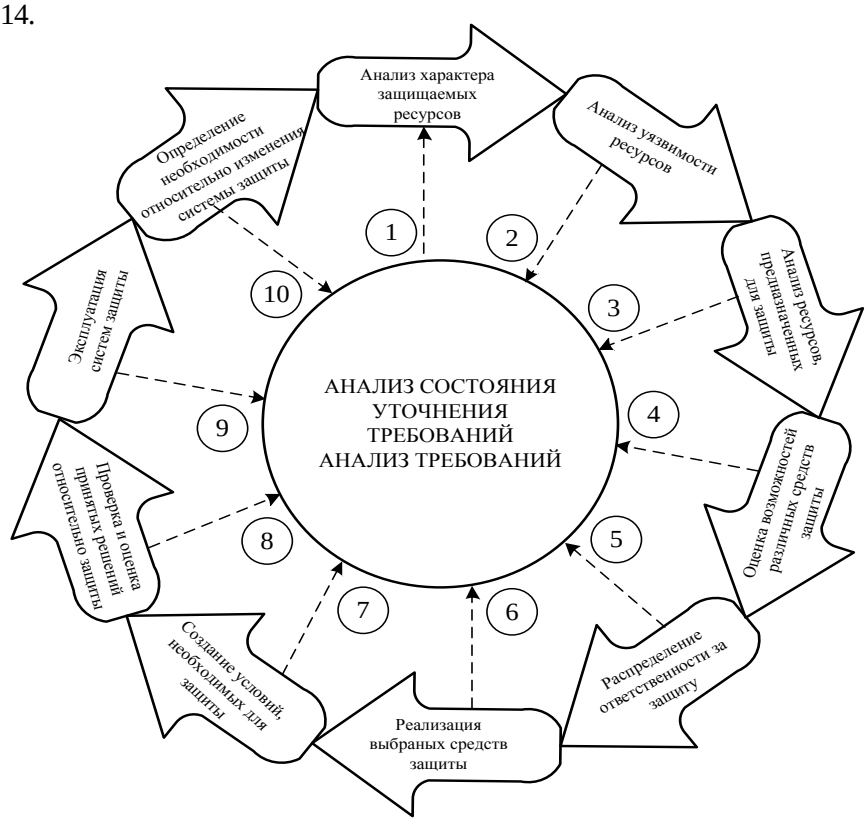


Рис. 10.14. Циклическая диаграмма структуры и содержания организации работ по защите информационных ресурсов

Ее основу составляют такие положения:

непрерывный сбор информации о функционировании механизмов защиты и проведенных работах (для этого осуществляется постоянный контроль защиты информации);

систематический анализ состояния защиты информации;

систематическое уточнение требований к защите информации;

проведение при необходимости (через неудовлетворительное состояние защиты или изменение требований) всего цикла работ по организации защиты.

Защита информации в современных информационных системах является масштабной и очень сложной проблемой. Потому должна быть программа реализации концепции защиты на общегосударственном уровне при наличии надежной системы органов, способных профессионально решать все проблемы защиты информации. При этом многоаспектность и разноплановость унифицированной концепции защиты информации определяют необходимость наличия и функционально разноплановых органов защиты. Чтобы работы по защите информации выполнялись планомерно и целенаправленно, необходимы административные органы. Учитывая сложность, многовекторность, широкий спектр факторов и высокий уровень неопределенности всего комплекса проблем защиты, должны регулярно проводиться специальные научные исследования и разработки, для чего необходимы специализированные научно-исследовательские и проектно-конструкторские органы, а в информационной системе создается соответствующая служба защиты. Для комплектации всех органов кадрами-профессионалами в сфере защиты информации необходимы учебные заведения (осуществляющие подготовку по соответствующим направлениям) с кадрами высочайшей квалификации, которые проводят научно-методическую работу.

Общую организационную структуру органов, ответственных за защиту информации в государстве, можно показать на трех уровнях (рис. 10.15). В структуре предусмотрены органы пяти категорий: административные, научно-организационные, центры защиты информации, службы защиты информации на объектах и учебные центры. Основные функции органов перечисленных категорий можно сформулировать так:

административные - формирование общего заказа на работы в сфере защиты информации, их общая организация, координация и финансирование, а также промышленное производство средств защиты;

научно-организационные - организация научно-исследовательских и опытно-конструкторских работ, координация всех исследований и разработок в сфере защиты информации;

центры защиты информации - проведение научно-исследовательских и опытно-конструкторских работ в сфере защиты информации, предостав-

ление широкого спектра услуг по вопросам защиты информации предприятиям, учреждениям и организациям;

службы защиты информации - решение всего комплекса вопросов, связанных с непосредственной защитой информации на предприятиях, учреждениях и организациях;

учебные заведения и центры — подготовка, переподготовка и повышение квалификации специалистов для решения задач во всех органах вышеперечисленных категорий, разработка научных и других пособий.

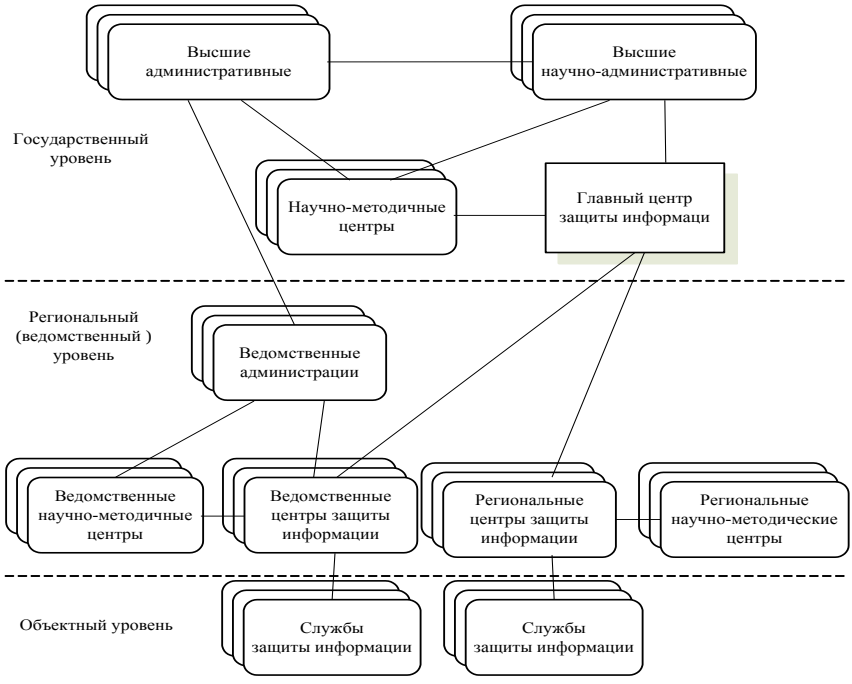


Рис. 10.15. Структура органов, отвечающих за защиту информации

Основу общегосударственной системы органов защиты, как правило, составляют специализированные организации (так называемые центры защиты), ориентированные на решения базовой совокупности вопросов эффективной защиты информации в любых информационных системах разного назначения и разной архитектуры. Другими словами, центры защиты должны взять на себя основную часть работ по защите информации, причем от эффективности их работы в значительной степени будет зависеть эффективность защиты информации.

В структуре органов защиты информации предусмотрены центры защиты двух уровней: главный центр, а также территориальные и ведомственные. Спектр работ по защите информации очень

широкий - от фундаментальных исследований к разработке и внедрению систем защиты информации для конкретных информационных систем, а объем их очень большой и беспрерывно возрастает, что предопределяется интенсивным внедрением вычислительной техники в разные сферы деятельности и резким расширением границ защиты информации. Поэтому сеть территориальных и ведомственных центров защиты развивается и беспрерывно расширяется за счет создания новых центров. Основные принципы построения и функции центров иллюстрирует рис. 10.16.

