А. А. Харкевич

ТЕОРИЯ ИНФОРМАЦИИ

ОПОЗНАНИЕ ОБРАЗОВ

А. А. Харкевич ИЗБРАННЫЕ ТРУДЫ 3

АКАДЕМИЯ НАУК СССР

ИНСТИТУТ
ПРОБЛЕМ ПЕРЕДАЧИ
ИНФОРМАЦИИ

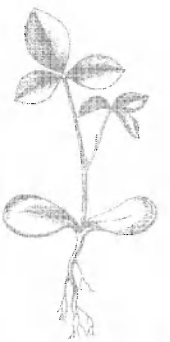
А. А. Харкевич

ИЗБРАННЫЕ ТРУДЫ

в трех томах

3

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА
1973



А. А. Харкевич

ТЕОРИЯ ИНФОРМАЦИИ

ОПОЗНАНИЕ ОБРАЗОВ

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА
1973

УДК 62-50 : 621.391 + 621.391.19

Теория информации. Опознание образов. Харкевич А. А. Избранные труды в трех томах. Т. III. М., «Наука», 1973 г., стр. 524.

В третьем томе собраны работы, посвященные теории информации и опознанию образов.

Теоретико-информационные исследования связаны главным образом с выделениями сигналов из смеси сигнал-помеха, построением помехоустойчивых кодов, передачей некоторых видов сообщений.

Работы по опознанию образов связаны как с общими вопросами (постановка задачи опознания, выбор признаков), так и с решением прикладных задач (построение читающих машин). Сюда же вошли работы философского характера: о ценности информации, о механизме творческого процесса.

Таблиц 42. Иллюстраций 291. Библиогр. 119 назв.

Х $\frac{03314-0213}{042\,(02)-73}$ 1130-72

ОБНАРУЖЕНИЕ СЛАБЫХ СИГНАЛОВ

1. Прием тех или иных сигналов при наличии помех возможен, если сигнал больше помех. Условия приема можно характеризовать отношением средней мощности сигнала к средней мощности помех. Условимся измерять это отношение логарифмической мерой и называть *превышением* сигнала над помехой.

Когда превышение приближается к нулю или даже становится отрицательным (т. е. когда помехи по мощности превосходят сигнал), прием обычными средствами становится невозможным.

2. Существует по меньшей мере два специальных метода обнаружения слабых сигналов, т. е. приема сигналов в условиях отрицательного превышения: метод накопления и корреляции. Ниже излагаются сущность и теоретические основания обоих методов.

3. Общая идея метода накопления известна: она состоит в том, что для повышения надежности связи в неблагоприятных условиях сигнал несколько раз повторяется. К этому методу мы постоянно бессознательно прибегаем, повторяя и переспрашивая при плохой слышимости во время разговора по телефону. В телеграфии подобный же прием применялся давно в форме так называемой системы Вердана. В последнее время метод накопления путем повторения сигнала был использован З. Баем в его опытах по радиолокации Луны [1, 2]. Этот же метод был применен и описан К. В. Владимирским [3].

Не входя в технические детали, опишем метод накопления в его современной форме. Суть дела заключается в том, что смесь *периодического* сигнала и шума нарезается на куски, длительность которых равна периоду сигнала, и эти куски накладываются друг на друга. Эти операции выполняются схемой рис. 1, действие которой состоит в том, что синхронизированный с сигналом коммутатор поочередно подключает на вход устройства один из нескольких накопителей.

Суммирование в накопителях происходит для сигнала и для шума по разным законам; сигнал когерентен, а шум некогерентен. Поэтому накопленное значение энергии сигнала пропорционально n^2, а накопленное значение энергии шума пропорционально n, если через n обозначить число повторений. Следовательно, отно-

шение мощностей накопленных сигнала и помехи растет как $n^2/n=n$.

Остановимся на некоторых подробностях. Дело в том, что мы описали эффект накопления в терминах с р е д н и х мощностей. Это само по себе не вызывает возражений, но не следует забывать, что сумма отрезков шума, попадающих в данный накопитель, есть случайная величина, подверженная флюктуациям. Нужно оценить возможность появления больших флюктуаций, могущих исказить результат накопления. Такую оценку можно сделать, только опираясь на теорию вероятностей.

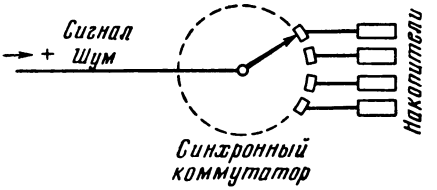


Рис. 1

Поставим прежде всего задачу. При работе схемы рис. 1 в каждый накопитель через равные промежутки времени (равные периоду сигнала и периоду обращения коммутатора) попадает случайное значение шума, имеющееся в момент, когда щетка коммутатора вступает на данную ламель. Возможно, что мы получим значение, усредненное за время пребывания щетки на данной ламели; это, однако, не играет принципиальной роли, так как получаемое накопителем значение шума и при таких условиях остается случайной величиной. Обозначим эту случайную величину через ξ_k и составим среднее арифметическое n слагаемых ξ_k

$$Y_n=\frac{1}{n}\sum_{k=1}^{n}\xi_k.$$

Эта величина при малом числе слагаемых сильно флюктуирует около постоянного среднего значения ξ, с увеличением же числа слагаемых вероятность больших флюктуаций убывает.

Обозначим теперь значение сигнала через a. Среднее арифметическое n сигналов есть, очевидно, a. Теперь следует сделать вероятностное сравнение средних величин накопленных шума и сигнала. Это приводит к следующей постановке вопроса: какова вероятность того, что разность между Y_n и ξ не превзойдет по абсолютной величине заданной величины a? Постановка вопроса пояснена графически на рис. 2.

Ответ на поставленный вопрос дает известное неравенство Чебышева

$$P\{|Y_n-\xi|<a\}>1-\frac{D\{\xi\}}{na^2}. \qquad (1)$$

В формуле (1) $D(\xi)$ означает дисперсию, т. е. средний квадрат величины ξ. Обозначим эту величину через σ^2 и заметим, что она непосредственно выражает среднюю мощность шума, тогда как a^2 выражает мощность сигнала. Число слагаемых n есть число повторений. Разрешая (1) относительно n, получим

$$n < \frac{1}{1-P} \cdot \frac{\sigma^2}{a^2}. \qquad (2)$$

Таким образом, верхний предел требуемого числа повторений определяется задаваемой нами вероятностью и исход-

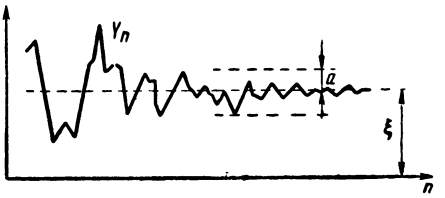


Рис. 2

ным соотношением мощностей сигнала и помехи, т. е. исходным превышением.

К сожалению, мы получили для n оценку в форме неравенства, ограничивающего n сверху. Оценки снизу теория вероятностей не дает; однако она дает больше: она позволяет выразить n приближенным равенством. Это равенство основывается на теореме Ляпунова, утверждающей, что при выполнении некоторого общего условия (условия Линдеберга—Феллера) распределение для суммы случайных величин при увеличении числа слагаемых сходится к нормальному. Практически важно, что оно сходится довольно быстро. Вытекающее из теоремы Ляпунова приближенное равенство можно в наших обозначениях записать так:

$$P\{|Y_n - \xi| < a\} \approx \frac{2}{\sqrt{2\pi}} \int_0^{\frac{a}{\sigma}\sqrt{n}} e^{-x^2/2} dx = 2\Phi\left(\frac{a}{\sigma}\sqrt{n}\right). \qquad (3)$$

Здесь через Ф обозначена функция Гаусса. Задавшись P и a/σ, можно при помощи таблиц найти n.

Интересно сравнить численные результаты, получаемые с помощью приведенных соотношений. Положим, что исходное превышение равно 20 *дб*, т. е. $\sigma^2/a^2 = 100$. Чтобы довести превышение до нуля по средним мощностям, требуется 100 повторений. Положим, что одинаковый порядок величин шума и сигнала определяется в вероятностных формулах вероятностью 0,5. Тогда по формуле (2) найдем $n < \frac{1}{0{,}5} \cdot 100 = 200$; по формуле же (3) получим $\Phi\left(\frac{a}{\sigma}\sqrt{n}\right) = 0{,}25$; $\frac{a}{\sigma}\sqrt{n} = 0{,}67$; $n = 45$.

Задавшись большим значением вероятности во избежание искажения принимаемого сигнала флюктуациями шумовой суммы, мы увеличим, естественно, требуемое число повторений. Общие теоретические основания метода накопления, таким образом, установлены.

4. Второй метод обнаружения слабых сигналов основан на применении корреляционного анализа [4]. Функция корреляции двух стационарных случайных процессов $\xi(t)$ и $\eta(t)$ определяется как

$$R_{\xi\eta} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t)\,\eta(t-\tau)\,dt \tag{4}$$

и является, таким образом, функцией временного сдвига τ. В частности,

$$R_{\xi\xi} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t)\,\xi(t-\tau)\,dt \tag{5}$$

называется функцией автокорреляции для процесса $\xi(t)$.

Положим, что мы располагаем аппаратурой, выполняющей операцию (5), и подаем на вход этой аппаратуры смесь сигнала $x(t)$ и шума $\xi(t)$. На выходе мы получим

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T [\xi(t) + x(t)]\,[\xi(t-\tau) + x(t-\tau)]dt =$$

$$= \lim_{T\to\infty} \frac{1}{T} \left\{ \int_0^T \xi(t)\xi(t-\tau)dt + \int_0^T x(t)\,x(t-\tau)\,dt + \right.$$

$$\left. + \int_0^T \xi(t)\,x(t-\tau)\,dt + \int_0^T x(t)\,\xi(t-\tau)\,dt \right\} =$$

$$= R_{\xi\xi} + R_{xx} + R_{\xi x} + R_{x\xi}.$$

Сигнал и шум — взаимно независимые процессы; корреляция между ними отсутствует и функции взаимной корреляции — два последних слагаемых — равны нулю. Функция $R_{\xi\xi}$ — функция автокорреляции шума — является основной вероятностной характеристикой шума как случайного процесса. Общее ее свойство состоит в том, что она имеет наибольшее значение (равное дисперсии) при $\tau=0$ и убывает (может быть, немонотонно) с возрастанием τ, стремясь к нулю. Интервал, на котором функция корреляции имеет еще заметные значения, называется временем корреляции. Время корреляции — важная числовая характеристика случайного процесса, тесно связанная с шириной его спектра. Для оценки порядка величин можно пользоваться соотношением

$f_0\tau_0=1$, где τ_0 — время корреляции; f_0 — ширина спектра. Что касается функции корреляции сигнала, то нетрудно показать, что для периодической функции $x(t)$ функция корреляции есть также периодическая функция (аргумента τ) с тем же периодом.

Принимая все это во внимание, мы приходим к следующему заключению: на выходе нашей аппаратуры мы получим сумму двух функций $R_{\xi\xi}(\tau)$ и $R_{xx}(\tau)$. Первая будет убывать, вторая же имеет периодический характер. Поэтому для достаточно больших τ функция $R_{\xi\xi}$ исчезнет, и мы будем наблюдать лишь функцию R_{xx}, отображающую сигнал. Для оценки влияния исходного превышения удобно ввести нормированные (путем деления на соответствующие дисперсии) функции корреляции $r(\tau)$, удовлетворяющие условию $r(0)=1$. Тогда на выходе устройства будем иметь

$$y(\tau)=R_{\xi\xi}(\tau)+R_{xx}(\tau)=\sigma^2 r_{\xi\xi}(\tau)+a^2 r_{xx}(\tau),$$

и теперь ясно видно, что первое слагаемое перестанет играть роль при тем бóльших значениях τ, чем больше исходное отношение мощностей помехи и сигнала.

Таким образом показана, хотя и в общих чертах, возможность обнаружения под помехой сколь угодно слабого периодического сигнала корреляционным методом.

5. Метод накопления был применен на практике для приема очень слабых локационных сигналов, отраженных от Луны.

Интересно обсудить возможность использования этого метода для обычной связи, например радиотелеграфной.

В этом случае можно представить себе систему связи в следующем виде: передатчик посылает некоторую кодовую комбинацию, например пятизначную комбинацию кода Бодо. Эта комбинация непрерывно повторяется, пока на приемнике происходит процесс накопления. Процесс этот состоит в том, что синхронный коммутатор, выполняющий функции распределителя, посылает каждый элемент кодовой комбинации в соответствующий накопитель до тех пор, пока данная комбинация не будет уверенно прочитана. После этого передатчик переходит к следующей комбинации. Таким образом, возможна телеграфная работа, разумеется, в соответственно замедленном темпе. Осуществление описанной схемы наталкивается на существенное затруднение. Дело в том, что схема требует синхронной и синфазной работы распределителей передатчика и приемника. Между тем передача и прием синхронизирующих импульсов возможны только тем же методом, каким принимаются рабочие импульсы, т. е. методом накопления. Это обстоятельство либо потребует разработки соответствующей автоподстройки частоты и фазы, либо, быть может, удастся, сохранив идею метода накопления, придумать асинхронный вариант его осуществления.

Прием по методу корреляции свободен от этого недостатка — он не требует синхронизации. Зато он обладает своими, если не не-

достатками, то, во всяком случае, специфическими особенностями. Одна из них состоит в том, что функция корреляции не зависит от фазовых соотношений. Поэтому передача обычным телеграфным кодом с амплитудной модуляцией исключена.

6. В обоих описанных методах приема возможность приема слабого сигнала, покрытого помехой, достигается ценой увеличения времени, потребного на передачу данного сообщения. Это обстоятельство вытекает из общего представления о «несжимаемости» сигнала. Здесь имеется в виду то, что «объем» сигнала, т. е. произведение из длительности сигнала на ширину его спектра и на динамический диапазон, есть величина постоянная. «Несжимаемость» сигнала означает, что сокращение какого-либо измерения сигнала должно вызывать соответствующее увеличение другого (или других) измерения. В нашем случае сокращение динамического диапазона (т. е. уменьшение превышения сигнала над помехой) компенсируется соответствующим увеличением длительности передачи. Аналогично обстоит дело и в других случаях. Так, например, в принципе возможно выделить сколь угодно слабый периодический сигнал из смеси его с помехой посредством настроенных на частоты сигнала резонаторов с достаточно высокой добротностью; чем острее избирательность резонатора, тем больше соотношение между мощностями сигнала и шума в данной полосе. Однако процесс раскачивания резонатора требует времени; в сущности резонатор является также накопителем энергии в ее колебательной форме.

Итак, рассмотренные методы являются с изложенной точки зрения примером преобразования — «деформации» — сигнала. Желательность такого рода деформации, т. е. в нашем случае обмена динамического диапазона на время, должна устанавливаться на основе широкой оценки данной ситуации.

Литература

1. *Z. Bay.* Reflection of the microwaye from the Moon.—Hung. Acta Phys., 1947, v. 1, N 1.
2. *В. С. Вавилов.* Опыты по радиолокации Луны. — УФН, 1949, т. 39, № 3.
3. *К. В. Владимирский.* О синхронном фильтре. — ЖЭТФ, 1951, т. 21, № 3.
4. *Y. W. Lee, T. P. Chatham, J. B. Wiesner.* Detection of periodic signals in noise. — Proc. IRE, 1950, v. 38, N 10.

ОЧЕРКИ ОБЩЕЙ ТЕОРИИ СВЯЗИ

ПРЕДИСЛОВИЕ

Книга названа «очерками» потому, что развитие общей теории связи нельзя считать законченным. Это кладет и на изложение отпечаток некоторой отрывочности. Нерешенных вопросов еще очень много; читатель найдет здесь целый ряд более или менее ясно сформулированных тем дальнейших исследований.

Книга написана коротко. Я избегал второстепенных, но громоздких в изложении подробностей, чтобы не заслонить сути дела.

Теория сравнительно нова; новые понятия требуют и новой терминологии. Я не остановился перед введением целого ряда терминов, вполне понимая свою ответственность. Не приводя здесь доводов в защиту того или иного термина, скажу лишь, что я старался во всех случаях подобрать по возможности простые и выразительные русские слова.

Приношу сердечную признательность моим рецензенту и редактору: Н. А. Железнову и М. Д. Карасеву.

Москва, май 1954 А. А. Харкевич

ВВЕДЕНИЕ

Общая теория связи как самостоятельная теоретическая дисциплина возникла и развилась в недавнее время. Ее появление обусловлено, с одной стороны, наличием большого количества накопленных и ожидающих обобщения частных знаний в области теории и техники связи, с другой — необходимостью решать все более трудные проблемы связи, когда целесообразные пути да и самая возможность решения проблемы заранее совершенно неясны.

Сила общей теории связи, как и всякой обобщающей теоретической дисциплины, состоит в том, что она ставит и решает основные вопросы в общем виде, позволяя не только обозреть все ранее сделанное в области техники связи, но и предуказать заслуживающие внимания направления дальнейшего развития. С другой стороны, теория в ряде случаев достаточно ясно показывает, чего можно достичь, а чего нет.

Приступая к изложению предмета, нужно прежде всего очертить его границы. Общей теорией связи сейчас много занимаются, она возбуждает довольно широкий интерес. Общую теорию связи называют часто теорией информации, а также статистической теорией. Действительно, для новой теории характерно введение в рассмотрение статистических аспектов связи. Однако общая теория изучает также ряд вопросов, не имеющих статистического характера. Необходимо внести определенность в характеристику предмета; в дальнейшем общая теория связи рассматривается как *теоретическая основа техники связи*. Этим и определяются назначение и проблематика общей теории связи. Связь определяется как передача сообщений (подразумевается связь, осуществляемая при помощи тех или иных сигналов; в первую очередь электрическая связь).

Все задачи, стоящие перед техникой связи, сводятся к двум основным проблемам. Первая основная проблема — проблема эффективности связи. Проблема эта состоит в том, чтобы передать наибольшее количество сообщений наиболее экономным способом. Теория показывает, в каком специальном смысле истолковывается применительно к связи «экономность»; теория устанавливает количественную меру для сообщения как объекта связи; теория позволяет сравнивать между собой различные системы связи по эффективности; теория указывает резервы, за счет которых может быть осуществлено дальнейшее повышение эффективности.

Вторая основная проблема — проблема надежности[1] связи. Вследствие влияния помех принятое сообщение никогда не тождественно переданному. Надежность есть мера соответствия принятого сообщения переданному. При данных условиях связи, т. е. при заданной помехе, надежность зависит от свойств системы, от ее способности противостоять вредному действию помех. Это свойство системы называют *помехоустойчивостью*. Теория позволяет сравнить между собой различные системы связи по помехоустойчивости; теория указывает общие пути повышения помехоустойчивости.

Проблема эффективности и проблема надежности — вот основные проблемы связи. Постановка и разрешение этих проблем составляют содержание общей теории связи. Следует попутно отметить, что требования эффективности и надежности в известном смысле противоречивы. Подобного рода противоречия часты в технике; теория помогает отысканию приемлемого компромисса.

Нужно заметить, что, несмотря на быстрые темпы развития, общая теория связи не получила еще завершения в своих основных построениях. Обращает на себя внимание, в частности, отсутствие до настоящего времени системы основных законов типа законов сохранения характерных для многих сложившихся отраслей знания. Наличие подобного рода законов, специфичных для связи, интуитивно ощущается. Однако эти законы еще не найдены и не сформулированы.

От общей теории ожидают часто больше, чем она может дать; полагают, что теория содержит ответы на любые практические вопросы. Это, конечно, не так. Общая теория устанавливает определенный строй идей и предуказывает пути практической деятельности, но не ее результаты. Иными словами, для того чтобы теория могла принести пользу, важно, чтобы ею овладели инженеры, непосредственно занятые новыми разработками в области связи. Для этого необходимо, очевидно, распространение теоретических сведений. С этой целью и написана настоящая книга.

---

[1] В соответствии с более поздним предложением А. А. Харкевича эту характеристику в настоящее время называют верностью передачи.

Глава 1

ОСНОВНЫЕ ПОНЯТИЯ

§ 1. Связь

Связь состоит в передаче сообщений от отправителя к получателю. Во всем дальнейшем подразумевается лишь один вид связи — электрическая связь, при которой передача сообщения осуществляется электрическими сигналами, передаваемыми по проводам или без проводов — в виде электромагнитных волн.

Система связи есть система передачи. Но здесь нужно сразу же установить, что именно является объектом передачи, что транспортируется от отправителя к получателю. В системе электропередачи объектом передачи является энергия. Она должна быть передана с минимальными потерями потребителю. В системе же связи объектом передачи является не энергия, а сообщение. Конечно, передача сообщения сопровождается (и обусловливается) передачей энергии, но не в передаче энергии состоит назначение системы связи. Известно, что энергетический коэффициент полезного действия системы связи, например радиосвязи, исчезающе мал. Следовательно, для оценки эффективности системы связи нужен особый, специфический для связи критерий. Для установления же такого критерия нужно определить, что такое сообщение, и дать ему количественную меру.

Такой мерой может служить *количество сведений*, содержащееся в сообщении. Определением этой величины мы займемся позднее, а пока что попытаемся лишь дать о ней общее представление.

Рассмотрим различные виды электрической связи, начиная со старейшего — с телеграфии. В этом случае сообщение представляет собой некоторый текст. Мерой количества сведений, содержащегося в сообщении, может служить количество слов. Эта мера издавна и принята в телеграфии как наиболее естественная и удобная. Но она, к сожалению, не универсальна. Обратимся к телефонии. Здесь количество сведений не определяется только количеством слов: получатель черпает сведения также из интонации и ритма живой речи. Также и фототелеграф: он передает не только слова, но и почерк, и сопровождающие текст рисунки. Наконец, если мы перейдем к такому виду связи, как телевидение, то станет ясно, что количество сведений, содержащееся в телевизионном изображении, нужно определять как-то совсем по-другому. А между тем и текст телеграммы, и заполненный бланк фототелеграмм, и живая речь, и телевизионное изображение — все это *сообщения*, если мы кратко определим сообщение как то, что подлежит передаче.

Одно из основных положений общей теории связи как раз и состоит в том, что количество сведений, содержащееся в данном

сообщении, каково бы оно ни было, измеряется некоторым универсальным образом. Если установление такой универсальной меры возможно, то возможно и построение общего критерия эффективности системы связи. Одним из показателей эффективности может служить количество сведений, могущее быть переданным по системе в единицу времени. Определенную таким образом величину можно назвать *пропускной способностью* системы связи.

Нужно, однако, иметь в виду, что всякая реальная система связи работает при наличии разного рода *помех*, которые могут исказить передаваемое сообщение или даже сделать невозможным его прием. Поэтому к системе связи кроме требования эффективности следует предъявлять еще и требование *надежности связи*. Надежность связи есть мера достоверности принятых сообщений, т. е. мера соответствия принятых сообщений переданным. Надежность связи зависит от условий работы системы связи, определяемых характером и интенсивностью помех, условиями распространения сигналов, технической исправностью аппаратуры, а также от собственных свойств системы связи, определяемых способом передачи сигналов, т. е. видом модуляции и строением кода. Способность системы противостоять вредному влиянию помех, обусловленная ее собственными свойствами, называется *помехоустойчивостью*. В дальнейшем мы будем интересоваться непосредственно только помехоустойчивостью систем связи, оставляя в стороне все остальные факторы, влияющие на надежность связи.

Проблема связи в столь общей постановке приводит, естественно, к весьма общего характера выводам, находящим себе применение за пределами собственно связи; эти выводы полностью приложимы, в частности, к телеуправлению и к телеизмерениям. Мы будем, однако, в дальнейшем интересоваться применениями теории только к электрической связи в обычном понимании этого термина.

Высказанные соображения уже подводят нас к построению общей системы связи: мы будем входить постепенно во все большие подробности.

§ 2. Система связи

Связь осуществляется, вообще, следующим образом. Отправитель подает сообщение, которое поступает на передатчик. Здесь сообщение должно быть превращено в электрический *сигнал*. Сигнал не есть сообщение, но между сообщением и сигналом должно быть однозначное соответствие для того, чтобы на приемном конце сигнал мог быть снова превращен в сообщение, соответствующее поданному. Это обратное превращение выполняется приемником. Так, например, текст телеграммы (сообщение) превращается телеграфным аппаратом на передающем конце (передатчик) в определенную последовательность импульсов электри-

ческого тока (сигнал). Эти импульсы воздействуют на буквопечатающий аппарат на приемном конце (приемник), который восстанавливает сообщение в форме текста.

Естественно, что принятое сообщение должно быть в идеальном случае тождественным поданному. В действительности имеются искажения; степень этих искажений определяет надежность связи.

Передатчик и приемник связаны между собой *линией связи*. Линия может представлять собой, например, пару проводов или некоторую ограниченную зону пространства, по которой электро-

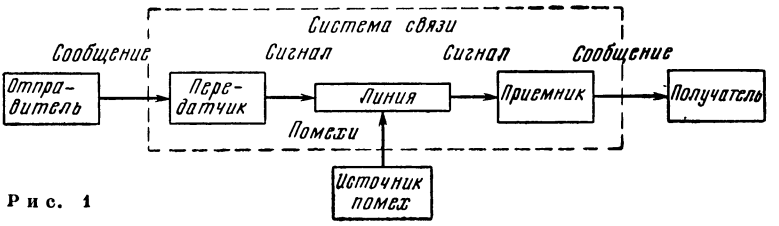


Рис. 1

магнитные волны распространяются от передатчика к приемнику («луч») в случае передачи без проводов.

В приемник попадает не только посланный передатчиком сигнал, но и помеха. Помехи могут быть как внешние, так и внутренние (например, шум электронных ламп). Однако для общего рассмотрения удобнее объединить все источники помех в один.

На основании всего сказанного можно представить себе систему связи, как показано схематически на рис. 1. Отправитель и получатель не включены в систему связи. Таким образом, система связи определена как совокупность технических средств (передатчика, линии и приемника) для передачи сообщения.

§ 3. Линия и канал связи

Для техники связи, начиная с самых первых ее шагов, характерно естественное стремление к увеличению пропускной способности системы связи. Стоимость сооружения системы связи велика; достаточно представить себе проводную (воздушную или кабельную) линию связи протяженностью в несколько тысяч километров. Поэтому возникает тенденция к «уплотнению» линии связи, к более эффективному ее использованию. Одна из возможностей состоит в одновременной передаче по линии нескольких сообщений. В этом случае каждое сообщение следует по своему *каналу связи*. Такая связь называется *многоканальной*. Сигналы всех каналов смешиваются на передающем конце и поступают в линию. На приемном конце сигналы снова разделяются и пре-

образуются в независимые сообщения. Таким образом, каналом связи мы называем совокупность технических устройств, обеспечивающую независимую передачу данного сообщения. Число каналов на одной физической линии может быть очень велико и измеряться десятками и даже сотнями.

Схематически система многоканальной связи представлена на рис. 2. Для того чтобы из смеси сигналов, поступающих с линии, выделить сигнал некоторого канала, необходимо произвести операцию *разделения* (селекции). Поэтому на схеме показаны подключенные к выходному концу линии разделители (селекторы),

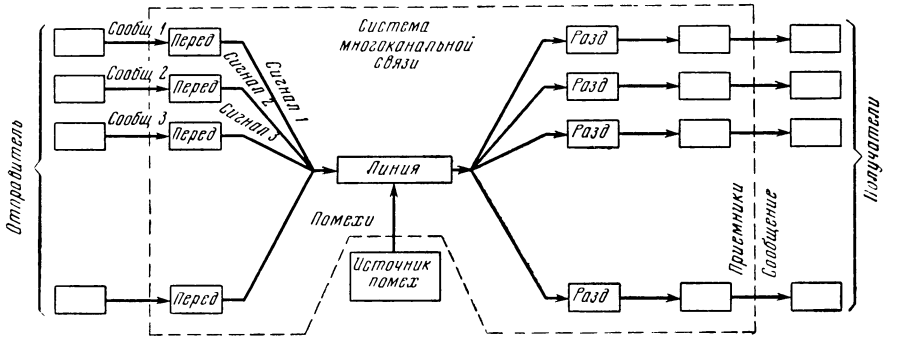


Рис. 2

за которыми уже включены приемники. Для того чтобы разделение было возможно, необходимо, чтобы сигналы различных каналов отличались друг от друга по некоторому определенному физическому признаку и чтобы разделители способны были, основываясь на этом признаке, реагировать на требуемый сигнал и не реагировать на все остальные. Условия, при которых достигается такой результат, выясняет общая теория разделения (селекции), основы которой будут изложены ниже. Пока заметим лишь, что возможны различные методы разделения, из которых только часть использована современной техникой.

По поводу многоканальной системы связи следует еще заметить, что в ней возникает новый вид помех; помехи от соседних каналов обусловлены несовершенством разделения. Влияние сигналов соседних каналов мы относим к помехам, определяя помехи как все то, что не есть желаемый сигнал.

§ 4. Сообщение и сигнал

Предыдущие параграфы дают краткую характеристику обстановки, в которой происходит основное действие — передача сообщений. Теперь мы займемся объектом передачи — сообщением — и его превращением в сигнал.

Выше сообщение было определено как то, что подлежит передаче, а сигнал — как некоторое электрическое возмущение, отображающее сообщение.

Превращение сообщения в сигнал состоит из трех операций, которые могут быть независимыми или совмещенными. Эти три операции следующие: преобразование, кодирование и модуляция.

Под преобразованием понимается просто перевод неэлектрических величин, определяющих первоначальное сообщение, в электрические. Так, например, в телефонии переменное звуковое давление, определяющее звук речи, преобразовывается в соответствующим образом изменяющийся электрический ток посредством микрофона. При передаче изображения оно ощупывается (сканируется) узким лучом света; отраженный свет попадает на фотоэлемент, преобразующий колебания светового потока в соответствующие колебания электрического тока. В обоих этих примерах микрофон и фотоэлемент являются преобразователями соответствующих переменных величин (звукового давления и светового потока) в электрическую величину (ток или напряжение). К такого рода преобразователям предъявляется обычно общего характера требование линейности, т. е. пропорциональности между воздействием и откликом [1]. Мы будем называть преобразованное в форму электрического возмущения сообщение прóсто сообщением (во всех тех случаях, когда в процессе образования сигнала участвует отдельная операция преобразования).

Под кодированием понимается построение сигнала по некоторому определенному принципу, имеющему, как мы увидим, простое математическое выражение.

И, наконец, под модуляцией понимается воздействие на некоторый параметр электрического тока — постоянного или переменного, — в результате чего в изменениях этого параметра оказывается, так сказать, заложенным передаваемый сигнал.

Можно сказать, что кодирование определяет математическую сторону, а модуляция — физическую сторону процесса превращения сообщения в сигнал. Оба эти вопроса — кодирование и модуляция — довольно обширны. Некоторые подробности даны в последующих параграфах.

§ 5. Дискретное сообщение

Начнем с простейшего случая сообщения, представляющего собой текст или, в шифрованном виде, последовательность чисел. Именно таковы сообщения в телеграфии. Мы назовем такое сообщение дискретным, так как оно состоит из отдельных элементов (букв, цифр).

Необходимо прежде всего освоиться с представлением о том,

[1] Впрочем, иногда это требование излишне, как, например, для фототелеграфа предназначенного для передачи черно-белого изображения.

что передача дискретного сообщения может быть сведена всегда к передаче последовательности чисел. В самом деле, передавая некоторое слово по буквам, мы передаем, конечно, не самые буквы, а некоторые символы, которые могут, например, рассматриваться как порядковые номера букв или вообще как некоторые, условно приписанные им числа. К этому и сводится любая телеграфная азбука, т. е. телеграфный код.

Здесь кажется на первый взгляд неочевидной необходимость связывать тот или иной символ телеграфного кода с числом. Как будто можно было бы просто установить, что в коде Морзе букве А соответствует символ точка—тире, и не говорить вовсе о числах. Но понятие числа необходимо нам для того, чтобы уяснить себе принцип построения кода и обозреть возможные разновидности кодов.

§ 6. Системы счисления

Приступая к изучению кодов, нужно прежде всего отказаться от представления об исключительности привычной нам десятичной системы счисления и вспомнить, что возможно множество других систем счисления и представления чисел. Десятичное счисление, очевидно, возникло только вследствие наличия у нас на руках десяти пальцев, которые использовались в качестве счетов. Никаких преимуществ эта система перед другими не имеет, а с точки зрения кода, как мы увидим, больший интерес представляют другие системы счисления.

Сущность десятичной системы состоит в том, что, располагая десятью цифрами (от 0 до 9), мы можем записать одной цифрой любое число в пределах первого десятка. Десять — уже двузначное число, которое мы записываем единицей в разряде десятков и нулем в разряде единиц.

Но по этому же принципу можно построить систему из любого числа цифр, например из пяти (0, 1, 2, 3 и 4). Такая *пятеричная* система отличается тем, что число пять, записанное по этой системе, будет уже двузначным числом, изображаемым единицей в разряде пятков и нулем в разряде единиц. Запись различных чисел по пятеричной системе представлена ниже:

Десят. система	1	2	3	4	5	6	7	8	9	10	11	12
Пятер. система	1	2	3	4	10	11	12	13	14	20	21	22

Аналогичным образом можно построить *троичную* систему из трех цифр (0, 1, 2). Число три в этой системе будет двузначным; оно записывается единицей в разряде троек и нулем в разряде единиц:

Десят. система	1	2	3	4	5	6	7	8	9	10	11	12
Троичн. система	1	2	10	11	12	20	21	22	100	101	102	110

Числа по *двоичной* системе, имеющей преимущественный интерес с точки зрения кода, записываются при помощи всего лишь двух цифр (скажем, 0 и 1):

Десят. система	1	2	3	4	5	6	7	8	9	10	11	12
Двоичн. система	1	10	11	100	101	110	111	1000	1001	1010	1011	1100

Возможна, наконец, *единичная* система, располагающая одной единственной цифрой, например 1. В этой системе число знаков равно числу единиц в числе. Она отражает самую примитивную систему счета [1].

Десят. система	1	2	3	4	5
Единич. система	1	11	111	1111	11111

Называя число цифр основанием системы счисления, заметим, что мы рассматривали пока системы с основанием меньше десяти, системы, низшие по отношению к десятичной. С таким же успехом мы могли бы строить и высшие системы, например двенадцатиричную (счет на дюжины), сториную (счет на сотни) и т. п. Однако в случае двенадцатиричной системы нам пришлось бы ввести две новые цифры для обозначения чисел десять и одиннадцать (двенадцать — уже двузначное число в этой системе). Сториная система должна располагать сотней различных цифр.

Сопоставляя различные системы, можно заметить, что чем ниже основание системы, т. е. чем меньше число цифр, которым она оперирует, тем больше знаков требуется для записи данного числа по данной системе. Связь между этими величинами очень проста. Всякое число N можно записать в форме

$$N = b^n,$$

где b — основание системы счисления. Показатель n, округленный до ближайшего большего целого числа (т. е. целая часть $n+1$), дает число знаков в записи числа N.

§ 7. Код и его элементы

Код служит для передачи чисел. Этим определяется его связь с системой счисления. Код представляет собой набор комбинаций, составленных из различных элементов. Под элементами кода понимаются различные элементарные сигналы. Элементы кода играют ту же роль, какую играют цифры в системе счисления.

Говоря о различимости элементов кода, мы имеем в виду не какие-либо формальные различия, а реальные, ощутимые, даже грубые различия, которые позволяют приемной аппаратуре уверенно отличить один элемент кода от другого.

[1] В дальнейшем станет ясно, что для построения кода единичная система непригодна.

Обращаясь за примером снова к телеграфии, возьмем случай телеграфирования током одного направления. В этом случае наиболее грубо отличными элементами кода могут служить *посылка* тока, т. е. включение тока на некоторое вполне определенное время, и отсутствие посылки. Из этих двух элементов может быть построен *двоичный* код.

Рассмотрим код Бодо. Он является двоичным, так как состоит из двух элементов. Он относится к числу равномерных кодов, т. е. все его комбинации составлены из одинакового числа элементов и имеют одинаковую длительность. Благодаря этому каждый

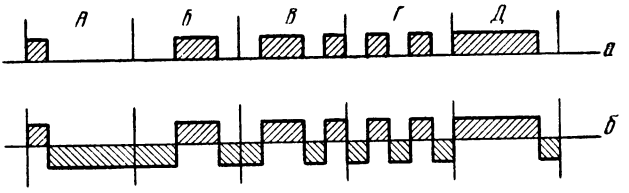


Рис. 3

элемент кода занимает вполне определенное положение во времени, находясь на определенном месте внутри комбинации. Комбинации кода Бодо составляются из пяти элементов каждая; такой код называется пятизначным. Пятизначным двоичным кодом можно передать 32 различные буквы, так как $2^5=32$ *.

Несколько комбинаций кода Бодо приведено ниже (комбинации записаны в виде чисел по двоичной системе, причем 1 означает посылку, 0 — ее отсутствие):

Буква	А	Б	В	Г	Д	Е	Ж
Код	10000	00110	01101	01010	11110	01000	00011

Форма тока в телеграфной линии показана на рис. 3, *а*. Если имеется возможность менять направление тока, то разумно взять в качестве элементов двоичного кода положительную и отрицательную посылки (элементы +1 и —1) (рис. 3, *б*).

Но, в сущности говоря, различие элементов кода в обоих рассмотренных случаях (0 и 1, +1 и —1) есть просто различие в силе тока посылки. Можно представить себе троичный код с элементами +1, 0 и —1 (кстати именно такой код применил в 1832 г. изобретатель электрического телеграфа П. Л. Шиллинг). Далее. если приемный аппарат улавливает с уверенностью различие между посылкой 0 и посылкой 1, то он сможет констатировать точно такое же различие между посылкой 1 и посылкой 2 (между прочим, четверичный код с элементами +1, —1, +3 и —3 применяется в телеграфной системе «квадруплекс»). Таким образом, мы можем построить коды со сколь угодно высоким основанием. Основание, т. е. число элементов кода, определяется только

* В аппаратах Бодо это число почти удваивается путем перевода регистра.

числом различимых в указанном выше смысле элементов, которые можно реализовать в данных условиях. Однако для связи предпочтительны коды низшего типа. В частности, двоичный код, введенный в практику очень давно, прочно в ней удерживается [1].

Мы рассматривали до сих пор в качестве элементов кода лишь посылки, отличающиеся друг от друга по силе. Но строение кода определяется только числом элементов кода, а не физическими различиями между ними. Можно было бы различать посылки не по силе, а по другому признаку, например по длительности. Именно по такому принципу построен код Морзе, элементами которого являются точка (короткая посылка) и тире (втрое более длинная посылка). Код Морзе также является двоичным. Различие в параметре, которым определяется набор элементов кода, — это различие уже не в коде, а в способе модуляции. На коде Морзе мы сейчас не останавливаемся, так как он относится к числу неравномерных, к обсуждению связанных с этим вопросов мы еще не подготовлены.