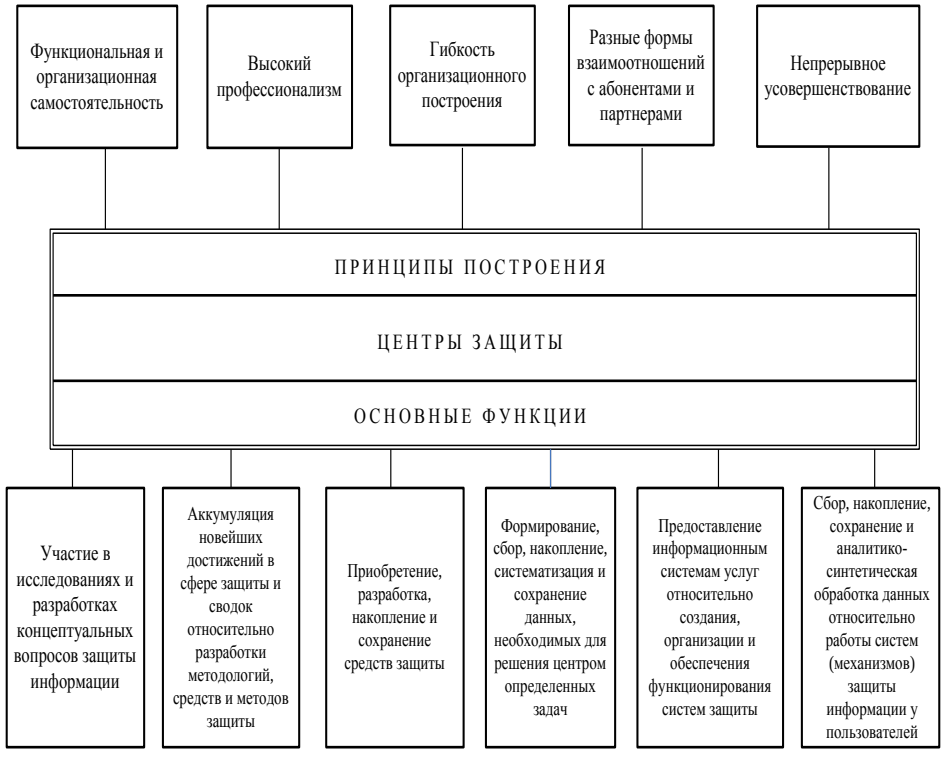


Рис. 10.16. Основные принципы построения и функции центров защиты

Главный центр защиты информации имеет научно-исследовательский профиль с соответствующими высококвалифицированными кадрами-профессионалами и оснащением, необходимым для эффективного выполнения таких базовых функций: организация и проведения фундаментальных исследований в сфере защиты информации и поддержка на этой основе концепций защиты на уровне новейших достижений науки и техники; разработка, формирование и беспрерывное усовершенствование методологической и ин-

струментальной баз защиты информации; научно-методическое и инструментальное обеспечение для создания новых территориальных и ведомственных центров защиты информации; научное обоснование организационно-административных решений в сфере защиты информации; предоставление текущей повседневной помощи территориальным и ведомственными центрам защиты.

Фактически центр защиты информации (территориальный, ведомственный) является специализированным научно-производственным предприятием (объединением), профессионально ориентированным на разработку, практическую реализацию и внедрение различных средств, методов и мероприятий защиты.

Для непосредственной организации создания и функционирование системы защиты информации в информационных системах формируется ***специальная штатная служба защиты*** (служба безопасности).

Служба защиты информации является специальным штатным или внештатным подразделением информационных систем, предназначенным для квалифицированной разработки системы защиты и обеспечение ее функционирование. Функции такой службы иллюстрирует рис. 10.17.

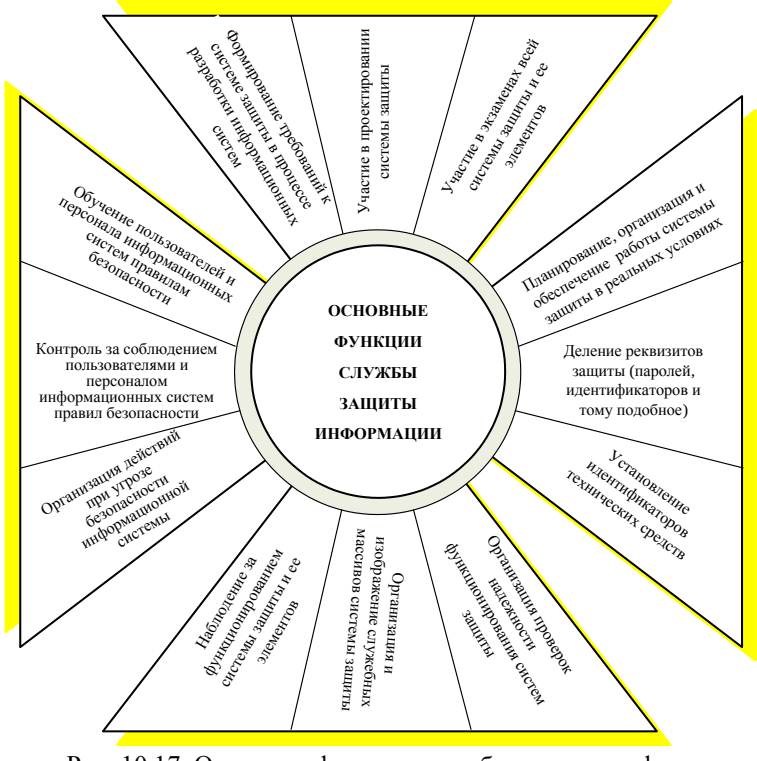


Рис. 10.17. Основные функции службы защиты информации

Определение организационно-правового статуса службы защиты информации зависит от таких позиций:

численность службы должны быть достаточной для выполнения всех заданных функций;

руководитель несет персональную ответственность за соблюдение правил обращения с защищенной информацией;

персонал должен выполнять только обязанности, связанные с безопасностью информационных систем, иметь доступ ко всем ее ресурсам и осуществлять их блокирование при наличии угроз безопасности;

руководитель службы может запретить введение новых ресурсов, если они не отвечают требованиям безопасности;

службе защиты должны быть обеспеченные все условия, необходимые для выполнения своих функций.

10.3. Научно-методическое обеспечение защиты информации

Научно-методическое обеспечение защиты информации — организация разработки, издания, распространения и целенаправленного использования официальных документов, научных работ, учебников и научно-методичных материалов, необходимых для однозначного понимания проблем защиты и обеспечения решения задач защиты информации.

Важный компонент научно-методологического базиса защиты информации - это система типовых документов, основное назначение которых:

обеспечение научно-методологического и концептуального единства при решении задач защиты;

создание условий для однозначного понимания и практической реализации основных положений унифицированной концепции защиты информации;

обеспечение необходимыми данными всех органов и лиц, причастных к защите информации;

обеспечение нормативно-правового регулирования процессов защиты.

По принципу максимальной унификации всех решений защиты информации документационное обеспечение образовывает единую систему типовых документов. Она включает справочно-информационные документы, стандарты, руководящие и методические материалы и инструкции.

К *справочно-информационной группе* относятся документы, которые содержат в систематизированном виде полную совокупность соответствующих сведений, которые, с одной стороны, необходимые и достаточные для получения четкого и однозначного представления о всех аспектах проблемы защиты информации, а с другой - были бы признанные подавляющим большинством специалистов-профессионалов в сфере информа-

ционной безопасности. Как основные подгруппы этой группы документов могут быть выделены словари, описания концепции защиты, справочники.

К *группе стандартов* относятся такие документы и соответствующие решения, которые удовлетворяют по крайней мере трем требованиям:

являются образцом (эталоном) по всем основным параметрам;

имеют точный сертификат по основным параметрам;

утвержденные полномочными органами, причем факт утверждения должны гарантировать пользователям соответствие стандарта своему сертификату.

К *руководящим методическим материалам* относится совокупность документов, имеющих рекомендательный характер и содержащих полное систематизированное описание соответствующих вопросов в сфере информационной безопасности и утвержденных полномочными органами. Такие материалы по обыкновению представляют наиболее представительную группу документов, которые регламентируют разные аспекты работ:

общеметодологические или концептуальные вопросы информационной безопасности;

уязвимости информации и методологии оценивания;

требования к защите и норме защищенности;

функции и задачи защиты;

средства защиты информации (общие вопросы);

технические средства защиты информации;

программные средства защиты информации;

организационные средства защиты информации;

криптографические средства защиты информации;

стенографические средства защиты информации;

механизмы защиты информации;

системы защиты информации (архитектура и технология функционирования);

методология проектирования систем защиты информации;

организация работ по защите информации;

условия, которые оказывают содействие повышению эффективности защиты информации;

нормативное и правовое обеспечение информационной безопасности и т.д.