§ 8. Модуляция

В передаче сигналов всегда принимает участие некоторый физический агент, переносящий в себе сигнал. Мы назовем этот агент *переносчиком*. В электросвязи переносчиком является электрический ток или электромагнитная волна. На передающем конце мы оказываем на переносчик некоторое воздействие, изменяющее в соответствии с сообщением тот или иной параметр переносчика. Это воздействие и называется *модуляцией*. Различные виды модуляции сводятся к различиям в переносчиках и подвергаемых изменению (модулируемых) параметрах.

При телеграфировании переносчиком служит постоянный ток, сила или направление которого могут изменяться в соответствии с сигналом. (Телеграфную модуляцию называют часто манипуляцией, для такого названия в наше время нет достаточных оснований.) При телефонировании переносчиком является также постоянный ток. Модуляция осуществляется микрофоном. Говоря более подробно, процесс состоит в том, что изменения звукового давления вызывают изменения сопротивления микрофона (преобразование). Изменения же сопротивления вызывают модуляцию постоянного тока. Таким образом, в простейшей телефонной связи микрофон совмещает функции преобразователя и модулятора.

При телефонировании и телеграфировании высокой частотой (по проводам или без них) переносчиком является синусоидаль-

[1] Почти исключительное применение находит себе двоичный код и в современных электронных вычислительных машинах. Здесь оказываемое двоичному коду предпочтение связано с тем, что проще всего построить запоминающие и суммирующие элементы машины в форме триггеров, могущих принимать одно из двух возможных устойчивых состояний.

ное колебание высокой (несущей) частоты. Параметрами, определяющими это колебание, являются амплитуда, частота и фаза. Возможно модулировать каждый из этих параметров. Таким образом, получаются *амплитудная модуляция* (АМ), *частотная модуляция* (ЧМ) и *фазовая модуляция* (ФМ). На рис. 4 показаны сообщение и соответствующие формы колебаний для всех трех видов модуляции.

В новейших системах связи, в особенности в многоканальных системах с временны́м разделением, переносчиком является периодическая последовательность коротких импульсов. Такая последовательность определяется уже большим числом параметров, а именно высотой («амплитудой») импульсов, длительностью импульсов, их положением во времени (фазой), частотой следования. В соответствии с этим различают модуляцию последовательности импульсов по высоте — *амплитудно-импульсную модуляцию* (АИМ), модуляцию по *длительности* (ДИМ), *модуляцию по частоте следования* (ЧИМ). Виды импульсной модуляции показаны на рис. 5. Кроме того, в качестве самостоятельного вида модуляции упоминается *кодово-импульсная модуляция* (КИМ). На этом последнем названии следует остановиться подробнее. Дело в том, что КИМ вовсе не является самостоятельным видом модуляции с развитой выше точки зрения, с которой модуляция есть управление тем или иным параметром переносчика, и, следовательно, различие в видах модуляции есть различие в переносчиках и их параметрах. КИМ подразумевает лишь применение низшего кода, в частности двоичного. Но из предыдущего должно быть ясно, что тип кода и вид модуляции — вещи, совершенно независимые и могущие сочетаться в любых комбинациях.

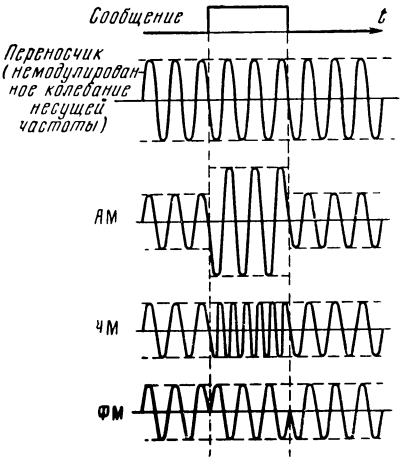
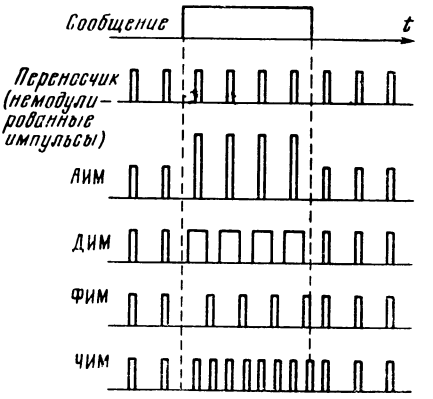


Рис. 4.

Рис. 5.



Кроме того, неправильно думать, что возможны модулированные, но некодированные сигналы. Всякий сигнал — хотим мы этого или нет — кодирован, и вопрос состоит лишь в том, по какой системе он закодирован.

§ 9. Непрерывное сообщение

Предыдущие параграфы вносят уже некоторую ясность в вопрос о механизме превращения дискретного сообщения в сигнал. Мы переходим теперь к сообщениям, не имеющим дискретной природы. Речь идет, в частности, о передаче звуков и изображений. В обоих случаях после преобразования мы получаем сообщение в форме непрерывной функции времени. При передаче звука эта функция воспроизводит временной ход звукового давления, при передаче изображения — распределение яркости по строке изображения.

На первый взгляд кажется, что случай непрерывного сообщения качественно отличен от случая дискретного сообщения: при непрерывном сообщении выражающая его функция вполне определяется несчетным множеством своих значений (т. е. бесконечным количеством чисел, выражающих мгновенные значения функции, на протяжении конечного промежутка времени).

Это, вообще, верно. Но нужно учесть, что на практике мы всегда имеем дело с функциями с ограниченным спектром, т. е. с функциями, спектральное разложение которых не содержит частот выше некоторой граничной частоты ω_c. Эта граничная частота определяется в конечном счете свойствами получателя, точнее, особенностями слуха и зрения. Так, например, общепринятой нормой для телефонии является ограничение полосы частот сверху на частоте 4—5 *кгц*. Ограничение полосы частот в телевидении определяется принятым стандартом четкости (числом строк). Аналогично обстоит дело и в других возможных случаях: мы всегда сознательным образом ограничиваем передаваемую полосу частот.

Но функции с ограниченным спектром обладают замечательным свойством: они вполне определяются конечным числом значений на протяжении конечного интервала времени. А раз так, то передача непрерывной функции с ограниченным спектром сводится опять-таки к передаче последовательности чисел. Это положение составляет содержание известной теоремы Котельникова, играющей фундаментальную роль в теории связи [10].

Поясним геометрический смысл высказанного утверждения. Всякая непрерывная кривая определяется на конечном интервале бесконечным множеством точек и для построения кривой нужно знать все ее точки. Кривая же, представляющая функцию с ограниченным спектром, может быть построена при задании на конечном интервале конечного числа точек. Через эти точки (при условии, что они расположены достаточно часто; как часто — мы

сейчас установим) кривая может быть проведена единственным образом. Это может показаться странным; но не следует забывать, что ограничение спектра представляет собой серьезное ограничение свойств функции, а следовательно, и свойств изображающей эту функцию кривой, стало быть, и способов соединения каждых двух соседних точек, принадлежащих кривой.

Доказательство теоремы Котельникова основано на разложении функции $\bar{f}(t)$ с ограниченным спектром в особого рода ряд. Ниже приводится это разложение. Для ограниченного спектра мы имеем

$$\bar{S}(\omega) = \begin{cases} \int\limits_{-\infty}^{\infty} \bar{f}(t)\, e^{-j\omega t} dt & \text{при } |\omega| < \omega_c, \\ 0 & \text{при } |\omega| \geqslant \omega_c, \end{cases}$$

На конечном интервале $(-\omega_c, \omega_c)$ функция $S(\omega)$ может быть представлена рядом Фурье

$$\bar{S}(\omega) = \sum_{-\infty}^{\infty} D_k e^{j\pi k\omega/\omega_c},$$

где $2\omega_c$ — период по частоте. Коэффициенты разложения D_k определяются по обычной формуле

$$D_k = \frac{1}{2\omega_c} \int\limits_{-\omega_c}^{\omega_c} \bar{S}(\omega)\, e^{-j\pi k\omega/\omega_c} d\omega.$$

Но $\pi/\omega_c = \Delta t$; интеграл выражает не что иное, как умноженное на 2π значение $\bar{f}(-k\Delta t)$. Таким образом,

$$\bar{S}(\omega) = \Delta t \sum_{-\infty}^{\infty} \bar{f}(-k\Delta t)\, e^{jk\omega\Delta t}.$$

Выражая функцию $\bar{f}(t)$ через ее спектр $\bar{S}(\omega)$, получим

$$\bar{f}(t) = \frac{1}{2\pi} \int\limits_{-\omega_c}^{\omega_c} \bar{S}(\omega)\, e^{j\omega t} d\omega = \frac{\Delta t}{2\pi} \int\limits_{-\omega_c}^{\omega_c} e^{j\omega t} d\omega \sum_{-\infty}^{\infty} \bar{f}(k\Delta t)\, e^{-jk\omega\Delta t},$$

где знак при k изменен на том основании, что суммирование производится все равно по всем как положительным, так и отрицательным значениям k.

Изменим порядок действий

$$\bar{f}(t) = \frac{\Delta t}{2\pi} \sum_{-\infty}^{\infty} \bar{f}(k\Delta t) \int\limits_{-\omega_c}^{\omega_c} e^{j\omega(t-k\Delta t)} d\omega =$$

$$= \sum_{-\infty}^{\infty} \bar{f}(k\Delta t)\, \frac{\sin \omega_c (t - k\Delta t)}{\omega_c (t - k\Delta t)}.$$

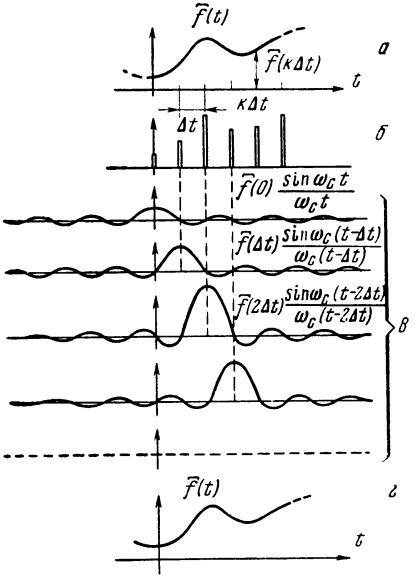


Рис. 6

Следовательно, функция с ограниченным спектром вполне определяется своими мгновенными значениями $\bar{f}(k\Delta t)$, отсчитанными через $\Delta t = \pi/\omega_c$. Доказанное положение и есть теорема Котельникова.

Наиболее важный для теории связи вывод из этой теоремы состоит в том, что передача непрерывного сообщения в том практически важном случае, когда сообщение может быть представлено функцией с ограниченным спектром, сводится к точно такой же ситуации, как и передача дискретного сообщения (см. добавление 1).

Итак, функция с ограниченным спектром может быть представлена разложением

$$\bar{f}(t) = \sum_{-\infty}^{\infty} \bar{f}(k\Delta t)\,\frac{\sin\omega_c(t-k\Delta t)}{\omega_c(t-k\Delta t)}. \qquad (1)$$

Каждое слагаемое этой суммы по физическому смыслу представляет отклик идеального фильтра нижних частот с граничной частотой ω_c на весьма короткий импульс, происходящий в момент $t=k\Delta t$ и имеющий площадь, равную мгновенному значению функции $\bar{f}(t)$ в тот же момент. Этим определяется и весь механизм передачи функции с ограниченным спектром по каналу связи, механизм, поясняемый еще рис. 6. На рис. 6, *а* представлена функция $\bar{f}(t)$. Через равные интервалы Δt берутся отсчеты мгновенных значений функции и в канал посылаются короткие импульсы, площади которых (т. е., например, высоты при неизменной длительности) пропорциональны соответствующим отсчетам (рис. 6, *б*). На приемном конце эти импульсы пропускаются через фильтр нижних частот. Отклик фильтра на каждый из импульсов представлен рядом кривых рис. 6, *в*[1].

[1] Может показаться непонятным, откуда на выходе фильтра берется предвестник входного импульса, начинающийся при $t=-\infty$. Но дело в том, что в фильтре нижних частот максимум выходного импульса запаздывает относительно входного на так называемое время пробега. Кривые рис. 6, *в* сдвинуты влево на время пробега; при таком изображении удобнее сопоставлять явления на входе и выходе фильтра. Время пробега возрастает с увеличением крутизны среза на граничной частоте. У нас предполагается идеальный фильтр, т. е. фильтр с бесконечной крутизной среза. Такой фильтр должен был бы обладать и бесконечным временем пробега. Вот

В сумме же на выходе фильтра получается снова исходная функция $\bar{f}(t)$ (рис. 6, *г*). Следует отметить замечательное свойство суммы (1): в моменты $k\Delta t$ значение суммы, т. е. значение $\bar{f}(t)$, определяется только одним k-м слагаемым суммы, так как все остальные слагаемые в этот момент обращаются в нуль. Действительно,

$$\frac{\sin \omega_c (t - k\Delta t)}{\omega_c (t - k\Delta t)} = \begin{matrix} 1 & \text{при } t = k\Delta t, \\ 0 & \text{при } t = i\Delta t \ [i \neq k] \end{matrix}$$

при любых целых значениях k и i, так как

$$\omega_c (i\Delta t - k\Delta t) = (i - k)\ \omega_c \Delta t = (i - k)\ \pi.$$

Таким образом, хотя выходные импульсы и перекрываются, но в моменты отсчета значение функции определяется только одним из них.

Теперь заметим, что в действительности мы имеем дело с сигналами, представляющими собой функции, ограниченные не только по частоте, но и во времени [1].

F — функция конечной длительности T с ограниченным спектром определяется $m = T/\Delta t = 2FT$ значениями. Но мы придем к тому же результату, если представим функцию на интервале T конечным тригонометрическим полиномом

$$\bar{f}(t) = \sum_{k=-n}^{n} C_k e^{j2\pi kt/T},$$

где n — номер наивысшей гармоники определен соотношением $2\pi n/T = \omega_c$. Действительно, так как каждая из n комплексных амплитуд C_k определяется двумя числами (как комплексная величина), то функция $\bar{f}(t)$ на интервале T определяется всего $m = 2n = 2FT$ числами.

Таким образом, получается, что функция с ограниченным спектром F и конечной длительностью T определяется $m = 2FT$ числами независимо от того, что представляют собой эти числа — мгновенные ли значения функции, отсчитанные через Δt, либо спектральные коэффициенты разложения в ряд Фурье.

Общий вывод из всего сказанного таков: передача как дискретного, так и непрерывного сообщений сводится в конечном счете к передаче последовательности дискретных чисел.

почему при вышеуказанном сдвиге начало явления оказывается перенесенным в бесконечность. Во всяком случае ясно, что на выходе фильтра сигнал не может появиться раньше момента подачи его на вход.

[1] Строго говоря, эти условия противоречивы. Поэтому о ширине спектра функций, ограниченных во времени, приходится говорить в духе более общих критериев. При этом и утверждение теоремы Котельникова принимает приближенный характер.

§ 10. Квантование

Итак, передача сообщения сведена к передаче чисел, отображаемых в сигнале, например, импульсами, высоты которых пропорциональны передаваемым числам. Однако, если бы даже эта пропорциональность могла быть строго выдержана на передающем конце, мы все равно не смогли бы правильно отсчитать переданное число на приемном конце из-за неизбежных помех. Случайный импульс помехи, накладываясь на передаваемый импульс, искажает результат. Для избежания возникающей неопределенности выбирают дискретную шкалу передаваемых уровней с та-

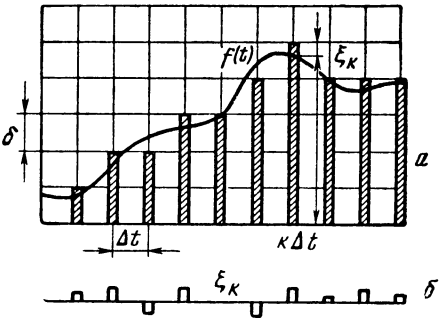


Рис. 7

ким расчетом, чтобы помеха не превосходила половины интервала между двумя соседними уровнями [1]. При таких условиях, приняв сигнал некоторой величины и отнеся его к ближайшему дискретному уровню установленной шкалы, мы заведомо не совершаем ошибки. Замена непрерывной шкалы уровней дискретной называется квантованием; сигнал, представляемый последовательностью дискретных значений, называется квантованным.

Механизм квантования на передающем конце сводится к тому, что вместо данного мгновенного значения передаваемой величины (при непрерывном сообщении) передается ближайшее значение по установленной шкале дискретных уровней. Графически процесс квантования непрерывного сообщения $f(t)$ можно представить при помощи рис. 7. Кривая $f(t)$ наложена на прямоугольную сетку с ячейками $\Delta t\,\delta$, где δ — шаг шкалы уровней. Квантование состоит в том, что вырабатываются импульсы, высота которых равна не ординате кривой, а высоте ближайшего уровня (рис. 7, *а*).

Само собой разумеется, что квантование сопровождается искажением, так как посылаемые импульсы воспроизводят функцию сообщения $f(t)$ неточно. Разность между квантованными импульсами и импульсами высотой $f(k\Delta t)$, которая обозначена на рис. 7, *а* через ξ_k и образует последовательность импульсов, представлен-

[1] Помеха есть случайный процесс, и приведенная формулировка нуждается в уточнении, которое будет сделано позднее (см. § 22).

ную на рис. 7, *б*, можно рассматривать как особого рода помеху. Она известна под названием «шум квантования».

Этот недостаток системы квантования компенсируется, однако, одним серьезным техническим преимуществом, состоящим в том, что при передаче квантованного сигнала по длинной линии связи можно избежать накопления помех вдоль линии путем периодического восстановления сигнала. Этот прием (в принципе известный уже давно в форме восстанавливающих телеграфных трансляций) основан на том, что истинный уровень передаваемого квантованного сигнала на конце некоторого отрезка линии в точности известен, несмотря на наличие помех (при условии, что помеха не превосходит половины шага шкалы уровней). Если так, то можно регенерировать сигнал, т. е. создать его заново, полностью очистив от посторонних помех, и послать дальше. Эта операция восстановления сигнала может быть повторена сколько угодно раз, и таким образом может осуществляться связь на произвольные расстояния, для чего нужно оборудовать на линии ряд ретрансляционных пунктов — повторителей. Именно так и устраиваются длинные радиорелейные линии. Единственная и притом постоянная помеха, которая при этом неизбежно остается, — это шум квантования.

Характеризуя положение кратко и в грубых чертах, можно сказать, что передача сигнала возможна с точностью до помехи [1].

§ 11. Количество сведений

Мы подготовлены теперь к тому, чтобы дать количественную оценку важной величине, названной в § 1 количеством сведений.

Передача сообщений сведена, как объяснено выше, к передаче последовательности чисел. Таким образом, число является универсальным элементом сообщения. На такие элементы можно разложить любое сообщение, дискретное и непрерывное: текст, речь, музыку, изображение и т. д. Естественно было бы выбрать число за единицу количества сведений. Но такая постановка вопроса лишена смысла. пока мы не рассматриваем всей совокупности возможных элементов данного рода передачи. Здесь возникает одно из важных понятий теории связи. Мы попытаемся начать его разъяснение со следующего рассуждения. Положим, что мы передаем словесный текст, которым мы желаем сообщить сведения о некоторой ситуации, и положим, что мы располагаем для этой цели всего двумя словами: «хорошо» и «плохо». Очевидно, что оценка ситуации при таких условиях может быть только очень грубой; мы скажем, что количество сведений о ситуации будет мало. Если же мы располагаем большим числом слов, образующим

---

[1] В дальнейшем, однако, будут указаны методы приема сигнала при наличии помехи, значительно его превосходящей.

богатый лексикон и позволяющим передать тонкие оттенки нашей оценки, то те же самые слова «хорошо» и «плохо», входя в состав этого лексикона, будут уже гораздо более точно выражать нашу оценку и, следовательно, доставят большее количество сведений. Рассматривая различные слова как возможные элементы сообщения, мы видим, что количество сведений, содержащееся в словесном тексте, зависит не только от числа слов, составляющих этот текст, но и от общего числа слов в лексиконе, из которого мы сделали выбор слов для нашего текста.

Другой пример подведет нас ближе к технической постановке вопроса. Возьмем телевизионное изображение. Его передача состоит в передаче распределения яркости по строке изображения. Сигнал предполагается квантованным, т. е. передается некоторая градация яркости с дискретными ступенями, занимающими интервал между белым и черным. Если бы мы располагали всего двумя ступенями, то могли бы передать лишь черно-белое изображение предмета. Количество сведений о предмете было бы в этом случае мало. Чем больше ступеней яркости мы передаем, тем большее количество сведений о предмете можно получить на приемном конце. Таким образом, количество сведений возрастает с числом ступеней шкалы уровней сигнала.

Общий вывод из подобных рассуждений состоит в том, что каждый элемент сообщения содержит тем большее количество сведений, чем больше общее число возможных элементов, из которого данный элемент выбран.

Теперь мы можем рассуждать дальше. Сообщение состоит не из одного, а из многих элементов. Пусть число возможных элементов есть m, а число элементов в сообщении n. Выбирая первый элемент сообщения, мы делаем выбор из m возможных элементов. Выбирая второй элемент, мы делаем выбор из того же числа m элементов, но число возможных комбинаций выбора двух элементов составляет уже m^2. Если же сообщение содержит n элементов, то число различных сочетаний этих элементов есть

$$N = m^n.$$

Это — *число возможных сообщений*. Оно и может служить мерой количества сведений в сообщении. Однако при выборе меры нужно подчинить ее естественному требованию аддитивности, состоящему в том, что количество сведений, содержащееся в сообщении, должно быть пропорционально n, или, попросту говоря, вдвое более длинное сообщение (например, телеграмма из вдвое большего числа слов) должно содержать вдвое большее количество сведений. Мы должны, стало быть, выбрать в качестве меры количества сведений не само число N, а некоторую функцию $I = f(N)$, удовлетворяющую условию аддитивности. Для отыскания функции f сформулируем условие аддитивности: приращение функции f должно быть пропорционально приращению числа

элементов сообщения n. В дифференциальной форме это условие имеет вид

$$df = kdn.$$

С другой стороны,

$$dN = \ln m \cdot Ndn.$$

Заменяя здесь dn на df, получаем

$$df = k_1 dN/N,$$

откуда искомая функция

$$f = k_1 \ln N = \log_a N.$$

Для выбора основания логарифма рассмотрим простейшую ситуацию, когда сообщение представляет собой один символ, обусловленный выбором одного из двух возможных («да» или «нет»), т. е. когда $n=1$, $m=2$. В этом случае будем иметь

$$f = \log_a N = \log_a m^n = \log_a 2.$$

Количество сведений, получаемое при таких условиях, мы примем за единицу. Это определяет выбор основания логарифмов: мы условливаемся выражать количество сведений *двоичным* логарифмом числа N. Итак, количество сведений в сообщении выражается так [1]:

$$I = \log_2 N = n \log_2 m. \qquad (1)$$

Таким образом, каждая посылка двоичного кода несет одну единицу количества сведений; кодовая группа Бодо, состоящая из пяти посылок (знаков), содержит пять единиц и т. д.

Нам понадобится еще определение количества сведений, приходящегося на один элемент сообщения. Мы назовем эту величину содержательностью сообщения и обозначим $I' = I/n$.

Концепция выбора, приводящая к важному соотношению (1) и общепринятая в настоящее время, была впервые предложена Хартли в 1928 г. [21]. В новейшее время эта концепция развита с вероятностной точки зрения Шэнноном [27].

Во избежание недоразумений следует пояснить, что понятие количества сведений не затрагивает смысла передаваемого сообщения и тем более действия, которое может полученное сообщение произвести на получателя. Для техники связи и для развиваемой здесь теории совершенно безразлично, содержит ли телеграмма извещение о смерти или о рождении или же представляет собой бессмысленный набор слов. Поэтому термин количество сведений и все производные термины не следует понимать в обычном житейском смысле; в теории связи этот термин означает скорее информацию, которую сообщение могло бы содержать, нежели ту ин-

[1] В дальнейшем знак log означает везде двоичный логарифм.

формацию, которую оно фактически содержит по нашей житейской оценке. Такой подход к проблеме можно иллюстрировать хотя бы тем, что плата за услуги связи взыскивается с клиентов, например, по числу слов телеграммы или по продолжительности телефонного разговора независимо от фактического содержания телеграммы или разговора.

§ 12. Физические характеристики сигнала

Нам нужно теперь ввести некоторые величины, которые позволили бы охарактеризовать сигнал с физической точки зрения. С одной стороны, это необходимо для установления количества сведений, которое может быть передано сигналом, с другой — для выяснения соотношения между характеристиками сигнала и свойствами канала связи, по которому этот сигнал должен быть передан.

Если сигнал представляется некоторой функцией времени, то, определив так или иначе эту функцию, мы определяем и сигнал. Для этого, если, например, сигнал выражается функцией с ограниченным спектром, достаточно знать мгновенные значения этой функции через интервалы $\Delta t = \pi / \omega_c$ (см. § 9).

Однако в таком полном описании сигнала нет надобности. Для построения ряда выводов общей теории достаточно гораздо более общего описания сигнала. Такое общее описание состоит в том, что сигнал определяется лишь небольшим числом обобщенных измерений. Дело обстоит примерно так, как если бы мы, не вдаваясь в детальное описание свойств некоторого предмета, указали бы лишь такие обобщенные его измерения, как габариты, вес и т. п. Конечно, выбор таких измерений должен быть целесообразен, т. е. эти измерения должны содержать именно те сведения о предмете, которые нас интересуют в первую очередь. Например, указывая габариты и вес предмета, мы наиболее целесообразно характеризуем свойства предмета с точки зрения условий его транспортировки; другие возможные свойства предмета (например, цвет, детали формы) с этой точки зрения могут не играть никакой роли.

Сигнал есть также в определенном смысле объект транспортировки, так как он должен быть передан по каналу связи от передатчика к приемнику. Техника связи и есть по существу техника транспортирования сигнала. Поэтому мы должны попытаться ввести для характеристики сигнала именно такие его измерения, которые определяют условия его передачи.

Всякий сигнал, рассматриваемый как явление во времени, имеет начало и конец. Поэтому первым наиболее естественным измерителем свойств сигнала является его *длительность*. Легко видеть, что длительность сигнала просто связана с количеством сведений, которое при прочих равных условиях должно быть пропорционально длительности. С другой стороны, длительность сиг-

нала также просто связывается с условиями работы канала связи: чем больше длительность сигнала, тем на большее время занимается канал.

Однако знания длительности сигнала еще недостаточно. Эта величина лишь определяет интервал времени, в пределах которого сигнал существует, т. е. не равен тождественно нулю. Но теперь нужно дать некоторую характеристику функции сигнала в интервале его существования, т. е. на протяжении его длительности.

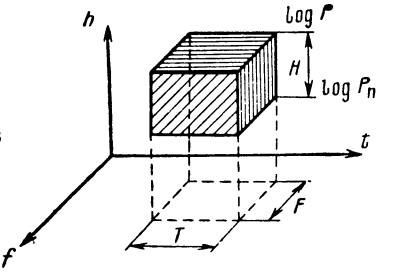


Рис. 8

Одним из характерных измерителей может служить энергия или средняя мощность сигнала как величина, характеризующая силу сигнала. Но мощность сигнала сама по себе не определяет свойства сигнала как переносчика сведений, так как мы не можем игнорировать реальные условия передачи сигнала, определяемые наличием помехи. Поэтому силу сигнала целесообразно характеризовать не абсолютной мощностью, а отношением мощности сигнала к мощности помех. Удобно, как это будет видно из дальнейшего, ввести в качестве измерителя, характеризующего силу сигнала, величину

$$H = \log \frac{P}{P_{\text{п}}},$$

где P и $P_{\text{п}}$ — соответственно средние мощности сигнала и помехи. Определенную таким образом величину мы будем называть превышением сигнала над помехой, или просто превышением. Легко видеть, что величина H выражает не что иное, как относительный средний уровень сигнала над помехой.

Но и этого мало. Мы ввели длительность сигнала и превышение. Но у нас отсутствует еще какая-либо характеристика поведения сигнала (на протяжении его длительности), которая показывала бы скорость его изменения.

Мы могли бы ввести в качестве третьего измерителя сигнала диапазон скоростей его изменения, т. е. интервал значений первой производной функции сигнала. Подобным же образом можно было было бы дать еще более подробное описание свойств функции, введя в рассмотрение вторую производную, и т. д.

В теории связи ограничиваются пока тремя измерителями свойств сигнала, вводя в качестве третьего *ширину спектра* сигнала. Эта величина дает представление о характере сигнала и является, как мы увидим, весьма удобной.

Мы примем для описания общих свойств сигнала три основных измерения: длительность T, ширину спектра F и превышение H. Эти измерения можно себе представить в виде отрезков определенной длины, отложенных параллельно трем координатным осям: времен, частот и уровней. Таким образом, возникает геометрическое представление сигнала как некоторого объема в трехмерном пространстве. Этот объем представляется как параллелепипед с ребрами T, F и H («габаритные размеры» сигнала). Этот геометрический образ изображен на рис. 8. Произведение трех измерений $V=TFH$ мы будем называть *объемом* сигнала.

§ 13. Сигнал и канал

Введя понятие об объеме сигнала, мы можем сравнительно просто представить соотношения между свойствами сигнала и свойствами канала связи.

Канал связи можно охарактеризовать также тремя параметрами: 1) временем $T_к$, в течение которого канал предоставлен для работы, 2) полосой частот $F_к$, которую канал способен пропустить, и 3) полосой уровней $H_к$, зависящей от допустимой нагрузки аппаратуры канала. Совершенно очевидно, что передача сигнала с измерениями T, E и H по каналу с параметрами $T_к$, $F_к$ и $H_к$ возможна при условии $T_к \geqslant T$; $F_к \geqslant F$; $H_к \geqslant H$.

Три параметра канала можно перемножить и назвать их произведение

$$V_к = T_к F_к H_к$$

емкостью канала. Сигнал может быть передан по каналу, если емкость канала не менее объема сигнала, или, образно говоря, если сигнал «вмещается» в канал. Это представление связывается с геометрическим образом двух параллелепипедов, из которых один должен поместиться в другом, что возможно, очевидно, если все три стороны вмещающего параллелепипеда больше соответствующих сторон вмещаемого. Впрочем, в дальнейшем выяснится возможность деформаций объема сигнала, позволяющих согласовать сигнал с каналом, так что условие возможности передачи можно смягчить и записать в более общем виде

$$V_к \geqslant V.$$

§ 14. Количество сведений и объем сигнала

Чем больше объем сигнала, тем большее количество сведений он может перенести. Важное соотношение между обоими этими величинами мы выведем пока для частного случая.

Пусть сигнал представляет собой последовательность модулированных по высоте импульсов (АИМ) со скважностью единица и пусть число ступеней шкалы уровней равно m. Если все уровни равновероятны и если общее число элементов сообщения есть n, то количество сведений равно

$$I = \log N = n \log m. \tag{1}$$

Обозначим шаг шкалы уровней через δ. Тогда ступени шкалы будут отвечать значениям $0, \delta, 2\delta, \ldots, i\delta, \ldots, (m-1)\delta$. Мгновенная мощность сигнала будет $(i\delta)^2$, а средняя мощность

$$P = \frac{1}{m}\sum_{i=0}^{m-1}(i\delta)^2 = \frac{\delta^2}{m}\sum_{i=0}^{m-1} i^2 = \frac{\delta^2}{6}(m-1)(2m-1).$$

Рассмотрим случай $m \gg 1$. В этом случае

$$P = \frac{1}{3}\delta^2 m^2,$$

откуда

$$m^2 = 3\frac{P}{\delta^2} = AP. \tag{2}$$

Теперь определим n. Если длительность сигнала есть T, то

$$n = T/\Delta t,$$

где Δt — длительность одного импульса. С другой стороны,

$$\Delta t = 1/2F$$

и, следовательно,

$$n = 2FT. \tag{3}$$

Подставив (2) и (3) в (1), получим

$$I = n \log m = \frac{1}{2} n \log m^2 = FT \log AP. \tag{4}$$

Введем теперь в (4) мощность помехи. Она может быть выражена среднеквадратичным значением помехи σ

$$P_{\text{п}} = \sigma^2. \tag{5}$$

С другой стороны, при выборе шага шкалы уровней мы сообразовываем величину шага с интенсивностью помехи и можем положить

$$\delta = k\sigma, \tag{6}$$

где k — коэффициент, зависящий от статистики помехи. Его значение будет определено позднее (§ 22). Из (5) и (6) находим

$$P_{\text{п}} = \delta^2/k^2, \tag{7}$$

сопоставляя этот результат с (2), можем записать

$$m^2 = \frac{3}{k^2} \cdot \frac{P}{P_{\text{п}}} = a \frac{P}{P_{\text{п}}}. \tag{8}$$

Следовательно,

$$I = FT \log a \frac{P}{P_{\text{п}}}. \tag{9}$$

В такой форме выражение для количества сведений удобно сопоставлять с выражением для объема сигнала. Для объема сигнала мы имели (§ 12)

$$V = FTH = FT \log \frac{P}{P_{\text{п}}}. \tag{10}$$

Введем определение *удельной содержательности сигнала*

$$\nu = 1/V. \tag{11}$$

Из (9) и (10) мы получаем

$$\nu = \frac{\log (aP/P_{\text{п}})}{\log (P/P_{\text{п}})} = 1 + \frac{\log a}{H}. \tag{12}$$

Величина удельной содержательности показывает, насколько эффективно используется сигнал данного объема для передачи сведений; она показывает, так сказать, плотность упаковки сведений в объеме сигнала. Задача состоит, очевидно, в увеличении удельной содержательности сигнала всеми имеющимися в нашем распоряжении средствами. Как мы увидим, удельная содержательность зависит от основания кода, от способа модуляции и, что наиболее существенно, от статистики сообщения. Использование этих зависимостей позволяет увеличить удельную содержательность сигнала довольно значительно, и развитие техники связи уже идет в настоящее время по пути разработки и реализации возможностей, указываемых теорией.

В качестве иллюстрации рассмотрим влияние на удельную содержательность сигнала основания кода при АИМ. Для этого нам нужно найти значение коэффициента a в зависимости от основания кода. Напомним, что соотношение $a=3/k^2$ относится к частному случаю $m \gg 1$. В общем же случае

$$P = \frac{1}{6} \delta^2 (m-1)(2m-1),$$

откуда

$$m^2 = \frac{m^2 P}{\frac{1}{6} \delta^2 (m-1)(2m-1)} = a_m \frac{P}{P_{\text{п}}}$$

и коэффициент a_m

$$a_m = \frac{6m^2}{k^2(m-1)(2m-1)}.$$

Коэффициент a_m монотонно убывает с возрастанием m; иначе говоря, чем основание кода ниже, тем удельная содержательность сигнала больше. Если положить $m=2$, т. е. перейти к двоичному коду, то получим

$$a_2 = 8/k^2,$$

т. е. более выгодные соотношения, чем при $m \gg 1$. В этом заключается одна из причин предпочтения, отдаваемого двоичному коду.

В заключение этого параграфа заметим еще, что хотя назначение системы связи и состоит в передаче сообщений, а не в передаче энергии, тем не менее эти две категории тесно связаны между собой, и оказывается, что количество сведений, переносимых сигналом в единицу времени, прямо пропорционально логарифму мощности сигнала, как это видно из формулы (4). Таким образом, передача сообщений непосредственно связана с передачей энергии от отправителя к получателю; задача техники состоит в рациональном использовании этой энергии.

§ 15. Пропускная способность системы связи

В действительных условиях работы системы связи в приемник поступает сигнал с примешанными к нему помехами. Это обстоятельство изменяет оценку количества сведений. Можно рассматривать положение при наличии помех, полагая, что помехи также приносят сведения, однако сведения ложные, подлежащие по возможности устранению. Преимущество такой точки зрения состоит в том, что мы применяем к оценке сигнала и помехи один и тот же количественный критерий — величину, называемую нами количеством сведений.

Действительное количество сведений при наличии помех мы определим следующим образом: пусть I — количество сведений в передаваемом сигнале; через $I_п$ обозначим количество «сведений», доставляемое помехой. На приемный конец поступает смесь сигнала с помехой, содержащая количество сведений

$$I_1 = I + I_п$$

(простое суммирование предполагает полную взаимную независимость сигнала и помехи, как оно обычно и есть в действительности).

Чтобы учесть на приемном конце количество только полезных сведений, нужно вычесть из I_1 количество «сведений» помехи, так что действительное количество сведений, передаваемое системой связи, есть

$$I = I_1 - I_п.$$

Положим теперь, что сигнал и помеха обладают одинаковыми свойствами (речь идет о вероятностных характеристиках сигнала и помехи, и эта формулировка будет позднее уточнена), т. е. что количество сведений может быть выражено как для сигнала, так и для помехи формулами вида (4) § 14 с одинаковым коэффициентом A. Тогда имеем

$$I_1 = FT \log A (P_c + P_п); \quad I_п = FT \log A P_п,$$

откуда

$$I = I_1 - I_п = FT \log (1 + P_c/P_п). \tag{1}$$

Предельная пропускная способность системы связи, т. е. наибольшее количество сведений, которое система может передать в единицу времени, есть

$$C = \frac{I}{T} = F \log \left(1 + \frac{P_c}{P_п}\right). \tag{2}$$

Эта формула имеет весьма общий характер и представляет собой одно из важных соотношений общей теории связи [1].

Заметим, что пропускная способность неограниченно возрастает при уменьшении мощности помехи. При достаточном превышении сигнала над помехой единицей в скобках можно пренебречь и количество сведений будет

$$I = FT \log \frac{P_c}{P_п} = FTH.$$

Это означает, что при неограниченном увеличении превышения удельная содержательность сигнала стремится к единице.

Формула (1) показывает, что передача сообщений возможна и при отрицательных превышениях, т. е. при $P_п > P_c$, однако пропускная способность убывает при возрастании мощности помех, стремясь в пределе к нулю.

Формулу (1) нельзя непосредственно сравнивать с формулами для количества сведений, приведенными в § 14. Дело в том, что формула (1) выражает теоретический предел количества сведений, могущего быть переданным со сколь угодно малой вероятностью ошибки [2]. В то же время формулы § 14 дают количество сведений при допущении вполне определенной вероятности ошибки; при этом допущении количество сведений, определенное по формуле (9) § 14, может и превзойти предел, устанавливаемый формулой (1). Однако это возможно лишь, если допускается большая вероятность ошибки.

---

[1] Само собой разумеется, что приведенное здесь рассуждение не может рассматриваться как формальное доказательство соотношения (2), см. по этому поводу [28].

[2] Это обстоятельство нашло в предшествующем рассуждении отражение в том, что мы предположили ложные «сведения» в количестве $I_п$ полностью отделимыми.