К *инструкциям* относятся систематизированные подборки типичных инструктивных материалов для разных категорий подразделов и лиц, которые имеют отношение к защите информации. Типовые инструкции утверждаются соответствующими полномочными органами, и на их основе в каждой конкретной организации разрабатываются и утверждаются рабочие инструкции, которые учитывают соответствующую специфику. Такими документами, например, могут быть инструкции о порядке ин-

формации с ограниченным доступом, антивирусной защиты, действий в нештатных ситуациях и т.п.

Классификационная структура системы типового документационного обеспечения защиты информации приведена на рис. 10.18.

Рис. 10.18. Классификационная структура типовых документов защиты информации

Для того чтобы это обеспечение имело системный характер, предполагается его однозначная, информативная и наглядная идентификация.

Однозначность идентификации может быть обеспечена при использовании многоуровневого, например четырехуровневого, идентификатора (см. рис. 10.18), содержащего элементы, последовательно идентифицирующие принадлежность к системе документов из защиты, группы документов, подгруппы и документов в подгруппе. Чтобы идентификаторы были информативными и наглядными, целесообразно первые три элемента (систему документов, группу документов и подгруппу документов) подать мнемонично (буквами), а документы в каждой из подгрупп независимо обозначать последовательными цифровыми номерами.

Условия, оказывающие содействие повышению эффективности защиты информации. Одним из принципиальных положений концепции защиты информации является наличие обратной связи от конструктивных компонентов концепции к ее начальной основе, т.е. к концепциям построения и организации функционирования информационных систем. Обратная связь устанавливает условия, соблюдение которых создает объектив-

ные предпосылки для наиболее эффективного решения задачи безопасности.

Совокупность таких условий (рис. 10.19) является разноплановой, разномасштабной и разновременной по реализации. Общеметодологическими являются условия, создающие общие предпосылки повышения эффективности защиты информации.

Рис. 10.19. Классификация условий, способствующих повышению эффективности защиты информации в информационной системе

Как показано на рис. 10.19, в этом классе выделено две группы условий: осознание проблемы и наличие предпосылок решения.

Основные предпосылки повышения эффективности защиты информации связаны с наличием такой совокупности условий:

надежной, обоснованной, разработанной и общепризнанной концепции информационной безопасности;

достаточно полно разработанных и проверенных методов и моделей защиты;

необходимого арсенала средств защиты;

полного ряда детально разработанного документационного обеспечения;

достаточного количества квалифицированных специалистов-профессионалов в сфере информационной безопасности.

Повышение эффективности безопасности связано с разработкой и реализацией сложной программы:

проведение всесторонних исследований сущности проблемы безопасности и путей ее решения;

разработка теоретических и практических основ информационной безопасности;

построение и реализация методов и моделей защиты информации;

разработка соответствующего арсенала разных средств защиты для решения поставленных задач;

экспериментальные исследования (с помощью специального полигона) разрабатываемых мероприятий, средств, методов и моделей;

создание методического обеспечения с помощью сформированных высококвалифицированных авторских коллективов и соответствующих полиграфических и материально-технических ресурсов;

подготовка достаточного количества профессионалов (от инженерно-технического персонала к специалистам высшей квалификации) в сфере информационной безопасности;

формирование в составе информационных систем служб безопасности, определение их статуса, создание соответствующих *условий* и обеспечение функционирования.

В классе организационных условий выделяются две группы: структурно-функциональная однозначность компонентов информационной системы и организационно-методологическое единство управления.

Структурно-функциональная однозначность компонентов информационной системы связана с тем, что для каждого ресурса системы строго и однозначно определено функциональное назначение, место в общей архитектуре, режимы функционирования, порядок использования и т.д.

Организационно-методологическое единство управления связано с тем, что все процедуры управления защитой информации во всех структурных элементах информационной системы всеми органами управления осуществляются лишь в пределах единой концепции безопасности.

Концептуальная стандартизация в сфере построения информационных систем представляет собой стандартизацию на уровне концепций, общих принципов и правил организации и обеспечение типа деятельности. Она предусматривает разработку стандартных принципов и правил организации и обеспечение обработки данных на индустриальной основе с комплексным применением информационных систем. При этом должно обеспечиваться структурирование ресурсов и технологий обработки информации с целью эффективного осуществления конкретного вида деятельности.

Создание законодательной основы, необходимой для обеспечения защиты информации. Абсолютная необходимость создания законодательной основы очевидна, поэтому в ведущих западных странах, а теперь и в Украине этому вопросу уделяется достаточно большое внимание.

Проблема законодательного регулирования процессов обработки информации впервые начала обсуждаться за рубежом в 60-е годы XX ст. и, в частности в США в связи с предложением создать общенациональный банк данных. В настоящее время на международном уровне сформировалась устойчивая система взглядов на информацию как на ценнейший ресурс жизнеобеспечения общества, правовое регулирование в сфере которого должно идти по следующим трем направлениям.

Защита прав личности на частную жизнь. Этот аспект не является новым для мирового сообщества. Основные принципы установления пределов вмешательства в частную жизнь со стороны государства и других субъектов определены основополагающими нормами ООН, а именно Декларацией прав человека. К концу 70-х годов сформулированы два принципа, нашедших впоследствии отражение в национальных законодательствах по информатике ряда стран Запада:

установление пределов вмешательства в частную жизнь с использованием компьютерных систем;

введение административных механизмов защиты граждан от такого вмешательства.

Примерами документов, относящихся к этому направлению, являются: резолюция Европарламента «О защите прав личности в связи с прогрессом информатики» (1979 г.) и Конвенция ЕС «О защите лиц при автоматизированной обработке данных персонального характера» (1980 г.).

Защита государственных интересов. Проблема решается с помощью достаточно разработанных национальных законодательств, определяющих национальные приоритеты в этой области. Интеграция стран-членов ЕС потребовала координации усилий в данной области, в результате чего общие принципы засекречивания информации отражены в Конвенции ЕС по защите секретности.

Защита предпринимательской и финансовой деятельности. Данный аспект проблемы решается путем создания законодательного механизма, определяющего понятие «коммерческая тайна» и устанавливающего условия для осуществления «добросовестной» конкуренции, квалификации промышленного шпионажа как элемента недобросовестной конкуренции.

К этому же направлению можно отнести создание механизмов защиты авторских прав, в частности прав авторов программной продукции. Последний аспект отражен в директиве ЕС «О защите программ для ЭВМ и баз данных» (1990 г.).

Разработанные на международном уровне концептуальные основы и принципы защиты информации нашли отражение в национальных законодательствах ведущих стран Запада. Ниже приведены некоторые примеры действующих в них законодательных актов:

Великобритания – Билль о надзоре за данными (1969 г.), Закон о защите информации (1984 г.);

Франция – Закон об информатике, картотеках и свободах (1978 г.);

ФРГ – Закон о защите данных персонального характера от злоупотреблений при обработке данных (1977 г.), Закон о защите информации (1978 г.);

США – Закон о тайне частной информации (1974 г.), Акт о злоупотреблениях в использовании ЭВМ (1986 г.), Закон о безопасности компьютерных систем (1987 г.);

Канада – Закон о компьютерных и информационных преступлениях (1985 г.).

Наиболее развитое законодательство в этой сфере действует в США (свыше сотни различных законодательных актов). Законодательство США охватывает:

определение и закрепление государственной политики в области информатизации;

обеспечение развитого производства, технологий;

борьбу с монополизмом и стимуляцию приоритетных направлений;

организацию информационных систем;

защиту прав потребителя, особенно прав граждан на информацию, защиту информации о гражданах;

регулирование прав разработчиков программ для ЭВМ.

В большинстве стран в законодательствах установлена ответственность за нарушение порядка обработки и использования персональных данных; компьютерные преступления расцениваются как преступления, представляющие особую опасность для граждан, общества и государства, и влекут за собой значительно более жесткие меры наказания, нежели аналогичные преступления, совершенные без применения компьютерной техники; как преступления рассматриваются также действия, создающие угрозу нанесения ущерба, например, попытка проникновения в систему, внедрение программы-вируса и т.п.

Говоря об отечественном опыте правового обеспечения информатизации и защиты информации, мы должны отметить, что вопрос этот был впервые поставлен в нашей стране в 70-х годах в связи с развитием АСУ различных уровней. Однако, нормативная основа в ту пору не вышла за рамки ведомственных актов, нескольких постановлений правительства и аналогичных актов республиканского уровня. Поэтому законодательное регулирование процессов информатизации к началу 90-х годов нельзя было назвать удовлетворительным. Необходимо было срочно создать правовую основу информатизации Украины, законодательно обеспечить эффективное использование информационного ресурса общества, урегулировать правоотношения на всех стадиях и этапах информатизации, защи-

тить права личности в условиях информатизации, сформировать механизм обеспечения информационной безопасности.

1991 г. может быть отмечен как начало активной законотворческой деятельности в этом направлении. Законодатели при этом справедливо сконцентрировали свое внимание на следующих наиболее острых для Украины проблемах: проблеме права на информацию; проблеме собственности на некоторые виды информации; проблеме признания информации объектом товарного характера.

В Конституции Украины (1996 г.) закреплено общее право граждан на информацию. Ограничения этого права могут устанавливаться законом только в целях охраны личной, семейной, профессиональной, коммерческой и государственной тайны, а также нравственности. Перечень сведений, составляющих государственную тайну, устанавливается законом.

Принят базовый закон Украины «Об информации, информатизации и защите информации», а также специальные законы «О государственной тайне», «О правовой охране программ ЭВМ и баз данных», «О правовой охране топологий интегральных микросхем», «О международном информационном обмене». Одной из целей закона Украины «О коммерческой тайне» явилось создание со стороны государства необходимых гарантий защиты субъектов путем предоставления им права засекречивать ценную информацию в качестве коммерческой тайны для защиты ее владельца от промышленного шпионажа и недобросовестной конкуренции.

10.4. Международные стандарты информационной безопасности

Стандарты информационной безопасности являются *нормативно-техническими документами, внедряющими комплекс норм, правил и требований, предназначенных для взаимодействия между производителями, потребителями и экспертами соответствующей сферы в процессе создания и эксплуатации защищенных объектов*. Такие стандарты утверждаются компетентными органами и являются важнейшими критериальными средствами, используемыми для решения прикладных задач информационной безопасности. Они могут делиться на критерии, методики, системы, требования к средствам защиты, цифровую подпись и т.п.

Известнейшими в мировой практике стандартами информационной безопасности, которые дали толчок соответствующим процессам стандартизации в международных организациях и государствах, являются:

Критерии безопасности компьютерных систем;

Критерии безопасности информационных технологий;

Федеральные критерии безопасности информационных технологий;

Канадские критерии безопасности компьютерных систем;

Общие критерии безопасности информационных технологий.

Критерии безопасности компьютерных систем (1983 г.) мини-

стерства обороны США («Оранжевая книга») предназначены для определения требований безопасности, которые выдвигаются к аппаратному, программному и специальному обеспечению компьютерных систем и изготовление соответствующей методологии анализа политики безопасности, которая реализуется в компьютерных системах военного назначения. В критериях предложены такие категории требований безопасности: политика безопасности, аудит и корректность.

Здесь предусмотрены четыре группы критериев: D (класс D1), C (классы C1, C2), B (классы B1, B2, B3) и A (класс A1), характеризующих степень защищенности, начиная от минимальной и заканчивая формально доказанной.

Укажем, что критерием оценки фактически является соответствие множества средств защиты данной системы множеству, указанному в одном из классов оценки, и в случае, если набор средств недостаточный, то систему защиты относят к первому низшему классу.

В ***Критериях безопасности информационных технологий*** (1991 г.), разработанных странами Европы (поэтому и названных «Европейские критерии»), общая оценка уровня безопасности системы состоит из функциональной мощности средств защиты и уровня адекватности их реализации. В этом стандарте наблюдается тесная связь с «Оранжевой книгой», но главное отличие «Европейских критериев» заключается в том, что здесь впервые введено понятие адекватности средств защиты и специальная шкала для его критериев, причем адекватности отводится значительно больше внимания, чем функциональным требованиям. Во время проверки адекватности анализируется весь жизненный цикл системы - от начальной стадии проектирования к эксплуатации и сопровождению. В документе определяется семь уровней адекватности - от E0 к E6. Уровень E0 - минимальный (аналог уровня D «Оранжевой книги»), на уровне E1 анализируется лишь общая архитектура системы, а адекватность средств защиты подтверждается функциональным тестированием, на уровне E3 к анализу привлекаются исходные тексты программ и схемы аппаратного обеспечения, а на уровне E6 нужно формальное описание функций безопасности, общей архитектуры и политики безопасности. Степень безопасности определяется самым слабым из критически важных механизмов защиты.

Заметим, что недостатком этого документа является отсутствие четкой взаимосвязи между процессом проектирования системы и оценкой ее безопасности, которая может привести к дополнительным расходам во время доработки информационной системы с целью повышения уровня защищенности.

Федеральные критерии безопасности информационных технологий (1990-е годы) практически охватывают весь спектр проблем, связанных с защитой, и впервые объявили концепцию профиля защиты, содержащую требования к проектированию и технологии разработки, и квали-

фикационный анализ продукта информационных технологий (IT-продукта).

Согласно «Федеральным критериям» процесс разработки систем обработки информации осуществляется в виде последовательности таких основных этапов:

разработка и анализ профиля защиты;

разработка и квалификационный анализ IT-продуктов;

компоновка и сертификация системы обработки информации.

«Федеральные критерии» регламентируют только первый этап этой схемы: разработку и анализ профиля защиты. Процесс создания IT-продуктов и компоновка систем обработки информации остаются за пределами этого стандарта.

Преимуществом этих критериев является то, что вместо обобщенной универсальной шкалы классов безопасности и жестких директив этот документ содержит согласованный с предыдущими стандартами ранжированный перечень функциональных требований, который разрешает разработчикам и пользователям подбирать наиболее пригодные требования для конкретного IT - продукта и среды эксплуатации.

Канадские критерии безопасности компьютерных систем (1990-е годы), равно как и «Федеральные критерии», содержат независимое ранжирование требований по отдельно взятым разделам, в результате чего определяется множество отдельных критериев, характеризующих работу отдельных подсистем обеспечения безопасности. В этом документе кроме функциональных критериев введены критерии адекватности реализации, отражающие уровень корректности реализации политики безопасности и определяющие требования к процессу проектирования, разработки и реализации компьютерных систем.

В критериях применен дуальный принцип представления требований безопасности в виде функциональных требований к средствам защиты и гарантиям их реализации. «Канадские критерии» являются хорошо сбалансированным конгломератом «Оранжевой книги» и «Федеральных критериев», усиленных требованиями гарантий реализации политики безопасности, и вместе с другими стандартами послужили основой для разработки «Общих критериев безопасности информационных технологий».

Общие критерии безопасности информационных технологий стали продуктом объединения Канадских, Федеральных и Европейских стандартов в единый согласованный документ. Эти критерии регламентируют все стадии разработки, квалификационного анализа и эксплуатации IT-продуктов. Они определяют множество типовых требований, вводят шкалы, разрешающие потребителям создавать отдельные требования, отвечающие их нуждам. Основными документами, описывающими все аспекты безопас-

ности IT-продукта с точки зрения пользователей и разработчиков, являются профиль защиты и проект защиты.

Общие критерии безопасности информационных технологий - это стандарт информационной безопасности, обобщающий содержание и опыт использования «Оранжевой книги». В нем развиты «Европейские критерии», воплощены в реальные структуры концепции типовых профилей защиты «Федеральных критериев» США и соответственно «Канадским критериям» представлена одинаковая основа для формулировки разработчиками, пользователями и оценщиками информационных технологий (квалифицированными экспертами и) требований, метрических свидетельств и гарантий безопасности. Версия 2.1 этого стандарта утверждена Международной организацией стандартизации (ISO) 1999 г. как международный стандарт информационной безопасности ISO/IES 15408. Материалы стандарта фактически являются энциклопедией требований и гарантий информационной безопасности, отбираемых и реализуемых в функциональных стандартах (профиль защиты) обеспечения информационной безопасности для конкретных систем, сетей и средств как пользователями, так и разработчиками и операторами сетей. Кроме того, ISO приняла еще ряд стандартов, регулирующих другие сфере деятельности в сфере информационной безопасности.