Мы можем сравнить на примере результаты, получаемые при определенных практических условиях, с предельным соотношением (1) [25]. Пусть система передачи есть система АИМ с равновероятными положительными и отрицательными значениями, образующими ряд

$$-\frac{m-1}{2}\delta,\ \ldots,\ -2\delta,\ -\delta,\ 0,\ \delta,\ 2\delta,\ \ldots,\frac{m-1}{2}\delta$$

— всего m значений. Мощность сигнала будет

$$P=\frac{2\delta^2}{m}\sum_{i=1}^{\frac{m-1}{2}} i^2=\frac{\delta^2}{12}(m^2-1),$$

откуда

$$m^2=1+\frac{P}{\delta^2/12}.$$

Подставляя это в выражение для количества сведений, получим

$$I=FT\log\left(1+\frac{P}{\delta^2/12}\right)=FT\log\left(1+\frac{12}{k^2}\cdot\frac{P}{P_{\text{п}}}\right). \qquad (3)$$

Сравнивая (3) и (1), мы видим, что для передачи данного количества сведений в рассматриваемой системе нужно затратить мощность

$$P=\frac{k^2}{12}P_c,$$

т. е. мощность, в $k^2/12$ раз большую, чем теоретически необходимая.

§ 16. Преобразование сигнала кодированием

Мы установили, что количество сведений определяется объемом сигнала, который выражается произведением длительности, ширины спектра и превышения. Соотношение этих величин зависит от того, как закодирован сигнал. Изменяя систему кодирования, можно получить иное соотношение измерений сигнала при том же объеме его и при том же количестве сведений.

Количество сведений есть

$$I=\log N=\log m^n.$$

В этом выражении m есть основание кода, n — число элементов сигнала.

Если каждое мгновенное значение функции сообщения передается только одним импульсом, то это вовсе не значит, что сигнал не кодирован, как иногда полагают. Это лишь означает, что при-

меняется код с относительно высоким основанием, равным числу уровней квантования. Сигнал может быть перекодирован любым образом с применением кода с бóльшим или меньшим основанием. Количество сведений при этом измениться не должно.

Запишем аналитические соотношения, отметив величины, относящиеся к новому коду, индексом 1. Мы имеем

$$I = \log m^n = \log m_1^{n_1}.$$

Отсюда

$$n_1 = n \frac{\log m}{\log m_1}.$$

Таким образом, понижение основания кода m влечет за собой соответствующее увеличение числа элементов n. Число элементов сигнала, т. е. число импульсов, равно

$$n = T/\Delta t = 2FT.$$

Здесь T — общая длительность сообщения, которую мы полагаем неизменной при изменении кода. Поясним эти соотношения примером. Пусть первоначальный сигнал получен квантованием функции сообщения при числе уровней $m=64$. Перейдем теперь от кода с основанием 64 к двоичному коду, т. е. возьмем $m_1=2$. Тогда

$$n_1 = n \frac{\log 64}{\log 2} = 6n.$$

Это означает, что каждый отсчет функции сообщения должен передаваться уже не одним, а шестью импульсами; мы перешли к шестизначному двоичному коду. Далее,

$$n = 2FT, \quad n_1 = 2F_1T, \quad F_1 = Fn_1/n = 6F.$$

Таким образом, уменьшив мощность сигнала путем сокращения шкалы уровней, мы соответственно расширили требуемую полосу частот. Это и понятно: ведь если за то же время, за которое раньше передавался только один импульс, теперь передается шесть, то длительность каждого из шести импульсов вшестеро меньше. А известно, что произведение из длительности импульса на ширину его спектра есть постоянная величина. Стало быть, сокращая длительность импульса, мы во столько же раз увеличиваем ширину его спектра.

Можно представить себе такое изменение кода, при котором произойдет обратное преобразование сигнала, т. е. сокращение спектра, получаемое ценой увеличения мощности сигнала. Положим, например, что мы передаем при помощи одного элемента кода (одного импульса) комбинацию двух отсчетов функции сообщения, так что число элементов сигнала сокращается вдвое, $n_1=n/2$. При этом импульсы сигнала будут следовать друг за другом через $2\Delta t$, т. е. через вдвое большее время. Это позволяет вдвое увеличить длительность импульса, а следовательно, вдвое

сократить ширину спектра сигнала. При этом, однако, соответственно возрастет мощность сигнала. Если первоначальное число уровней было равно m, то теперь число комбинаций значений уровней двух соседних отсчетов составит $m_1=m^2$. В нашем примере $m=64$; для сокращения полосы частот вдвое нужно перейти к коду с основанием $m_1=64^2=4096$. Количество сведений остается неизменным

$$I_1 = \log m_1^{\frac{n}{2}} = \log (m^2)^{\frac{n}{2}} = I.$$

Итак, путем выбора кода мы можем произвести такое преобразование сигнала, при котором ширина спектра F или превышение H изменяются в желаемое число раз, причем произведение этих величин сохраняет свой порядок. Инвариантом преобразования является величина $F \log m$. Общая длительность сигнала T остается неизменной.

Такого рода преобразование позволяет согласовать измерения сигнала с параметрами канала связи.

§ 17. Преобразование объема сигнала

В предыдущем параграфе показано, как можно сжать сигнал в направлении оси уровней с соответствующим расширением его в направлении оси частот, или наоборот. Так возникает наглядное представление о деформациях объема сигнала.

Мы рассмотрели частный вид деформации, при которой подвергались согласованному изменению измерения F и H; измерение T (длительность сигнала) предполагалось неизменным. Вообще же возможны и практически применяются и другие виды деформаций сигнала, в которых участвует любая пара из трех измерений сигнала или даже все три. Кроме того, применяются преобразования сигнала, при которых его объем не деформируется, но сдвигается без деформации вдоль одной из осей.

Приведем сперва примеры простейших преобразований без деформации — преобразований переноса. Перенос сигнала вдоль оси t на t_0 есть попросту задержка на время t_0, которая может быть осуществлена при помощи линии задержки или для произвольно большой задержки путем записи сигнала с последующим его воспроизведением (рис. 9, *а*).

Перенос сигнала без деформации вдоль оси частот на f_0 (рис. 9, *б*) осуществляется при *однополосной модуляции* сигналом несущей частоты f_0.

Перенос сигнала без деформации вдоль оси уровней означает просто усиление сигнала (при переносе вверх, как на рис. 9, *в*, перенос вниз означает ослабление).

Простейшим примером деформации сигнала, в которой участвуют измерения F и T, служит запись сигнала и воспроизведение с измененной скоростью. Если записанный при скорости v

сигнал воспроизвести со скоростью $v_1 = av$, то длительность сигнала сократится в a раз, но во столько же раз возрастут все частоты, а стало быть, и ширина спектра. Превышение сигнала остается при такой деформации без изменения; следовательно, неизменным останется и объем сигнала.

Что касается деформаций сигнала с участием измерений T и H, то и такие деформации возможны. Интересным примером служит прием сигнала *методом накопления*, который будет рассмотрен ниже.

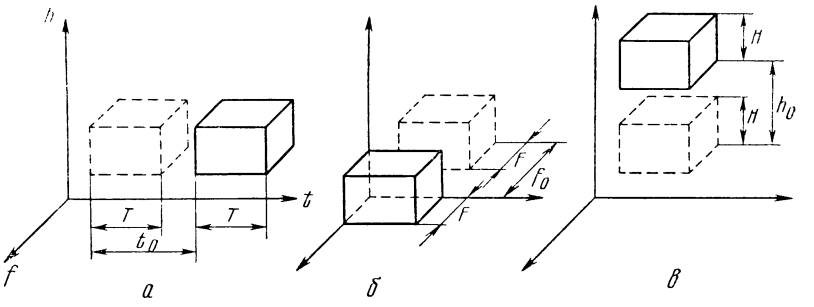


Рис. 9

Мы видим теперь, что возможность передачи данного сигнала по данному каналу определяется только соотношением между объемом сигнала и емкостью канала. Согласование измерений сигнала с параметрами канала всегда возможно путем соответствующих деформаций сигнала.

В примере с преобразованием путем записи и воспроизведения с измененной скоростью произведение FT остается неизменным. А так как и превышение сигнала не меняется, то объем сигнала $V = FTH$ сохраняет при описанном преобразовании свою величину; измерение F увеличивается во столько же раз, во сколько сокращается измерение T. Таким образом, описанное преобразование, характеризуемое сохранением объема, напоминает деформацию несжимаемого тела. Однако не при всех деформациях объем сигнала сохраняется. Как правило, при сокращении одного измерения увеличивается другое (или другие), но объем сигнала может при этом измениться. Пользуясь механической аналогией, можно уподобить ситуацию той, которую мы имеем при деформации упругого тела. В этом случае между измерениями тела существует связь, но более сложная; так, например, продольное сжатие сопровождается поперечным расширением в случае несжимаемого тела, так что при сжатии в одном каком-либо направлении объем тела уменьшается. Отношение поперечной деформации к продольной выражается, как известно, коэффициентом Пуассона.

Для того чтобы оценить выгоду той или иной деформации сигнала, полезно ввести понятие об *инварианте преобразования*,

под которым будет пониматься некоторая функция измерений сигнала, сохраняющая при данном виде преобразования постоянное значение.

Для всех преобразований переноса инвариантами преобразования являются, очевидно, как объем сигнала, так и каждое его измерение в отдельности.

Для преобразования путем записи и воспроизведения инвариантами являются величины H и FT, а следовательно, и объем сигнала V.

В качестве примера более сложных соотношений рассмотрим преобразование сигнала кодированием. Как установлено в § 16, инвариантом преобразования, состоящего в изменении основания кода, является величина

$$\log N = n \log m = \text{const}.$$

Выразим эту величину через измерения сигнала. Пусть речь идет об АИМ сигнале с высотой импульса

$$h_i = i\delta \, (i = 0, 1, 2, \ldots, n-1).$$

Для числа элементов n мы имеем (см. § 14) $n = 2FT$. Что же касается основания кода m, то его можно выразить через превышение сигнала при помощи соотношения

$$H = \log \frac{P}{P_{\text{п}}} = \log \frac{k^2}{6} (m-1)(2m-1).$$

Разрешая это квадратное относительно m уравнение, получаем

$$m = \frac{3}{4} + \sqrt{\frac{1}{16} + \frac{3}{k^2} 2^H}.$$

Таким образом, инвариантом преобразования в рассматриваемом случае является величина

$$F \log \left(\frac{3}{4} + \sqrt{\frac{1}{16} + \frac{3}{k^2} 2^H} \right) = \text{const}.$$

Длительность сигнала также остается постоянной. Таким образом, при данном преобразовании объем сигнала изменяется. Он равен

$$V = FTH = \log N \frac{\log \frac{k^2}{6} (m-1)(2m-1)}{\log m^2}$$

и изменяется от

$$V = \log N \frac{\log \frac{k^2}{3} m^2}{\log m^2}$$

при $m \gg 1$ до

$$V = \log N \frac{\log \frac{k^2}{8} m^2}{\log m^2} = \log N \left(\log k - \frac{1}{2} \right)$$

при $m = 2$.

Рассмотрим еще случай симметричной относительно нуля шкалы уровней, т. е. амплитудно-импульсную модуляцию, в которой высота импульсов определяется соотношением

$$h_i = i\delta\left(i = -\frac{m-1}{2}, \ldots, -2, -1, 0, 1, 2, \ldots, \frac{m-1}{2}\right).$$

В этом случае (см. § 15)

$$P = \frac{1}{m}\sum h_i^2 = \frac{2\delta^2}{m}\sum_{i=1}^{\frac{m-1}{2}} i^2 = \frac{\delta^2}{12}(m^2-1),$$

$$H = \log\frac{P}{P_{\text{п}}} = \log\frac{k^2}{12}(m^2-1),$$

и инвариантом преобразования является величина

$$F\log\left(1+\frac{12}{k^2}2^H\right) = \text{const}.$$

§ 18. Сравнение некоторых видов связи

В качестве иллюстрации общих соотношений интересно сравнить между собой численные характеристики различных видов связи. Ниже приведены ориентировочные цифры, выражающие количество сведений, могущее быть переданным в единицу времени при помощи различных видов связи. Мы пользуемся соотношением

$$I = n\log m.$$

Количество сведений в единицу времени

$$I/T = 2F\log m.$$

Для определения этой величины нужно знать ширину спектра сигнала F и число ступеней m, т. е. число различимых градаций силы сигнала. Все эти величины даны в табл. 1 для пяти видов связи [1].

В табл. 1 поражает огромная цифра, характеризующая производительность телевидения как системы связи. Однако картина изменится, если мы подсчитаем удельную содержательность сигнала. Для этого нужно сперва найти объем сигнала по формуле

$$V = FTH = FT\log\frac{P}{P_{\text{п}}}.$$

---

[1] При составлении нижеследующей сводки возникает затруднение с терминологией. Мы пользуемся ниже термином «сигнал» для всех рассматриваемых видов связи. Для телеграфа это не требует оговорок. Применительно к телефонии речь идет о низкочастотном сигнале (без модуляции); для телевидения подразумевается так называемый «видеосигнал».

Таблица 1

Вид связи	F, гц	m	$\log m$	I/T
Телеграф (Морзе, быстродейств.)	$4 \cdot 10^2$	2	1	$8 \cdot 10^2$
Телеграф (Бодо)	40	2	1	80
Фототелеграф	$3 \cdot 10^3$	12	3,6	$2{,}2 \cdot 10^4$
Телефон (импульсн.)	$4 \cdot 10^3$	128	7	$5{,}6 \cdot 10^4$
Телевидение	$6 \cdot 10^6$	30	4,9	$5{,}9 \cdot 10^7$

Положим, что передача ведется амплитудно-модулированными импульсами, тогда

$$P = \frac{\delta^2}{6}(m-1)(2m-1).$$

С другой стороны,

$$P_{\text{п}} = \delta_{\text{п}}^2 = \frac{1}{k^2}\delta^2.$$

Выберем коэффициент запаса k так, чтобы вероятность ошибки при помехе в форме белого шума составляла около 10^{-3} (см. § 22). Тогда можем записать

$$H = \log \frac{P}{P_{\text{п}}} \approx \log 7\,(m-1)(2m-1)$$

и составить табл. 2.

Таблица 2

Вид связи	F, гц	H	$FH = V/T$	$\nu = 1/V$
Телеграф (Морзе)	$4 \cdot 10^2$	4,4	$1{,}8 \cdot 10^3$	0,45
Телеграф (Бодо)	40	4,4	$1{,}8 \cdot 10^2$	0,45
Фототелеграф	$3 \cdot 10^3$	10,7	$3{,}2 \cdot 10^4$	0,68
Телефон	$4 \cdot 10^3$	17,8	$7{,}1 \cdot 10^4$	0,79
Телевидение	$6 \cdot 10^6$	13,6	$8{,}2 \cdot 10^7$	0,72

Величина $FH = V/T$ представляет объем сигнала, отнесенный к единице времени. Разделив на эту величину I/T из табл. 1, найдем удельную содержательность сигнала

$$\nu = 1/V$$

(последний столбец табл. 2). Как видим, удельная содержательность сигналов различных видов связи оказывается одного по-

рядка, как оно и должно быть, так как эта величина лишь слабо зависит от числа градаций, а именно

$$\nu = \frac{1}{V} = \frac{\log m^2}{\log \frac{k^2}{6}(m-1)(2m-1)}.$$

Мы должны были бы получить наибольшую удельную содержательность для телеграфа ($m=2$). Это у нас не получилось, так как для ширины спектра мы взяли практические цифры, а превышение сигнала подсчитали по теоретической формуле.

Цифры, относящиеся к ширине спектра, показывают, что телевизионная передача (по крайней мере в ее современном виде) не может быть размещена в обычном вещательном диапазоне и практически может быть осуществлена только в диапазоне укв. Что касается телеграфа, то следует заметить, что увеличение быстродействия не изменяет удельной содержательности сигнала, так как объем сигнала увеличивается за счет расширения спектра ровно во столько же раз, во сколько увеличивается количество сведений.

Интересно еще сравнить те же виды связи по числу слов, которое можно передать в минуту (см. табл. 3).

Таблица 3

Вид связи	A, *слов/мин*	FH	A/FH
Телеграф (Морзе)	500	$1{,}8 \cdot 10^3$	280
Телеграф (Бодо)	60	$1{,}8 \cdot 10^2$	330
Фототелеграф*	60	$3{,}2 \cdot 10^4$	1,9
Телефон	120**	$7{,}1 \cdot 10^5$	1,7
Телевидение***	$4 \cdot 10^5$	$8{,}2 \cdot 10^7$	4,9

Примечания. * Предполагается, что бланк фототелеграммы заполнен текстом с допустимой плотностью, т. е. мелкой машинописью через один интервал.

** Приведенная цифра выражает среднюю скорость чтения готового текста. При разговоре число слов в минуту может быть значительно меньше.

*** Речь идет об использовании телевидения следующим образом: передается изображение текста, покрывающее весь экран при такой же относительной разрешающей способности, какая принята для фототелеграфа. Передается 25 различных текстов в секунду (частота кадров). Эта система известна под названием «Ультрафакс».

В табл. 3 обращает на себя внимание громадная скорость передачи текста при помощи телевидения, достигающая полумиллиона слов в минуту. Однако цифры, выражающие число слов в минуту, сами по себе еще недостаточно характерны. Нужно сопоставить их с объемом соответствующего сигнала. Если разделить число слов в минуту на произведение FH, что даст содержательность сигнала, выраженную числом слов, отнесенным

к объему сигнала, то картина радикально меняется. Для телевидения, фототелеграфа и телефона цифры оказываются одного порядка, а для телеграфа — раз в 100 больше. Из этого следует, что для передачи словесного текста телеграф является пока наивыгоднейшим видом связи.

Нужно еще заметить, что такое сложное и дорогое сооружение, как система связи, должно быть полностью использовано.

Прежде всего линия связи не должна иметь простоев и по возможности должна эксплуатироваться все 24 *час* в сутки за вычетом минимально необходимого времени на проверки и ремонты. Но этого мало. Если линия обладает полосой пропускания F_k, то нужно полностью использовать эту полосу. «Простой» линии по частоте совершенно так же убыточен, как простой во времени. То же самое относится и к использованию полосы уровней линии. Короче говоря, объем сигнала должен по возможности приближаться к емкости линии.

Если все же остаются неиспользованные резервы емкости, то они должны быть употреблены для организации многоканальной связи. Эти меры называются обычно *уплотнением* линии связи. Так, например, полоса пропускания дальних проводных магистралей используется обычно следующим образом: 0—80 *гц* — телеграф; 0,3—2,4 *кгц* — телефон (на звуковой частоте); 3,2—5,2 *кгц* — фототелеграф; 6,6—30 *кгц* — трехкратный двусторонний телефон (на высокой частоте); 40—80 *кгц* и 100—140 *кгц* — по 12 телефонных каналов (на высокой частоте). Кроме того, в порядке так называемого вторичного уплотнения на месте одного высокочастотного телефонного канала можно разместить до 18 каналов тонального телеграфа [3]. Подобная система характеризует постоянную тенденцию техники связи к наилучшему использованию линий; однако это далеко не предел. Как мы увидим дальше, теория указывает обширные, до сих пор не использованные техникой возможности.

Глава 2

ВОПРОСЫ СТАТИСТИЧЕСКОЙ ТЕОРИИ

§ 19. Определения

Эта глава посвящена некоторым применениям теории вероятностей к теории связи. Прежде чем приступить к рассмотрению этих применений, следует напомнить определения понятий и величин, которыми оперирует теория вероятностей и которые имеют отношение к нашей теме. Попытка кратко изложить основы теории вероятностей здесь не предпринимается; в этом нет нужды при наличии хороших современных руководств по теории вероятностей [4, 8].

Объектом изучения в теории вероятностей является *случайная величина*. Основное свойство этой величины состоит в том, что невозможно в точности предсказать ее значение; можно лишь утверждать, что она с известной вероятностью примет одно из возможных значений. Эти значения могут образовать либо дискретный набор, либо же случайная величина изменяется непрерывно с изменением какой-либо независимой переменной.

Важной характеристикой случайной величины является *распределение вероятностей*, показывающее, с какой вероятностью случайная величина принимает одно из своих возможных значений. Если случайная величина ξ может принимать одно из дискретного набора N значений ξ_k, то распределение вероятностей выражается таким же дискретным набором N чисел p_k, причем p_k означает вероятность величине ξ принять значение ξ_k. Так как при любых обстоятельствах ξ будет иметь одно из возможных значений, то

$$\sum_N p_k = 1.$$

Среднее значение величины ξ называется *математическим ожиданием* и определяется как

$$M(\xi) = \sum_N p_k \xi_k. \tag{1}$$

Если все значения ξ_k равновероятны, то вероятность любого значения составляет $1/N$, и математическое ожидание переходит в простое среднее арифметическое

$$\xi_{\text{ср}} = \frac{1}{N} \sum_N \xi_k.$$

Если величина ξ может изменяться непрерывно и если переменная шкалы значений ξ обозначена через x, то распределение характеризуется вероятностью того, что величина ξ примет значение, заключенное между $x=x_1$ и $x=x_2$. Эта вероятность выражается формулой

$$p\{x_1 \leqslant \xi \leqslant x_2\} = \int_{x_1}^{x_2} \varphi(x)\, dx. \tag{2}$$

Стоящая под знаком интеграла функция $\varphi(x)$ называется *плотностью распределения вероятностей*. Применяется также функция

$$F(x) = \int_{-\infty}^{x} \varphi(z)\, dz,$$

называемая *функцией распределения*.

Плотность вероятностей $\varphi(x)$ имеет обычно максимум, соответствующий наивероятнейшему значению ξ. С возрастанием

абсолютной величины x плотность вероятностей убывает, стремясь к нулю. Однако часто $\varphi(x)$ стремится к нулю лишь в пределе при $|x| \to \infty$. Это означает, что возможность появления очень больших значений ξ не исключена, но что вероятность таких значений мала.

Математическое ожидание случайной величины с непрерывным изменением выражается через плотность вероятностей следующим образом:

$$M(\xi) = \int_{-\infty}^{\infty} x\varphi(x)\,dx. \tag{3}$$

Еще одной важной характеристикой случайной величины является *дисперсия*, определяемая как

$$D(\xi) = M(\xi^2) - M^2(\xi). \tag{4}$$

Дисперсия характеризует средний квадрат уклонения случайной величины ξ от ее среднего значения. Такие случайные беспорядочные уклонения называют в физике *флюктуациями*.

Если случайная величина изменяется с изменением некоторой независимой переменной, в частности времени, то говорят о *случайном* (или *стохастическом*) *процессе*. Связь между случайной величиной ξ и временем t записывают в виде $\xi(t)$, понимая под этим, что в моменты t_1, t_2,. . ., t_k величина ξ может с известной степенью вероятности принять значения

$$\xi_1 = \xi(t_1); \quad \xi_2 = \xi(t_2); \quad \ldots; \quad \xi_k = \xi(t_k).$$

Различают процессы с *последействием* и без *последействия*. В первых последующие значения случайной величины в некоторой мере зависят от предшествующих. Для случайных же процессов без последействия все значения случайной величины совершенно независимы.

Случайный процесс называют *стационарным*, если его вероятностные характеристики не зависят от времени. Теория этого вида процессов наиболее разработана; в то же время именно эти процессы представляют большой интерес для техники.

Для многих стационарных процессов среднее значение, определяемое как математическое ожидание, в силу так называемой эргодической гипотезы совпадает со средним во времени, т. е.

$$M(\xi) = \int_{-\infty}^{\infty} xf(x)\,dx = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t)\,dt = \bar{\xi}. \tag{5}$$

При этом в силу стационарности начало отсчета времен безразлично. Аналогично определяется и средний квадрат

$$M(\xi^2) = \int_{-\infty}^{\infty} x^2\varphi(x)\,dx = \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi^2(t)\,dt = \bar{\xi^2}. \tag{6}$$

Для дисперсии имеем

$$D(\xi) = \overline{\xi^2} - (\xi)^2 = \lim_{T=\infty} \frac{1}{T} \left[\int_0^T \xi^2(t)\, dt - \left(\int_0^T \xi(t)\, dt \right)^2 \right]. \qquad (7)$$

Если ξ может принимать как положительные, так и отрицательные значения, так что, как это часто бывает, среднее значение равно нулю, то дисперсия совпадает со средним квадратом.

Основные свойства случайного процесса описываются *функцией корреляции*. Она устанавливает вероятностную связь между значениями ξ в зависимости от их удаления друг от друга (во времени).

Функция корреляции определяется как

$$B(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \xi(t)\, \xi(t-\tau)\, dt = \overline{\xi\xi}_\tau. \qquad (8)$$

При сдвиге $\tau=0$ под знаком интеграла получаем квадрат величины ξ и, следовательно,

$$B(0)^\xi = \xi^2. \qquad (9)$$

Часто пользуются нормированной функцией корреляции, определяемой соотношением

$$b(\tau) = \frac{1}{\xi^2} B(\tau). \qquad (10)$$

Случайный процесс можно также характеризовать его статистическим спектром $G(\omega)$.

Спектр $G(\omega)$ связан с функцией корреляции $B(\tau)$ парой косинус-трансформаций Фурье (обе функции четны):

$$G(\omega) = \frac{2}{\pi} \int_0^\infty B(\tau) \cos \omega\tau d\tau, \qquad (11)$$

$$B(\tau) = \int_0^\infty G(\omega) \cos \omega\tau d\omega. \qquad (12)$$

Величины ξ^2 и $G(\omega)$ имеют простой физический смысл. Средний квадрат величины ξ выражает среднюю мощность процесса (если, например, ξ означает ток или напряжение). Спектр $G(\omega)$ есть не что иное, как средняя спектральная плотность мощности.

Функция корреляции отражает степень последействия. Чем больше интервал корреляции, т. е. интервал значений τ, в котором функция корреляции имеет еще заметную величину, тем более удаленные значения случайной величины имеют между собой связь. Для процессов без последействия функция корреляции имеет разрывно́й характер: она равна единице (речь идет о норми-

рованной функции $b(\tau)$) при $\tau=0$ и нулю при всех других значениях τ. Соответственно этому спектр процессов без последействия имеет бесконечную протяженность и однороден. Вообще же, как это следует из теории преобразований Фурье, интервал корреляции τ_0 и ширина спектра F связаны между собой соотношением

$$\tau_0 F = \mu,$$

где μ — постоянная порядка единицы. Таким образом, всякое ограничение спектра увеличивает корреляцию.

Возвращаясь к функциям распределения, заметим, что наибольшее значение в теории имеет так называемое *нормальное распределение*, при котором плотность вероятностей выражается функцией

$$\varphi(x) = \frac{\alpha}{\sqrt{\pi}} e^{-\alpha^2 (x-a)^2}. \tag{13}$$

Точка $x=a$, отвечающая максимуму $f(x)$, называется *центром распределения*. Величина α называется *мерой точности*; чем α больше, тем круче спадает кривая $f(x)$ по обе стороны от $x=a$, т. е. чем более «кучно» распределены вероятности, тем быстрее убывает вероятность значений, отличных от наивероятнейшего.

Вычислим дисперсию случайной величины с нормальным распределением, положив для простоты $a=0$, т. е. полагая распределение симметричным

$$\varphi(x) = \frac{\alpha}{\sqrt{\pi}} e^{-\alpha^2 x^2}. \tag{14}$$

При симметричном распределении математическое ожидание равно нулю и дисперсия совпадает со средним квадратом. Для этой величины мы имеем общее выражение

$$M(\xi^2) = \int_{-\infty}^{\infty} x^2 \varphi(x)\, dx.$$

Подставляя сюда (14) и выполняя интегрирование, находим

$$M(\xi^2) = D(\xi) = 1/2\alpha^2.$$

Введем среднеквадратичное значение величины ξ

$$\sigma = \sqrt{M(\xi^2)} = 1/\sqrt{2}\alpha.$$

Теперь мы можем записать выражение для плотности вероятностей при нормальном распределении, выразив меру точности α через среднеквадратичное значение σ

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}. \tag{15}$$

Наблюдаемые в действительности распределения случайных величин очень часто приближаются к нормальному. Это обстоятельство находит теоретическое обоснование в знаменитой теореме Ляпунова, согласно которой (при весьма общего характера допущениях) распределение среднего арифметического суммы независимых случайных величин в пределе (при увеличении числа слагаемых) сходится к нормальному распределению.

Одним из важнейших положений теории вероятностей является так называемый *закон больших чисел*. Сущность этого закона состоит в том, что хотя среднее арифметическое суммы n случайных величин (или n значений случайной величины) отличается от среднего по всему ансамблю (т. е. от математического ожидания), но с увеличением n вероятность того, что эта разница будет меньше сколь угодно малой величины, стремится в пределе к единице. Это утверждение выражается теоремой Чебышева

$$\lim_{n\to\infty} p\{|Y_n - M(\xi)| < \varepsilon\} = 1, \tag{16}$$

где

$$Y_n = \frac{1}{n}\sum_n \xi_k,$$

n — число слагаемых; ε — произвольное положительное число. Оценка разности в формуле (16) дается неравенством Чебышева

$$p\{|Y_n - M(\xi)| < \varepsilon\} > 1 - \frac{D(Y_n)}{\varepsilon^2}. \tag{17}$$

Для правильного понимания закона больших чисел нужно иметь в виду, что разность $Y_n - M(\xi)$ не стремится к нулю с увеличением n; она колеблется (флюктуирует) около нуля. Закон больших чисел утверждает, что с увеличением n вероятность больших флюктуаций убывает, но возможная величина флюктуации при этом не ограничивается.

Кроме простых вероятностей нам придется применять *условные вероятности*. Понятие условной вероятности связывает вероятность некоторого события с условием осуществления другого события, предполагая, что оба события не независимы. Обозначим через $p(j)$ вероятность того, что случайная величина примет значение j (простая вероятность). Если же мы пожелаем определить вероятность появления значения j при условии, что предыдущее значение было i, то условная вероятность такого события будет записана как $p(j/i)$. Вероятность последовательного появления двух значений i и j, обозначаемая $p(ij)$, будет равна вероятности появления значения i, умноженной на условную вероятность появления значения j при условии наличия i, т. е.

$$p(ij) = p(i)\,p(j/i). \tag{18}$$

Полная вероятность появления значения j определится как сумма $p(ij)$ для всех возможных значений i

$$p(j) = \sum_i p(ij) = \sum_i p(i)\, p(j/i). \tag{19}$$

Аналогично

$$p(i) = \sum_j p(ij) = \sum_j p(j)\, p(i/j). \tag{20}$$

Таким образом, условную вероятность появления j после i можно записать в виде

$$p(j/i) = \frac{p(ij)}{p(i)} = \frac{p(ij)}{\sum\limits_j p(ij)}. \tag{21}$$

Если появления значений i и j — события независимые, то от условных и полных вероятностей мы возвращаемся к простым

$$p(j/i) = p(j), \quad p(ij) = p(i)\, p(j),$$

$$p(j) = \sum_i p(i)\, p(j/i) = p(j) \sum_i p(i) = p(j)$$

и т. д.

Мы применим теперь методы и определения теории вероятностей к исследованию некоторых проблем теории связи, начиная с простейших.

§ 20. Оптимальный неравномерный код

В § 7 упоминался код Бодо. Этот код относится к числу равномерных кодов, т. е. каждая его комбинация состоит из одинакового числа элементов (из пяти) и имеет, следовательно, одинаковую длину.

Однако возможны и применяются неравномерные коды, составленные из комбинаций различной длины. Таков код Морзе, являющийся, как и код Бодо, двоичным кодом (элементы — точка и тире), но применяющий для букв комбинации с числом элементов от одного до четырех, для цифр — комбинации из пяти элементов и, наконец, шестиэлементные комбинации для разного рода знаков. Некоторые из этих комбинаций приведены ниже (0 означает точку, 1 — тире):

Буква (знак) . . .	А	Б	В	Г	Д	Е	Ж	2	зпт	?
Код	01	1000	011	110	100	0	0001	00111	010101	001100

Нам желательно, чтобы средняя длина комбинации была наименьшей. Но для этого, очевидно, нужно присвоить более короткие кодовые комбинации более часто встречающимся буквам и сохранить более длинные комбинации для редких букв. Именно по такому принципу и был в свое время составлен код Морзе.

Наша постановка задачи — типично вероятностная постановка. Средняя длина комбинации выражается непосредственно математическим ожиданием. Если длина комбинации (число элементов в комбинации) есть l_k и если вероятность соответствующей этой комбинации k-й буквы есть p_k, то средняя длина комбинации будет

$$l_{\text{ср}} = \sum_{k=1}^{m} p_k l_k.$$

Если бы все буквы имели одинаковую вероятность, то вместо математического ожидания мы выразили бы среднюю длину комбинации простым средним

$$l_0 = \frac{1}{m} \sum_{k=1}^{m} l_k$$

(так как в этом случае вероятность каждой буквы составляла бы $1/m$).

Для получения наименьшего значения $l_{\text{ср}}$ нужно поступить так: расположить возможные кодовые комбинации в порядке возрастающей длины, а все подлежащие кодированию буквы и знаки — в порядке убывающей вероятности. Нужно заметить, что вероятности букв в разных языках существенно различны. Например, в английском языке чаще всего встречается буква Е. Код Морзе присваивает ей наиболее короткое кодовое обозначение — точку. В русском же языке чаще всего встречается буква О ($p \approx 0{,}11$), имеющая в принятом коде довольно длинное обозначение — три тире. Таким образом, принятый у нас код Морзе можно было бы усовершенствовать. Это дало бы сокращение средней длины комбинации примерно на 8% (см. добавление 2). Экономия такого порядка едва ли оправдала бы необходимость переучивания всех телеграфистов, работающих кодом Морзе.

§ 21. Неравномерный код без разделительных знаков

Код Морзе состоит из точек и тире. Для отделения друг от друга точек и тире употребляется пауза, длительность которой принята равной длительности точки. Так как пауза обязательно следует за каждой точкой и за каждым тире, то код Морзе, собственно говоря, состоит не из элементов точка и тире, а из элементов точка—пауза и тире—пауза. Но существенно не это, а то, что каждая кодовая комбинация должна отделяться от другой специальным разделительным знаком. В коде Морзе таким знаком служит длинная пауза, длительность которой равна длительности тире (т. е. тройной длительности точки). Таким образом, в составе кода Морзе появляется третий элемент, и код из двоичного превращается в троичный. Для отдельных букв требуются два элемента (двоичный код), а для текста — уже три элемента (троич-

ный код). Эти три элемента изображены на рис. 10. Мы записываем их как 0, 1, —.

В равномерном коде, например в коде Бодо, надобности в разделении нет, так как все комбинации имеют одинаковую длину, что позволяет безошибочно принять отдельную комбинацию, как это происходит хотя бы в старт-стопных аппаратах. В коде Морзе такое выделение комбинации невозможно. Пусть, например, передано слово «ДА», кодированное в виде 100—01. Если разделительный знак выпадет, то получившаяся последовательность 10001 может быть прочитана самым различным образом: НУ

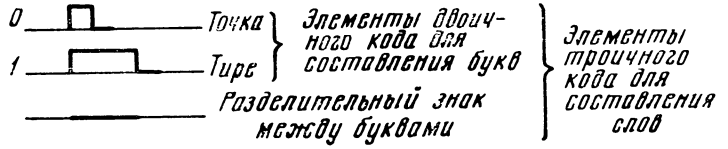


Рис. 10

(10—001), ТЖ (1—0001), БТ (1000—1), ТСТ (1—000—1) и даже НИТ — почти НЕТ! (10—00—1) и т. д.

Тем не менее возможно построить двоичный код, не требующий разделительных знаков. Общая идея состоит в том, что не должны применяться комбинации, начальная часть которых уже использована в качестве самостоятельной комбинации. Так, например, можно применить комбинации 10 и 001, но нельзя применить комбинации 10 и 100, так как если передано 10, то неизвестно, передана ли полностью комбинация 10 или первые два элемента комбинации 100.

Способ составления кода по этому принципу очень наглядно иллюстрируется диаграммой рис. 11. Эта диаграмма подобна обычному ключу для чтения кода Морзе. Однако в коде Морзе каждая узловая точка соответствует определенной комбинации, использованной в коде. Код же без разделительных знаков строится по следующему правилу: на пути от вершины пирамиды к любой использованной в коде комбинации не должна встречаться никакая другая комбинация. Таким образом, выбор некоторой комбинации «запирает» все дальнейшие разветвления диаграммы и исключает все отвечающие этим разветвлениям комбинации. На рис. 11 кружками отмечены использованные комбинации: 1, 010, 011, 0000, 0001, 00100, 00101, 00110, 00111. Любая непрерывная последовательность этих комбинаций разделяется вполне определенным и единственным образом. Так, например, последовательность

00100111001110100110001

разделяется следующим образом:

00100—1—1—1—00111—010—011—0001.

Для разделения удобно пользоваться диаграммой рис. 11 в качестве ключа. Пользование диаграммой состоит в том, что последо-

вательными шагами от вершины пирамиды вправо (1) и влево (0), начиная с первой цифры последовательности, мы доходим до кружка, обозначающего одну из кодовых комбинаций. Поставив разделительный знак, возвращаемся к вершине пирамиды для отыскания следующей комбинации и т. д. Описанную операцию нетрудно механизировать.

Если требуется построить код без разделительных знаков, учитывающий статистику сообщений, т. е., например, вероятность различных букв, то порядок построения такого кода (имеется в виду двоичный код) состоит в следующем. Все буквы (или вообще

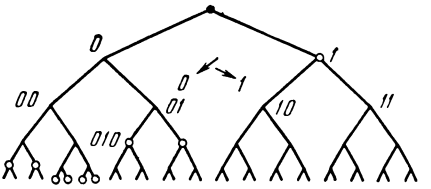


Рис. 11

все сообщения, подлежащие кодированию) записываются в порядке убывающей вероятности. Записанная последовательность разбивается на две группы так, чтобы суммарные вероятности распределились между группами по возможности поровну. Затем каждая группа разбивается на две подгруппы с соблюдением того же условия равенства вероятностей. Это деление продолжается до тех пор, пока в подгруппах не останется лишь по одному сообщению. Процесс деления представлен графически на рис. 12 для некоторого произвольного ансамбля сообщений. Римскими цифрами обозначены номера последовательных этапов деления. Когда деление закончено и каждая буква нашла свое место на диаграмме, то кодовое обозначение определяется, как указано раньше: шаг влево дает цифру 0, шаг вправо — цифру 1. Получаемый в результате код, соответствующий построению рис. 12, дан в табл. 4.

Таблица 4

Сообщение k	Вероятность p_k	Код	Число знаков l_k	$p_k l_k$	Сообщение k	Вероятность p_k	Код	Число знаков l_k	$p_k l_k$
1	0,3	00	2	0,6	9	0,03	11011	5	0,15
2	0,2	01	2	0,4	10	0,02	111000	6	0,12
3	0,1	100	3	0,3	11	0,02	111001	6	0,12
4	0,1	1010	4	0,4	12	0,02	11101	5	0,10
5	0,05	1011	4	0,2	13	0,02	11110	5	0,10
6	0,05	11000	5	0,25	14	0,01	111110	6	0,06
7	0,03	11001	5	0,15	15	0,01	1111110	7	0,07
8	0,03	11010	5	0,15	16	0,01	1111111	7	0,07

Первые два столбца воспроизводят данные рис. 12: номер сообщения и соответствующую вероятность p_k. В третьем столбце дано полученное кодовое обозначение. В четвертом столбце дана длина сообщения l_k, выраженная числом знаков в кодовой комбинации. Наконец, в последнем столбце дано произведение $p_k l_k$. Средняя длина комбинации равна математическому ожиданию; в нашем примере

$$l_{\text{ср}} = \sum_{1}^{16} p_k l_k = 3{,}24$$

(при наименьшей длине, равной двум, и наибольшей, равной семи).

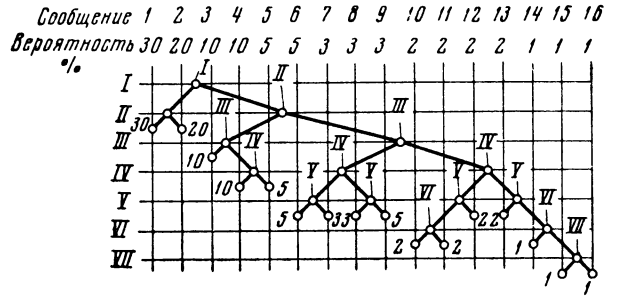


Рис. 12

Можно очень просто доказать, что построенный таким способом код является оптимальным, т. е. что не может быть построен код с меньшей средней длиной комбинации. Доказательство основано на следующем рассуждении: положим, что в последовательности произведений $p_k l_k$ имеются два члена $p_m l_m$ и $p_n l_n$, причем $p_m > p_n$ и $l_m > l_n$, т. е. более вероятному сообщению присвоена более длинная кодовая комбинация. Тогда выгодна перемена местами кодовых комбинаций, так как

$$p_m l_m + p_n l_n > p_m l_n + p_n l_m,$$

и, следовательно, перестановка уменьшает сумму $p_k l_k$, т. е. среднюю длину комбинации [1]. Но, как легко видеть, при описанном

---

[1] Справедливость этого неравенства можно доказать следующим образом. Перепишем его в виде

$$1 + \frac{p_n}{p_m} \cdot \frac{l_n}{l_m} > \frac{l_n}{l_m} + \frac{p_n}{p_m}.$$

Так как $p_n/p_m < 1$ и $l_n/l_m < 1$, то, введя обозначения $p_n/p_m = 1 - \varepsilon_1$, $l_n/l_m = 1 - \varepsilon_2$, где ε_1 и ε_2 — положительные числа, меньшие единицы, получим

$$1 + (1 - \varepsilon_1)(1 - \varepsilon_2) > 2 - \varepsilon_1 - \varepsilon_2,$$

или, после приведения, $\varepsilon_1 \varepsilon_2 > 0$, и неравенство доказано.

способе построения кода такой случай невозможен. Возможны лишь случаи, когда одной и той же длине комбинации соответствуют разные вероятности (например, № 6 и 9) или, наоборот, одной и той же вероятности соответствуют разные длины (например, № 10 и 13). Но в этих случаях, очевидно, перестановка не изменяет средней длины.

Итак, мы построили оптимальный неравномерный код без разделительных знаков. Этот код называют иногда кодом Шэннона—Фэно.

§ 22. Белый шум

Мы переходим теперь к вопросу о статистических характеристиках помех. Помехи, с которыми приходится сталкиваться в связи, довольно разнообразны и по своему происхождению, и по характеристикам. Однако один вид помех выделяется как своей распространенностью, так и тем, что он принципиально не может быть устранен. Речь идет о помехах, порождаемых разного рода флюктуациями: флюктуациями проводимости, флюктуациями эмиссии и тому подобными явлениями, происходящими в элементах аппаратуры связи. Эти флюктуации обусловлены дискретной микроструктурой ряда физических явлений, проявляющейся в определенных условиях. Возьмем, к примеру, поведение газа в некотором объеме. Статистические параметры, как, например, давление, испытывают в обычных условиях незначительные флюктуации, так как ансамбль, определяемый числом молекул или ионов, необычайно велик. Но в условиях высокого вакуума число молекул настолько сокращается, что флюктуации становятся уже заметными. В известных условиях, например, тепловое движение может произвести заметный механический эффект, вроде раскачивания легкого маятника. Аналогичным образом проявляется и дискретная структура электрического тока: она обнаруживается в электронных лампах в виде так называемого дробового эффекта.

Все подобного рода явления создают помеху определенного типа, называемую «белым шумом». Ему, как случайному процессу, приписываются определенные вероятностные характеристики, а именно: предполагается, что он описывается нормальным (гауссовым) распределением и что он относится к числу процессов без последействия. Последнее означает, что белый шум обладает неограниченным однородным спектром [1]. Функция корреляции $b(\tau)$ белого шума равна единице при $\tau=0$ и нулю при всех ненулевых значениях τ.

---

[1] Этим и объясняется эпитет «белый». Речь идет об аналогии с белым светом, имеющим сплошной и более или менее однородный (в пределах видимой части) спектр.

В теоретических исследованиях, связанных с влиянием случайных помех, рассматривается преимущественно именно помеха типа белого шума. Это оправдывается, с одной стороны, распространенностью белого шума, о чем уже говорилось, с другой — простотой его математического описания. Кроме того, нужно учесть, что все другие виды помех имеют не равную нулю корреляцию и, следовательно, ограниченный спектр; поэтому они перестают, как это всем известно, оказывать влияние в диапазоне очень коротких волн (от метровых и ниже), который как раз в современной технике связи приобретает все большее значение. Белый же шум в силу неограниченности своего спектра мешает на любых частотах и является поэтому особо зловредным видом помехи. Исходя из всех этих соображений, мы ограничимся рассмотрением помехи типа белого шума.

Белый шум можно представить себе как последовательность бесконечно коротких импульсов, имеющих случайную высоту и следующих друг за другом через случайные промежутки времени. Если бы импульсы следовали друг за другом беспорядочно во времени, но имели бы одинаковую высоту, то это была бы помеха со сплошным спектром, но без всякого распределения вероятностей по высоте импульсов. Если же импульсы имели бы случайную высоту, но следовали бы друг за другом через равные промежутки времени, то спектр такой последовательности был бы неоднороден [2].

Белый шум в вышеприведенном определении представляет собой асбстракцию. Если бы спектр его был действительно неограничен, то это означало бы, что мощность шума, приходящаяся на конечную полосу частот, бесконечно мала. С другой стороны, импульсы, возникающие в результате реального физического процесса, не могут быть бесконечно короткими. Поэтому реальный белый шум имеет ограниченный, хотя и далеко простирающийся, спектр и состоит из импульсов конечной, но очень малой длительности [2]. Ширина спектра и длительность импульсов связаны между собой известным соотношением: их произведение имеет порядок единицы. При ограниченной ширине спектра мощность шума в конечной полосе частот конечна; поэтому оперируют спектральной плотностью мощности $G(\omega)=dP/d\omega$, полагая ее для белого шума постоянной. Эту постоянную мы будем обозначать буквой ρ. Таким образом, мощность белого шума в полосе частот F, *гц*, будет

$$P=\rho 2\pi F.$$

---

[1] На сплошной спектр был бы наложен линейчатый спектр периодической составляющей процесса, отсутствующий лишь при условии, что среднее значение последовательности равно нулю.

[2] Обычно полагают, что спектр реального белого шума простирается до частот порядка 10^{12}—10^{1} *гц*. Напомним, что колебания видимой части светового спектра имеют частоту порядка 10^{14} *гц*.

Для нас имеет первостепенное значение вопрос о связи между средним значением (средним квадратом, т. е. мощностью или среднеквадратичным значением) шума и вероятностью больших пиковых значений. Эта связь полностью определяется распределением вероятностей, и мы сейчас выведем нужные нам соотношения.

Средний квадрат (мощность) белого шума выражается через плотность вероятностей формулой

$$\sigma^2 = M(\xi^2) = \int_{-\infty}^{\infty} x^2 \varphi(x)\, dx.$$

Мы предполагаем нормальное распределение в виде

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}.$$

Определим вероятность больших флюктуаций, т. е. вероятность появления мгновенных значений, превосходящих по абсолютной величине некоторую заданную величину ε. Эта вероятность равна

$$p\{|\xi| > \varepsilon\} = 2\int_{\varepsilon}^{\infty} \varphi(x)\, dx = \\ = \frac{2}{\sqrt{2\pi}\,\sigma} \int_{\varepsilon}^{\infty} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}} dx = 1 - \Phi\left(\frac{\varepsilon}{\sqrt{2}\,\sigma}\right),$$

где Φ — символ функции Лапласа (интеграл вероятностей).

Нас интересует сравнение случайных выбросов помехи с величиной сигнала. В § 10, где разбирались вопросы квантования, говорилось о том, что ошибки при передаче сигнала не произойдет, если помеха не превосходит половины шага шкалы квантования δ. Теперь мы можем уточнить эту формулировку: под помехой понимается возможное пиковое значение помехи. Появление сколь угодно большого значения не исключено; нам следует лишь позаботиться о том, чтобы такое событие было маловероятно, т. е. чтобы число возможных ошибок было достаточно мало. Это есть по существу требование надежности связи.

Определим произвольное число ε, введенное в предыдущих вычислениях, как

$$\varepsilon = \delta/2,$$

так как при таком значении помехи уже возможна ошибка при приеме. Тогда аргумент функции Лапласа запишется в виде

$$\frac{\varepsilon}{\sqrt{2}\,\sigma} = \frac{1}{2\sqrt{2}} \cdot \frac{\delta}{\sigma}.$$

Итак, вероятность ошибки зависит от отношения δ/σ. Увеличивая это отношение, мы уменьшаем вероятность ошибки и повышаем надежность связи.

Подсчитаем зависимость вероятности ошибки от отношения высоты ступеньки δ к среднему значению помехи σ. При малых значениях аргумента можно воспользоваться таблицами[1] функции $\Phi(x)$, для больших же аргументов удобно применить асимптотическое разложение функции $\Phi(x)$, сохранив только первый член разложения. Это дает[2]

$$\Phi(x) \approx 1 - \frac{1}{\sqrt{\pi}} \cdot \frac{e^{-x^2}}{x} \quad [x \gg 1]$$

и, следовательно,

$$p\left\{|\xi| \geqslant \frac{\delta}{2}\right\} = 1 - \Phi\left(\frac{1}{2\sqrt{2}} \cdot \frac{\delta}{\sigma}\right) \approx 2\sqrt{\frac{2}{\pi}} \cdot \frac{e^{-\frac{1}{8}(\delta/\sigma)^2}}{\delta/\sigma}.$$

Вычисление дает следующие результаты:

δ/σ	1	2	3	4	5	6	7
p	0,62	0,32	0,13	$4{,}6 \cdot 10^{-2}$	$1{,}2 \cdot 10^{-2}$	$3{,}0 \cdot 10^{-3}$	$4{,}9 \cdot 10^{-4}$

δ/σ	8	9	10	11	12	13
p	$6{,}6 \cdot 10^{-5}$	$7{,}1 \cdot 10^{-6}$	$6{,}0 \cdot 10^{-7}$	$3{,}8 \cdot 10^{-8}$	$2{,}0 \cdot 10^{-9}$	$8{,}3 \cdot 10^{-11}$

Отсюда видно, что вероятность ошибки чрезвычайно быстро убывает с увеличением отношения δ/σ; так, при увеличении этого отношения на одну единицу с девяти до десяти, т. е. на 10%, вероятность ошибки убывает вдесятеро. Связь при отношении δ/σ порядка единицы невозможна, так как вероятность ошибки слишком велика. Для наглядности подсчитаем еще, как часто в среднем может произойти ошибка в приеме одного импульса, если в секунду передается 10 000 импульсов, как в современной импульсной телефонии. Средний интервал t_0 безошибочной связи, определяемый как

$$t_0 = \Delta t/p,$$

---

[1] При пользовании таблицами следует соблюдать осторожность, так как встречаются различные определения функции $\Phi(x)$. У нас

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz.$$

[2] *И. М. Рыжик, И. С. Градштейн.* Таблицы интегралов, сумм, рядов и произведений. Гостехиздат, 1951, стр. 327.

составит при этих условиях

δ/σ	1	2	3	4	5	6
t_0			$8 \cdot 10^{-4}$ *сек*	$2 \cdot 10^{-3}$ *сек*	$8 \cdot 10^{-3}$ *сек*	$3 \cdot 10^{-2}$ *сек*

δ/σ	7	8	9	10	11	12	13
t_0	0,2 *сек*	1,5 *сек*	14 *сек*	2,8 *мин*	44 *мин*	14 *час*	14 *сут*

Как видно, при отношении δ/σ порядка десяти можно уже считать связь достаточно надежной. Значение $\delta/\sigma=10$ очень критично в том смысле, что при $\delta/\sigma < 10$ вероятность ошибки недопустимо велика, а при увеличении отношения δ/σ сверх десяти вероятность ошибки быстро становится так исчезающе мала, что такое увеличение не имеет практически никакого смысла.

Итак, при наличии помехи в виде белого шума со среднеквадратичным значением σ для обеспечения надежной связи шаг шкалы уровней при квантовании следует выбирать исходя из условия $\delta \geqslant 10\sigma$. Мы определили, таким образом, численное значение коэффициента запаса $k=\delta/\sigma$, введенного в § 14.

§ 23. Вероятностные характеристики сигналов

Сигнал связи должен рассматриваться как случайный процесс, так как никому, кроме отправителя, сообщение в целом неизвестно, а следовательно, нельзя предвидеть течение во времени функции, представляющей соответствующий сообщению сигнал. Можно лишь говорить о вероятностных характеристиках сигнала, так как он строится из элементов некоторого ансамбля. Такими характеристиками могут являться функции корреляции и спектр (не считая средней мощности сигнала).

Мы рассмотрим некоторые типичные виды сигналов и выведем выражения для названных характеристик.

Начнем с обобщенного телеграфного сигнала. Предположим, что передача происходит посредством посылок тока разного знака и произвольной длительности $\xi(t)$. Тогда ток в цепи имеет форму, как, например, на рис. 13. Форма эта характерна тем, что ток может принимать только два значения $+h$ и $-h$, равные по абсолютной величине, но моменты перемены знака случайны [1].

Введем в рассмотрение число перемен знака или число нулей, как часто выражаются. Если на протяжении времени τ число нулей четно, то по истечении этого времени знак ξ сохранится; если же нечетно, то знак ξ изменится на обратный.

Функция корреляции определяется средним значением произведения $\xi=\xi(t)$ на $\xi_\tau=\xi(t-\tau)$. Это произведение в нашем случае равно либо h^2, либо $-h^2$ в зависимости от того, четно или нечетно число нулей $\xi(t)$ на интервале τ (ясно, что число нулей может быть только целым).

[1] К такому же виду приводится сигнал в случае так называемой ограниченной речи.

Но различные числа нулей на данном интервале неравновероятны. Здесь мы встречаемся с понятием о распределении вероятностей нулей или плотности нулей. Появление нуля в данном интервале есть событие независимое. Можно показать [14], что вероятность появления в интервале τ числа нулей, равного k, подчиняется так называемому распределению Пуассона и равна

$$p(k)=\frac{(\mu\tau)^k}{k!}e^{-\mu\tau}, \qquad (1)$$

где μ — среднее количество нулей в единицу времени. Очевидно, что в случае телеграфного сигнала эта величина определяется частотой манипуляции.

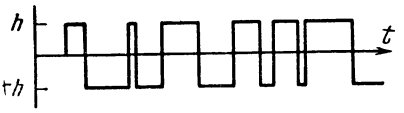


Рис. 13

Теперь мы можем записать выражение для функции корреляции, суммируя вероятности четных и нечетных чисел нулей отдельно. Мы получаем

$$B(\tau)=\overline{\xi\xi_\tau}=h^2\left[\sum_{k=0}^{\infty}p(2k)-\sum_{k=0}^{\infty}p(2k+1)\right].$$

Подставляя сюда распределение (1) и суммируя ряды, находим

$$B(\tau)=h^2e^{-2\mu|\tau|}.$$

Отсюда находим и спектр

$$G(\omega)=\frac{2}{\pi}\int_0^{\infty}B(\tau)\cos\omega\tau d\tau=$$

$$=\frac{2}{\pi}h^2\int_0^{\infty}e^{-\mu\tau}\cos\omega\tau d\tau=\frac{4}{\pi}\frac{h^2\mu}{\omega^2+4\mu^2}.$$

Графики функции корреляции и спектра изображены на рис. 14, *а* и *б* соответственно.

Вырожденным случаем предыдущего является телеграфный сигнал, передаваемый равномерным двоичным кодом (например, кодом Бодо). В этом случае сигнал состоит из посылок $+h$ и $-h$ равной длительности Δt. Вероятность появления очередной посылки $+h$ составляет $^1/_2$, как и вероятность появления посылки $-h$. Поэтому вероятность любой комбинации из k посылок составляет $(^1/_2)^k$, т. е. все комбинации (их общее число равно 2^k) предполагаются равновероятными, а стало быть, равновероятны четные и нечетные числа нулей. Применяя это заключение в предыдущем выводе, находим, что функция корреляции в рассматриваемом

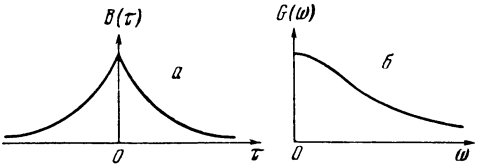


Рис. 14

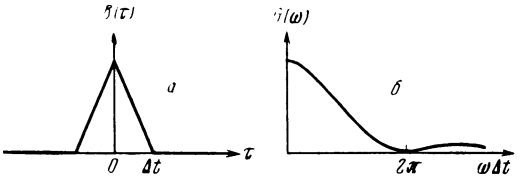


Рис. 15

случае должна равняться нулю. Однако очевидно, что при $\tau=0$ функция корреляции равна h^2; она линейно убывает до нуля на интервале $0 < \tau < \Delta t$. Таким образом, функция корреляции имеет треугольный характер, представленный на рис. 15, *а*. Соответствующий спектр будет

$$G(\omega)=\frac{2}{\pi}h^2\int_0^{\Delta t}\left(1-\frac{\tau}{\Delta t}\right)\cos\omega\tau d\tau=\frac{h^2}{\pi}\Delta t\,\frac{1-\cos\omega\Delta t}{\frac{1}{2}(\omega\Delta t)^2}$$

(рис. 15, *б*).

Рассмотрим еще последовательность импульсов заданной формы, но случайной величины, следующих друг за другом через равные промежутки времени. Это — случай амплитудно-импульсной модуляции (рис. 16). Положим для упрощения, что среднее значение последовательности равно нулю.

Импульсы следуют друг за другом через равные интервалы Δt. Форма импульсов одинакова, т. е. выражается одной и той же функцией времени $f(t)$, но величина различна и изменяется случайно. Следовательно, аналитически k-й импульс выражается как

$$h_k f(t-k\Delta t),$$

где h_k — масштабный коэффициент, флюктуирующий около нуля.

Положим, что длительность каждого импульса ограничена и равна τ_0. Тогда, как нетрудно сообразить, функция корреляции будет равна нулю вне промежутка $-\tau_0 < \tau < \tau_0$.

Действительно, по определению функция корреляции равна среднему значению произведения данной последовательности на такую же последовательность, но сдвинутую относительно первой на τ. Если $|\tau| > \tau_0$, то импульсы второй последовательности либо попадут в промежутки между импульсами первой, либо произойдет частичное наложение импульсов. В первом случае само произ-

ведение будет, очевидно, нуль — это тривиально, — во втором же случае среднее значение произведения даст нуль в силу того, что положительные и отрицательные значения импульсов, а следовательно, и их произведений равновероятны по нашему предположению.

Разбивая ось времени на отрезки длиной Δt, мы можем записать для функции корреляции

$$B(\tau) = \lim_{n\to\infty} \frac{1}{n\Delta t} \sum_{k=0}^{n} h_k^2 \int_{k\Delta t}^{(k+1)\Delta t} f(t - k\Delta t) f(t - k\Delta t - \tau)\, dt.$$

Но интеграл в правой части не зависит от переменной суммирования k; он представляет собой некоторую характеристику одиночного импульса, выражаемого функцией $f(t)$. Обозначая

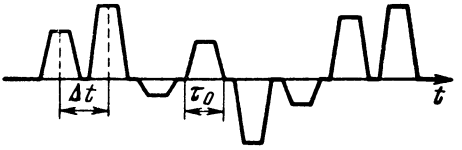


Рис. 16

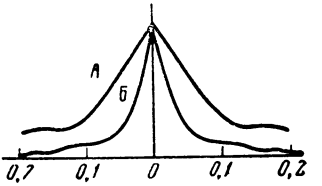


Рис. 17

$$B_0(\tau) = \frac{1}{\Delta t} \int_0^{\Delta t} f(t) f(t - \tau)\, dt,$$

получим

$$B(\tau) = B_0(\tau) \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} h_k^2 = \bar{h}^2 B_0(\tau).$$

Средний квадрат множителя h находится по заданному распределению вероятностей.

Спектр мощности рассматриваемой последовательности с точностью до постоянного множителя совпадает со спектром одиночного импульса.

Мы рассмотрели некоторые примеры, в которых функция корреляции может быть найдена аналитически. Однако в действительности часто приходится иметь дело с сигналами более сложной природы, для которых аналитическое определение корреляции невозможно. Таковы, например, сигналы вещания и телевидения, т. е. сигналы при передаче речи и музыки или движущихся изображений самого различного характера. В этих случаях статистические характеристики сигналов, и в частности функции корреляции, могут определяться экспериментально. В качестве примера

приведем графики функций корреляции для телевизионного сигнала, точнее, для подлежащего передаче изображения. Функция корреляции, разумеется, существенно зависит от характера изображения. Чем мельче детали изображения, тем быстрее спадает функция корреляции (и тем шире соответствующий спектр). На рис. 17 изображены полученные экспериментально функции корреляции A и $Б$, относящиеся к сдвигу в вертикальном направлении для изображений: A — лицо крупным планом, $Б$ — зрители на трибуне общим планом [22]. По оси ординат отложена нормированная функция корреляции, по оси абсцисс — относительный сдвиг двух изображений, выраженный в долях соответствующего (в данном случае вертикального) размера кадра. Можно отметить значительно большую корреляцию для изображения A.

§ 24. Количество сведений при неравновероятности элементов сообщения

В § 11 введено определение количества сведений. Эта величина определена как логарифм числа сообщений, которые можно составить из m различных элементов при общем числе их, равном n. При этом предполагалось, что все m элементов равновероятны, т. е. что на протяжении сообщения каждый элемент встречается число раз, стремящееся в среднем к n/m, когда длина сообщения (т. е. общее число элементов n) неограниченно возрастает [1].

Однако в действительности элементы сообщения неравновероятны, и это обстоятельство влияет на количество сведений, содержащееся в сообщении. Именно при неравновероятности элементов число возможных сообщений меньше, чем m^n; соответственно уменьшается и количество сведений. Нам нужно теперь обобщить определение количества сведений, распространив его на случай, когда элементы сообщения неравновероятны. Пусть имеется набор из m элементов

$$h_1,\ h_2,\ \ldots,\ h_i,\ \ldots,\ h_m,$$

вероятности появления которых неравны и выражены соответственно как

$$p_1,\ p_2,\ \ldots,\ p_i,\ \ldots,\ p_m.$$

Составим из этих элементов сообщение, содержащее всего n элементов. Среди них пусть будет n_1 элементов h_1, n_2 элементов h_2, . . ., n_m элементов h_m.

Вероятность каждой данной комбинации из n элементов выразится произведением вероятностей отдельных элементов, так как мы предполагаем, что появление каждого данного элемента есть

---

[1] Выражение «стремится в среднем» нужно понимать в духе закона больших чисел.

независимое событие. Учитывая же наличие повторяющихся элементов, мы получим для вероятности некоторого сообщения

$$p = p_1^{n_1} p_2^{n_2} \dots p_m^{n_m} = \prod_{i=1}^{m} p_i^{n_i}. \qquad (1)$$

Положим теперь, что n достаточно велико для того, чтобы можно было считать

$$n_i = p_i n.$$

С другой стороны, при достаточно большом n можно считать все перестановки, т. е. все возможные сообщения, равновероятными. Тогда

$$p = \frac{1}{N} = \prod_{i=1}^{m} p_i^{np_i},$$

откуда для N — числа возможных сообщений — находим

$$N = 1/p = 1/\Pi p_i^{np_i}. \qquad (2)$$

Логарифмируя, получим количество сведений в сообщении при неравновероятности его элементов

$$I = \log N = -n \sum p_i \log p_i. \qquad (3)$$

Использованные при выводе равенства, справедливые с вероятностью единица при $n \to \infty$, остаются в силе в среднем и для конечного n.

Формула (3) выражает основное определение количества сведений [1].

Если все элементы равновероятны, т. е. если все p_i равны между собой и равны $1/m$, то, как легко видеть, (3) переходит в (1) § 11. Если вероятность какого-либо элемента равна единице, то вероятности всех остальных элементов равны нулю; в этом случае количество сведений равно нулю. Это обстоятельство часто поясняют тем соображением, что количество сведений связано с неопределенностью ситуации и непредвиденностью сообщения; если вероятность появления какого-либо сигнала, извещающего нас о ситуации, равна единице, т. е. если появление данного сигнала есть достоверное событие, то не надо этот сигнал и посылать, так как ситуация и без того ясна и сигнал никаких сведений в себе не несет. Наоборот, функция I имеет максимум при равновероятности всех элементов сообщения, т. е. при $p_i = 1/m$, тогда

$$I = n \log m.$$

[1] Количество сведений, отнесенное к одному элементу сообщения («информация на один символ»), т. е. величину $I' = I/n = -\Sigma p_i \log p_i$, в зарубежной литературе обозначают буквой H и, по почину Шэннона, называют *энтропией*.

Этот случай соответствует наиболее неопределенной ситуации. Все действительные случаи передачи сообщений лежат между этими двумя предельными.

Формула (3) относится к дискретному распределению вероятностей, какое мы получаем в результате квантования функции сигнала. Но первоначальный — неквантованный — сигнал имеет непрерывное распределение, выражаемое плотностью вероятностей $\varphi(x)$. Введя эту функцию, мы можем заменить в формуле (3) суммирование интегрированием. Мы имеем при квантовании

$$p_i = \int\limits_{(i-1/2)\delta}^{(i+1/2)\delta} \varphi(x)\,dx$$

или, приближенно [1],

$$p_i \approx \varphi(i\delta)\,\delta = \varphi(h_i)\,\delta.$$

Соответственно

$$\log p_i \approx \log[\varphi(h_i)\,\delta].$$

Таким образом,

$$I = -n\sum p_i \log p_i \approx -n\sum \varphi(h_i)\,\delta \log[\varphi(h_i)\,\delta].$$

Суммирование можно приближенно заменить интегрированием, беря x вместо h_i и dx вместо δ (только под знаком суммы):

$$I \approx -n\left\{\int\limits_{x_1}^{x_2} \varphi(x)\log\varphi(x)\,dx + \log\delta\right\},$$

где пределы интегрирования определены соотношением $x_2 - x_1 = l = m\delta$; l означает протяженность шкалы уровней. Вводя $\log\delta$ под знак интеграла, получаем окончательно

$$I \approx -n\int\limits_{x_1}^{x_2} \varphi(x)\log[\delta\varphi(x)]\,dx. \qquad (4)$$

Эта формула при предельном переходе для $\delta \to 0$ дает $I \to \infty$, как оно и должно быть, так как число ступеней шкалы уровней становится при таком переходе бесконечно большим даже в случае конечной протяженности шкалы [2]. Но как приближенная формула

---

[1] Условие приближения состоит в том, что в разложении $\varphi(x)$ в ряд Тэйлора можно пренебречь членами, начиная с квадратичного, или, проще говоря, можно считать функцию $\varphi(x)$ изменяющейся линейно на интервале δ.

[2] Все это совершенно игнорирует Шеннон, который пишет для «энтропии» непрерывного распределения ([9], стр. 54)

$$H = -\int p(x)\log p(x)\,dx,$$

не смущаясь тем, что плотность вероятностей $p(x)$ имеет размерность $1/x$.

для вычисления количества сведений в квантованном сигнале, т. е. при конечном δ, формула (4) будет нам очень полезна, так как с интегралами легче обращаться, чем с суммами.

§ 25. Количество сведений и распределение вероятностей

Как показано в предыдущем параграфе, количество сведений зависит от распределения вероятностей для функции сигнала. Возникает естественный вопрос о том, при каком распределении вероятностей сигнал заданной мощности переносит наибольшее количество сведений. Задача ставится так: нужно найти распределение, т. е. совокупность значений p_i, при котором величина

$$I' = -\sum p_i \log p_i$$

имеет максимум при двух дополнительных условиях

$$\sum p_i = 1$$

— условие нормировки — и

$$\sum h_i^2 p_i = P = \text{const},$$

или, так как $h_i = i\delta$,

$$\sum i^2 p_i = P/\delta^2 = c$$

— условие постоянства мощности. Поставленная таким образом задача есть задача на отыскание условного экстремума. Для ее решения можно применить способ множителей Лагранжа [1].

По этому способу мы составляем функцию

$$\Phi = -\sum p_i \log p_i + \lambda_1 \left(\sum p_i - 1\right) + \lambda_2 \left(\sum i^2 p_i - c\right)$$

и ищем ее абсолютный экстремум. Постоянные λ_1 и λ_2 определяются из дополнительных условий.

Мы имеем

$$\frac{\partial \Phi}{\partial p_i} = -\log p_i - 1 + \lambda_1 + \lambda_2 i^2 = 0,$$

откуда

$$p_i = 2^{\lambda_1 - 1} 2^{\lambda_2 i^2} = A 2^{\lambda_2 i^2}.$$

Постоянная λ_2 не может быть положительной; таким образом, найденное распределение есть не что иное, как дискретное нормальное распределение. Однако довести решение задачи до конца, т. е. определить постоянные A и λ_2, затруднительно, так как эти постоянные нужно определить из уравнений

$$A \sum 2^{\lambda_2 i^2} = 1; \quad A \sum i^2 2^{\lambda_2 i^2} = P/\delta^2.$$

[1] См., например, *В. И. Смирнов.* Курс высшей математики, т. 1. Гостехиздат, 1954, стр. 392.

Выход из затруднения предоставляет нам формула (4) § 24; вместо отыскания оптимального дискретного распределения мы можем искать оптимальное непрерывное распределение неквантованного сигнала. Для упрощения будем считать шкалу квантования неограниченной. Тогда задача ставится следующим образом: найти функцию $\varphi(x)$, дающую максимум интегралу

$$\int_{-\infty}^{\infty} \varphi(x) \log \varphi(x)\, dx$$

при определенных дополнительных условиях, в частности, при условии, что задана мощность сигнала $P=\sigma^2$.

Решение этой вариационной задачи (данное в добавлении 4) показывает, что интересующими нас экстремальными свойствами обладает симметричное нормальное распределение

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}.$$

Итак, сигнал заданной средней мощности переносит наибольшее количество сведений, если функция сигнала подчинена нормальному распределению.

Из этого положения следует ряд важных выводов. Самый прямой вывод состоит в том, что желательна такая обработка сигнала, которая приближает распределение к нормальному. В результате такой обработки можно либо передать большее количество сведений, располагая определенной мощностью сигнала, либо сэкономить мощность при передаче данного количества сведений.

Соотношение между мощностью и количеством сведений для наивыгоднейшего случая нормального распределения нетрудно вычислить. Мы имеем

$$I=-n\int_{-\infty}^{\infty} \varphi(x) \log [\delta\varphi(x)]\, dx,$$

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}, \quad \log[\delta\varphi(x)]=\frac{1}{\ln 2}\left(-\ln\sqrt{2\pi}\,\frac{\sigma}{\delta}-\frac{x^2}{2\sigma^2}\right),$$

$$I=\frac{2n}{\ln 2}\left[\frac{1}{\sqrt{2\pi}\,\sigma}\left(\ln\sqrt{2\pi}\,\sigma\int_0^{\infty} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}dx+\right.\right.$$

$$\left.\left.+\frac{1}{2\sigma^2}\int_0^{\infty} x^2 e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}dx\right)\right]=\frac{n}{\ln 2}\left(\ln\sqrt{2\pi}\,\frac{\sigma}{\delta}+\frac{1}{2}\right)=$$

$$=n\log\sqrt{2\pi e}\,\frac{\sigma}{\delta}.$$

Можно выразить количество сведений через мощность, введя $P=\sigma^2$. Это даст

$$I=\frac{n}{2}\log 2\pi e\frac{P}{\delta^2}.$$

Итак, связь между количеством сведений и мощностью приведена к виду

$$I=\frac{n}{2}\log AP=\frac{n}{2}\log a\frac{P}{P_{\text{п}}}$$

(см. § 14). Мы имеем для рассмотренного случая нормального распределения

$$I=\frac{n}{2}\log\frac{2\pi e}{k^2}\cdot\frac{P}{P_{\text{п}}}.$$

При всяком ином распределении количество сведений оказывается при прочих равных условиях меньше. Пусть, например, имеется равномерное в конечных пределах распределение

$$\varphi(x)=\begin{cases}\dfrac{1}{l} & \text{при } |x|\leqslant l/2,\\ 0 & \text{при } |x|>l/2.\end{cases}$$

Для количества сведений получаем

$$I=-\frac{n}{l}\int\limits_{-l/2}^{l/2}dx=n\log\frac{l}{\delta}=\frac{n}{2}\log\frac{l^2}{\delta^2}.$$

Средняя мощность равна

$$P=\frac{1}{l}\int\limits_{-l/2}^{l/2}dx=n\log\frac{l}{\delta}=\frac{n}{2}\log\frac{l^2}{\delta^2}.$$

Подставляя это значение в выражение для I, получаем

$$I=\frac{n}{2}\log 12\frac{P}{\delta^2},$$

тогда как для нормального распределения

$$I=\frac{n}{2}\log 2\pi e\frac{P}{\delta^2}=\frac{n}{2}\log 17\frac{P}{\delta^2}.$$

Другой существенный вывод из того, что нормальное распределение имеет экстремальные свойства, состоит в том, что, приписывая белому шуму такого рода распределение, мы наделяем белый шум, как помеху, наиболее зловредными свойствами. В самом деле, обладая нормальным распределением, белый шум при заданной средней мощности несет наибольшее количество вредящих передаче «сведений». Это обстоятельство оправдывает пользование белым

шумом при разного рода расчетах как «стандартной» помехой; не говоря уже об упрощении анализа, мы выполняем при этом свои расчеты для наихудшего случая с запасом.

С точки зрения этих соображений мы можем теперь уточнить смысл выведенной в § 15 формулы (1)

$$I = FT \log\left(1 + \frac{P_c}{P_п}\right).$$

Если подразумевать помеху в виде белого шума, то эта формула выражает теоретический максимум количества сведений, передаваемого сигналом при наличии шума. Этот максимум может быть достигнут лишь в том случае, когда распределение вероятностей для сигнала является нормальным. Мы можем теперь отбросить сделанное в § 15 допущение об одинаковости вероятностных свойств сигнала и помехи и сделать вывод заново. Мы имеем

$$I = \frac{n}{2} \log A_c P_c,$$

$$I_п = \frac{n}{2} \log A_п P_п,$$

$$I_1 = \frac{n}{2} \log A_1 (P_c + P_п),$$

$$I = I - I_п = \frac{n}{2} \log \frac{A_1}{A_п}\left(1 + \frac{P_c}{P_п}\right).$$

Коэффициент A_1 подсчитывается для распределения, отвечающего сумме сигнала и помехи. Если помеха есть белый шум, то отношение $A_1/A_п$ никогда не может превзойти единицы и принимает это предельное значение в том уже упомянутом случае, когда вероятностные характеристики сигнала совпадают с характеристиками белого шума.

§ 26. Понятие избыточности

При рассмотрении действительных сообщений мы замечаем, что не все то, что содержится в сообщении, необходимо для его полного восстановления на приемном конце и что, следовательно, сообщение могло бы передаваться без ущерба для дела в сокращенном виде. Это обстоятельство давно используется в том сокращенном телеграфном языке, на котором составляется текст обычной телеграммы. Всеобщей традицией является устранение из текста предлогов и союзов, так как они легко могут быть восстановлены при чтении телеграммы по общей конструкции фразы, по падежным окончаниям и т. п.[1] Можно было бы пойти по этому

[1] Для экономии в письменном тексте давно уже узаконен целый ряд сокращений, например: «т. к.», «т. е.», «и т. п.», а также сокращенных обозначений организаций и учреждений по начальным буквам их полных наименований.

пути еще значительно дальше. Специальные исследования показывают, что возможность безошибочного восстановления текста еще сохраняется при сокращении первоначального объема текста вдвое. Так возникает важное понятие об *избыточности* сообщения.

Избыточность текстового сообщения обусловлена тем, что отдельные элементы текста — слова, а также элементы слов — буквы — не независимы. Между этими элементами существуют вероятностные связи, определяемые статистической структурой языка. Именно знание — пусть безотчетное — этой структуры позволяет нам восполнить сокращения текста. При этом надо понимать, что, говоря о безошибочном восстановлении текста, мы имеем в виду малую вероятность ошибки восстановления. Возьмем, к примеру, следующую фразу: «Это, может (быть), и не играет (роли), так (как), с другой (стороны), оба явления относятся друг к (другу) как причина и (следствие)».

Если взятые в скобки слова были бы пропущены, то они могли бы быть восстановлены с высокой степенью вероятности, т. е. с малой вероятностью ошибки, на том основании, что такие последовательности, как «друг к другу», «причина и следствие», являются типичными для языка, и условная вероятность того, что за словом «причина» последует слово «следствие», очень велика.

Аналогичные соображения относятся и к последовательности букв. Так, например, если первая буква есть «ч», то из гласных букв следующей не может быть «ы» или «я», наиболее вероятны буквы «и» и «е». Из согласных букв в качестве следующей за «ч» исключены «с», «ц», «ф» и т. д. Буква «т» весьма вероятна, но если первые две буквы образуют комбинацию «чт», то вероятнее всего, что третья буква будет «о» («что»), хотя не исключено появление в качестве третьей буквы «е» (например, «чтение») или «и» (например, «чтица», «чтить»). Появление же в качестве третьей буквы «у» не исключено, но очень маловероятно, так как это соответствует немногочисленным и малоупотребительным в русском языке комбинациям («чтут», «чту»).

Условные вероятности букв в данном языке определяются статистикой многобуквенных сочетаний — полиграмм, т. е. двухбуквенных сочетаний — диаграмм, трехбуквенных — триграмм и т. д.

Наличие таких статистических закономерностей сильно сокращает число буквенных сочетаний, употребляемых в качестве слов. Так, например, если бы все комбинации были возможны, то, имея круглым счетом 30 букв, мы могли бы составить из них 30 однобуквенных слов, $30^2=900$ двухбуквенных, $30^3=27\,000$ трехбуквенных, $30^4=810\,000$ четырехбуквенных и т. д. Между тем в действительности язык содержит примерно 50 000 слов. Если принять среднее число букв в слове равным семи, то окажется, что лишь около 0,0002% всех возможных комбинаций букв являются словами. Можно себе представить новый язык с алфавитом из

30 букв, в котором все возможные комбинации разрешены; средняя длина слова в этом «языке» составила бы только 3,5 буквы. Такого рода искусственный язык мог бы служить кодом, но кодом не буквенным, а словесным.