Международная организация гражданской авиации (ICAO) имеет свои подходы к стандартизации, в основу которых положены Стандарты и Рекомендованная практика, отображенные в приложениях к Чикагской конвенции. Вопрос безопасности рассматривается в Приложении 17 этой конвенции, направленной на защиту международной гражданской авиации от актов незаконного вмешательства, и базовые вопросы связаны с защитой от кибер-терроризма.

Базовые подходы и методы оценки состояния безопасности. Важным элементом оценки состояния безопасности является экспертиза как техническое подтверждение того, что мероприятия безопасности и контроля, подобранные для определенного средства защиты информации, отвечают стандартам и нормально функционируют. Эффективность стандартов в значительной мере связана с обеспечением нужного уровня состояния безопасности и зависит от реализации ряда мероприятий, которые можно поделить на группы:

анализ угроз;

разработка, выбор и применение мероприятий и средств безопасности, адекватных угрозам;

сертификация и аккредитация средств безопасности;

планирование и организация действий в непредвиденных обстоятельствах.

Разработка методов и средств, обеспечивающих эффективность этих мероприятий, связана с реализацией различного вида оценок, например

для измерения уровней риска, защищенности, гарантий или выбора оптимального варианта системы защиты и т.п.

Рассмотрим методы и средства, используемые в теории и практике информационной безопасности для решения таких задач.

Известна обобщенная последовательность измерения безопасности, включающая следующие шаги.

Шаг 1. Формулирование требований и подходов, включая требования обеспечения необходимого уровня безопасности.

Шаг 2. Использование избранных методов измерения.

Шаг 3. Интерпретация результатов.

Шаг 4. Определение соответствия измеренного уровня безопасности и необходимого ее уровня.

Из приведенных шагов видим, что данная последовательность нуждается в использовании конкретных методов, моделей и средств.

Существует вероятностная модель оценки защищенности, основывающаяся на соответствующих средствах защиты, перекрывающих заранее известные множества возможных каналов несанкционированного доступа, а соответствующий показатель зависит от «прочности» наиболее слабого звена. Вероятность непреодоления нарушителем препятствия с учетом возможного отказа системы определяют по формуле

$$P = P_{\text{в бл}}(1-P_{\text{отк}}) \wedge (1-P_{\text{обх}_1}) \wedge (1-P_{\text{обх}_2}) \wedge \ldots \wedge (1-P_{\text{обх}_j}),$$

где $P_{\text{в}_{\text{бл}}} = (1-P_{\text{пр}})$ — вероятность выявления и блокирования несанкционированных действий нарушителя; $P_{\text{отк}}(t) = e^{-\lambda t}$ — вероятность отказа системы; $P_{\text{обх}}$ — вероятность обхода препятствия нарушителем; j — количество путей обхода препятствия; $P_{\text{пр}}$ — вероятность преодоления препятствия нарушителем.

Для неконтролируемых возможных каналов несанкционированного доступа расчет ведется по выражению

$$P_{\text{СЗИ}} = (1-P_{\text{пр}}) \wedge (1-P_{\text{обх}_1}) \wedge (1-P_{\text{обх}_2}) \wedge \ldots \wedge (1-P_{\text{обх}_j}).$$

В случае, если каналы закрыты двумя и больше средствами защиты, расчет выполняют по формуле

$$P_{\Sigma} = 1 - \prod_{i=1}^{m}(1-P_i),$$

где i — порядковый номер препятствия; m — количество дублирующих препятствий; P_i — «прочность» i-го препятствия.

Такой подход к оценке защищенности предусматривает начальные условия, которые, например, задаются в техническом задании на компьютерную систему обработки данных, где и обсуждена модель нарушителя, т.е. средства защиты от нарушителей определенного класса уже определены на этапе проектирования и фактически выполняется оценка их «прочности».

Также используют метод экспертных оценок безопасности, рассматривающий систему обеспечения защиты n характеристик средств защиты. Если эти средства используются совместно, то *непременно* возрастает *степень обеспечения* безопасности компьютерных систем. Для количественной оценки вводится некоторая мера G_i характеристики F_i. Считается, что когда $G_i = 0$, то характеристики F_i система не имеет.

Далее используется субъективный весовой коэффициент важности W_i, присвоенный характеристике F_i некоторым экспертом (экспертами). При этом должны выполняться условия $0 \le G_i \le 1$ и $W_i > 0$ для $1 \le i \le n$.

Для определения степени безопасности (SB) компьютерных систем на основе уже определенных параметров используется линейный метод «взвешивания и подсчета», представленный уравнением

$$SB = \frac{1}{n}\sum_{i=1}^{n} W_i G_i .$$

В методе предусматривается, что приведенная формула определенной мерой противоречит положению об определении «прочности системы защиты прочностью ее самого слабого звена», и показывает, что для идеально безопасной системы $SB = 1$, а для целиком незащищенной $SB = 0$.

Для оценки рисков в теории информационной безопасности, например для ранжирования угроз, используют дельфийские списки, отображающие группу экспертов, собирающих информацию в пределах проблемной сферы.

Дельфийская команда - это основа компьютеризированных экспертных систем, поскольку на базе их знаний формируются продукционные правила, моделирующие принятие решения человеком. Команды формируются исходя из компетентности в исследуемой области знаний конкретной системы, уровня информации о состоянии дел, практического опыта и т.п. с целью объединения суждений экспертов для достижения определенного консенсуса.

Для оценки риска команда определяет множество угроз с целью их ранжирования по степени опасности. Следующим шагом является формирование системы угроз, упорядоченной по убыванию их опасности. В основу формирования такого списка могут быть положены разные принципы, например, наибольший риск, уровень секретности, стоимость, трудоемкость, следствия, ущерб, вероятность возникновения. Часто наведение порядка выполняет один человек, действуя как дельфийская команда.

Используют также простое, кардинальное и относительное ранжирование риска.

По *методу простого ранжирования* все возможные угрозы, пораженные места и другие характеристики, которые составляют основу критериев при принятии решений, расставляются в нисходящий ряд, т.е. опаснейшие или важнейшие элементы находятся в начале списка, а менее значимые - в его конце.

Метод кардинального ранжирования основывается на том, что каждой угрозе в результате ее влияния присваивается конкретное числовое значение, определяемое суммой убытка. Такой метод по обыкновению использует категории высокого, среднего и низкого рисков.

В *относительном ранжировании* любой список заносят в таблицу, по которой строят треугольную матрицу. После этого переходят к простому ранжированию, а полученная матрица становится моделью решения.

Такой метод существенным образом упрощает консолидацию суждений благодаря возможности сравнения отдельно взятой угрозы с теми, что остались. Для облегчения принятия решения полезной оказывается возможность сравнения двух угроз, игнорируя все другие. Важное преимущество метода относительного ранжирования риска заключается в том, что нет потребности в принятии единого решения, т.е. эксперт группы может отдать голос за одну из двух угроз или поделить его, например, на равные части (0,5 голоса на угрозу).

Безопасность является качественной характеристикой системы, вследствие чего возникают трудности относительно ее измерения в любых единицах и потом сравнение безопасности, например двух систем.

Нельзя также не учитывать тот факт, что принятие решения при экспертизе зависит от субъективных суждений эксперта, его знаний и опыта. Уменьшение отрицательного влияния этого фактора (особенно когда нет полной информации о системе, а данные, подлежащие обработке, заданы нечетко и часто связаны с суждениями и интуицией человека) достигается применением математического аппарата теории нечеткости и мягких вычислений, которая оперирует такими понятиями, как нечеткие или лингвистические переменные, нечеткие отношения и т.д. Также одним из перспективных подходов построения систем реализации экспертизы в сфере защиты информации и определение состояния безопасности является использование нейронных сетей и генетических алгоритмов.

Основные выводы

Информация - это сведения, представленные в любой организационной форме в произвольном виде, на любых носителях, о событиях и явлениях, которые имели или имеют место в обществе, государстве и окружающей среде.

Информационная система - организованная совокупность предприятий, подразделов и специалистов, нормативно-правового обеспечения комплекса организационных и технических мероприятий, информационных технологий и информационных ресурсов, предназначенных для обеспечения информацион-

ных процессов, в частности, создания, распространения, использования, хранения и утилизации информации.