Для уяснения понятия избыточности рассмотрим еще один пример. Представим себе, что художник заготовляет рисунки последовательных кадров мультипликационного фильма. Пусть сцена представляет, как действующее лицо идет по лесу. Само собой разумеется, что художник рисует лишь последовательные положения персонажа на фоне леса, нарисованного один раз для всей сцены. Было бы крайне нелепо, если бы лес рисовался для каждого кадра заново. Но именно так обстоит дело при передаче телевизионного изображения современным методом: каждый кадр передается заново, как совершенно независимое изображение. Ясно, что избыточность телевизионного сигнала очень велика, и ясно, что можно было бы существенно сократить объем телевизионного сигнала, использовав вероятностные связи между соседними кадрами [1].

Вообще объем сообщения, а следовательно, и сигнала можно сократить за счет имеющейся избыточности. Однако нужно иметь в виду и другую сторону вопроса. Дело в том, что избыточность, увеличивая объем сообщения, играет и важную положительную роль: она уменьшает возможность ошибки в принятом сообщении при наличии помех. Если, например, какое-либо слово телеграммы искажено при передаче, то в большинстве случаев мы можем восстановить его правильный вид, основываясь на остаточной избыточности, т. е. на избыточности, которая содержится в тексте телеграммы и не устранена применением телеграфного языка. Если бы текст телеграммы был сокращен до предела, т. е. если бы избыточность была полностью устранена, то исправление ошибок стало бы невозможным и сообщение стало бы очень «хрупким»: любая помеха искажала бы его непоправимо [2]. Поэтому при относительно сильных помехах не только не сокращают избыточность сообщения, но, напротив, искусственно увеличивают избыточность

[1] Для числа возможных сообщений в телевидении получаются потрясающие цифры [22]. Считая, что телевизионное изображение состоит круглым счетом из 500 000 элементов, яркость каждого из которых может принимать одно из 100 различимых значений, мы получим, что число различных «изображений» составит $100^{500000}=10^{10^6}$. Между тем, если бы телецентр работал круглосуточно, передавая по 25 различных изображений в секунду, то он передавал бы в год всего около 10^9 изображений. Для того чтобы передать все возможные комбинации, потребовалось бы $10^{9999991}$ лет (!). Эти цифры приведены, чтобы показать, как мала доля действительно используемых комбинаций, т. е. комбинаций, на самом деле являющихся изображениями.

[2] Это можно пояснить на примере гипотетического «языка», о котором говорилось выше. Так как в этом языке все комбинации букв являются «словами», то ошибка в приеме одной буквы заменяет одно слово другим, и если передается одно слово, то нет никакой возможности узнать, правильно ли оно принято. Однако ошибка может быть обнаружена и исправлена за счет избыточности в конструкции целой фразы из нескольких слов.

для того, чтобы повысить надежность связи, т. е. вероятность правильного приема. Практическим примером применения этой идеи может служить метод накопления, состоящий в том, что передаваемый сигнал повторяется несколько раз и несколько принятых экземпляров одного и того же сигнала, по-разному искаженных помехами, сличаются между собой. Вероятность получения таким методом правильного сигнала (т. е. сигнала, соответствующего переданному) возрастает с числом повторений. Одной из ранних форм осуществления этого принципа является телеграфная система Бодо—Вердана. Другим примером могут служить так называемые корректирующие коды, принцип которых состоит в том, что к кодовой комбинации добавляются дополнительные знаки, увеличивающие избыточность, но зато позволяющие обнаружить ошибку передачи.

Для того чтобы ввести понятие избыточности в теорию, нужно дать избыточности количественное определение. К этому мы и переходим в последующих параграфах.

§ 27. Количество сведений при взаимозависимости элементов

В § 24 было получено выражение для содержательности, т. е. для количества сведений в сообщении, приходящегося на один элемент при m различных элементах (код с основанием m),

$$I' = -\sum_{i=1}^{m} p_i \log p_i.$$

Это соотношение предполагает все элементы сообщения независимыми, т. е. появление каждого данного элемента не связано с предшествующими элементами. Положим теперь, что такая связь существует и что она выражается условной вероятностью $p\,(h_j/h_i)$ появиться элементу h_j, если предшествующий элемент был h_i. Для АИМ, которую мы все время имеем в виду, $h_i = i\delta$, $h_j = j\delta$. Поэтому вышеуказанную условную вероятность мы будем в дальнейшем записывать как $p\,(j/i)$.

Для каждого значения i мы будем иметь для содержательности, как логарифма числа возможных сообщений,

$$I'_i = -\sum_{j=1}^{m} p\,(j/i) \log p\,(j/i),$$

а всего, т. е. для всех возможных i,

$$I' = -\sum_{i=1}^{m} p\,(i)\, I_i = -\sum_{i=1}^{m} p\,(i) \sum_{j=1}^{m} p\,(j/i) \log p\,(j/i).$$

Это и есть общее выражение для содержательности сообщения при наличии вероятностной взаимосвязи между соседними элементами сообщения.

Мы видим теперь, что содержательность сообщения, зависящая от числа элементов сообщения и основания кода, определяется также, во-первых, распределением вероятностей и, во-вторых, наличием вероятностных связей между элементами сообщения, или, короче, от полных и условных вероятностей элементов сообщения.

Можно подытожить все предыдущие результаты, представив их следующей сводкой:

1. Общий случай — элементы взаимозависимы и неравновероятны

$$I' = -\sum_i p(i) \sum_j p(j/i) \log p(j/i).$$

2. Элементы взаимозависимы и равновероятны

$$I'_2 = -\frac{1}{m} \sum_i \sum_j p(j/i) \log p(j/i).$$

3. Элементы независимы и неравновероятны $[p(j/i) = p(j)]$

$$I'_1 = -\sum_j p(j) \log p(j).$$

4. Элементы независимы и равновероятны $[p(j) = 1/m]$

$$I'_0 = \log m.$$

В пояснение этих соотношений рассмотрим пример. Пусть имеется всего два элемента: a и b, так что $m=2$. Подсчитаем количество сведений для четырех вышеперечисленных случаев.

1. Пусть две полные и четыре условные вероятности имеют следующие значения: $p(a)={}^3/_4$, $p(b)={}^1/_4$, $p(a/a)={}^2/_3$, $p(b/a)={}^1/_3$, $p(a/b)=1$, $p(b/b)=0$ (т. е. после b всегда следует a). Тогда

$$\begin{aligned} I' &= -\sum_i p(i) \sum_j p(j/i) \log p(j/i) = \\ &= -\{p(a)[p(a/a) \log p(a/a) + p(b/a) \log p(b/a)] + \\ &\quad + p(b)[p(a/b) \log p(a/b) + p(b/b) \log p(b/b)]\} = \\ &= -\frac{3}{4}\left(\frac{2}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{1}{3}\right) = 0{,}685. \end{aligned}$$

3. $p(a) = {}^3/_4$, $p(b) = {}^1/_4$,

$$\begin{aligned} I'_1 &= -\sum_i p(i) \log p(i) = -[p(a) \log p(a) + \\ &\quad + p(b) \log p(b)] = -\left(\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4}\right) = 0{,}815. \end{aligned}$$

4. $I_0 = \log m = \log_2 2 = 1$.

Что касается второго случая, который мы пока пропустили, то он, пожалуй, наиболее поучителен. Беря

$$p(a) = p(b) = {}^1/_2,$$

мы должны подчинить условные вероятности дополнительным связям, вытекающим из формулы полной вероятности, а именно:

$$p(a/a) = p(b/b) = p_1; \quad p(a/b) = p(b/a) = p_2,$$

причем, конечно, $p_1+p_2=1$. Условная вероятность p_1 выражает вероятность повторения, а p_2 — вероятность чередования элементов. Содержательность резко зависит от соотношения p_1 и p_2, как показано ниже,

p_1, p_2	$1/2$, $1/2$	$1/4$, 3,4	1,8, 7,8	0,1
I'	1	0,815	0,541	0

Когда условные вероятности p_1 и p_2 равны, то это равносильно независимости a и b и мы получаем такие же соотношения, как в случае 4. Когда же, например, $p_1=0$, то это означает, что мы имеем вполне определенную последовательность: непрерывное чередование a и b. Так как характер этой последовательности наперед известен, то никаких сведений такого рода сигнал не несет.

§ 28. Количественное определение избыточности

Как видно из предыдущего, наличие вероятностных взаимозависимостей между элементами сообщения уменьшает количество сведений, приходящееся на каждый элемент. Теперь можно дать количественное определение избыточности, основываясь на нижеследующем очень простом рассуждении.

Положим, что мы передаем сообщение из n элементов, содержащее количество сведений

$$I = I'n = -n \sum_i p(i) \sum_j p(j/i) \log p(j/i).$$

Но если бы мы устранили внутренние вероятностные связи, то содержательность возросла бы до максимального значения $I'_{\max}$. При этом, очевидно, то же самое количество сведений могло бы содержаться в сообщении, состоящем из меньшего числа элементов, скажем n_0. Из равенства

$$nI' = n_0 I'_{\max}$$

находим $n_0/n = I'/I'_{\max}$.

Так как $n_0 < n$, то передача сообщения в его первоначальном виде сопровождается затратой $n-n_0$ лишних, избыточных элементов. Мерой R избыточности может поэтому служить относительное число лишних элементов, т. е.

$$R = (n - n_0)/n = 1 - n_0/n$$

или

$$R = 1 - I'/I'_{\max}.$$

Остается условиться относительно того, что понимать под I'_{max}. Если принять, что избыточность определяется только взаимосвязью элементов, то под I'_{max} следует понимать величину

$$I'_1 = -\sum p_i \log p_i.$$

Если же учитывать, что на количество сведений влияет также и распределение вероятностей, которое мы можем изменять, то в качестве I'_{max} следует взять величину

$$I'_0 = \log m,$$

так как эта величина выражает максимум—максиморум сведений, который может содержать один элемент сообщения.

Таким образом, можно ввести три определения избыточности:

частная избыточность, обусловленная взаимосвязью

$$R_p = 1 - I'/I'_1;$$

частная избыточность, зависящая от распределения

$$R_\varphi = 1 - I'_1/I'_0;$$

полная избыточность

$$R = 1 - I'/I'_0.$$

Между этими тремя величинами существует следующая зависимость:

$$R = R_p + R_\varphi - R_p R_\varphi.$$

При небольших R_p и R_φ можно считать приближенно, что полная избыточность равна сумме частных

$$R \approx R_p + R_\varphi.$$

Для иллюстрации возьмем цифры из примера предыдущего параграфа. Мы имели

$$I' = 0{,}685; \quad I'_1 = 0{,}815; \quad I'_0 = 1.$$

Для избыточностей получаем

$$R_p = 1 - 0{,}685/0{,}815 \approx 0{,}16; \quad R_\varphi = 1 - 0{,}815 \approx 0{,}18;$$
$$R = 1 - 0{,}685 \approx 0{,}31 \ (R_p + R_\varphi = 0{,}34; \quad R_p R_\varphi = 0{,}03).$$

§ 29. Оптимальное распределение вероятностей

Сопоставляя предыдущие результаты, мы обнаруживаем противоречие в вопросе об оптимальном распределении вероятностей. С одной стороны, было найдено, что оптимальным распределением является симметричное нормальное (§ 25), так как при таком распределении сигнал несет при данной средней мощности наибольшее количество сведений. С другой стороны, наибольшее коли-

чество сведений получается при равномерном распределении (§ 24).

Противоречие это, конечно, только кажущееся. Дело заключается в том, что два разных результата получаются при разных взаимоисключающих дополнительных условиях, налагаемых на функцию распределения.

Если искать функцию, дающую максимум содержательности

$$I' = -\int_{-\infty}^{\infty} \varphi(x) \log [\delta\varphi(x)]\, dx, \tag{1}$$

требуя при этом неизменности мощности, т. е. постоянства

$$P = \int_{-\infty}^{\infty} x^2 \varphi(x)\, dx, \tag{2}$$

то оптимальным распределением оказывается (как показано в добавлении 4) симметричное нормальное распределение.

Если же это дополнительное условие не налагается, то оптимальное распределение оказывается равномерным (что также показано в добавлении 4). Предполагается, что в обоих случаях выполняется обязательное условие нормировки

$$\int_{-\infty}^{\infty} \varphi(x)\, dx = 1. \tag{3}$$

Таким образом, применение равномерного распределения вместо нормального сопряжено с неизбежным увеличением средней мощности сигнала.

Мы сделаем сейчас в подробностях вывод, относящийся к равномерному распределению. При этом выводе, как сказано, ничего не говорится о мощности; условие, которое предполагается выполненным в рассматриваемом случае, состоит в том, что интервал изменения сигнала как случайной величины конечен и неизменен. Иначе говоря, функция распределения существует только на конечном интервале $(-l/2 < x < l/2)$, вне которого она равна нулю. Величина l — это конечная длина шкалы уровней сигнала. Таким образом, дополнительное условие, учитываемое в рассматриваемом случае, записывается в виде

$$\varphi(x) = 0 \quad \text{при} \quad |x| > l/2. \tag{4}$$

Это условие лишь меняет пределы интегрирования, но не меняет уравнения Эйлера (см. добавление 4), и мы получаем оптимальное распределение

$$\varphi(x) = 1/l \quad \text{при} \quad |x| \leqslant l/2. \tag{5}$$

Содержательность при этом

$$I'_p = -\int_{-l/2}^{l/2} \varphi(x) \log [\delta\varphi(x)]\, dx = \log \frac{l}{\delta}. \qquad (6)$$

Для того чтобы сравнить получаемый результат с тем, который мы имели бы при нормальном распределении, следует подчинить это распределение условию (4). Это дает

$$\varphi(x) = C \frac{\alpha}{\sqrt{\pi}} e^{-\alpha^2 x^2} \quad \text{при} \quad |x| \leqslant l/2, \qquad (7)$$

где C — нормирующий множитель, определяемый при помощи условия (3). Мы имеем

$$C \frac{\alpha}{\sqrt{\pi}} \int_{-l/2}^{l/2} e^{-\alpha^2 x^2} dx = 1,$$

откуда

$$C = \frac{1}{\Phi(\alpha l/2)},$$

где

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz$$

— функция Лапласа. Множитель C больше единицы; он увеличивает масштаб нормальной кривой так, чтобы площадь, ограни-

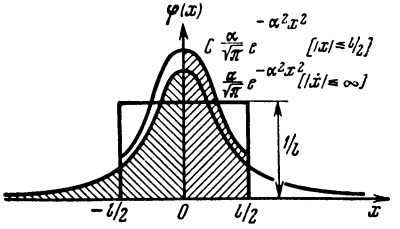


Рис. 18

чиваемая ею в конечных пределах, равнялась площади, ограничиваемой обычной нормальной кривой в бесконечных пределах. Иначе говоря, $1/C$ представляет собой долю площади нормальной кривой при усечении на значениях $|x| = l/2$. Все это поясняет рис. 18, на котором показаны все вышеупомянутые распределения. Множитель C подобран так, чтобы площади, заштрихованные в правой и левой частях рисунка, были равны. При $l \to \infty$ $C \to 1$. Итак, распределение определяется параметром

$$s = \alpha l/2,$$

который нужно выбрать. После этого можно вычислить содержательность

$$I'_{\text{н}} = -\int_{-l/2}^{l/2} \varphi_{\text{н}}(x) \log [\delta\varphi_{\text{н}}(x)]\, dx =$$

$$= -\frac{2Ca}{\sqrt{\pi}\ln 2}\left[\ln\frac{Ca}{\sqrt{\pi}}\int_0^{l/2} e^{a^2x^2}dx - a^2\int_0^{l/2} x^2 e^{-a^2x^2}dx\right] - \log\delta.$$

Второй интеграл вычисляется интегрированием дважды по частям с применением формулы [1]

$$\int_0^x \Phi(z)\, dz = x\Phi(x) + \frac{e^{-x^2}-1}{\sqrt{\pi}}.$$

В конце концов получается

$$I'_{\text{н}} = \log\frac{\sqrt{\pi e}}{2Cs}\cdot\frac{l}{\delta} - \frac{C}{\sqrt{\pi}\ln 2}\, s e^{-s^2}. \tag{8}$$

Второе слагаемое относительно мало (порядок малости определяется множителем e^{-s^2}), так что приближенно можно положить

$$I'_{\text{н}} \approx \log\frac{\sqrt{\pi}l}{2Cs}\cdot\frac{l}{\delta}. \tag{9}$$

Этот результат следует сравнить с тем, который был получен для равномерного распределения

$$I'_p = \log\frac{l}{\delta}.$$

Пусть, например, $C=1{,}01$. Тогда

$$\Phi(s) = 1/1{,}01 = 0{,}99,$$

откуда при помощи таблиц находим

$$s = 1{,}81$$

и

$$I'_{\text{н}} \approx \log 0{,}8\,\frac{l}{\delta}.$$

Для окончательного разъяснения вопроса рассмотрим еще численный пример на дискретное распределение. Пусть имеется четыре дискретных уровня: $x=-2, -1, 1, 2$ и пусть распределение вероятностей симметрично. Обозначим вероятность уровней ± 1 через p_1 и вероятность уровней ± 2 через p_2. В силу симметрии распределения условие нормировки имеет вид

$$p_1 + p_2 = {}^1/_2.$$

[1] *И. М. Рыжик* и *И. С. Градштейн*. Таблицы интегралов, сумм, рядов и произведений. Гостехиздат, 1951, стр. 255, формула (4.241).

Средняя мощность равна

$$P = \sum x_i^2 p_i = 2(p_1 + 4p_2).$$

Содержательность

$$I' = -\sum p_i \log p_i = p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2}.$$

Удельная содержательность

$$\nu = \frac{1}{V} = \frac{2I'}{\log P/P_{\text{ш}}} = 2\,\frac{p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2}}{\log 200\,(p_1 + 4p_2)}.$$

Построим зависимости величин P, I' и ν от p_1 (рис. 19). Из рисунка видно следующее: мощность убывает по линейному закону по мере того, как увеличивается вероятность меньшего уровня и соответственно уменьшается вероятность большего; содержательность имеет максимум при $p_1 = p_2 = 1/4$, т. е. при равномерном распределении. Что же касается удельной содержательности, то эта величина имеет максимум при $p_1 \approx 0{,}29$.

Таким образом, без учета мощности оптимум получается при равномерном распределении (точка *а*), а с учетом мощности — при неравномерном (точка *б*). Соответствующие распределения изображены на том же рис. 19.

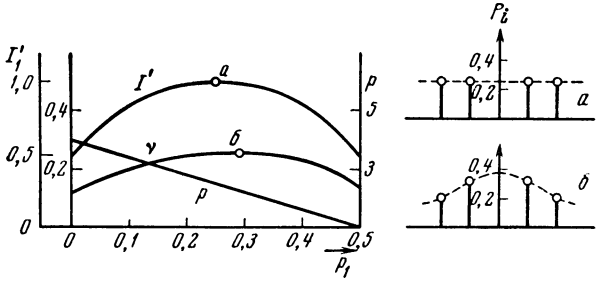


Рис. 19

Вывод из всего сказанного таков: несмотря на некоторую избыточность, свойственную всякому неравномерному распределению вероятностей, целесообразно применять распределения, приближающиеся к нормальному, так как такие распределения обеспечивают наибольшую удельную содержательность, т. е. передачу наибольшего количества сведений при заданной средней мощности. Таким образом, обсуждая вопрос о возможности уменьшения избыточности, мы будем в дальнейшем иметь в виду только частную избыточность R_p, обусловленную вероятностной взаимосвязью элементов сообщения.

§ 30. Перераспределение вероятностей

Одна из технических задач теперь поставлена: она состоит в таком преобразовании сообщения (или сигнала), чтобы имеющееся первоначальное распределение вероятностей было изменено, а именно приближено к симметричному нормальному. Нам нужно теперь рассмотреть вопрос о том, какими средствами эта задача может быть разрешена.

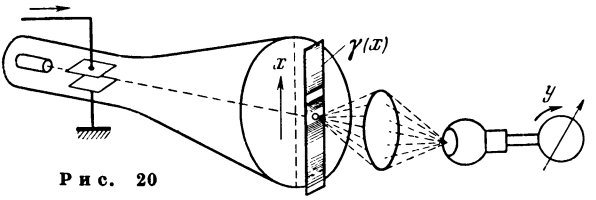


Рис. 20

Начнем с простейшего примера. Пусть имеется сигнал с нормальным, но несимметричным распределением

$$\varphi(x) = \frac{\alpha}{\sqrt{\pi}} e^{-\alpha^2 (x-a)^2}.$$

Средняя мощность такого сигнала равна

$$P = \int_{-\infty}^{\infty} x^2 \varphi(x)\, dx = \frac{\alpha}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2 e^{-\alpha^2 (x-a)^2} dx = \frac{1}{2\alpha^2} + a^2.$$

Для того чтобы сделать распределение симметричным, нужно просто вычесть из сигнала постоянную составляющую a (добавление постоянной величины смещает распределение вдоль оси x). Количество сведений при этом, конечно, не изменится, мощность же сигнала уменьшится на a^2. Поэтому, например, гораздо выгоднее с энергетической точки зрения телеграфировать двоичным кодом с элементами $+h$, $-h$, нежели с элементами 0, $2h$. Во втором случае средняя мощность вдвое больше, если элементы равновероятны.

Теперь обратимся к общему случаю, когда имеется сигнал с некоторым данным распределением и задача состоит в преобразовании сигнала таким образом, чтобы данное распределение заменилось желаемым, в частности нормальным.

Один из возможных способов такого преобразования описан Оливером [26]. Идея этого способа поясняется рис. 20. Подлежащий преобразованию сигнал подается на отклоняющие пластины электронно-лучевой трубки. Луч отклоняется; величина отклонения пропорциональна уровню сигнала в первоначальной, неизменной шкале. Перед экраном трубки располагается маска с переменной прозрачностью, закон изменения которой выражается

функцией $\gamma(x)$. Световой поток L, проходящий через маску, попадает на фотоэлемент, дающий ток

$$I = kL,$$

где k — чувствительность фотоэлемента. Выходной ток фотоэлемента представляет собой новый сигнал с измененным распределением вероятностей. Действительно,

$$I = kL = k_1\gamma(x),$$
$$I/k_1 = y = \gamma(x),$$

и если раньше вероятность уровня x определялась плотностью $\varphi(x)$, то теперь та же самая плотность отвечает уровню $y=\gamma(x)$. Таким образом, новая плотность вероятностей может быть выражена как

$$\varphi_1(x_1) = \varphi[\gamma^{-1}(y)].$$

Из этого функционального уравнения находим γ

$$\gamma^{-1}(y) = \varphi^{-1}[\varphi_1(y)].$$

Таким образом, зная первоначальное распределение φ и задаваясь желаемым новым распределением φ_1, определяем закон изменения прозрачности маски. Вышеприведенные функциональные уравнения поясняются рис. 21 и 22, из которых первый относится к дискретному, а второй к непрерывному распределению. Показанное на этих рисунках построение является, кстати говоря, удобным методом решения функциональных уравнений.

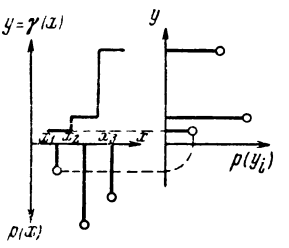


Рис. 21

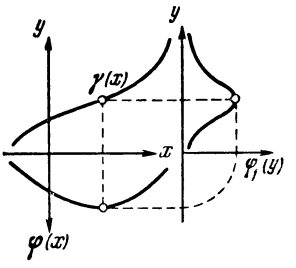


Рис. 22

Само собой разумеется, что описанное преобразование может осуществляться и без помощи электронной трубки и фотоэлемента. Ведь речь идет в сущности о нелинейном четырехполюснике с заданной характеристикой $y = \gamma(x)$. Всякий нелинейный элемент схемы изменяет распределение вероятностей. Так, например, можно рассматривать обычный ограничитель как преобразователь распределения, сводящий к нулю вероятность уровней, превосходящих порог ограничения. Четырехполюсник с требуемой характеристикой можно осуществить многими способами, не говоря уже о применении вычислительных машин, могущих выполнить любые функциональные преобразования.

Преобразованный сигнал, прошедший нелинейный четырехполюсник, будет, конечно, искажен. На приемном конце он должен быть подвергнут для восстановления первоначального вида обратной нелинейной обработке.

Описанное преобразование относилось к амплитудно-модулированному сигналу с большим числом уровней или с большой длиной шкалы уровней. Иначе говоря, преобразованию подвергался сигнал, кодированный при большом основании кода. Полученный преобразованный сигнал может быть перемодулирован и перекодирован любым образом.

§ 31. Декорреляция сигнала укрупнением

Очередная задача состоит в устранении вероятностных взаимосвязей между элементами сигнала. Такого рода связь может быть выражена через корреляцию сигнала; отсутствие взаимосвязи можно определить как отсутствие корреляции. Поэтому операцию устранения взаимосвязей мы будем кратко называть *декорреляцией.*

В настоящее время известны два метода декорреляции сигнала. Один из этих методов, описываемый ниже, может быть назван *методом укрупнения.*

Общая идея метода состоит в том, что сигнал разбивается не на элементы, а на отрезки, каждый из которых содержит по нескольку элементов. Такие отрезки могут рассматриваться как элементы некоторого нового сигнала, и можно показать, что вероятностные связи между такими укрупненными элементами слабее, чем между элементами первоначального, неукрупненного сигнала. Следует иметь в виду, что при укрупнении сигнала происходит его преобразование, состоящее в переходе к коду с более высоким основанием

$$m_1 = m^r,$$

где m — первоначальное основание, а r — число элементов в отрезке. Рассмотрим некоторые соотношения, относящиеся к методу укрупнения.

При наличии взаимосвязей количество сведений выражается через условные вероятности появления данного элемента при наличии определенных значений предшествующих элементов. Ограничимся рассмотрением случая, когда учитывается связь только двух соседних элементов, т. е. когда появление данного элемента обусловливается только значением одного предшествующего элемента. В этом случае (см. § 27) содержательность равна

$$I' = -\sum_{i=1}^{m} p(i) \sum_{j=1}^{m} p(j/i) \log p(j/i). \qquad (1)$$

Преобразуем это выражение при помощи формулы полной вероятности (см. § 19), из которой мы получаем для условной вероятности

$$p(j/i) = p(ij)/p(i).$$

Здесь $p(ij)$ — вероятность появления определенной пары значений ij. Подставляя (2) в (1), находим

$$I' = -\sum_{i=1}^{m}\sum_{j=1}^{m} p(ij)\log\frac{p(ij)}{p(i)} =$$
$$= -\left[\sum_i\sum_j p(ij)\log p(ij) - \sum_i\sum_j p(ij)\log p(i)\right].$$

Вторая сумма сразу упрощается

$$-\sum_i\sum_j p(ij)\log p(i) = -\sum_i \log p(i)\sum_j p(ij) =$$
$$= \sum_i p(i)\log p(i) = I'_1.$$

Первая сумма также упростится, если мы будем рассматривать каждую пару значений ij как новый элемент k. Число таких элементов будет, очевидно, m^2. Таким образом, для первой суммы можно записать

$$-\sum_i\sum_j p(ij)\log p(ij) = -\sum_{k=1}^{m^2} p(k)\log p(k) = I'_k.$$

Итак,

$$I' = I'_k - I'_1.$$

Мы можем действительно образовать новый, укрупненный сигнал, в котором роль элементов будут играть пары значений ij.

Содержательность укрупненного сигнала

$$I'_k = -\sum_{k=1}^{m^2} p(k)\log p(k).$$

Таким образом, описанное преобразование увеличивает содержательность на

$$I'_1 = \sum_{i=1}^{m} p(i)\log p(i),$$

объем же сигнала остается неизменным. С одной стороны, он возрастает вдвое из-за увеличения числа элементов, т. е. из-за повышения основания кода; с другой стороны, убывает во столько же раз вследствие того, что один элемент нового сигнала представляет пару элементов первоначального сигнала.

Пусть n и m — число элементов и основание кода для первоначального, а n_1 и m_1 — те же величины для укрупненного сигнала. Тогда

$$V_1 = n_1 \log m_1 = \frac{n}{2} \log m^2 = n \log m = V.$$

Так как в укрупненном сигнале $n_1 = n/2$, то полные количества сведений до преобразования и после преобразования выражаются соответственно как

$$I = nI_1, \quad I_k = n_1 I'_k = \frac{n}{2} I'_k.$$

Полезный результат укрупнения можно характеризовать относительным увеличением количества сведений, т. е. величиной

$$\frac{I_k - I}{I} = \frac{I'_k}{2I'} - 1.$$

Если бы пары были совершенно независимы, то укрупненный сигнал был бы полностью декоррелирован [1]. В действительности корреляция может существовать между любыми соседними элементами. Поэтому корреляция имеется и между парами, так что мы должны принять

$$I'_k = -\sum_{k=1}^{m^2} p(k) \sum_{l=1}^{m^2} p(l/k) \log p(l/k),$$

где l — данная пара; k — предшествующая. Но корреляция между парами, очевидно, слабее, чем между элементами первоначального сигнала. Это можно объяснить тем, что лишь последний элемент предшествующей пары оказывает влияние на первый элемент последующей пары. Поэтому, объединив элементы в пары, мы частично декоррелировали сигнал, и избыточность R_p для укрупненного сигнала меньше, чем для первоначального.

Процесс укрупнения отрезков сигнала можно было бы продолжить. Так, например, пары — диаграммы — можно было бы соединить в четырехэлементные группы — тетраграммы. Это дало бы дальнейшую декорреляцию. Если сообщение разделено на отрезки l_s, состоящие из r элементов каждый, и если общее возможное число таких отрезков есть $m_1 = m^r$, то количество сведений на отрезок для полностью декоррелированного сигнала есть

$$-\sum_{s=1}^{m_1} p(l_s) \log p(l_s),$$

---

[1] Так обстояло бы дело, если бы корреляция существовала только между элементами 1 и 2, 3 и 4, . . ., l и $l+1$, . . . Тогда между парами 1, 2; 3, 4; . . .; l, $l+1$; . . . корреляция отсутствовала бы.

а количество сведений на элемент

$$I' = -\frac{1}{r}\sum_{s=1}^{m_1} p(l_s)\log p(l_s).$$

Итак, один из возможных методов декорреляции сигнала состоит в том, что формируется новый сигнал, элементами которого являются отрезки из того или иного числа элементов первоначального сигнала. Эта операция уменьшает избыточность в тем большей мере, чем длиннее берутся отрезки.

Поясним сказанное на примере, для которого возьмем цифры из примера § 27, а именно $m=2$ (элементы a и b) $p(a)={}^3/_4$. $p(b)={}^1/_4$, $p(a/a)={}^2/_3$, $p(b/a)={}^1/_3$, $p(a/b)=1$, $p(b/b)=0$.

Вычислим вероятности пар, т. е. $p(ij)$

$$p(ab) = p(a)\,p(b/a) = \frac{3}{4}\cdot\frac{1}{3} = \frac{1}{4},$$

$$p(ba) = p(b)\,p(a/b) = \frac{1}{4}\cdot 1 = \frac{1}{4},$$

$$p(aa) = p(a)\,p(a/a) = \frac{3}{4}\cdot\frac{2}{3} = \frac{1}{2},$$

$$p(bb) = p(b)\,p(b/b) = \frac{1}{4}\cdot 0 = 0.$$

Теперь подсчитаем содержательность для сигнала, составленного из пар элементов

$$I'_k = -\sum_{k=1}^{m^2} p_k \log p_k =$$

$$= [p(ab)\log p(ab) + p(ba)\log p(ba) + p(aa)\log p(aa) +$$

$$+ p(bb)\log p(bb)] = 2\cdot\frac{1}{4}\log 4 + \frac{1}{2}\log 2 = 1{,}5.$$

Полное количество сведений

$$I_k = n_1 I'_k = \frac{n}{2} I'_k = 0{,}75n.$$

Для первоначального сигнала

$$I = -n\sum_{i=1}^{m} p(i)\sum_{j=1}^{m} p(j/i)\log p(j/i) = 0{,}685n.$$

Приращение количества сведений

$$I_1 = I_k - I = 0{,}065n.$$

Относительное увеличение количества сведений составляет

$$\frac{I_1}{I}\cdot 100 = \frac{0{,}065}{0{,}685}\cdot 100 = 9{,}5\ \%.$$

Полученный частично декоррелированный сигнал можно подвергать дальнейшей обработке — перекодированию и перемодуляции. Но предварительно нужно нормализовать распределение вероятностей.

Оливер [26] предложил устройство, названное им «диаграммер», в котором совмещены функции формирования укрупненного сигнала и перераспределения вероятностей. В принципе это устройство представляет собой нелинейный шестиполюсник, на два входа которого подаются значения данного и предшествующего элементов первоначального сигнала. С выхода снимается новый

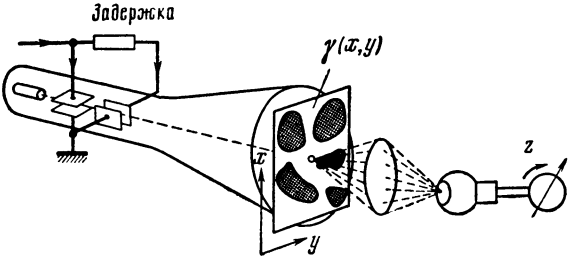


Рис. 23

сигнал с надлежащим распределением вероятностей. Устройство показано схематически на рис. 23. Первоначальный сигнал подается на одну пару отклоняющих пластин электронно-лучевой трубки. На другую пару пластин подается тот же сигнал, но с задержкой на время, равное длительности одного элемента. В результате этого луч отклоняется одновременно в одном направлении — скажем, по вертикали (x) — значением данного элемента, а в другом направлении — по горизонтали (y) — значением предшествующего элемента. Положение луча на экране трубки определяется двумя координатами, т. е. парой значений ij. Остается получить сигнал с желаемым распределением вероятностей. Для этого служит маска с соответствующим двумерным распределением прозрачности $\gamma(x, y)$. Световой поток, проходящий через маску, попадает на фотоэлемент и создает соответствующий ток, изменение которого и представляет собой выходной сигнал (z). В принципе возможно подобным же образом учесть взаимовлияние и более удаленных элементов сигнала, но это ведет к чрезвычайному усложнению аппаратуры.

§ 32. Предсказание

Другой общий метод декорреляции сигнала — это метод предсказания.

Предположим, что характер вероятностных взаимосвязей таков, что, приняв некоторую часть сообщения, мы можем по свойствам этой части предсказать недостающую часть сообщения. Если бы такое предсказание могло быть сделано точно, то в пе-

редаче остальной части сообщения не было бы надобности и, следовательно, время передачи, а с ним и объем сигнала могли бы быть соответственно сокращены. Иногда это возможно; примерами могут служить хотя бы общепринятые сокращения слов и целых фраз. Но в общем случае точное предсказание невозможно; возможно лишь приближенное предсказание, основанное на знании статистики сообщения. Однако и приближенное предсказание полезно с точки зрения сокращения объема сигнала. Мы можем передавать не данный элемент сообщения, а лишь разность между предсказанным и действительным значениями этого элемента. Тогда и на приемном конце можно производить предсказание данного элемента сообщения и, прибавляя к предсказанному значению вышеуказанную разность, поступающую по каналу связи, восстанавливать истинное значение данного элемента. Легко предвидеть, что таким путем можно сократить сигнал, так как разностный сигнал — мы будем называть его сигналом ошибки — в значительной мере декоррелирован и обладает более выгодным распределением вероятностей. Следовательно, и мощность сигнала ошибки будет меньше, чем мощность сигнала, полностью передающего сообщение.

Прежде чем приступить к оценке выигрыша, получаемого при применении этого принципа, напомним некоторые общие положения. С математической точки зрения задача предсказания есть задача экстраполяции некоторой случайной последовательности. Обозначим элементы этой последовательности через h_k, и пусть h_0 означает данный элемент, h_1 — предыдущий, так что во времени последовательность располагается (от прошлого к настоящему) в следующем порядке:

$$h_n, \ldots, h_k, \ldots, h_2, h_1, h_0.$$

При наличии вероятностных взаимосвязей следует полагать, что наблюдаемое данное значение h_0 зависит от предшествующих значений h_k ($k=1, 2, 3, \ldots, n$). Можно предсказать значение

$$h_п = \psi(h_1, h_2, \ldots, h_n) \tag{1}$$

и потребовать, чтобы функция ψ была подобрана так, чтобы ошибка предсказания (ошибка экстраполяции)

$$\varepsilon = h_0 - h_п \tag{2}$$

была наименьшей. Это и есть задача экстраполяции.

Мы ограничимся частным случаем *линейного предсказания*, т. е. случаем, когда функция ψ выражается просто взвешенной суммой предшествующих значений

$$\psi = a_1h_1 + a_2h_2 + \ldots + a_nh_n = h_п. \tag{3}$$

Ошибка линейного предсказания равна

$$\varepsilon = h_0 - h_п = h_0 - \sum_{k=1}^{n} a_kh_k. \tag{4}$$

Средний квадрат ошибки

$$\bar{\varepsilon}^2 = \bar{h}_0^2 - 2\sum_{k=1}^{n} a_k \overline{h_0 h_k} + \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l \overline{h_k h_l} \tag{5}$$

или, вводя коэффициенты корреляции,

$$B(0) = \bar{h}_0^2,\ B(k) = \overline{h_0 h_k},\ B(k-l) = \overline{h_k h_l},$$

$$\bar{\varepsilon}^2 = B(0) - 2\sum_{k=1}^{n} a_k B(k) + \sum_{k=1}^{n}\sum_{l=1}^{n} a_k a_l B(|k-l|). \tag{6}$$

Будем теперь подбирать коэффициенты a_k так, чтобы средний квадрат ошибки был наименьшим. Для этого нужно продифференцировать $\bar{\varepsilon}^2$ по a_k и определить все a_k из полученной системы уравнений. Мы получим

$$\frac{\partial \bar{\varepsilon}^2}{\partial a_k} = -2B(k) + 2\sum_{l=1}^{n} a_l B(|k-l|) = 0. \tag{7}$$

В развернутом виде эта система уравнений выглядит так:

$$\begin{aligned}
&a_1 B(0) + a_2 B(1) + a_3 B(2) + \ldots + a_n B(n-1) = B(1),\\
&a_1 B(1) + a_2 B(0) + a_3 B(1) + \ldots + a_n B(n-2) = B(2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1 B(n) + a_2 B(n-1) + a_3 B(n-2) + \ldots + a_n B(1) = B(n).
\end{aligned}$$

Рассмотрим частный случай, когда коэффициент корреляции есть показательная функция разности $k-l$, т. е.

$$B(|k-l|) = C\alpha^{|k-l|}, \tag{8}$$

где $\alpha < 1$. Именно такого характера корреляция была нами получена в § 23 для телеграфного сигнала. К такой же форме приближается и корреляция телевизионных сигналов.