Информационные ресурсы - это любая совокупность информации, включая документы, независимо от содержания, времени и места создания, а если они находятся под юрисдикцией государства и доступны для использования лицом, обществом и государством, то они являются национальными информационными ресурсами.

Безопасность информации - это такое состояние системы, в которой она циркулирует, при котором минимизируется ущерб от ее несанкционированного распространения, и/или использования, и/или нарушения целостности информации.

Информационная безопасность — это такое состояние защищенности жизненно важных интересов лица, общества и государства, при котором сводится к минимуму ущерб из-за неполноты, несвоевременности и недостоверности информации, которая используется; отрицательное информационное влияние; отрицательные последствия функционирования информационных технологий; несанкционированное распространение, использование и нарушение целостности информации.

Конфиденциальная информация - это сведения, которые находятся во владении, пользовании или распоряжении отдельных физических или юридических лиц и распространяются по их желанию соответственно предусмотренным ими условий. Кроме того, к конфиденциальной информации можно отнести информацию, которая принадлежит государству.

Информация ограниченного использования - это творческая информация, на которую распространяется авторское право и право на интеллектуальную собственность, несанкционированное использование которой наносит ущерб авторам этой информации.

Право собственности на информацию - урегулированные законом общественные отношения относительно владения, пользования и распоряжения информацией.

Законодательные средства защиты информации — это множество нормативно-правовых актов (конвенции, законы, указы, постановления, нормативные документы и т.п.), которые действуют в определенном государстве и обеспечивают юридическую поддержку для решения задач защиты информации.

Организационные средства защиты информации - множество процессов и действий (контроль за утилизацией носителей информации с ограниченным доступом, планирование мероприятий по восстановлению утраченной информации, аудит систем защиты, реализация экспертиз и т.п.), осуществляемые на всех технологических этапах (проектирование, изготовление, модификация, эксплуатация, утилизация и т.п.) существования соответствующих ресурсов и ведут к созданию, усовершенствованию,

упорядочению и согласованности взаимосвязей и взаимодействия их компонент с целью решения задач защиты информации.

Стандарты информационной безопасности - нормативно-технические документы, которые вводят комплекс норм, правил и требований, предназначенных для взаимодействия между производителями, потребителями и экспертами соответствующей сферы в процессе создания и эксплуатации защищенных объектов.

Вопросы для самоконтроля

1. Охватывает ли термин «информация» сведения, которые документированы или публично объявлены?

2. Какое общее требование выдвигается для процесса распространения информации?

3. Что должны обеспечивать системы хранения информации?

4. Какие существуют условия реализации процесса использования информации?

5. Чем заканчивается жизненный цикл информации?

6. Как классифицируется информация по режиму доступа?

7. На какие две группы делятся известные подходы оценки безопасности информации?

8. Какие виды сведений относятся к информации с ограниченным доступом?

9. По каким требованиям сведения относят к той или другой категории конфиденциальной информации?

10. Какие сведения зачисляются к информации ограниченного использования?

11. Приведите перечень основных вопросов организации и обеспечение защиты информации и раскройте их общее содержание.

12. Приведите структуру органов, ответственных за защиту информации.

13. Назовите и раскройте основные положения концепции построения и работы центров защиты информации.

14. Раскройте содержание работы центров защиты информации по формированию методологического базиса и инструментальных средств защиты.

15. Раскройте назначение и основные задачи служб защиты информации.

16. Приведите структуру системы типичных документов по защите информации и дайте короткую характеристику основных типов документов.

17. Приведите структуру и дайте короткую характеристику условий, которые оказывают содействие повышению эффективности защиты информации.

18. Какие известнейшие в мировой практике стандарты информационной безопасности?

19. Какие стандарты объединяет в себя документ «Единые критерии безопасности информационных технологий»?

20. Какая концепция впервые предложена в Федеральных критериях безопасности информационных технологий?

The main conclusions

Information is the data, given in any organizational form and kind, on any media, about events and the phenomena that took or take place in society, state and environment.

Information system is organized collection of enterprises, subdivisions and experts, normative-legal support of the complex of organizational and technical measures, information technologies and information resources intended for providing of information processes, in particular creation, distribution, using, storage and utilization of the information.

Informational resources are any collections of the information, including documents, regardless of content, time and place of creation, and if they are under jurisdiction of the state and accessible for the use by a person, society and the state, they are national informational resources.

Safety of information is a state of the system, where it circulates, when harm is minimized through its unauthorized distribution and/or use, and/or violation of integrity of the information.

Information safety is a state of protectability of the vital interests of person, society and state, when doing harm because of incompleteness, impertinence and unauthenticity of the using information; negative informational influence; negative consequences of functioning of information technologies; unauthorized distribution, using and violation of integrity of information is reduced to minimum.

Confidential information is the data that are in possession, using or at disposal of separate physical or legal persons and extend at their request according to the conditions foreseen by them. Besides, the information that belongs to the state can be considered to the confidential information.

The information of limited use is the creative information, on which the copyright and the right on intellectual property extend, unauthorized usage of which harms the authors of this information.

The property on information is public relations, regulated by the law, about possession, using and disposing of the information.

Legislative facilities of information protection are the plural of normative-legal acts (conventions, laws, decrees, decisions, normative documents and so on) that operate in the certain state and provide legal support for solving of the problems of information protection

Standards of informational safety are normative and technical documents that introduce the complex of norms, rules and requirements, intended for interaction between producers, consumers and experts of corresponding sphere in the process of creation and maintenance of the protected objects.

Ключевые слова

Русский	Английский
документ	document
ведомости	information
режим доступа	access mode
угроза безопасности	security threat
стандарты	standards
конфиденциальность	confidentiality
право собственности	law of propert

Предметный указатель **Subject index**

Автокорреляция 322, 323, 438 Autocorrelation
Автоматизированная система 24, 31, 33 Automatized system
Авторизация (санкция, разрешение) 710, 733 Authorization
Авторизованный пользователь 522 Authorized user
Адаптивная фильтрация 335 Adaptive filtration
Администратор 431, 587, 650, 660 Administrator, administrative user
Акустический канал 345 Acoustic channel
Алгоритм декодирования 508, 554, 576 Decoding algorithm
Алгоритм кодирования 509, 517, 524 Encoding algorithm
Алгоритм хеширования 608, 609 Hashing algorithm
Алгоритм шифрования данных 593 Data encryption algorithm
Амплитуда 110, 118, 122 Amplitude
Амплитудная модуляция 244, 335 Amplitude modulation
Анализ риска 761 Risk analysis
Аналого-цифровой преобразователь 280, 335 Analogy-digital transformer
Аппаратные средства 46, 575 Hardware
Аппроксимация 76, 168, 170 Approximation
Асимметричная криптосистема 583 Asymmetrical cryptosystem
Асимметричный шифр 589 Asymmetrical cipher
Атака 634 Attack
Атрибут доступа 709 Tag, access mediation information
Аутентификация 598, 697 Authentication
Аутентификация пользователя 697 User authentication

База данных 26 Database
Байт 75 Byte
Безопасность 19 Safety, security
Безопасность входа Login security
Безопасность данных 19 Data security
Безопасность информации 19, 63, 728 Information security
Безопасность информационной системы 732 Information system security
Безопасность связи 762 Communications security
Белый шум 149 White noise
Беспроводная сеть 443 Wireless network
Бит 77, 79 Bit
Бит/с 110 Bits per second; bps
Блок данных 613, 615 Data block