Подставляя (8) в (7), находим

$$-\alpha^k + \sum a_l \alpha^{|k-l|} = 0.$$

Очевидным решением этой системы уравнений являются следующие значения коэффициентов a_l:

$$a_1 = \alpha,\quad a_2 = a_3 = \ldots = a_m = 0.$$

Таким образом, в рассматриваемом случае, т. е. когда коэффициент корреляции выражается показательной функцией (8), предсказание основывается только на непосредственно предшествующем элементе последовательности. Знание значений всех более удаленных элементов ничего не изменяет в предсказании и не увеличивает его точности.

Рассмотрим подробнее предсказание по одному лишь предшествующему значению. Мы имеем: предсказанное значение

$$h_{\text{п}} = a_1 h_1, \tag{9}$$

ошибку предсказания

$$\varepsilon = h_0 - a_1 h_1 \tag{10}$$

и средний квадрат ошибки

$$\bar{\varepsilon}^2 = \bar{h}_0^2 - 2a_1\overline{h_0 h_1} + a_1^2\bar{h}_1^2.$$

Но так как

$$\bar{h}_0^2 = \bar{h}_1^2 = B(0),$$

то

$$\bar{\varepsilon}^2 = B(0)(1 + a_1^2) - 2a_1 B(1). \tag{11}$$

Наименьший средний квадрат ошибки получился бы при условии

$$d\bar{\varepsilon}^2/da_1 = 2a_1 B(0) - 2B(1) = 0,$$

т. е. при

$$a_1 = \frac{B(1)}{B(0)} = b_1. \tag{12}$$

При этом условии мы имели бы

$$\bar{\varepsilon}^2_{\min} = B(0) - B^2(1)/B(0) = B(0)(1 - b_1^2). \tag{13}$$

Теперь можно оценить выигрыш в мощности. $\bar{\varepsilon}^2$ есть мощность сигнала ошибки. Мощность же первоначального сигнала есть

$$\bar{h}^2 = B(0).$$

Следовательно, передавая сигнал ошибки вместо основного сигнала, мы уменьшаем мощность в отношении $\bar{\varepsilon}^2/B(0)$. Для этого отношения мы имеем из (11)

$$\bar{\varepsilon}^2/B(0) = 1 + a_1^2 - 2a_1 b_1, \tag{14}$$

а в случае, когда $a_1 = b_1$,

$$\left.\frac{\bar{\varepsilon}^2}{B(0)}\right|_{\min} = 1 - b_1^2. \tag{15}$$

Если бы корреляция отсутствовала ($b_1 = 0$), то выигрыша в мощности мы не получили бы. Если же нормированный коэффициент корреляции b_1 между соседними элементами был бы равен единице, то это означало бы возможность точного предсказания и средний квадрат ошибки, а следовательно, и мощность сигнала ошибки были бы равны нулю.

Положим, что в качестве предсказываемого значения мы берем предыдущие, т. е. принимаем

$$a_1 = 1; \quad h_п = h_1.$$

Тогда

$$\varepsilon = h_0 - h_1; \quad \bar{\varepsilon}^2 = \bar{h}_0^2 - 2\overline{h_0 h_1} + \bar{h}_1^2 = 2\,[B(0) - B(1)],$$

отношение мощностей

$$\bar{\varepsilon}^2 / B(0) = 2\,(1 - b_1). \tag{16}$$

Как видим, в этом случае выигрыш в мощности может быть получен только при условии, что нормированный коэффициент

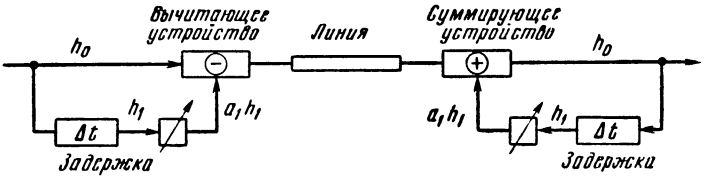


Рис. 24

корреляции между соседними элементами не менее половины.

При наилучшем же предсказании по формуле

$$h_п = b_1 h_1$$

мы получаем выигрыш в мощности при любом $b_1 < 1$, как это видно из формулы (15).

Скелетная схема системы связи с использованием вышеописанного принципа представлена на рис. 24. Предсказание в этой системе производится только по предшествующему значению. В передатчике имеется вычитающее устройство, на которое подаются одновременно данный элемент h_0 и умноженный на a_1 предшествующий элемент h_1. Для этого предшествующий элемент задерживается на интервал Δt, разделяющий два соседних элемента. На выходе вычитающего устройства получается ошибка ε, которая и посылается в линию. В приемнике имеется суммирующее устройство, складывающее поступившую с линии ошибку ε с предыдущим элементом $a_1 h_1$. Этот элемент получается также путем задержки на Δt. На выходе суммирующего устройства получаем снова требуемый элемент h_0.

Подобным же образом, но с применением цепочки задерживающих звеньев может быть построена и система с линейным предсказанием по нескольким предшествующим значениям.

Возможность предсказания не исчерпывается статистическим предсказанием. Так, например, для некоторой реализации может быть использован ряд Тэйлора

$$f(t_1 + \Delta t) = f(t_1) + f'(t_1)\frac{\Delta t}{1!} + f''(t_1)\frac{\Delta t^2}{2!} + \ldots + f^{(n)}(t_1)\frac{\Delta t^n}{n!} + R_n.$$

При этом подразумевается, что последовательность значений $\ldots, h_m, \ldots, h_2, h_1, h_0$ получается путем взятия отсчетов мгновенных значений непрерывной функции сигнала $f(t)$ через интервалы Δt. Таким образом, $h_1=f(t_1)$, $h_0=f(t_1+\Delta t)$. Кроме того, за функцией $f(t)$ ведется непрерывное наблюдение, так что в требуемые моменты могут быть известны и значения ее производных. Ограничившись двумя первыми членами ряда, мы будем иметь

$$f(t+\Delta t)=f(t_1)+f'(t_1)\Delta t+R_1.$$

Определим предсказанное значение как

$$h_п=h_1+f'_1\Delta t,$$

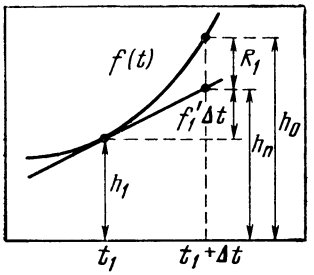


Рис. 25

где $f'_1=f'(t_1)$ — значение первой производной функции $f(t)$ в момент взятия отсчета h_1, и, принимая во внимание

$$h_0=h_1+f'_1\Delta t+R_1,$$

найдем, что ошибка предсказания выражается непосредственно остаточным членом ряда Тэйлора (рис. 25). Такой способ предсказания представляет известный интерес с технической точки зрения.

Мы рассматриваем систему с предсказанием, в которой передается сигнал ошибки предсказания, а истинное значение восстанавливается на приемном конце путем добавления ошибки предсказания к предсказанному значению. Как мы видели, применение этой системы позволяет сократить объем сигнала. Но в описанном виде эта система обладает одним существенным недостатком.

Дело в том, что сигнал ошибки содержит погрешность, обусловленную помехой. Либо помеха налагается на сигнал в линии, либо же при применении квантования помеха есть шум квантования. Помеха имеет случайный характер, и среднее значение ее можно считать равным нулю. Но флюктуации, обусловленные помехой, растут с течением времени и могут привести к недопустимому искажению принятого сигнала. Нижеследующие выкладки поясняют это положение.

Пусть первый переданный отсчет функции сообщения был $h_m=\varepsilon_m$. Второй отсчет h_{m-1}, передана ошибка $\varepsilon_{m-1}=h_{m-1}-h_m$. Третий отсчет h_{m-2}, ошибка ε_{m-2} и т. д. На приемном конце восстановление истинных отсчетов производится в следующем порядке:

$$h_m=\varepsilon_m,\quad h_{m-1}=h_m+\varepsilon_{m-1},\quad h_{m-2}=$$

$$=h_{m-1}+\varepsilon_{m-2}=h_m+\varepsilon_{m-1}+\varepsilon_{m-2}=\varepsilon_m+\varepsilon_{m-1}+\varepsilon_{m-2},\ \ldots,\quad h_k=\sum_{i=m}^{k}\varepsilon.$$

Так обстояло бы дело, если бы не было помехи. Но в действительных условиях на приемный конец поступает сигнал, представляющий собой сумму сигнала ошибки ε и помехи ξ. Таким образом, на приемном конце функция сообщения восстанавливается по следующей схеме:

$$h_m^* = \varepsilon_m + \xi_m,\ h_{m-1}^* = h_m^* + \varepsilon_{m-1} + \xi_{m-1},$$
$$h_{m-2}^* = h_{m-1}^* + \varepsilon_{m-2} + \xi_{m-2} + \ldots$$
$$\ldots,\ h_k^* = \sum_{i=m}^{k} (\varepsilon_i + \xi_i) = h_k + \sum_{i=m}^{k} \xi_i = h_k + \xi_k.$$

Итак, значение h_k^* функции сообщения, восстанавливаемое на приемном конце путем последовательного суммирования, отличается от истинного значения h_k на величину $\Sigma\, \xi_i$. Среднее значение этой величины равно нулю. Но флюктуации суммы возрастают с увеличением числа слагаемых. Мы имеем (в предположении независимости ξ_i)

$$D(\xi) \sum_{i=1}^{n} D(\xi_i) = nD(\xi).$$

Считая, что помеха обусловлена квантованием и что, следовательно, ξ может принимать с равной вероятностью все значения на промежутке $(-\delta/2,\ \delta/2)$, где δ — шаг шкалы квантования, получим

$$D(\xi) = \frac{1}{\delta} \int_{-\delta/2}^{\delta/2} x^2 dx = \frac{1}{12}\, \delta^2.$$

Среднеквадратичное уклонение от истинного значения для n-го по порядку отсчета составит

$$\sqrt{\overline{\zeta^2}} = \sqrt{\frac{n}{12}}\, \delta.$$

Такое накопление погрешности может, однако, быть устранено введением самопроверки. Нужно дублировать процесс восстановления функции сообщения на передающем конце и сообразовывать передаваемые сигналы с результатом этого процесса. Таким образом, накопление погрешностей может быть предотвращено. Ясно, что для того, чтобы образование погрешности могло быть взято под контроль на передающем конце, необходимо прибегнуть к квантованию.

Пусть h_k — истинное значение функции сообщения в k-й момент; h_{k+1} — значение в предшествующий момент; $\varepsilon_k = h_k - h_{k+1}$ — истинная разность. Введем разность

$$\varepsilon_k^* = h_k - h_{k+1}^*,$$

где

$$h_k^* = \sum_{i=m}^{k} (\varepsilon_i^* + \xi_i)$$

есть результат суммирования квантованных разностей ξ^*; ξ есть погрешность квантования. Проследив процесс суммирования, можно найти соотношение

$$\varepsilon_k^* = \varepsilon_k - \xi_{k+1}.$$

Подставив это соотношение в выражение для h_k^*, получим

$$h_k^* = \sum_{i=m}^{k} (\varepsilon_i^* + \xi_i) = \sum_{i=m}^{k} (\varepsilon_i + \xi_i - \xi_{i+1}) = \xi_k + \sum_{i=m}^{k} \varepsilon_i = h_k + \xi_k.$$

Таким образом, все погрешности, кроме последней, исключаются и h_k^* воспроизводит истинное значение h_k с точностью до $\delta/2$.

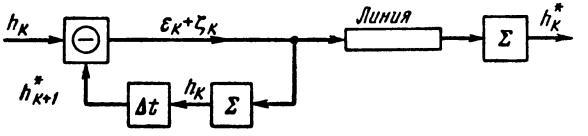


Рис. 26

Скелетная схема системы с самопроверкой показана на рис. 26. Значение h_k, поданное на вход, попадает в вычитающее устройство. Сюда же поступает значение h_{k+1}^*, получаемое путем суммирования и задержки на Δt квантованных разностей $\varepsilon_k^* + \xi_k$. Суммирующее устройство дублирует совершенно аналогичное устройство на приемном конце системы.

Применяя систему передачи с предсказанием, мы образуем сигнал ошибки, представляющий собой разность между предсказанным и истинным значениями передаваемой функции. Сигнал ошибки, вообще говоря, имеет более слабую автокорреляцию, нежели исходная функция; предсказание приводит к декорреляции сигнала. Разберем возникающие при этом соотношения, ограничившись случаем линейного предсказания по одному предшествующему значению.

Пусть передаваемая функция есть $f(t)$. Будем вести непрерывное предсказание на Δt вперед, основываясь на значении функции в данный момент. Процесс предсказания выразится соотношением

$$f(t + \Delta t) = af(t) + \varepsilon(t),$$

где $\varepsilon(t)$ — ошибка предсказания. Найдем функцию автокорреляции ошибки. Имеем

$$\varepsilon(t) = f(t + \Delta t) - af(t),$$

$$B_\varepsilon(\tau) = \overline{\varepsilon(t)\,\varepsilon(t-\tau)} =$$

$$= \overline{[f(t+\Delta t) - af(t)]\,[f(t+\Delta t-\tau) - af(t-\tau)]} =$$

$$= \overline{f\,(t+\Delta t)\,f(t+\Delta t-\tau)} + a^2\overline{f(t)\,f(t-\tau)} -$$

$$- a\overline{f(t)\,f(t+\Delta t-\tau)} - a\overline{f(t+\Delta t)\,f(t-\tau)}.$$

Введя автокорреляцию функции $f(t)$

$$B_f(\tau) = \overline{f(t) f(t-\tau)},$$

получим

$$B_\varepsilon(\tau) = (1+a^2) B_f(\tau) - a[B_f(\tau+\Delta t) + B_f(\tau-\Delta t)].$$

Для нормировки разделим это выражение на $B_\varepsilon(0)$

$$b_\varepsilon(\tau) = \frac{B_\varepsilon(\tau)}{B_\varepsilon(0)} = \frac{(1+a^2) b_f(\tau) - a\ [b_f(\tau+\Delta t) + b_f(\tau-\Delta t)]}{1+a^2-2ab_f(\Delta t)}.$$

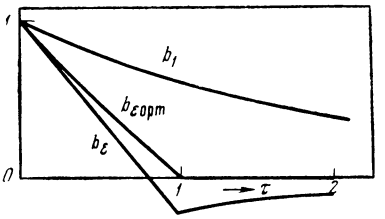


Рис. 27

Если $a=1$, т. е. если за ожидаемое (предсказываемое) значение берется просто предшествующее значение, то

$$b_\varepsilon(\tau) = \frac{b_f(\tau) - \frac{1}{2}[b_f(\tau+\Delta t) + b_f(\tau-\Delta t)]}{1-b_f(\Delta t)}. \qquad (17)$$

Если же берется оптимальное предсказание, т. е.

$$a = b_f(\Delta t),$$

то

$$b_\varepsilon(\tau) = \frac{[1+b_f^2(\Delta t)]\, b_f(\tau) - b_f(\Delta t)\,[b_f(\tau+\Delta t) + b_f(\tau-\Delta t)]}{1-b_f^2(\Delta t)}. \qquad (18)$$

Рассмотрим пример. Пусть

$$b_f(\tau) = e^{-\mu|\tau|}.$$

Подставляя это в формулу (17), получим

$$b_\varepsilon(\tau) = \begin{cases} \dfrac{1}{1-e^{-\mu\Delta t}}\left[e^{-\mu\tau} - \dfrac{1}{2}e^{-\mu\Delta t}(e^{-\mu\tau}+e^{\mu\tau})\right] & [0<\tau<\Delta t], \\ \dfrac{e^{-\mu\tau}}{1-e^{-\mu\Delta t}}\left[1-\dfrac{1}{2}(e^{-\mu\Delta t}+e^{\mu\Delta t})\right] & [\Delta t<\tau<\infty]. \end{cases} \qquad (20)$$

Для случая же оптимального предсказания по формуле (18) найдем

$$b_\varepsilon(\tau)_{\text{opt}} = \begin{cases} \dfrac{1}{1-e^{-2\mu\Delta t}}(e^{-\mu\tau} - e^{-2\mu\Delta t}e^{\mu\tau}) & [0<\tau<\Delta t], \\ 0 & [\Delta t<\tau<\infty]. \end{cases} \qquad (21)$$

Графики функций (19)—(21) при $\Delta t=1$, $\mu=0{,}5$ изображены на рис. 27. Интересно отметить, что в случае оптимального предска-

зания функция автокорреляции равна тождественно нулю при $\tau > \Delta t$. Это связано с тем обстоятельством, что в случае экспоненциальной функции автокорреляции (19) предсказание по одному предшествующему (т. е. отстоящему на Δt) значению является исчерпывающим.

§ 33. «Дельта-модуляция»

Недавно предложена новая система передачи сигналов, получившая название «*дельта-модуляция*». Это название обусловлено тем, что передается не функция сообщения, а лишь ее приращение. Таким образом, Δ-модуляция относится к системам с предсказа-

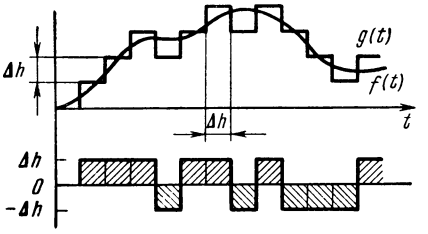


Рис. 28

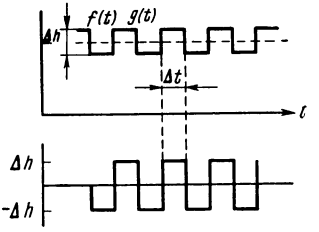


Рис. 29

нием по предыдущему значению. Для устранения накопления ошибок в схеме Δ-модуляции применена самопроверка, общий принцип которой описан в предыдущем параграфе.

Но Δ-модуляция имеет специфическую особенность, состоящую в том, что передается не величина приращения функции сообщения, а только знак этого приращения.

Таким образом, сигнал при Δ-модуляции оказывается автоматически кодированным по двоичной системе и представляет собой, например, импульсы всегда одинаковой величины, но различного знака (рис. 28). Восстановление функции сообщения на приемном конце осуществляется простым суммированием Δ-импульсов.

При Δ-модуляции заданным является стандартное приращение Δh. Постоянная величина передается посредством Δ-модуляции, как показано на рис. 29. В нижней части рисунка показаны Δ-импульсы, а в верхней — результат их суммирования. Линейное нарастание функции сообщения воспроизводится, как показано на рис. 30, *а*.

Скелетная схема Δ-модуляции изображена на рис. 31. Суммирование Δ-импульсов дает ступенчатую функцию $g(t)$. Эта функция сравнивается с функцией сообщения $f(t)$ в компараторе. Действие компаратора состоит в том, что в тактовые моменты, задаваемые импульсным генератором, он сравнивает $g(t)$ и $f(t)$ и выдает Δ-импульс стандартной величины с тем или иным знаком в зависимости от знака разности $f(t)-g(t)$.

Ясно, однако, что воспроизвести любое изменение функции сообщения суммированием импульсов стандартной величины Δh, следующих друг за другом с неизменным интервалом Δt, невозможно. На рис. 30, *б* изображен предельный случай передачи линейно нарастающей функции сообщения.

Очевидно, что для того, чтобы $g(t)$ могла следовать за изменениями $f(t)$, необходимо выполнить условие

$$f'_{\max}\Delta t \leqslant \Delta h. \tag{1}$$

С другой стороны, для того чтобы помеха квантования была не слишком велика, необходимо задать минимальное число m ступеней шкалы квантования. Вытекающее отсюда условие можно выразить неравенством

$$\Delta h \leqslant f_{\max}/m. \tag{2}$$

Беря в (1) и (2) знаки равенства, получим

$$m\Delta t = f_{\max}/f'_{\max}.$$

Число уровней m задается качественными требованиями, а отношение

$$M = f_{\max}/f'_{\max}$$

определяется статистикой сообщения и, следовательно, также является заданным. Таким образом, оказывается, что интервал Δt, а стало быть, и частота следования Δ-импульсов определяются однозначно

$$f_0 = 1/\Delta t = m/M.$$

Для нахождения M нужно знать спектр и распределение. Если известен спектр мощности $G(\omega)$ функции f, то спектр производной будет

$$G_1(\omega) = \omega^2 G(\omega).$$

Средний квадрат функции f

$$\overline{f^2} = \int_0^\infty G(\omega)\, d\omega,$$

а средний квадрат производной

$$\overline{f'^2} = \int_0^\infty G_1(\omega)\, d\omega = \int_0^\infty \omega^2 G(\omega)\, d\omega.$$

Положим, например, что спектр функции сообщения равномерен в полосе от нуля до ω_c и спектральная плотность имеет в пределах этой полосы постоянное значение ρ.

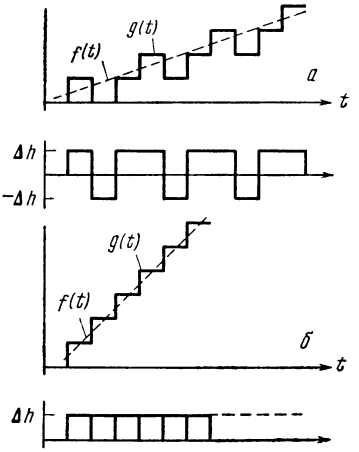


Рис. 30

Тогда

$$\overline{f^2} = \rho\omega_c, \quad \overline{f'^2} = \frac{1}{3}\rho\omega_c^3, \quad \frac{\overline{f^2}}{\overline{f'^2}} = \frac{1}{3}\omega_c^2.$$

Пусть далее

$$f_{\max} = Kf_{\text{ср}} = K\sqrt{\overline{f^2}}, \quad f'_{\max} = K_1 f'_{\text{ср}} = K_1\sqrt{\overline{f'^2}}.$$

Здесь K — коэффициент, называемый в технике *пикфактором*; соответственно K_1 — пикфактор для производной. Коэффициенты K и K_1 зависят уже только от распределения.

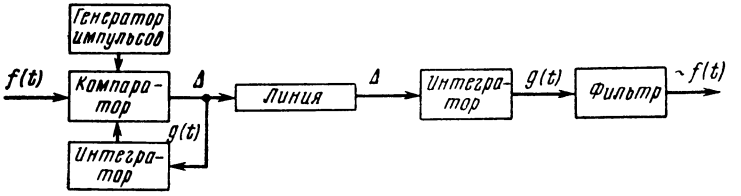


Рис. 31

Итак,

$$M = \frac{f_{\max}}{f'_{\max}} = \frac{K}{K_1} \cdot \frac{\omega_c}{\sqrt{3}},$$

и частота следования Δ-импульсов

$$f_0 = \frac{2\pi}{\sqrt{3}} \cdot \frac{K}{K_1} mf_c.$$

Если взять для телефонии $m=100$, $f_c=5$ *кгц* и положить, что K и K_1 — величины одного порядка, то получится частота порядка 1 *Мгц*. В действительности можно обойтись значительно меньшими частотами следования, что обусловлено тем, что спектр речи неравномерен и имеет острый максимум в области сравнительно низких частот (около 300 *гц*). Так, например, в описанном устройстве [24] частота следования составляла 100 *кгц*.

Мощность сигнала при Δ-модуляции равна, очевидно, Δh^2. Для импульсов в линии можно выбрать величину

$$\delta = k\sigma,$$

и тогда мощность будет

$$P = k^2 P_{\text{п}},$$

так что система работает всегда с неизменным превышением

$$H = \log\frac{P}{P_{\text{п}}} = 2\log k.$$

Система Δ-модуляции дает результаты, сравнимые с обычной системой КИМ, но выгодно отличается от последней исключительной простотой аппаратуры, что и обусловливает внимание к Δ-модуляции.

§ 34. Возможности сокращения телефонного сигнала

При передаче сообщений нет никакого смысла передавать большее количество сведений, чем то, которое способен воспринять получатель. В частности, если конечным звеном системы связи являются органы чувств человека, то количество сведений может быть без ущерба для дела ограничено с учетом свойств этих органов. Известно, например, что человеческий слух не воспринимает колебаний с частотами ниже примерно 30 *гц* и выше 12—15 *кгц*. Поэтому для совершенной (в смысле получения полной иллюзии) передачи звука можно ограничиться указанным диапазоном, хотя естественные звуки имеют гораздо более широкий спектр, захватывающий неслышимые ультразвуковые частоты. Для обычной телефонной связи субъективно совершенное воспроизведение звука вовсе не требуется. К телефонной связи предъявляются обычно два требования (не всегда, впрочем, ясно формулируемые): разборчивость речи и возможность узнать собеседника по голосу. Для выполнения этих требований оказывается достаточным передавать полосу всего лишь 2,5—3 *кгц*, что и является общепринятой нормой для каналов телефонной связи. Но и это далеко не предел возможного сокращения объема телефонного сигнала. Вот одно из простейших соображений по этому поводу. При передаче хотя и с ограниченной полосой количество сведений таково, что позволяет воспроизвести форму телефонного тока на приемном конце независимо от способа передачи (т. е. независимо от того, производится ли передача на постоянном токе, или на несущей частоте, или любым импульсным методом). Между тем, как известно, слух практически не реагирует на фазовые соотношения. Следовательно, количество сведений, получаемое при существующих методах телефонии, излишне велико. Можно было бы отбросить все то, что характеризует фазовые соотношения (т. е. передавать не форму кривой, а текущий спектр амплитуд). Это дало бы сокращение количества сведений, а стало быть, и объема сигнала вдвое. Само собой разумеется, что осуществление этой возможности требует разработки совершенно новых способов передачи.

Дальнейшая экономия могла бы быть получена за счет того, что величина различаемой ступени силы звука резко разнится для различных частот и т. д.

Если отбросить требование узнаваемости и сохранить лишь требование разборчивости, то может быть допущено еще большее сокращение количества сведений, а стало быть, и объема сигнала. Однако отыскание решений затрудняется тем, что нам неизвестны пока объективные характеристики речевого сигнала, которые однозначно определяли бы разборчивость как субъективную характеристику передачи. Поэтому в настоящее время применяется эмпирический метод поисков, состоящий в том, что речевое сообщение самым различным образом видоизменяется в целях сокращения объема сигнала, а о позволительности того или иного видоизме-

нения с точки зрения сохранения достаточной разборчивости судят по результатам субъективных испытаний.

Интересный пример такого изменения речевого сообщения представляет собой так называемая «*ограниченная речь*». Суть дела состоит в том, что допускаются только два фиксированных значения функции сообщения; единственный общий признак, который эта функция сохраняет от первоначальной, — это расположение нулей, т. е. точек, в которых функция меняет знак. Технически подобного рода преобразование можно осуществить путем значительного усиления звукового напряжения с последующим ограниче-

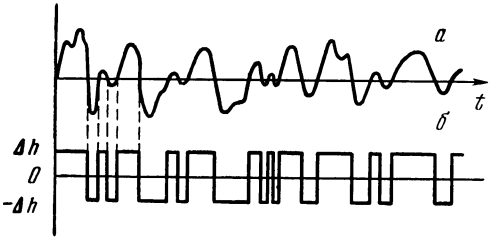


Рис. 32

нием сверху и снизу на заданных уровнях. На рис. 32, *а* изображена осциллограмма исходного звукового напряжения, а на рис. 32, *б* — «ограниченная речь». Положения нулей на рис. 32, *а* и *б* совпадают. Как ни мало похожи друг на друга графики рис. 32, *а* и *б*, оказывается, что при таком видоизменении речевого сообщения сохраняется высокая степень разборчивости. Между тем ограничение речи дает очень существенную экономию количества сведений и объема сигнала. Ведь результат ограничения превращает речевое сообщение в сигнал, кодированный по двоичной системе. Если взять для сравнения число уровней в первоначальном сообщении, равным 128 (что соответствует нормам, принятым в импульсной телефонии), то объем сигнала сокращается при ограничении в семь раз ($\log 128/\log 2=7$). Возможно, что другие, пока не известные способы видоизменения речевого сообщения позволят, сохранив требуемую разборчивость, еще больше сократить объем сигнала. В связи с этим можно заметить, что по имеющимся данным можно обеспечить лучшую разборчивость, если сохранить в видоизмененном сообщении не положение нулей, а положение экстремальных точек первоначального сообщения. Для получения такого результата нужно, очевидно, перед ограничением продифференцировать исходное сообщение.

Говоря о возможных путях существенного сокращения телефонного сигнала, следует упомянуть и о принципе так называемой «*синтетической телефонии*». Общую идею можно предварительно пояснить следующим соображением: пусть требуется передать сообщение, состоящее из примыкающих друг к другу отрезков синусоид различной амплитуды, частоты и длительности. Так как

синусоидальное колебание полностью определяется амплитудой, частотой и начальной фазой, то вовсе не нужно передавать вышеописанное сообщение полностью; достаточно лишь в моменты, соответствующие стыкам каждых двух отрезков, передавать три числа, определяющих амплитуду, частоту и начальную фазу очередного отрезка, или даже только два числа, если фаза нас не интересует. Ясно, что таким образом можно очень существенно сократить объем сигнала.

Обратимся теперь к речи. Механизм образования звуков речи вкратце состоит в том, что богатый гармониками звук голосовых связок, изменяющий свою силу и основную частоту, подвергается дальнейшей обработке в ротовой полости. Во-первых, ротовая полость работает как резонатор и, перестраиваясь, выделяет определенные частоты — *форманты*, определяющие различия между гласными звуками. Во-вторых, движение языка, зубов и губ модулирует звук, производя различные согласные. Звуки речи могут образовываться и без участия голосовых связок — за счет широкополосного шума, получающегося при вдувании воздуха в ротовую полость (как, например, при разговоре шепотом). Это довольно грубое описание достаточно для наших целей.

Мы видим, что имеется возможность производить синтез речи на приемном конце системы связи. Для этого нужен генератор звуковой частоты с богатым спектром (например, релаксационный), генератор белого шума, набор формантных фильтров (число их невелико, так как гласных звуков немного, а каждый из них достаточно хорошо определяется двумя формантами) и модулирующие органы. Располагая таким комплектом аппаратуры на приемном конце, мы можем передавать по каналу связи не речевой сигнал, а лишь командные сигналы, управляющие процессом синтеза. Именно, мы должны передавать сигналы, задающие основную частоту и силу звука генератора, имитирующего звук голосовых связок, силу звука генератора шума, сигналы, включающие те или иные формантные фильтры, и, наконец, сигналы, управляющие работой модулирующих органов. Если представить себе, что все эти сигналы передаются по каналу связи в форме импульсов, то ясно, что для каждого рода сигналов период следования не будет превосходить по порядку средней длительности фонетического элемента речи, т. е. величины порядка одной десятой секунды. Таким образом, ясно, что возможно очень значительное сокращение объема сигнала. И действительно, в описанном устройстве подобного рода общая ширина полосы, требуемая для передачи всех командных сигналов, составляла всего лишь 250 *гц* [19]. Само собой разумеется, что при такой системе передачи тонкие оттенки голоса теряются; если бы мы пожелали их сохранить, то это означало бы, что мы желаем передать большее количество сведений; выполнение этого пожелания неизбежно сопровождалось бы увеличением объема сигнала.

Глава 3

БОРЬБА С ПОМЕХАМИ

§ 35. Вводные замечания

В предыдущей главе обсуждалась одна из основных проблем связи, состоящая в достижении максимальной эффективности связи, т. е. в передаче данного количества сведений при минимальном объеме сигнала (или максимального количества сведений при данном объеме сигнала).

Но не меньшее значение имеет и вторая проблема — проблема надежности связи. Выше отмечалось, что сокращение избыточности увеличивает эффективность, но зато уменьшает надежность связи. Более внимательное изучение положения показывает, что мы имеем здесь один из примеров, характерных для техники противоречий между качественными и количественными требованиями. Можно, по-видимому, сказать, что вообще всякий метод повышения надежности связан с увеличением объема сигнала, т. е. с понижением эффективности. Это утверждение пока просто постулируется и будет иллюстрировано ниже рядом примеров.

При таком положении возникает задача нахождения приемлемого компромисса между находящимися в прямом противоречии требованиями: требованием высокой эффективности, для удовлетворения которого мы должны всячески сокращать объем сигнала, и требованием высокой надежности, для удовлетворения которого мы вынуждены увеличивать объем сигнала.

Мы условились не рассматривать проблему надежности в целом: мы будем лишь рассматривать влияние свойств самой системы связи, т. е. то, что называется помехоустойчивостью системы связи.

Общая характеристика системы связи может быть сведена к двум показателям: эффективности и помехоустойчивости. Первый показатель дает количественную, второй — качественную характеристику системы. Произведение (или вообще какая-либо возрастающая функция) обоих показателей могло бы служить единой мерой качества системы связи.

Для количественного выражения названных показателей нам нужны соответствующие критерии. Для эффективности системы связи критерий нами уже выбран и использован в предыдущей главе: это — удельная содержательность сигнала, величина, показывающая степень использования объема сигнала и равная отношению содержащегося в сигнале количества сведений к объему сигнала.

Нам нужно ввести критерий и количественную меру помехоустойчивости. Заметим предварительно, что помехоустойчивость может определяться как надежность при заданных помехе, условиях распространения и т. д. (и, разумеется, в предположении технической исправности системы связи). Таким образом, для оценки

надежности и помехоустойчивости может применяться одинаковый критерий; разница будет состоять лишь в том, что на надежность влияет ряд изменяющихся факторов, при определении же помехоустойчивости эти факторы считаются заданными, так что помехоустойчивость отражает лишь влияние на надежность свойств самой системы, т. е. примененного в данной системе способа передачи.

Надежность определена как мера достоверности принятых сообщений, т. е. мера соответствия принятых сообщений переданным. Несоответствие между принятыми и переданными сообщениями, обусловленное влиянием вышеуказанных факторов, выражается ошибками в принятом сообщении. Среднее относительное количество ошибок определяется вероятностью ошибки. Эта вероятность может служить мерой несоответствия; величину, обратную вероятности ошибки, можно взять в качестве меры соответствия, т. е. надежности, а следовательно, и помехоустойчивости. Однако в хорошей системе связи вероятность ошибки выражается малой дробью. Поэтому удобнее определять надежность и помехоустойчивость через логарифм величины, обратной вероятности ошибки. Выбор основания логарифма не имеет значения; практически удобно пользоваться десятичными логарифмами.

Итак, в качестве количественной меры надежности и помехоустойчивости можно взять величину

$$S = \lg \frac{1}{p_0} = -\lg p_0, \tag{1}$$

где p_0 — вероятность ошибки.

Для уточнения положения следует указать, что мы исходим из представления о квантованном сигнале, отображающем функцию с ограниченным спектром, т. е. о сигнале, представляющем последовательность дискретных чисел. Как было установлено в главе 1, такое представление является достаточно общим. При принятом дискретном характере сигнала ошибки имеют также дискретную природу. Дело сводится к тому, что вследствие наложения помехи отдельные числа заменяются другими — неверными.

В качестве примера определим помехоустойчивость АИМ при помехе в виде белого шума. Ошибка возникает, когда мгновенное значение помехи превосходит половину шага шкалы уровней δ. Вероятность такого события для помехи с нормальным распределением и со среднеквадратичным значением σ равна (см. § 22)

$$p\left\{|\xi| > \frac{\delta}{2}\right\} = 1 - \Phi\left(\frac{1}{2\sqrt{2}} \cdot \frac{\delta}{\sigma}\right)$$

или, пользуясь для малых вероятностей асимптотическим разложением функции Лапласа,

$$p_0 \approx \frac{2\sqrt{\frac{2}{\pi}} e^{-\frac{1}{8}\left(\frac{\delta}{\sigma}\right)^2}}{\frac{\delta}{\sigma}}.$$

Помехоустойчивость по определению (1) будет

$$S = \lg \sqrt{\frac{\pi}{8}} + \lg \frac{\delta}{\sigma} + \frac{0{,}434}{8}\left(\frac{\delta}{\sigma}\right)^2 = -0{,}203 + \lg \frac{\delta}{\sigma} + 0{,}054\left(\frac{\delta}{\sigma}\right)^2.$$

График зависимости S от δ/σ представлен на рис. 33. Помехоустойчивость порядка 6 единиц может считаться уже достаточно высокой, так как такая помехоустойчивость соответствует вероятности ошибки, равной 10^{-6}. При рассматриваемых условиях такая помехоустойчивость обеспечивается при δ/σ около 10 — значение, которое уже раньше было указано как рекомендуемое.

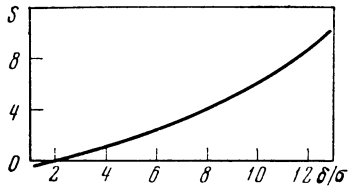


Рис. 33

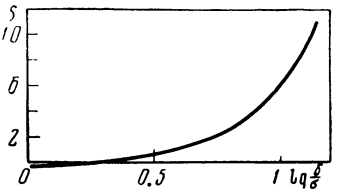


Рис. 34

Теперь посмотрим, какой ценой покупается повышение помехоустойчивости. Положим, что система работает кодом с высоким основанием ($m \geqslant 1$) и что все уровни равновероятны. При этих условиях мощность сигнала

$$P = \frac{1}{m}\Sigma h_i^2 = \frac{\delta^2}{m}\Sigma i^2 \approx \frac{1}{3}\delta^2 m^2,$$

мощность же помехи

$$P_{\text{п}} = \sigma^2.$$

Превышение сигнала над помехой

$$H = \log \frac{P}{P_{\text{п}}} = \log \frac{m^2}{3}\left(\frac{\delta}{\sigma}\right)^2.$$

Ширина спектра F и длительность T не зависят от отношения δ/σ. Поэтому объем сигнала растет пропорционально $\log(\delta/\sigma)$. Эта величина отложена по оси ординат на рис. 34; по оси абсцисс отложена надежность S (рис. 34 в сущности представляет ту же зависимость, что и рис. 33, но в другом масштабе).

Итак, повышение помехоустойчивости вызывает увеличение объема сигнала за счет увеличения превышения, т. е. за счет увеличения мощности сигнала. Рассмотренный пример является, таким образом, простейшей иллюстрацией к высказанному выше положению о связи между помехоустойчивостью и объемом сигнала. Но помехоустойчивость может быть повышена путем увеличения других измерений сигнала, т. е. ширины спектра или длительности. Примером могут служить широкополосные помехо-

устойчивые системы и в первую очередь система с применением частотной модуляции.

К вопросу об оценке помехоустойчивости можно подойти и с другой точки зрения. Заметим, что сигнал на выходе детектора приемного устройства, будь то обычный АМ детектор, или дискриминатор ЧМ приемника, или декодирующее устройство приемника ИМ, всегда оказывается преобразованным к амплитудно-модулированной форме. Это и естественно, так как в конечном счете на выходе приемника мы должны получить сообщение в его первоначальном виде, т. е. в виде некоторой функции времени, в изменениях мгновенного значения которой заложено сообщение. Преимущество систем с высокой помехоустойчивостью перед обычной системой АМ состоит в том, что при одном и том же превышении сигнала на входе приемника мы получаем большее превышение в преобразованной к АМ форме сигнала на выходе детектора. Таким образом, если определение (1) дает нам некоторую абсолютную оценку помехоустойчивости, то сравнение превышений на выходе детектора дает нам удобную относительную оценку, которой очень часто и пользуются.

В следующем параграфе дана именно такая сравнительная оценка помехоустойчивости АМ и ЧМ.

§ 36. Сравнение амплитудной и частотной модуляции

Преимущество частотной модуляции перед амплитудной в отношении помехостойкости общеизвестно. Мы проделаем вывод, показывающий это преимущество, с тем, чтобы обсудить результат с интересующей нас точки зрения.

Прежде всего напомним, что наложение на колебание несущей частоты некоторого постороннего — пока пусть синусоидального — колебания порождает модуляцию обоих родов, т. е. как амплитудную, так и частотную. Это положение легче всего уясняется с помощью векторной диаграммы рис. 35. На этом рисунке изображен вектор колебания несущей частоты с амплитудой X_0 и частотой ω_0 и вектор постороннего колебания с амплитудой A и частотой ω. Разложим вектор A на радиальную компоненту B и тангенциальную компоненту C. Компонента B дает амплитудную модуляцию

$$X_0 + B = X_0 + A\cos(\omega_0 - \omega)t = X_0\left[1 + \frac{A}{X_0}\cos(\omega_0 - \omega)t\right],$$

так что глубина амплитудной модуляции

$$m_A = A/X_0. \tag{1}$$

Компонента C дает модуляцию фазы. Мы имеем

$$\varphi \approx \operatorname{tg}\varphi = \frac{C}{X_0} = \frac{A}{X_0}\sin(\omega_0 - \omega)t.$$

Отсюда мгновенное изменение частоты

$$\dot{\varphi} = \frac{A}{X_0}(\omega_0 - \omega)\cos(\omega_0 - \omega)t,$$

глубина частотной модуляции

$$m_\omega = \frac{A}{X_0} \cdot \frac{\omega_0 - \omega}{\omega_0} \tag{2}$$

и отклонение частоты

$$\delta\omega = m_\omega \omega_0 = \frac{A}{X_0}(\omega_0 - \omega). \tag{3}$$

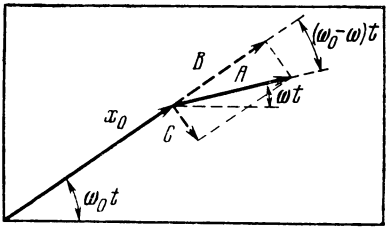


Рис. 35

Вызываемый этим отклонением сигнал на выходе дискриминатора

$$y = k\delta\omega = k\frac{A}{X_0}(\omega_0 - \omega), \tag{4}$$

где k — крутизна дискриминатора.

Теперь предположим, что постороннее колебание, представляющее помеху, есть случайный процесс $\xi(t)$, характеризующийся спектром мощности $G(\omega)$. Бесконечно малая мощность, приходящаяся на полосу $d(\omega)$, может быть выражена как

$$dP = G(\omega)\,d\omega = \frac{1}{2}\,dA^2. \tag{5}$$

На выходе же дискриминатора будем иметь согласно (4)

$$dP_{\text{п}} = \frac{1}{2}\,dy^2 = \frac{k^2}{X_0^2}(\omega_0 - \omega)^2 \frac{1}{2}\,dA^2 = \frac{k^2}{X_0^2}(\omega_0 - \omega)^2 G(\omega)\,d\omega. \tag{6}$$

Для получения полной мощности помехи на выходе нужно проинтегрировать это выражение по всей полосе пропускания

$$P_{\text{п}} = \frac{k^2}{X_0^2}\int_{\omega_0-\Omega}^{\omega_0+\Omega}(\omega_0 - \omega)^2 G(\omega)\,d\omega = \frac{2k^2}{X_0^2}\int_0^{\Omega}\omega_1^2 G(\omega_0 - \omega_1)\,d\omega_1, \tag{7}$$

здесь Ω — полуширина полосы пропускания, принимаемая равной ширине спектра сообщения.

Определим теперь мощность полезного сигнала, предположив для простоты, что он представляет собой синусоидальное колебание с амплитудой, которая дает наибольшее частотное отклонение $\Delta\omega$. Для полезного сигнала после дискриминатора будем иметь

$$y_1 = k\Delta\omega$$

и мощность сигнала

$$P_c = \frac{1}{2} y_1^2 = \frac{1}{2} k^2 \Delta\omega^2. \tag{8}$$

Сопоставляя (7) и (8), получим следующее выражение для отношения мощностей сигнала и помехи:

$$\left(\frac{P_c}{P_п}\right)_\omega = \frac{\Delta\omega^2 X_0^2}{4\int_0^\Omega \omega_1^2 G d\omega_1}. \tag{9}$$

Индекс ω указывает, что полученное выражение относится к частотной модуляции.

Формулу (9) мы сравним с аналогичным выражением для наиболее глубокой (стопроцентной) амплитудной модуляции, для которой

$$P_c = \frac{1}{2} X_0^2, \quad P_п = 2\int_0^\Omega G d\omega_1$$

и

$$\left(\frac{P_c}{P_п}\right)_A = \frac{X_0^2}{4\int_0^\Omega G d\omega_1}. \tag{10}$$

Пусть помеха представляет собой белый шум с однородным спектром. Обозначая

$$dP/d\omega = G(\omega) = \rho = \text{const},$$

получим вместо (10)

$$(P_c/P_п)_A = X_0^2/4\Omega\rho, \tag{11}$$

а вместо (9)

$$(P_c/P_п)_\omega = 3X_0^2\Delta\omega^2/4\Omega^3\rho = 3(\Delta\omega/\Omega)^2 (P_c/P_п)_A. \tag{12}$$

Эта формула показывает, что при частотной модуляции получается выигрыш (по сравнению с амплитудной модуляцией) в отношении мощностей сигнала и помехи, зависящий от $\Delta\omega/\Omega$.

Величина $\Delta\omega/\Omega$ есть так называемый *индекс частотной модуляции*. В современной технике применяются значения индекса порядка десяти. Так, например, общепринятая величина частотного отклонения составляет 75 *кгц*. При наивысшей частоте модуляции $\Omega = 5$ *кгц* (телефония) индекс частотной модуляции ра-

вен 15. При таких данных мы получаем, что отношение мощностей сигнала и помехи при частотной модуляции возрастает в $3 \cdot 15^2 = 675$ раз по сравнению с амплитудной модуляцией. Другими словами, применение частотной модуляции с индексом, равным 15, эквивалентно увеличению мощности при амплитудной модуляции в 675 раз. Очевидно, что надежность связи должна при этом соответственно возрасти, как оно и есть в действительности [1].

Но нам следует теперь выяснить цену получаемого выигрыша. При переходе от амплитудной модуляции к частотной затрачиваемая мощность не только не возрастает, но даже несколько сокращается, потому что, как известно, мощность при частотной модуляции остается равной мощности немодулированной несущей, а при стопроцентной амплитудной модуляции мощность увеличивается в полтора раза (при синусоидальной модуляции). Но зато ширина спектра сигнала при частотной модуляции существенно возрастает. Увеличение ширины спектра дает непосредственно величина индекса модуляции, так как при больших значениях индекса действительная ширина спектра частотно-модулированного колебания приближается к удвоенной величине частотного отклонения. Следовательно, увеличение отношения мощностей сигнала и помехи и связанное с этим повышение надежности связи при частотной модуляции непосредственно связаны с расширением спектра сигнала и, стало быть, с увеличением его объема. Количественные соотношения выражаются формулой (12).

Итак, частотная модуляция дает нам пример повышения надежности путем увеличения объема сигнала за счет ширины его спектра.

В последующих параграфах рассматривается возможность повышения надежности путем увеличения длительности сигнала.

§ 37. Метод накопления

Существует один издавна известный и применяемый в самых различных формах метод борьбы с помехами. Метод этот состоит в многократном повторении сигнала. Несколько принятых образцов или экземпляров сигнала оказываются по-разному искаженными помехой, так как сигнал и помеха — процессы независимые. Поэтому, сличая на приемном конце несколько экземпляров одного и того же сигнала, можно восстановить истинную форму переданного сигнала с тем большей уверенностью, чем большим числом экземпляров сигнала мы располагаем. Так как дело сводится в конечном счете к некоторому суммированию отдельных образцов сигнала, то метод этот может быть назван *методом накопления*.

---

[1] Надо, однако, заметить, что весь наш вывод относится к случаю малой помехи, т. е. к случаю значительного превышения сигнала над помехой. При малых превышениях помехоустойчивость ЧМ резко падает, о чем говорится ниже (§ 46).

В простейшей форме метод накопления применяется при телефонном разговоре в условиях плохой слышимости, когда мы переспрашиваем и повторяем одно и то же слово по нескольку раз.

Чтобы пояснить сущность метода накопления, можно сослаться на пример, не относящийся к связи, а именно на кинопроекцию. Отдельные кадры фильма грубо искажены разного рода дефектами. Однако на каждом кадре эти дефекты различны. Поэтому при быстром следовании кадров друг за другом дефекты почти не воспринимаются глазом, тогда как изображения на соседних кадрах тесно коррелированы между собой и общее зрительное впечатление получается накоплением впечатлений от последовательности кадров. Эта же идея применяется в криминалистике при наложении одного на другой нескольких слабых негативов.

Рассмотрим одну из форм применения метода накопления в телефонной связи. Суть дела состоит в том, что каждая комбинация двоичного телеграфного кода из элементов 0 и 1 передается несколько раз. Помеха не может изменить посылку 1 (наличие тока), но может дать ложный сигнал вместо посылки 0 (отсутствие тока). Несколько принятых комбинаций суммируются в накопителе. По окончании повторений накопитель срабатывает, выдавая нули на тех местах, где нуль фактически появился хотя бы один раз. Это иллюстрируется нижеследующей примерной табличкой:

Переданная комбинация	01001
Первая принятая комбинация	01011
Вторая » »	11011
Третья » »	01101
Комбинация на выходе накопителя . . .	01001

Легко видеть, что помехоустойчивость повышается этим методом. Действительно, вероятность ошибки есть p_0. Тогда, считая, что отдельная ошибка есть событие независимое, получим, что вероятность одной и той же ошибки при каждом из n повторений есть p_0^n. Если помехоустойчивость при однократной передаче есть

$$S = \lg \frac{1}{p_0},$$

то при n-кратном повторении помехоустойчивость будет

$$S = \lg \frac{1}{p_0} = nS,$$

т. е. возрастает в n раз. Если, например, вероятность ошибки равна 0,1 ($S=1$), то при трехкратном повторении вероятность ошибки упадет до 0,001 ($S_3=3$).

Для технического осуществления метода накопления необходимо, во-первых, устройство для повторения сигнала на передающем конце, а во-вторых, суммирующее устройство—накопитель — на приемном конце. В качестве накопителя могут при известных условиях использоваться органы чувств человека (слух и зрение), обладающие, как известно, определенными постоянными вре-

мени и, следовательно, интегрирующими свойствами. Для накопления могут использоваться также различные физические явления, в которых проявляется свойство инерционности, характеризуемое также некоторой постоянной времени. Так, например, в радиолокации используется в качестве накопителя электронно-лучевая трубка с послесвечением. Локационный сигнал представляет собой периодическую последовательность коротких импульсов; поэтому метод накопления в радиолокации особенно естествен и удобен. Применяются также трубки с накоплением в форме потенциального рельефа. Возможностям метода накопления в радиолокации посвящен ряд серьезных работ, но мы не будем заниматься этой областью.

За последнее время появился ряд разновидностей метода накопления, интересных как в принципиальном, так и в практическом отношениях.

Некоторые из этих разновидностей будут рассмотрены в последующих параграфах. Пока что можно отметить лишь следующее обстоятельство: метод накопления по основной своей идее связан с повторением сигнала. Следовательно, общая длительность сигнала при n повторениях возрастает в n раз, и мы видим, что повышение надежности при применении метода накопления покупается ценой увеличения объема сигнала путем увеличения его длительности. Таким образом, мы располагаем теперь примерами повышения надежности в результате увеличения всех трех измерений сигнала.

Принципиальную возможность повышения надежности связи методом накопления можно пояснить следующим общим рассуждением. Задача борьбы с помехами есть в сущности задача отличения сигнала от помехи или разделения сигнала и помехи. Отличить сигнал от помехи можно было бы по некоторому определенному физическому признаку, которым сигнал обладал бы, а помеха нет, или наоборот. Но таких признаков нет; и сигнал, и помеха представляют собой случайные процессы, существующие одновременно и обладающие перекрывающимися спектрами, так что единственная как будто бы возможность состоит в применении квантования с достаточно большим отношением δ/ω. Так вот, если мы будем сигнал периодически повторять, то мы сообщим ему новое свойство, существенно отличающее его от помехи: периодичность. Дело сводится теперь к тому, чтобы использовать новое качество сигнала для отличения его от помехи, что и осуществляется накопителем: периодический сигнал когерентен, а помеха некогерентна, а поэтому сигнал и помеха суммируются по разным законам, и превышение сигнала над помехой на выходе накопителя растет неограниченно с увеличением времени накопления.

Не нужно, конечно, думать, что сигнал должен стать периодическим в точном смысле этого слова, т. е. навечно. В этом случае сигнал не нес бы никаких сведений, т. е. не был бы сигналом. Речь идет о том, что свойство периодичности сообщается сигналу

на некоторое конечное время, достаточное для того, чтобы в результате накопления превышение сигнала над помехой достигало желаемой величины.

По поводу метода накопления следует еще заметить, что этот метод позволяет в принципе осуществлять связь в тяжелых условиях, т. е. при малых и даже отрицательных превышениях сигнала над помехой, когда мощность помех превосходит (и даже значительно превосходит) мощность сигнала.

§ 38. Фильтрация периодического сигнала

В этом и следующих параграфах мы рассмотрим возможность приема периодического сигнала, покрытого помехой. Теоретически оказывается возможным обнаружить и зарегистрировать сигнал при сколь угодно малом отношении мощности сигнала к мощности помехи, т. е. при сколь угодно большом отрицательном превышении. При этом, однако, возникают специфические соотношения, которые мы и будем разбирать.

Первая, наиболее очевидная возможность выделения периодического сигнала из смеси его с помехой состоит в применении фильтрации, т. е. частотного разделения смеси.

Идея фильтрации основана на том, что спектр помехи — имеется в виду, в частности, белый шум — однороден и характеризуется мощностью, приходящейся на единицу ширины полосы частот; в то же время периодический сигнал имеет спектр из дискретных линий, имеющих бесконечно малую протяженность по шкале частот. Поэтому можно себе представить приемное устройство, состоящее из набора фильтров, пропускающих очень узкие полосы частот, включающие в себя частоты гармоник периодического сигнала. При таких условиях мощность шума на выходе каждого фильтра будет мала и тем меньше, чем уже пропускаемая фильтром полоса; компонента же сигнала будет проходить через фильтр без ослабления при условии, что частота этой компоненты лежит в полосе пропускания. В дальнейшем для простоты мы будем считать сигнал синусоидальным.

Вышеприведенное рассуждение можно записать в аналитической форме следующим образом. Пусть ρ означает среднюю мощность помехи на единицу частоты; пусть, далее, $\Delta\omega_0$ — полоса пропускания фильтра. Тогда мощность помехи на выходе будет

$$P_{\text{п}} = \rho\Delta\omega_0, \tag{1}$$

и отношение мощностей сигнала и помехи на выходе

$$\frac{P_c}{P_{\text{п}}} = \frac{P_c}{\rho\Delta\omega_0} = \frac{P_c}{2\pi\rho\Delta f_0} \tag{2}$$

может быть сделано сколь угодно большим за счет выбора соответственно малой полосы пропускания Δf_0, т. е. путем применения высокоизбирательных фильтров.

Но это рассуждение требует существенного уточнения. Дело в том, что в действительности мы не имеем периодического сигнала в строгом смысле определения периодической функции. Сигнал, который мы называем периодическим, в действительности появляется в некоторый момент. Его спектр есть не дискретная линия, а сплошной спектр, отвечающий отрезку синусоиды от момента включения сигнала до текущего момента. С течением времени длина отрезка синусоиды возрастает и спектр этого отрезка соответственно меняется. Мы имеем дело с так называемым текущим спектром.

Спектр отрезка синусоиды неограничен. Однако можно, приняв тот или иной критерий, оценить его ширину Δf. Из теории спектров известно, что ширина спектра Δf связана с длительностью отрезка Δt соотношением

$$\Delta f \Delta t = \mu, \tag{3}$$

где μ — постоянная порядка единицы. Для выделения отрезка синусоиды необходимо, очевидно, выполнение условия

$$\Delta f_0 \geqslant \Delta f.$$

Приняв крайнее значение $\Delta f_0 = \Delta f$, можем переписать (2) в виде

$$\frac{P_c}{P_п} = \frac{P_c}{2\pi\rho\Delta t}.$$

Если же принять во внимание (3), то получится

$$\frac{P_c}{P_п} = \frac{P_c}{2\pi\rho\mu}\, t \approx \frac{P_c}{2\pi\rho}\, t. \tag{4}$$

Здесь и дальше вместо Δt записано t — время, протекавшее с момента включения. Смысл последнего соотношения состоит в том, что для увеличения отношения мощностей сигнала и помехи нужно затратить время, тем большее, чем больше желаемое отношение.

К этому заключению можно прийти и иным путем. Уточним обстановку и предположим для начала, что в качестве фильтра применен простой контур без затухания. При включении на такой контур синусоидального напряжения с частотой, совпадающей с резонансной частотой контура, амплитуда напряжения на индуктивности или емкости растет по линейному закону

$$U_m = \frac{1}{2} E_m \omega_0 t,$$

где E_m — амплитуда э. д. с.; ω_0 — резонансная частота. Таким образом, мощность колебаний в контуре, вызванных приходящим сигналом, пропорциональна t^2.

Теперь положим, что одновременно с синусоидальным сигналом на контур начал воздействовать белый шум. Его можно себе

представить как беспорядочную последовательность весьма коротких импульсов, т. е. записать помеху в виде

$$\sum_{k=0}^{n} \sigma_1 (t - t_k),$$

где t_k — случайные моменты, в которые появляются отдельные импульсы; n — общее число импульсов, равное $n=n_0 t$, если через n_0 обозначить среднее число импульсов в единицу времени, а через t — время, протекавшее с момента включения.

Каждый единичный импульс вызывает в контуре реакцию вида $\sin \omega_0 (t-t_k)$, так что отклик контура на воздействие в форме белого шума можно представить суммой

$$\sum_{k=1}^{n} \sin \omega_0 (t - t_k).$$

Мощность шумовых колебаний в контуре будет пропорциональна квадрату этой суммы, усредненному по периоду, т. е.

$$P_{\text{п}} = k \frac{1}{T} \int_0^T \left[\sum_{k=1}^{n} \sin \omega_0 (t - t_k) \right]^2 dt.$$

Раскрывая квадрат суммы, получим

$$P_{\text{п}} = k \frac{1}{T} \int_0^T \left[\sum_{k=1}^{n} \sin^2 \omega_0 (t - t_k) + \right.$$

$$\left. + \sum_{l,\, i=1}^{n} \sin \omega_0 (t - t_l) \sin \omega_0 (t - t_i) \right] dt,$$

где вторая сумма берется по всем $l \neq i$ от 1 до n. Преобразуя тригонометрические функции, получим

$$P_{\text{п}} = k \frac{1}{2T} \int_0^T \left\{ \sum_{k=1}^{n} [1 - \cos 2\omega_0 (t - t_k)] + \right.$$

$$\left. + \sum_{l,\, i=1}^{n} [\cos \omega_0 (t_l - t_i) - \cos \omega_0 (2t - t_l - t_i)] \right\} dt,$$

$$(P_{\text{п}})_{\text{ср}} = \frac{k}{2} n = \frac{k}{2} n_0 t,$$

так как вторая сумма флюктуирует около нуля вследствие случайности интервалов $(t_l - t_i)$, а следовательно, и случайности знаков косинусов. Эту сумму мы отбросили, не интересуясь пока флюктуациями мощности. Для нас важно, что в среднем мощность, обусловленная помехой, растет в контуре со временем по линейному закону. А так как мощность сигнала растет по квадратичному за-

кону, то отношение мощностей сигнала и помехи растет со временем по линейному закону в соответствии с ранее полученным результатом.

Мы рассматривали в качестве фильтрующей системы идеальный контур без потерь. Нетрудно видеть, что все сказанное сохраняет силу и для начальной стадии процесса в реальном контуре. Длительность этой начальной стадии тем больше, чем больше добротность контура.

Заключения, к которым мы пришли, интересны вот с какой точки зрения: фильтрация сигнала дает нам новый пример преобразования сигнала.

В рассмотренном случае преобразование состоит в том, что мы изменяем соотношение мощностей сигнала и помехи, т. е. превышение сигнала. Оказывается, что увеличение превышения покупается ценой увеличения длительности сигнала. Таким образом, перед нами пример преобразования с участием сигнала H и T (см. § 16).

С физической точки зрения можно толковать возможность обнаружения слабого сигнала при помощи фильтрации следующим образом: в контуре с течением времени происходит накопление эффекта сигнала и эффекта помехи. Однако процесс накопления для сигнала и для помехи протекает по-разному в силу того, что сигнал, будучи периодическим, когерентен, а помеха некогерентна. Поэтому эффективное значение сигнала растет со временем линейно, а эффективное значение помехи растет лишь как $\sqrt{t}$, по закону энергетического суммирования.

Очевидно, что принцип отличения сигнала от помехи по признаку когерентности имеет весьма общий характер. В следующем параграфе рассматривается еще одна форма использования этого принципа.

§ 39. Синхронное накопление

В недавнее время разработана специальная форма метода накопления, приспособленная для обнаружения очень слабого периодического сигнала.

Суть дела состоит в том, что смесь сигнала со значительно превосходящей его помехой нарезается на куски, длительность которых равна T/m, где T — период сигнала, а m — целое число. Это осуществляется синхронным коммутатором с ламелями и с периодом коммутации T, подключающим один из m накопителей на вход приемного устройства, как показано на принципиальной схеме рис. 36. Таким образом, отрезки длительностью T/m через интервал T и со сдвигом относительно начала на kT/m попадают на k-й накопитель, где и суммируются. На рис. 37 первая строка изображает сигнал в виде малого короткого импульса, периодически повторяющегося. Вторая строка изображает помеху: так же выглядела бы и смесь сигнала с помехой (непосредственно распо-

знать сигнал под помехой было бы невозможно). Третья строка представляет ритм работы устройства. Число накопителей взято равным пяти; на рисунке показано, что сигнал попадает всякий раз в один и тот же накопитель, а именно в четвертый. Это возможно только при условии, что период обращения коммутатора в точности равен периоду сигнала, т. е. при условии строгой синхронизации передающего и приемного устройств. Поэтому описываемый метод и назван методом синхронного накопления.

Описанное устройство позволяет не только обнаружить в результате накопления слабый импульс, но и определить его поло-

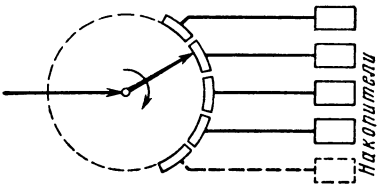


Рис. 36

жение во времени с тем большей точностью, чем больше число накопителей. Ясно, что все эти свойства лучше всего приспособлены для приема радиолокационных сигналов. В частности, проблема синхронизации в этом случае решается совершенно просто, так как и передающая и приемная части находятся в одном месте или, во всяком случае, под единым управлением.

Подобного рода схема была с успехом применена Баем в его известных опытах по радиолокации Луны [18]. Отсылая за подробностями к оригинальной работе или к опубликованному на русском языке обзору [5], отметим лишь оригинальное устройство накопителей.

Бай воспользовался электролизом. Его накопители представляли собой набор стеклянных трубочек, заполненных электролитом. Электролит разлагался выпрямленным током, и в трубочках выделялся газ. Высота газового столбика (т. е. уровень жидкости в трубочке) непосредственно дает интеграл тока, т. е. количество электричества, протекшее через данный накопитель за время опыта. С течением времени высота газового столбика растет одинаково во всех накопителях, кроме того, в который попадает сигнал. В этом единственном накопителе рост столбика происходит быстрее, и опыт продолжается до тех пор, пока это различие не сможет быть констатировано с достаточной уверенностью.

Само собой разумеется, что накопление может осуществляться многими другими способами, с применением интегрирующих устройств самой различной физической природы. Можно представить себе интеграторы электрические, механические, тепловые и др. Самое естественное — это применить для интегрирования цепь с емкостью. При этом, однако, нужно позаботиться о том, чтобы накапливающаяся емкость отличалась от схемы в то время, когда

на нее не подается подлежащий интегрированию ток; в противном случае накопленный заряд будет стекать. Изящное решение этой задачи дано Владимирским [6].

Теперь нам нужно перейти к рассмотрению количественных соотношений. Простейшее рассуждение, показывающее выигрыш, даваемый методом накопления, сводится к рассмотрению накопленных энергий. В силу некогерентности помехи накопленная энергия помехи будет возрастать как n (n — число повторений), тогда как накопленная энергия сигнала будет возрастать как n^2. В результате этого отношение накопленных энергий сигнала и помехи

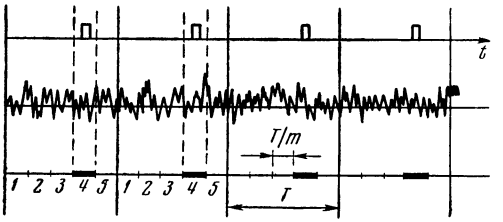


Рис. 37

будет расти пропорционально $n^2/n=n$. От энергий можно перейти к средним мощностям, определив их как отношение соответствующих энергий ко времени накопления.

Итак, в принципе можно получить при сколь угодно малом начальном превышении сигнала над помехой сколь угодно большое превышение на выходе накопителя — стоит только дать достаточно большое число повторений. Таким образом, повышение помехоустойчивости путем накопления сопряжено с затратой времени, т. е. с увеличением объема сигнала путем увеличения его длительности.

При всей простоте и бесспорности вышеприведенного рассуждения оно не содержит прямого ответа на естественно возникающий вопрос: каково должно быть число повторений для обеспечения заданной помехоустойчивости при данном начальном превышении сигнала над помехой (может быть отрицательном)? Ответ на этот вопрос можно получить только из вероятностных соображений. Уточним прежде всего постановку задачи. Предположим для простоты, что коммутатор подключает каждый накопитель на вход приемного устройства лишь на весьма короткий промежуток времени, настолько короткий, что ни сигнал, ни помеха не успевают заметно измениться за этот промежуток, и при каждом включении накопитель получает импульс, пропорциональный мгновенному значению помехи или сигнала в момент включения. Обозначим мгновенно значения помехи в эти моменты через ξ_i. Иначе говоря,

$$\xi_i = \xi(t + k\tau + iT),$$

где t — текущее время; k — номер накопителя; $\tau = T/m$ — длитель-

ность шага коммутатора; T — период коммутации (равный периоду сигнала)[1].

За n повторений (т. е. за n периодов) в накопителе накапливается сумма n слагаемых ξ_i. Введем среднее арифметическое

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i. \tag{1}$$

Так как ξ_i — случайная величина, то при малом числе слагаемых Y_n сильно флюктуирует около среднего значения ξ. С увеличением числа слагаемых вероятность больших флюктуаций убывает; в этом, собственно, и состоит закон больших чисел. С другой стороны, если обозначить через a значение сигнала, то после n повторений мы получим в накопителях значение na, а среднее за n повторений будет, конечно, a. Нам надлежит сравнить возможную величину флюктуаций накопленной помехи с накопленным значением сигнала. Ошибка при приеме возможна до тех пор, пока флюктуации могут превзойти сигнал по абсолютной величине.

Мы можем теперь сформулировать стоящую перед нами вероятностную задачу следующим образом: какова в зависимости от числа слагаемых вероятность того, что отклонение Y_n от среднего значения не превзойдет заданной величины a? Ответ на этот вопрос дает известное неравенство Чебышева

$$p\{|Y_n - \xi| < a\} > 1 - D(\xi)/na^2. \tag{2}$$

В этой формуле $D(\xi)$ означает *дисперсию*, т. е. средний квадрат уклонения случайной величины ξ от ее среднего значения. Величина $D(\xi)$ выражает, таким образом, среднюю мощность помехи $P_п = \sigma^2$. Что же касается a^2, то эта величина выражает мощность сигнала P_c. Число слагаемых n есть число повторений. Разрешая неравенство (2) относительно n, получаем

$$n < \frac{1}{1-p} \cdot \frac{P_п}{P_c} = \frac{1}{1-p} \left(\frac{\sigma}{a}\right)^2. \tag{3}$$

Таким образом, верхний предел числа повторений определяется заданной нами вероятностью p и исходным превышением, т. е. отношением мощностей сигнала и помехи.

К сожалению, мы получили для n оценку в форме неравенства, ограничивающего n сверху. Оценки снизу теория вероятностей

---

[1] В действительности время включения конечно, и в накопитель каждый раз попадает случайная величина

$$\eta_i = \int_{t+k\tau+iT}^{t+k\tau+iT+\Delta t} \xi(t)\, dt,$$

где Δt — продолжительность включения. Это обстоятельство, конечно, усложняет дело, так как статистика η отличается от статистики величины ξ. Однако общий принцип накопления сохраняет силу.

не дает. Однако она дает больше: она позволяет выразить p, а следовательно, и n приближенным равенством. Это равенство основывается на теореме Ляпунова, утверждающей, что при выполнении некоторого общего условия (условия Линдеберга—Феллера) распределение для суммы случайных величин при увеличении числа слагаемых сходится к нормальному. Практически важно, что оно сходится довольно быстро.

Вытекающее из теоремы Ляпунова приближенное равенство в наших обозначениях записывается в виде

$$p\{|Y_n - \xi| < a\} \approx \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{\sigma}\sqrt{\frac{n}{2}}} e^{-z^2} dz = \Phi\left(\frac{a}{\sigma}\sqrt{\frac{n}{2}}\right). \qquad (4)$$

Здесь Φ — символ функции Лапласа. Записанная в левой части величина есть вероятность того, что флюктуация накопленной помехи не превзойдет накопленного сигнала, т. е. вероятность ошибки будет

$$p_0 = p\{|Y_n - \xi| > a\} \approx 1 - \Phi\left(\frac{a}{\tau}\sqrt{\frac{n}{2}}\right). \qquad (5)$$

Для достаточно надежной связи вероятность ошибки должна быть мала, а число повторений велико. В этих условиях можно воспользоваться асимптотическим разложением функции Лапласа (см. § 22), что дает для вероятности ошибки

$$p_0 \approx \frac{1}{\sqrt{\pi}} e^{-\frac{n}{2}\cdot\frac{a^2}{\sigma^2}} \frac{1}{\sqrt{\frac{n}{2}}\cdot\frac{a}{b}}. \qquad (6)$$

Переходя к помехоустойчивости, определенной в § 35 как

$$S = \lg \frac{1}{p_0},$$

получим из (6)

$$S = \frac{1}{2}\lg \pi + \frac{1}{2}\lg \frac{n}{2}\cdot\frac{P_c}{P_п} + \frac{0{,}434}{2} n \frac{P_c}{P_п} \approx$$

$$\approx 0{,}25 + 0{,}5 \lg n \frac{P_c}{P_п} + 0{,}22 n \frac{P_c}{P_п}.$$

Таким образом, помехоустойчивость тем больше, чем больше исходное превышение сигнала над помехой и чем больше число повторений. Обе эти величины входят в формулу (7) на равных правах, и очевидно, что уменьшение исходного превышения может быть компенсировано соответствующим увеличением числа повторений. Для нахождения требуемого числа повторений при данном превышении и заданной помехоустойчивости можно воспользоваться графиком рис. 38, на котором изображена представляемая форму-

лой (7) зависимость. На том же рисунке представлено и предельное соотношение, даваемое неравенством Чебышева (штриховая линия), которое можно записать в виде

$$S = \lg \frac{1}{p_0} > \lg n \frac{P_c}{P_п}.$$

Этому неравенству удовлетворяет область справа от штриховой линии. График наглядно показывает, насколько грубую оценку дает в рассматриваемом случае неравенство Чебышева.

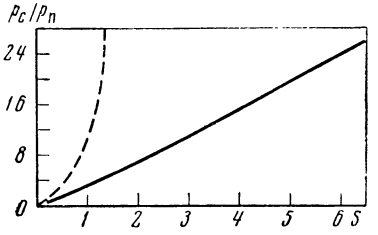


Рис. 38

Пользование графиком рис. 38 состоит в том, что, задавшись помехоустойчивостью S, находим соответствующее значение $nP_c/P_п$; затем, зная $P_c/P_п$, находим n. Так, например, задавшись помехоустойчивостью $S=3$ (вероятность ошибки $p_0=10^{-3}$), найдем $nP_c/P_п \approx 10$. Пусть сигнал лежит на 10 *дб* ниже помехи ($P_c/P_п = 0{,}1$). Отсюда получаем требуемое число повторений

$$n = 10/0{,}1 = 100.$$

§ 40. Корреляционный метод приема

Возможен специальный метод приема покрытого помехой периодического сигнала, основанный на измерении функции автокорреляции, получаемой на приемном конце смеси сигнала с помехой [23]. Идея метода заключается в использовании того факта, что функция корреляции беспорядочной помехи всегда убывает с возрастанием аргумента τ, тогда как функция корреляции периодического процесса сама периодична и имеет тот же период. Докажем прежде всего последнее положение.

Если имеем

$$x(t) = x(t + nT_1)$$

(определение периодической функции), то функции автокорреляции для $x(t)$

$$B_{xx}(\tau) = \frac{1}{T_1} \int_{-T_1/2}^{T_1/2} x(t)\, x(x+\tau)\, dt = \overline{x(t)\, x(t+\tau)} =$$

$$= \overline{x(t)\, x(t + nT_1 + \tau)} = B_{xx}(\tau + nT_1), \qquad (1)$$

т. е. функция автокорреляции B_{xx} оказывается периодической функцией аргумента τ с периодом T_1. Нетрудно найти коэффициенты разложения B_{xx} в ряд Фурье. Пусть

$$x(t)=\sum_{k=-\infty}^{+\infty} C_k e^{jk\omega_1 t},$$

где $\omega_1=2\pi/T_1$ — основная круговая частота периодической функции $x(t)$; C_k — комплексные амплитуды гармоник. Для функции автокорреляции имеем

$$B_{xx}=\frac{1}{T_1}\int_{-T_1/2}^{T_1/2}(\sum C_k e^{jk\omega_1 t})(\sum C_l e^{jl\omega_1(t+\tau)})\,dt.$$

Здесь, как и выше, в силу периодичности $x(t)$ среднее берется за период; индексы k и l принимают все значения между $-\infty$ и $+\infty$. Далее,

$$B_{xx}=\frac{1}{T_1}\sum_l\sum_k C_l C_k e^{jl\omega_1\tau}\int_{-T_1/2}^{T_1/2} e^{j(k+l)\omega_1 t}dt=$$

$$=\sum\sum C_l C_k\frac{\sin(k+l)\,\omega_1\frac{T_1}{2}}{j(k+l)\,\omega_1\frac{T_1}{2}}e^{jl\omega_1\tau}.$$

Но $\omega_1 T_1/2=\pi$. Таким образом, содержащий $(k+l)$ множитель равен нулю при любых комбинациях k и l, кроме $k+l=0$, т. е. $k=-l$. При этом указанный множитель обращается в единицу, и мы имеем, следовательно,

$$B_{xx}=\sum_k C_k C_{-k}e^{jk\omega_1\tau}=\sum_{k=-\infty}^{\infty}|C_k|^2 e^{jk\omega_1\tau}=$$

$$=2\sum_{k=1}^{\infty}|C_k|^2\cos k\omega_1\tau, \qquad (2)$$

т. е. амплитуда гармоники функции корреляции для $x(t)$ равна удвоенному квадрату амплитуды соответствующей гармоники $x(t)$; как четная функция B_{xx} разлагается в ряд по косинусам.

Рассмотрим теперь функцию автокорреляции для смеси сигнала $x(t)$ и помехи $\xi(t)$, т. е.

$$B(\tau)=\overline{[x(t)+\xi(t)]\,[x(t+\tau)+\xi(t+\tau)]}. \qquad (3)$$

Раскрывая произведение в этой формуле, перепишем ее в виде

$$B=\overline{x(t)\,x(t+\tau)}+\overline{\xi(t)\,\xi(t+\tau)}+\overline{x(t)\,\xi(t+\tau)}+$$

$$+\overline{\xi(t)\,x(t+\tau)}=B_{xx}+B_{\xi\xi}+B_{x\xi}+B_{\xi x}.$$

Первые два члена представляют собой функции автокорреляции соответственно сигнала и помехи. Два последних члена — функции взаимной корреляции сигнала и помехи. Но так как сигнал и помеха взаимно независимы, т. е. корреляция между ними отсутствует, то эти два члена равны нулю. Таким образом,

$$B(\tau) = B_{xx}(\tau) + B_{\xi\xi}(\tau). \qquad (4)$$

Функция $B_{\xi\xi}$ убывает с возрастанием τ, стремясь к пулю. Функция же B_{xx} по доказанному выше периодична. Стало быть, при достаточно большом сдвиге τ влиянием $B_{\xi\xi}$ можно пренебречь, и функция корреляции B отображает сигнал. Это поясняется рис. 39, на котором изображены оба слагаемых правой части (4). Величина сдвига, при котором можно уже пренебречь слагаемым $B_{\xi\xi}$, зависит от интервала корреляции для помехи, который связан с шириной спектра помехи; произведение интервала корреляции на ширину спектра имеет порядок единицы. Величина сдвига зависит также от превышения сигнала над помехой. Эта зависимость становится очевидной, если нормировать функции корреляции путем деления их на соответствующие дисперсии. Мы полагаем обычно средние значения как для сигнала, так и для помехи равными нулю. При этом условии дисперсия совпадает со средним квадратом и непосредственно выражает мощность.

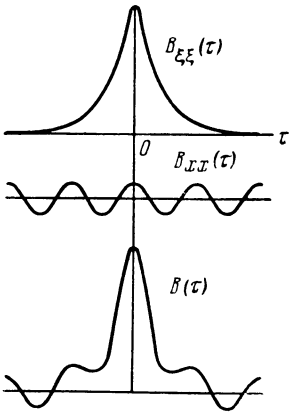


Рис. 39

Мы имеем

$$B(\tau) = a^2 b_{xx}(\tau) + \sigma^2 b_{\xi\xi}(\tau) = P_c b_{xx} + P_{\text{п}} b_{\xi\xi},$$

откуда видно, что второе слагаемое может быть отброшено при тем меньшем значении τ, чем больше $P_c/P_{\text{п}}$, т. е. чем больше превышение сигнала над помехой.