Блокирование 698	Blocking
Брандмауэр 699	Firewall
Ведомости 541	Information
Вероятность 26, 27	Probability
Взаимодействие открытых систем 415	Open system interconnection
Виртуальная информационно-коммуникационная сеть 431	Virtual information communication network
Виртуальная частная сеть 432	Virtual private network
Виртуальное соединение 715	Virtual connection
Владелец 463, 624, 721	Owner
Внешняя защита 13, 674	External security
Внутренняя защита 674	Internal security
Восстановление 156, 276, 340, 488, 512, 713	Recovery
Восстановление данных 340, 488, 512, 713	Data recovery
Восстановленный открытый текст 582	Reconstructed plaintext
Выборка 103, 198, 274	Sampler
Вычислительная система 580	Computer system
Гаммирование 589	Output feedback mode
Гарантии 756	Assurance, guarantee
Гарантия защиты 642	Security accreditation
Главный узел 745	Head end
Данные 73, 111	Data
Двоичная последовательность 503	Binary sequence
Двоичное кодирование 489	Binary encoding
Двоичный код 286, 492, 539	Binary code
Двоичный код с исправлением ошибок 459	Binary error-correction code
Двоичный код с определением ошибок 553	Binary error-detection code
Декодер источника 406	Decoder of the source
Декодирование 342, 539	Decode
Демодулятор 349	Demodulator
Демодуляция 342	Demodulation
Детектирование 262	Detection
Дешифрация 581	Decipherment
Дискретизация 117, 170, 188, 272, 302	Discretisation
Дискретный канал 356	Discrete channel

Диспетчер доступа — Reference monitor
Дистанционный доступ 24 — Remote access
Длина сообщения 75 — Message length
Документ 25, 53 — Document
Домен компьютерной системы 424 — Domain
Достоверность 12 — Authenticity
Достоверность данных 12 — Data validity
Достоверность передачи информации 340 — Data transmission validity
Доступ 12, 442, 665 — Access
Доступ к информации 12, 24 — Access to information
Доступность 411, 522 — Availability, accessibility
Доступность данных 424 — Data accessibility
Древоподобный код 494 — Tree code

Запрос доступа 476, 683 — Access request
Защита 12 — Protection; security
Защита информации 12 — Information protection
Защита информации в автоматизированной системе 44 — Information protection, information security, computer system security
Защита от несанкционированного доступа 12 — Protection from unauthorized access
Защита системы 12 — System security
Идентификация 285, 675, 697 — Identification
Избыточность кода 62, 439, 524 — Code surplus
Измерение 62 — Measurement
Имитовставка 589 — Data authentication code
Импорт информации — Information import
Импульсная модуляция 263 — Impulse modulation
Инициализация 653 — Initialization
Информации с ограниченным доступом 650 — Limited access information
Информационная система 13, 106 — Information system
Информационный поток 204, 354 — Traffic
Информационный сигнал 116 — Information signal
Информация 62 — Information
Исправление 342 — Correction
Исправление ошибок 342 — Error correction
Источник информации 35, 74 — Sours of information

Канал несанкционированного доступа 12	Channel of unauthorized Division
Канал передачи данных 343, 345	Data link
Канал связи 39, 87, 337, 347	Communication channel
Категория доступа 682	Security level
Ключ 580	Key
Код 73, 120, 486	Code
Кодер источника 340, 488	Coder of the source
Кодирование 282, 337	Coding; encoding
Кодирование источника сообщений 487	Encoding of source of reports
Кодирование со сжатием 487, 488	Compression coding
Кодовая избыточность 439, 502	Code redundancy
Коды без памяти 497, 515	Codes without memory
Коды с памятью 514	Codes with memory
Кольцевая сеть 410	Ring network
Коммутатор 413	Switch
Комплекс средств защиты 13, 45, 47	Trusted computing base
Комплексная система защиты информации 40	Complex system of information protection
Компрометация 539	Compromise
Компьютерная сеть 11	Computer network
Компьютерный вирус 666, 667	Computer virus
Контейнер 587, 626	Container
Конфиденциальность 12, 712	Confidence
Конфиденциальность информации 12	Information confidentiality
Концентратор 408, 409, 463, 479	Concentrator
Корректирующая способность кода 523, 524	Correcting ability of code
Криптоалгоритм 643	Crypto algorithm
Криптографическое преобразование 579	Crypto graphical transformation
Криптография 579	Cryptography
Критерии оценки защищенности 759	Security evaluation criteria
Линейные коды 531	Linear codes
Линия связи 15	Communication link
Локальная сеть 409, 444	Local network
Локальный доступ 651	Local access
Маршрутизатор 414	Router
Маршрутизация 419, 430	Routing
Матрица доступа 690	Access matrix
Межсетевой экран 697, 708	Firewall

Мера информации 77, 80 — Measure of information
Меры обеспечения безопасности 13 — Safeguards
Метка 565 — Label
Методы доступа 411 — Access methods
Методы защиты информации 45 — Information security methods
Механизмы защиты 19, 40 — Security mechanism
Модель Белла-Лападула 689 — Bell-Lapadula model
Модель нарушителя 760 — User violator model
Модель угроз 682 — Model of threats
Модификация 20, 607 — Modification
Модулятор 14, 250, 340 — Modulator
Мощность 60, 92, 106, 131, 149 — Power

Наблюдаемость — Accountability
Нарушитель 26 — User violator
Незашифрованный текст 585 — Clear text; plaintext
Непрерывный канал 356 — Continuous channel
Несанкционированный доступ 12, 648 — Unauthorized access
Обработка информации 24, 25 — Information processing
Оптимальный приемник 62 — Optimum receiver
Отказ 21 — Fault, failure
Отказ в обслуживании 654 — Denial of service
Отклонение 57, 145 — Deviation
Открытый ключ 583 — Public key
Открытый ключ шифрования 583 — Public-key encryption
Открытый текст 580 — Clear text, exposed text
Отношение сигнал/шум 480 — Signal-to-noise ratio
Отправитель 411 — sender
Отсчет 90, 91, 92 — Reading, counting
Оценка безопасности информации 729 — Information security evaluation
Оценка уязвимости 677 — Vulnerability assessment

Пароль 656, 660 — Password
Персональный идентификационный номер 643 — Personal identification number, pin
Плотность 87, 146, 147 — Density
Политика безопасности информации 689 — Information security policy
Полномочия 623, 655 — Privilege

Пользователь 11, 12	User
Помехи 15	Hindrances
Помехоустойчивость 17, 57	Noise immunity
Помехоустойчивый корректирующий код 522	Antigambling correcting code
Поток информации 354	Information flow
Право доступа 411	Access right
Право собственности 22, 629, 721	Property
Предел 24, 126	Boundary
Программная закладка 25, 659	Program bug
Проникновение 586, 659, 714	Penetration
Расстояние Хемминга 529	Hamming distance
Расшифровка данных 580, 581	Data decryption
Регистрация 11, 31	Audit, auditing
Режим доступа 722	Access routine
Рейтинг	Rating
Риск 293, 377, 761	Risk
Санкционированный доступ к информации 665	Authorized access to information
Сверточный двоичный код 556	Binary displacing code
Сетевая технология 410	Network technology
Сеть 388	Network
Сигнал 12, 14, 15	Signal
Сканирование портов 654, 657	Port scanning
Сообщение 11	Report, communication
Средства защиты 37, 671, 763	Protection facility
Стандарты 54	Standards
Стеганография 625	Steganography
Телекоммуникации 456	Telecommunications
Топология информационно-коммуникационной сети 479	Topology of information communication network
Трафик 411, 413	Traffic
Троянский конь 25	Trojan horse
Угроза 19, 20	Threat
Угроза безопасности 19, 20	Security threat
Удаленный доступ	Remote access

Удаленный контроль — Remote control
Управление безопасностью — Security management
Управление доступом 683 — Access control
Управление потоками 471 — Flow control
Управление риском 676 — Risk management
Уровень доступа 690 — Access level
Услуга безопасности 53, 54 — Security service
Установление связи 605 — Communications settings
Утечка информации 31, 32 — Information leakage
Уязвимость информации 19 — Vulnerability of information
Уязвимость системы 666 — System vulnerability

Фазовая модуляция 254 — Phase modulation
Функциональный профиль 757, 758 — Functionality profile

Хаб 449 — Hub
Хост 423, 428, 430 — Host
Хранение и поиск информации 11, 285, 722 — Information storage and retrieval
Хэш-функция 619 — Hash function

Целостность информации 12, 522 — Information integrity
Целостность системы 12, 46, 672 — System integrity
Циклический код 547 — Cyclic redundancy check
Цифровая подпись 606 — Digital signature
Цифровой фильтр 171 — Digital filter

Частотная модуляция 244, 256 — Frequency modulation

Шифр 610 — Cipher
Шифрование данных 63 — Data encryption
Шум 57, 60 — Noise

Электрический фильтр — Electric filter
Элементы защиты 673 — Elements of protection
Эллиптическая криптосистема 603 — Elliptic curve cryptosystem
Энергия 130, 131 — Energy
Эффективность защиты информации 40 — Efficiency of information protection

СПИСОК ЛИТЕРАТУРЫ

1. *Алфёров А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В.* Основы криптографии: Учеб. пособие. - М.: Гелиос АРВ, 2002. - 480 с.

2. *Бабак В.П.* Теоретичні основи захисту інформації: Підручник. - Книжкове вид-во НАУ, 2008. - 752 с.

3. *Бабак В.П., Белецкий А.Я., Гуржий А.Н.* Сигналы и спектры: Учебник. - К.: Кн. изд-во НАУ, 2005. – 520 с.

4. *Бабак В.П., Марченко Б.Г., Фриз М.Є.* Теорія ймовірностей, випадкові процеси та математична статистика: Підручник. - К.: Техніка, 2004. - 288 с.

5. *Бабак В.П.*, *Хандецький В.С., Шрюфер Е.* Обробка сигналів: Підручник. – К.: Либідь, 1999. - 392 с.

6. *Баскаков С.И.* Радиотехнические цепи и сигналы: Учебник. - М.: Высш. шк., 2000. - 462 с.

7. *Бузов Г.А., Калинин С.В., Кондратьев А.В.* Защита от утечки информации по техническим каналам: Учеб. пособие. - М.: Горячая линия - Телеком, 2005. - 416 с.

8. *Гашков С.Б. , Применко Э.А., Черепнев М.А.* Криптографические методы защиты информации. Учебное пособие. - М.: Academia, 2010. - 412 с.

9. *Герасименко В.А., Малюк А.А.* Основы защиты информации: Учеб. пособие. - М.: МГИФИ, 1997. - 538 с.

10. *Домарев В.В.* Безопасность информационных технологий. Системный подход. - К.: ООО «ТИД «ДС», 2004. - 992 с.

11. *Зегжда Д.П., Ивашко А.М.* Основы безопасности информационных систем. - М.: Горячая линия - Телеком, 2000. - 452 с.

12. *Інформаційна* безпека та сучасні мережеві технології: Англо-українсько-російський словник термінів / В. П. Бабак, О. Г. Корченко. - К.: НАУ, 2003. - 670 с.

13. *Клейменов С.А., Мельников В.П., Петраков А.М.* Информационная безопасность и защита информации. Учебное пособие. - М.: Academia, 2009. - 336 с.

14. *Конахович Г.Ф., Бабак В.П.*, *Фисенко В.М.* Специальный радио-мониторинг. - К.: МК Пресс, 2007. - 384 с.

15. *Конахович Г.Ф., Пузыренко А.Ю.* Компьютерная стеганография: Теория и практика. - К.: МК - Пресс, 2006. - 288 с.

16. *Корт С. С.* Теоретические основы защиты информации. - М.: Гелиос АРВ, 2004. - 240 с.

17. *Кузьмин И.В., Кедрус В.А.* Основы теории информации и кодирования: Учебник. - К.: Вища шк., 1986. - 238 с.

18. *Куприянов А.И., Сахаров А.В., Шевцов В.А.* Основы защиты информации. Учеб. пособие. - М.: Academia, 2008. - 256 с.

19. *Малюк А.А.* Информационная безопасность: концептуальные и методологические основы защиты информации: Учеб. пособие. - М.: Горячая линия - Телеком, 2004. - 280 с.

20. *Методы* и средства защиты информации. В 2-х томах / Ленков С.В., Перегудов Д.А., Хорошко В.А. / под ред. В. А. Хорошко. - К.: Арий, 2008. - Т. 1. Несанкционированное получение информации. - 464 с. Т. 2. Информационная безопасность. - 344 с.

21. *Мещеряков Р.В., Скрыль С.В., Шелупанов А.А. и др.* Техническая защита информации: Учебник. - М.: Горячая линия -Телеком, 2009. - 616 с.

22. *Орнатский П.П.* Теоретические основы информационно-измерительной техники: Учебник. - К.: Вища шк., 1983. - 455 с.

23. *Петраков А.П.* Основы практической защиты информации: Учеб. пособие. - М.: Радио и связь, 2000. - 368 с.

24. *Рабинер Л., Гоулд Б.* Теория и применение цифровой обработки сигналов / Пер. с англ. - М.: МИР, 1978. - 848 с.

25. Русско-украинско-английский словарь терминов по информационным технологиям / В. П. Бабак, О. Г. Байбуз, А. П. Приставка. - К.: НАУ, 2006. - 252 с.

27. *Скляр Б.* Цифровая связь. Теоретические основы и практическое применение / Пер. с англ. - М.: Изд. дом Вильямс, 2004. - 1104 с.

28. *Соколов А.В., Шаньгин В.Ф.* Защита информации в распределенных корпоративних сетях и системах. - М.: ДМК Пресс, 2002. - 656 с.

29. *Темников Ф.Е., Афонин В.А., Дмитриев В.И.* Теоретические основы информационной техники: Учеб. пособие. - М.: Энергия, 1979. - 512 с.

30. *Хоффман Л. Дж.* Современные методы защиты информации / Пер. с англ. - М.: Сов. радио, 1980. - 480 с.

31. *Шаньгин В.Н.* Защита компьютерной информации. - М.: ДМК Пресс, 2008. - 362 с.

32. *Ярочкин В.И.* Информационная безопасность: Учебник. - М.: Академпроект: Трикста, 2005. - 544 с.

33. *Bellare M., Rogaway P.* Introduction to Modern Cryptography. - University of California, 2005.

34. *Blackley J., Peltier J.* Information Security Fundamentals. – Peltier&Associates, Michigan, USA, 2004.

35. *Herold R., Robers M.* Encyclopedia of Information Aussurance. – Indiana, USA, 2010.

36. *Layton T.* Information Security: Design, Implementation, Measurement and Compliance. - Missouri, USA, 2006.

37. *Lin S., Costello D.J.* Error Control Coding: Fundamentals and Applications. - Prentice-Hall, Inc., Englewood Cliffs, N. J., 2003.

38. *Pritchard W.L., Sciulli J.A.* Satellite Communication Systems Engineering. - Prentice-Hall, N. J., 2006.

39. *Schneier B.* Applied Cryptography. - John Wiley & Sons, New Jork, 2008.

40. Stackpole B., Oksendahl E. Security Strategy: From Reguirements to Reality. - Washington, USA, 2010.

41. *Stallings W.* Cryptography and Network Security. - Prentice Hall, Upper Saddle River, NL, 2008.

42. *Stinson D.* Cryptography Theory and Practice. - CRC Press, Boca Raton, FL, 2005.

43. *Tipton H., Krause M.* Information Security Management Handbook. - California, USA, 2010.

У підручнику викладено основні поняття та методи захисту інформації, що базуються на математичному апараті перетворення та дослідження інформаційних сигналів, технологіях вимірювання, передачі та обробки інформації, сигналів і даних, на завадостійкому кодуванні, використанні сучасних інформаційних каналів передачі інформації, на алгоритмах шифрування і дешифрування, стегано- та криптографії, цифрового підпису тощо. Поряд з алгоритмічними та технічними методами розглянуто захист конфіденційної та комерційної інформації, захист від несанкціонованого доступу, захист інтелектуальної власності, законодавче забезпечення захисту інформації, а також міжнародні стандарти у сфері захисту інформації.

Для студентів технічних спеціальностей вищих навчальних закладів, аспірантів, наукових та інженерно-технічних працівників, зайнятих у сфері захисту інформації.

Навчальне видання

Бабак Віталій Павлович, **Ключников** Олександр Олександрович

ТЕОРЕТИЧНІ ОСНОВИ ЗАХИСТУ ІНФОРМАЦІЇ

Підручник

(Російською мовою)

Формат 70×100/16. Умов. друк. арк. 63. Тираж 500 пр. Зам. № 12-15.

Інститут проблем безпеки АЕС НАН України
Київська обл., 07270, м. Чорнобиль, вул. Кірова, 36-а
Свідоцтво суб'єкта видавничої справи ДК № 2114 від 25.02.2005 р.

Друк. ЗАТ "ВІПОЛ"
03151, Київ, вул. Волинська, 60.
Свідоцтво суб'єкта видавничої справи ДК № 752 від 27.12.2001 р.