Таким образом, исследование функции корреляции позволяет выделить периодический сигнал из смеси его с помехой.

В состав приемного устройства должен входить коррелометр — прибор, автоматически измеряющий функцию корреляции. В настоящее время существует уже большое число систем такого рода приборов.

Ясно, что для извлечения периодического сигнала из смеси его с помехой требуется время. Это следует уже из того, что для получения функции корреляции достаточно большого аргумента τ нужно действительный процесс задержать на время τ. Чем меньше исходное превышение сигнала над помехой, тем большая

требуется задержка для обнаружения сигнала. Но нужно еще принять во внимание, что функция корреляции определяется усреднением произведения $x(t)\,x(t+\tau)$ по всей бесконечной оси времени. В действительности выполнить такое усреднение, очевидно, невозможно.

Практически мы имеем дело с текущей функцией корреляции, определяемой как

$$B_T(\tau)=\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt.$$

Текущая функция корреляции зависит не только от сдвига τ, но и от интервала усреднения T. Это, конечно, сильно усложняет все соотношения. Во всяком случае ясно, что интервал усреднения T должен быть больше сдвига τ, и этим определяется время, потребное для уверенного обнаружения сигнала.

Корреляционный метод является лишь одним представителем обширной группы методов, основанных на исследовании вероятностных характеристик смеси сигнала с помехой, поступающей на приемный конец линии связи. Для любого метода этой группы характерно то, что для исследования каких бы то ни было вероятностных характеристик указанного сложного процесса нужно располагать достаточно большим образцом или отрезком процесса, т. е., иначе говоря, нужно вести наблюдение за процессом в течение достаточно длительного времени.

Таким образом, все вероятностные методы этой группы позволяют повысить надежность связи за счет увеличения длительности передачи сигнала.

§ 41. Корректирующие коды

Интересной иллюстрацией общих принципов и в то же время важным практическим средством борьбы с помехами являются корректирующие коды, т. е. коды, строение которых обеспечивает возможность обнаружения и исправления ошибок. Теория и практика подобных кодов развиты главным образом в связи с проблемами вычислительных машин, но ясно, что эти коды могут иметь применение и в связи.

Общая идея построения корректирующих кодов состоит в том, что к кодовой комбинации обычного кода добавляются дополнительные знаки, служащие для обнаружения и исправления ошибки. Таким образом, и здесь повышение надежности связи оплачивается либо увеличением длительности сигнала, либо расширением его спектра (если увеличенная за счет дополнительных знаков комбинация передается за то же время, что и обычная).

Первоначальное представление о возможности обнаружения ошибки из-за дополнительных знаков в кодовой комбинации можно составить при помощи следующего простого примера. Положим,

что мы ведем передачу обычным двоичным кодом с элементами 0 и 1. Припишем теперь к каждой кодовой комбинации еще одну двоичную цифру, выбрав ее так, чтобы, например, общее число единиц в комбинации было всегда четным. Легко видеть, что при этом условии ошибка в любом знаке изменит четное число единиц на нечетное и будет таким образом обнаружена. Способ составления дополненных комбинаций обычного пятизначного кода Бодо представлен ниже:

Буква	А	Б	В	Г	Д	Е
Обычный код	10000	00110	01101	01010	11110	01000
Дополненный код	100001	001100	011011	010100	111100	010001

Понятно, что такой простой дополненный код хотя и позволяет обнаружить отдельную ошибку, но не дает возможности ее локализовать и исправить; кроме того, такой код не в состоянии обнаружить двойную ошибку. Однако возможно построить более мощные корректирующие коды, для чего потребуется, конечно, большее количество дополнительных знаков. Мы постараемся теперь рассмотреть вопрос в несколько более общем виде.

Пусть имеется n_0-значный двоичный код. Общее число комбинаций этого кода составляет

$$N = 2^{n_0}.$$

Это число выражает число размещений, отличающихся друг от друга хотя бы в одном знаке. Дополним теперь код еще одним знаком, так что общее число знаков будет

$$n = n_0 + 1,$$

число же кодовых комбинаций оставим неизменным. Тогда

$$N = 2^{n_0} = 2^{n-1} = \frac{1}{2} 2^n$$

и можно подобрать кодовые комбинации так, что любые две комбинации будут различаться уже двумя знаками. При этом будет использована только половина возможного числа комбинаций 2^n; вторая половина образует запрещенные в данном коде комбинации. Ясно, что любая одиночная ошибка в разрешенной комбинации превратит ее в запрещенную, и таким образом ошибка обнаруживается.

Именно этим свойством обладает дополненный код в вышеприведенном примере.

Положим теперь, что мы добавили в исходную кодовую комбинацию столько дополнительных знаков, что стало возможным образовать прежнее число комбинаций $N = 2^{n_0}$, но так, что любые две комбинации различаются уже тремя знаками. Такой код позволит исправить одиночную ошибку. Действительно, в результате ошибки неверная комбинация будет отличаться от верной только

в одном знаке, а от любой другой — не менее чем в двух знаках. Поэтому искаженная одной ошибкой комбинация будет ближе[1] к истинной, чем к какой-либо другой, и ошибка может быть не только обнаружена, но и исправлена.

Что касается числа дополнительных знаков, которое необходимо для составления N комбинаций, различающихся не менее чем тремя знаками, то это число может быть определено на основании неравенства

$$N = 2^{n_0} \leqslant \frac{2^n}{n+1}.$$

Учитывая, что n_0 и n — целые числа, получим:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_0	0	0	1	1	2	3	4	4	5	6	7	8	9	10	11
$n-n_0$	1	2	2	3	3	3	3	4	4	4	4	4	4	4	4

Число $n-n_0$ выражает количество дополнительных знаков. Так, например, для того чтобы можно было обнаружить и исправить одиночную ошибку в пятизначном коде ($n_0=5$), нужно добавить еще 4 дополнительных знака ($n-n_0=4$), т. е. составить девятизначный корректирующий код ($n=9$).

Если применением корректирующего кода возможное влияние одиночной ошибки исключено, то опасность грозит теперь со стороны двойных ошибок и ошибок более высокой кратности. Если ошибки независимы и если считать для простоты ошибки кратности выше двух исключенными, то надежность должна была бы возрасти при применении описанного кода ровно вдвое при условии увеличения вероятности ошибки вследствие возрастания длины комбинации. Приняв это во внимание, можем записать

$$p_0 \approx \frac{n}{n_0} p^2,$$

где p — вероятность одиночной ошибки в исходных условиях. Помехоустойчивость при использовании корректирующего кода равна

$$S = \lg \frac{1}{p_0} \approx 2 \lg \frac{1}{p} - \lg \frac{n}{n_0} = 2S_0 - \lg \frac{n}{n_0}.$$

Для рассмотренного примера

$$\frac{n}{n_0} = \frac{9}{5} = 1{,}8; \quad \lg \frac{n}{n_0} \approx 0{,}25.$$

Повышение помехоустойчивости сопровождается увеличением объема сигнала, которое можно выразить через избыточность. Если определить избыточность R относительным числом излиш-

[1] Термину «ближе» придается прямой смысл при геометрической трактовке вопроса, о чем речь будет ниже.

них знаков, то получим для корректирующих кодов вообще

$$R = (n - n_0)/n_0.$$

Для корректирующего кода, обнаруживающего одиночную ошибку (различие в двух знаках),

$$n = n_0 + 1, \quad R = 1/n_0.$$

Для корректирующего кода, исправляющего одиночную ошибку (различие в трех знаках),

$$n - n_0 \geqslant \log(n+1); \quad R \geqslant \frac{1}{n_0}\log(n+1).$$

Естественно, возникает вопрос о построении оптимального кода, т. е. кода с наименьшей избыточностью, или, иначе говоря, о нахождении наименьшего n при заданном n_0. Не вдаваясь в анализ вопроса, заметим лишь, что рассматриваемые здесь коды являются оптимальными [20].

Для уяснения всего сказанного приведем пример. Пусть дан двузначный код; построим корректирующие коды, обнаруживающие и исправляющие ошибку:

№ комбинации	Исходный код	Обнаруживающий код	Исправляющий код
1	00	000	00000
2	01	011	01110
3	10	101	10101
4	11	110	11011

При пользовании обнаруживающим кодом пусть принята комбинация 010. Такой комбинации нет — это ошибка. При пользовании исправляющим кодом пусть принято 10111. Эта комбинация ошибочна. Она отличается от первой четырьмя знаками, от второй — тремя, от третьей — одним, от четвертой — двумя. Значит, правильная комбинация третья, т. е. 10101, и ошибка исправлена.

Корректирующую способность кода можно повышать и далее увеличением числа различающихся знаков в кодовых комбинациях. При этом, конечно, будут расти число дополнительных знаков и общая длина кодовой комбинации (при неизменном $N = 2^{n_0}$). Повышенная корректирующая способность может быть использована по желанию как для обнаружения, так и для исправления кратных ошибок. Если обозначить через d наименьшее расстояние между кодовыми комбинациями, т. е. наименьшее число знаков, которыми различаются между собой любые две комбинации, а через r и s — кратности соответственно обнаруживаемой и исправляемой ошибок, то, принимая во внимание, что для обнаружения ошибки требуется лишняя единица в d, а для исправления — две лишние единицы, можно записать (считая,

что исправление подразумевает обнаружение)

$$d = 1 + r + s \qquad (r \geqslant s).$$

Это соотношение поясняется следующей схемой:

d	r	s	Возможность, даваемая кодом
1	0	0	Отличение одной комбинации от другой
2	1	0	Обнаружение одиночной ошибки
3	1	1	Исправление (с обнаружением) одиночной ошибки
	2	0	Обнаружение двойной ошибки
	2	1	Исправление одиночной ошибки и обнаружение двойной
4	3	0	Обнаружение тройной ошибки
	2	2	Исправление (с обнаружением) двойной ошибки
5	3	1	Исправление одиночной и обнаружение тройной ошибки
	4	0	Обнаружение четверной ошибки и т. д.

§ 42. Геометрическое представление сигнала

Пусть имеется сигнал, состоящий из трех знаков двоичного кода. Число возможных различных сигналов равно восьми; вот возможные комбинации

000, 001, 010, 011, 100, 101, 110, 111

Построим теперь геометрическую модель сигнала. Для этого возьмем систему прямоугольных координат x_1, x_2, x_3 и будем откладывать по осям значения каждого из трех знаков сигнала. Таким образом, каждый сигнал, определенный тремя своими координатами, будет представлен точкой в трехмерном пространстве. Легко видеть, что точки, изображающие различные сигналы, расположатся на вершинах единичного куба, как показано на рис. 40. Координаты каждой вершины куба записаны тремя цифрами, порядок которых соответствует нумерации осей.

Действие помехи может проявиться в том, что какой-либо из знаков сигнала будет искажен, т. е. что вместо 0 появится 1, или наоборот. Изменение одного знака соответствует изменению одной координаты сигнала; это есть одиночная ошибка.

Теперь заметим, что наименьшее расстояние между любыми двумя сигналами равно длине ребра куба (т. е. единице в нашем масштабе). Это значит, что если один из знаков сигнала будет при передаче искажен помехой, то переданный сигнал заменится другим и ошибка не сможет быть обнаружена. Для обнаружения одиночной ошибки необходимо, чтобы точки, изображающие различные сигналы, отстояли друг от друга не менее чем на две единицы, т. е. чтобы нужно было пройти два ребра куба, чтобы попасть из одной точки в другую.

Такому условию удовлетворяют только четыре точки с координатами, например,

$$000, \quad 011, \quad 101, \quad 110$$

Изменение одной какой-либо координаты каждой из четырех перечисленных точек переводит ее в положение, не соответствующее ни одному из возможных четырех сигналов, и ошибка таким образом обнаруживается.

Для того чтобы одиночная ошибка могла быть не только обнаружена, но и исправлена, нужно чтобы расстояние между точками сигналов составляло не менее трех единиц (три ребра куба). Тогда в результате одиночной ошибки, т. е. смещения на одну единицу, точка расположится ближе к истинному сигналу, чем к любому другому, и, следовательно, истинное значение сигнала может быть восстановлено. Такому условию в нашей модели отвечают только две точки, например, с координатами

$$000, \quad 111$$

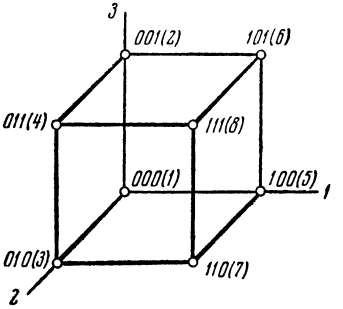


Рис. 40

Легко видеть, что все сказанное представляет собой геометрическую интерпретацию рассуждений предыдущего параграфа; мы рассмотрели трехзначный код ($n=3$) при $d=1$, 2 и 3 (см. стр. 128).

Мы видим на примере модели кода, что надежность связи определяется с геометрической точки зрения расстоянием между точками сигналов. Так обстоит дело и в более общем случае, к рассмотрению которого мы и переходим.

Пусть сигнал длительностью T представляется функцией со спектром, ограниченным наивысшей частотой ω_0. При таких условиях сигнал полностью определяется $m=\frac{1}{\pi}\omega_0 T$ числами (см. § 9).

Эти числа могут быть взяты в качестве координат в m-мерном пространстве, и они определят, таким образом, положение точки сигнала в этом пространстве. Здесь возникает прежде всего вопрос о том, какие именно числа следует взять в качестве координат сигнала. Выбор зависит от нашего усмотрения, так как мы можем представить функцию в виде разложения из m членов произвольно. Мы имели в § 9 следующие два разложения: разложение в ряд Фурье

$$f(t)=\sum_{-n}^{n} C_k e^{j\pi k \frac{t}{T}} \qquad (1)$$

и разложение в ряд по запаздывающим импульсам

$$\tilde{f}(t)=\sum_{k=1}^{m} f(k\Delta t)\,\frac{\sin\omega_c(t-k\Delta t)}{\omega_c(t-k\Delta t)}. \qquad (2)$$

Можно взять в качестве координат коэффициенты Фурье из (1). Но можно также взять для этой цели и значения

$$f_k=f(k\Delta t)$$

из (2). Обе возможности теоретически равноценны. Мы выберем вторую.

Итак, функция сигнала $\tilde{f}(t)$ определена у нас на интервале T посредством m мгновенных значений f_k. Можно сказать, что каждый сигнал определяется точкой в m-мерном пространстве с координатами f_k. Но можно также представить сигнал как m-мерный вектор; тогда f_k имеют смысл составляющих вектора; иначе говоря, f_k есть проекция вектора f на ось под номером k. Оба представления совпадают, если под точкой понимать конец вектора.

Различные сигналы представляются геометрически различными векторами в m-мерном пространстве — пространстве сигналов. Число измерений пространства сигналов всегда конечно. Противное означало бы либо, что спектр сигнала неограничен, как это следует из (1), либо, что сигнал имеет бесконечную длительность, как это следует из (2). Реальный же сигнал обладает конечной длительностью и ограниченным спектром [1]. Заметим, что квадрат длины (нормы) вектора сигнала, т. е.

$$\|f\|^2=\sum_{k=1}^{m} f_k^2,$$

выражает общую энергию сигнала (и притом независимо от того, какие величины — спектральные коэффициенты из (1) или мгновенные значения из (2) — выбраны в качестве составляющих вектора. Последнее утверждение вытекает из равенства Парсеваля).

Введем теперь в рассмотрение помеху. Если спектр помехи ограничен той же полосой, что и спектр сигнала, то и помеха представляется m-мерным вектором [2]. Этот вектор добавляется к вектору сигнала. Но так как помеха есть случайный процесс, не коррелированный с сигналом, то при каждом данном сигнале вектор помехи может иметь любую величину и направление; все направления равновероятны, а длина вектора определяется распределением вероятностей. В результате при наложении помехи на сигнал конец результирующего вектора не занимает какого-либо определенного положения; область возможных его положений представляется в виде облака, плотность

[1] См. по этому поводу сноску на стр. 27.

[2] Это условие, конечно, всегда будет выполнено, так как расширение полосы пропускания вызовет только уменьшение превышения сигнала над помехой.

которого выражает плотность вероятностей попадания конца результирующего вектора в данную точку[1].

Плотность вероятностей зависит от удаления от точки сигнала с координатами f_k; поверхности равных плотностей представляются m-мерными сферическими поверхностями. Все это представлено схематически на рис. 41 для двумерной модели, где изображены два сигнала, каждый из которых определен всего двумя координатами.

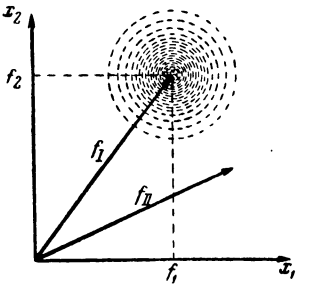


Рис. 41

Конечно, может случиться, что в результате наложения помехи сигнал *I* превратится в сигнал *II*. Вероятность такого события должна быть мала, и ясно, что она тем меньше, чем больше расстояние между точками *I* и *II*, изображающими два разных сигнала.

Подобного рода геометрические представления, широко распространенные в настоящее время, были развиты в работе В. А. Котельникова [11]. В основе содержания четырех последующих параграфов также лежат идеи этой работы.

§ 43. Геометрическая модель системы связи

Передача сообщений при помощи электрической связи состоит в том, что передаваемое сообщение отображается некоторым сигналом; на приемном конце по сигналу воспроизводится переданное сообщение. Сообщение и сигнал связаны однозначным соответствием; каждому сообщению соответствует единственный сигнал и обратно. Но при передаче на сигнал налагается помеха, так что принятый результирующий сигнал может уже отображать не переданное сообщение, а некоторое другое.

Эти общие соотношения представлены в геометрической форме на рис. 42. Передаваемые сообщения обозначены через $u(t)$; пространство U представлено двумерной моделью, в которой u_1 и u_2 изображают два различных передаваемых сообщения.

---

[1] Длина вектора помехи выражает энергию помехи. Если обозначить помеху через $\xi(t)$, то

$$\|\xi\|^2 = \sum_{k=1}^{m} \xi_k^2.$$

При конечном числе измерений m эта величина является случайной и, будучи усреднена по m, флюктуирует около средней мощности помехи. Флюктуации убывают с увеличением числа слагаемых. В пределе облако вырождается в бесконечномерный шаровой слой, заключающий вероятность единицы в бесконечно малой толщине, т. е. имеющий бесконечную плотность вероятностей.

Сообщения $u(t)$ преобразуются в сигналы $f(t)$ посредством функционального оператора ψ

$$f=\psi(u).$$

Это преобразование переводит пространство U в пространство F; каждому сообщению u соответствует свой сигнал f. На рис. 42 изображены сигналы f_{I} и f_{II}, соответствующие сообщениям u_{I} и u_{II}. Если помеха отсутствует, то принятый сигнал преобразуется в сообщение обратным оператором

$$v=\psi^{(-1)}(f),$$

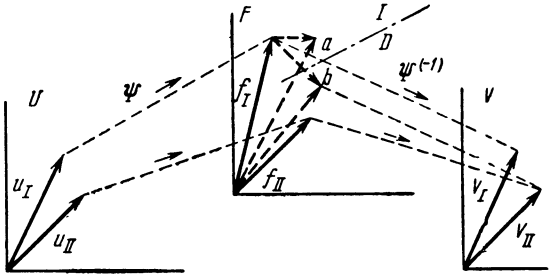


Рис. 42

т. е. переводом пространства F в пространство V. При однозначном соответствии сообщений и сигналов и при отсутствии помех принятое сообщение тождественно переданному, т. е.

$$v=\psi^{(-1)}(f)=\psi^{(-1)}\psi(u)=u.$$

Однако помеха существует. В средней части рис. 42 в пространстве F показано, что приложение вектора помехи к вектору сигнала f_1 дает некоторый новый сигнал, не отвечающий уже ни одному из двух сообщений u_{I} и u_{II}. Конец результирующего вектора оказывается в некоторой точке, которая может быть ближе либо к концу вектора f_{I} (точка a), либо к концу вектора f_{II} (точка b). Можно построить приемник, который будет воспроизводить сообщение $v_{\text{I}}=u_{\text{I}}$ всякий раз, когда конец результирующего вектора ближе к концу вектора f_{I}, чем к концу вектора f_{II}.

Такой приемник называется по Котельникову идеальным [1] [11, 12]. Его действие характеризуется тем, что пространство F оказывается разбитым на области, границы которых представляют собой места точек, равноотстоящих от точек (концов вектора) различных сигналов. Ошибка при приеме на идеальный приемник, т. е. замена переданного сообщения другим, возможным, происходит лишь тогда, когда результирующая точка, представляющая сигнал с наложенной на него помехой, переходит границу данной области и оказывается в соседней (как, например, точка b на рис. 42). Вероятность такого события определяет помехоустойчивость связи.

[1] По современной терминологии такой приемник называется оптимальным.

В описанных условиях, т. е. при применении идеального приемника, вероятность ошибки оказывается наименьшей; а следовательно, помехоустойчивость — наибольшей возможной. Предельно достижимую помехоустойчивость В. А. Котельников называет *потенциальной помехоустойчивостью*.

Совершенно ясно, что помехоустойчивость тем больше, чем больше расстояние d между соседними сигналами (т. е. между концами представляющих эти сигналы векторов). Но это расстояние зависит не только от расстояния r между соседними сообщениями, но и от вида оператора ϕ. Этот последний определяется способом преобразования сообщения в сигнал, в частности способом модуляции. Следовательно, помехоустойчивость системы связи может зависеть (и на самом деле зависит) от способа модуляции, причем большую помехоустойчивость обеспечивает тот способ модуляции, который при данном r дает большее d. Анализ этих соотношений составляет одну из основных задач общей теории помехоустойчивости.

§ 44. Общая теория помехоустойчивости

Задача, которая должна быть решена, ставится в общем виде так: найти вероятность ошибки при приеме, т. е. вероятность замены действительно переданного сообщения другим возможным. Вероятность ошибки определяет помехоустойчивость системы связи.

Рассмотрим в пространстве сигналов F два различных сигнала $\boldsymbol{f}_{\mathrm{I}}$ и $\boldsymbol{f}_{\mathrm{II}}$. Пусть передается сигнал $\boldsymbol{f}_{\mathrm{I}}$; при передаче на него накладывается помеха ξ (рис. 43), что дает результирующий вектор $\boldsymbol{X}$. Ошибки не произойдет (при применении идеального приемника), если конец вектора $\boldsymbol{X}$ будет лежать ближе к концу вектора $\boldsymbol{f}_{\mathrm{I}}$, чем к концу вектора $\boldsymbol{f}_{\mathrm{II}}$, представляющего ближайший возможный сигнал. Это условие можно записать в виде

$$\|\boldsymbol{X}-\boldsymbol{f}_{\mathrm{I}}\|<\|\boldsymbol{X}-\boldsymbol{f}_{\mathrm{II}}\|. \tag{1}$$

Используя очевидные векторные равенства

$$\boldsymbol{X}-\boldsymbol{f}_{\mathrm{I}}=\xi, \quad \boldsymbol{X}-\boldsymbol{f}_{\mathrm{II}}=\xi-\Delta\boldsymbol{f},$$

где

$$\Delta\boldsymbol{f}=\boldsymbol{f}_{\mathrm{II}}-\boldsymbol{f}_{\mathrm{I}},$$

можем представить (1) так:

$$\|\xi\|<\|\Delta\boldsymbol{f}-\xi\|. \tag{2}$$

Спроектируем теперь векторы, входящие в неравенство (2), на направление $\Delta\boldsymbol{f}$, получим алгебраическое неравенство

$$\xi_{\Delta}<d-\xi_{\Delta}$$

или

$$\xi_\Delta < d/2, \tag{3}$$

где ξ_Δ — проекция ξ на Δf.

Теперь заметим, что мы всегда можем выбрать координатные оси так, чтобы вектор Δf был параллелен одной из них, скажем, оси за номером k. Тогда вместо (3) получим

$$\xi_k < d/2. \tag{4}$$

Вероятность выполнения этого неравенства есть вероятность правильного приема; вероятность ошибки есть вероятность обратного неравенства. Заметим, кроме того, что ξ_k есть просто одно из мгновенных значений помехи. Таким образом, вероятность ошибки может быть выражена формулой

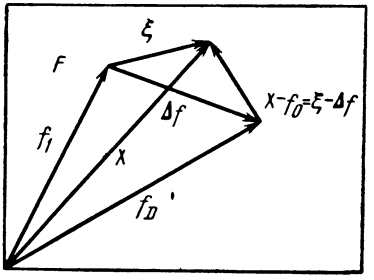


Рис. 43

$$p_{\text{ош}} = p\{|\xi| > d/2\} = \\ = 1 - \Phi(d/2\sqrt{2\sigma}). \tag{5}$$

Эта формула выражает вероятность ошибки через отношение d/σ; формула показывает, что вероятность ошибки при помехе заданной интенсивности тем меньше, чем больше расстояние d между ближайшими сигналами. Очередная задача состоит в определении d.

Вернемся к соотношениям предыдущего параграфа. Сигнал $\boldsymbol{f}$, соответствующий передаваемому сообщению $\boldsymbol{u}$, получается путем применения к $\boldsymbol{u}$ функционального оператора ψ

$$\boldsymbol{f} = \psi(\boldsymbol{u}).$$

Мы будем рассматривать лишь такие способы передачи, при которых ψ есть просто символ некоторой функции [1].

Если два сообщения, $\boldsymbol{u}_{\text{I}}$ и $\boldsymbol{u}_{\text{II}}$, различаются на Δu, т. е. если

$$\boldsymbol{u}_{\text{II}} = \boldsymbol{u}_{\text{I}} + \Delta \boldsymbol{u},$$

то два соответствующих сигнала, $\boldsymbol{f}_{\text{I}}$ и $\boldsymbol{f}_{\text{II}}$, связаны аналогичным соотношением

$$\boldsymbol{f}_{\text{II}} = \boldsymbol{f}_{\text{I}} + \Delta \boldsymbol{f}.$$

Если $\Delta \boldsymbol{u}$ и $\Delta \boldsymbol{f}$ можно рассматривать как малые приращения, то можно записать

$$\Delta \boldsymbol{f} = \frac{d\psi}{du} \Delta \boldsymbol{u} \tag{6}$$

[1] В общем же случае ψ означает любое функциональное преобразование. Так, например, для системы передачи, упомянутой на стр. 102 и 104, символ ψ будет означать преобразование Фурье.

или

$$\|\Delta f\|^2 = d^2 = \sum \Delta f_k^2 = \sum \left(\frac{\partial \psi}{\partial u_k} \Delta u_k\right)^2. \qquad (7)$$

Теперь заметим, что частные производные ψ по координатам и приращения, которые могут получить эти координаты, — величины совершенно независимые. Поэтому

$$d^2 = \sum \left(\frac{\partial \psi}{\partial u_k} \Delta u_k\right)^2 = n \overline{\left(\frac{\partial \psi}{\partial u_k} \Delta u_k\right)^2} =$$

$$= n \overline{\left(\frac{\partial \psi}{\partial u_k}\right)^2} \overline{\Delta u_k^2} = \overline{\left(\frac{\partial \psi}{\partial u_k}\right)^2} \sum \Delta u_k^2,$$

$$d^2 = \bar{\psi}'^2 r^2, \qquad (8)$$

где $r^2 = \sum \Delta u_k^2 = \|\Delta \boldsymbol{u}\|^2$ — квадрат расстояния между $\boldsymbol{u}_{\text{I}}$ и $\boldsymbol{u}_{\text{II}}$ в пространстве U.

Нам нужно найти наименьшее значение расстояния r. Для этого положим, что сообщение представлено квантованной функцией времени и что шаг шкалы квантования равен δ. Тогда приращение каждой координаты сообщения может равняться

$$\Delta u_k = i\delta,$$

где i — любое целое число. Если все Δu_k равны нулю, кроме одного, которое равно δ, то

$$r_{\min} = (\sqrt{\sum \Delta u_k^2})_{\min} = \delta \qquad (9)$$

и это и будет наименьшее возможное значение r. Вводя соотношения (9) и (8), получим наименьшее значение и для d

$$d_{\min}^2 = \bar{\psi}'^2 \delta^2. \qquad (10)$$

Таким образом, формула (5) для вероятности ошибки может быть переписана в виде

$$p_{\text{ош}} = 1 - \Phi\left(\frac{\sqrt{\bar{\psi}'^2}}{2\sqrt{2}} \cdot \frac{\delta}{\sigma}\right). \qquad (11)$$

Следующий этап исследования состоит в рассмотрении функции ψ для различных способов передачи и в сравнении их по помехоустойчивости. Этому посвящен следующий параграф.

§ 45. Удельная помехоустойчивость

Оценивая достоинства различных систем передачи и сравнивая их между собой, мы, естественно, должны были бы отдать предпочтение системе, обеспечивающей бóльшую помехоустойчивость. Точно так же в процессе разработки и совершенствования некоторой данной системы мы должны были бы стремиться к повышению ее помехоустойчивости. Но такая постановка вопроса,

не учитывающая прочих характеристик системы, была бы однобокой.

Дело в том, что повышение помехоустойчивости сопровождается, как правило, увеличением объема сигнала. Поэтому сравнение систем интереснее производить не по абсолютной помехоустойчивости, а по некоторому показателю, который позволил бы судить о том, ценой какого увеличения объема сигнала покупается повышение помехоустойчивости. Простейшей формой такого показателя является отношение

$$X = S/V,$$

где S — помехоустойчивость; V — объем сигнала. Назовем это отношение *удельной помехоустойчивостью*. Совокупность двух показателей — удельной помехоустойчивости и удельной содержательности сигнала (см. § 14) — образует достаточно полную общетехническую характеристику системы связи (если не касаться экономических моментов). Удельная содержательность дает количественную оценку работы системы, тогда как удельная помехоустойчивость является оценкой качества работы.

Постановку вопроса, приводящую к введению понятия удельной надежности, мы поясним на примере. Рассмотрим для этой цели два вида импульсной модуляции — АИМ и ФИМ. Мы определим сперва, пользуясь выводами теории § 44, помехоустойчивость связи, получаемую при применении этих двух систем, а затем сопоставим соответствующие объемы сигналов.

Положим, что функция сообщения $u(t)$ квантована (число ступеней m) и изменяется в пределах $0 < u < u_{\max}$. Тогда для сигнала АИМ можем записать

$$f = \sum f_k = \sum \mu \frac{u_k}{u_{\max}} Q(t - k\Delta t), \qquad (1)$$

где $u_k = u(k\Delta t)$; $\mu \leqslant 1$ — глубина модуляции, а $Q(t)$ — функция, описывающая отдельный импульс. Так как функция Лапласа $\Phi(x)$ есть возрастающая функция своего аргумента, то вероятность ошибки будет тем меньше, а надежность, следовательно, тем больше, чем больше аргумент функции Лапласа в формуле (14) § 44. Считая отношение δ/σ заданным, мы будем рассматривать только величину $\bar{\psi}'^2$. Дифференцируя (1) по u_k, найдем

$$\psi' = \frac{\mu}{u_{\max}} Q, \qquad \bar{\psi}'^2 = \frac{\mu^2}{u_{\max}^2} \bar{Q}^2. \qquad (2)$$

Положим для определенности, что передача ведется треугольными импульсами длительностью τ и высотой h.
При этом

$$\bar{Q}^2 = \frac{1}{3} h^2 \frac{\tau}{\Delta t} = \frac{E}{\Delta t}, \qquad (3)$$

где

$$E = \frac{1}{3} \tau h^2 \qquad (4)$$

энергия одного импульса. Подставляя (3) в (2), получим

$$\bar{\psi}'^2 = \frac{\mu^2}{u_{\max}^2} \cdot \frac{E}{\Delta t}. \tag{5}$$

Наибольшее возможное значение μ равно единице; стало быть,

$$\bar{\psi}'^2_{\max} = \frac{E}{u_{\max}^2 \Delta t}. \tag{6}$$

Эта величина и характеризует помехоустойчивость АИМ при наибольшей глубине модуляции.

Обратимся к ФИМ. В этом случае выражение для сигнала имеет вид

$$f = \sum f_k = \sum Q\left(t - k\Delta t + \lambda \frac{u_k}{u_{\max}} \tau_0\right). \tag{7}$$

Здесь $\lambda \leqslant 1$ — глубина фазовой модуляции; τ_0 — наибольшее смещение импульсов, которое можно принять равным Δt (пренебрегая длительностью импульса по сравнению с периодом следования Δt и не оставляя запасов). Дифференцируя (7), найдем

$$\psi' = \frac{\lambda \Delta t}{u_{\max}} Q', \qquad \bar{\psi}'^2 = \frac{\lambda^2 \Delta t^2}{u_{\max}^2} \bar{Q}'^2. \tag{8}$$

При треугольных импульсах

$$\bar{Q}'^2 = 12 \frac{E}{\tau^2 \Delta t}. \tag{9}$$

Подставляя (9) и (8), получим

$$\bar{\psi}'^2 = 12 \frac{\lambda^2 \Delta t^2}{u_{\max}^2 \tau^2} \cdot \frac{E}{\Delta t} = 3 \frac{\lambda^2 E}{u_{\max}^2 \Delta t} \left(\frac{F_f}{F_u}\right)^2 \tag{10}$$

или для $\lambda = 1$ (наибольшая глубина модуляции)

$$\bar{\psi}'^2_{\max} = 3 \frac{E}{u_{\max}^2 \Delta t} \left(\frac{F_f}{F_u}\right)^2, \tag{11}$$

где $F_f \approx 1/\tau$ — ширина спектра сигнала; $F_u = 1/2\Delta t$ — ширина спектра функции сообщения $u(t)$. Сравнивая (11) с (6), мы видим, что фазовая модуляция дает тем больший выигрыш в помехоустойчивости по сравнению с амплитудной модуляцией, чем больше отношение F_f/F_u. Но так как ширина спектра сообщения задана, то помехоустойчивость повышается с увеличением ширины спектра сигнала.

Таким образом, за повышение помехоустойчивости мы расплачиваемся соответствующим расширением спектра. Правда, увеличение ширины спектра еще не означает пропорционального увеличения объема сигнала.

Мы имеем вообще

$$V = FTH.$$

Длительность сигнала T предполагается во всех случаях неизменной. Ширину можно выразить как

$$F = 1/\tau,$$

где τ — длительность импульса. Но при ФИМ мы имеем $\tau = \Delta t/m$, где m — основание кода, т. е. число ступеней шкалы квантования, а при АИМ длительность импульса τ может равняться периоду следования Δt, т. е. импульсы могут иметь скважность единица. Отсюда и вытекает расширение спектра сигнала при ФИМ. Что же касается величины

$$H = \log \frac{P}{P_{\text{п}}},$$

то средняя мощность сигнала АИМ составляет (при $m \geqslant 1$)

$$P = \frac{1}{3} \cdot \frac{E}{\Delta t} m^2,$$

а для ФИМ

$$P = E/\Delta t,$$

где E — по-прежнему означает энергию одного импульса.

Итак, при переходе от АИМ к ФИМ спектр сигнала расширяется, а мощность уменьшается, но так, что объем сигнала (при принятых определениях) в общем возрастает. Отсюда вытекает целесообразность введения такой оценки, как удельная помехоустойчивость.

Весьма возможно, что в дальнейшем удастся, пересмотрев соответствующим образом определения, установить некоторый критерий помехоустойчивости, который был бы инвариантным по отношению к изменениям способа передачи. Другими словами, можно, вероятно, подобрать такую функцию помехоустойчивости и объема сигнала, которая была бы константой для всех способов передачи. Введение такой константы сильно облегчило бы формулировку общих положений теории связи. Более того, разыскание подобной константы было бы равносильно установлению одного из неизвестных пока законов сохранения, специфичных для теории связи. То же самое относится и к инвариантным преобразованиям, в которые входят удельная содержательность и объем сигнала.

§ 46. Порог помехоустойчивости

При дальнейшем обсуждении теории помехоустойчивости будет отброшено одно из принятых ранее упрощений.

Пусть сообщение u возрастает так, что все u_k увеличиваются в одинаковое число раз. На осциллограмме сообщения это соответствует изменению вертикального масштаба без изменения формы кривой. В пространстве же сообщений пропорциональное

изменение всех составляющих вектора сообщения выразится в изменении длины вектора без изменения его направления. Следовательно, при непрерывном увеличении сообщения точка, представляющая сообщение, будет перемещаться по прямой, проходящей через начало координат.

Обратимся теперь к пространству сигналов. Каждому приращению сообщения соответствует определенное приращение сигнала. При непрерывном увеличении сообщения конец вектора сигнала, т. е. точка сигнала, описывает некоторую траекторию, которая называется линией сигнала. Вид линии сигнала зависит

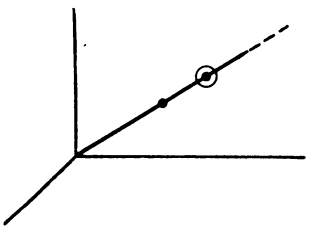


Рис. 44

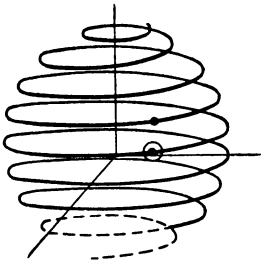


Рис. 45

от способа модуляции. Если приращение сигнала пропорционально приращению сообщения, т. е. если сигнал связан с сообщением линейной зависимостью, то линия сигнала есть прямая (рис. 44). Так обстоит, например, дело при амплитудной модуляции. Но если применяется модуляция, при которой связь между Δu и Δf нелинейна, то линия сигнала принимает более сложную форму.

Рассмотрим, к примеру, частотную модуляцию. Частотно-модулированный сигнал характеризуется тем, что энергия сигнала данной длительности остается неизменной при любых изменениях передаваемого сообщения. С точки зрения геометрии пространства F это означает, что длина вектора сигнала остается неизменной. Таким образом, линия ЧМ сигнала лежит на поверхности сферы соответствующего числа измерений. Нет возможности изобразить графически такое положение вещей; условно трехмерная картина дана на рис. 45.

Если сообщение и сигнал квантованы, т. е. могут иметь только конечные приращения, то непрерывные линии в пространствах U и F заменяются соответствующими совокупностями отдельных точек.

Мы не можем, к сожалению, дать на чертеже действительный вид многомерной линии сигнала для таких видов модуляции, как ЧМ. Но, может быть, небесполезно построить точки, представляющие квантованный импульсный сигнал, определяемый всего тремя мгновенными значениями и представимый в обычном трехмерном пространстве.

Рассмотрим сигналы АИМ, ДИМ и ФИМ. Относительно последних двух предположим, что ширина импульса для ДИМ и сдвиг импульса для ФИМ квантованы так, что приращения указанных величин равны длительности импульса ФИМ. Этим условием определяется длительность импульса по отношению к периоду следования Δt. Что касается АИМ, то длительность импульса мы выберем такую же, как и длительность импульса ФИМ и как минимальная длительность импульса ДИМ, так как при этом условии ширина спектра сигнала при всех трех видах модуляции будет одинакова. Ширина спектра определяет частоту отсчетов. Вообще

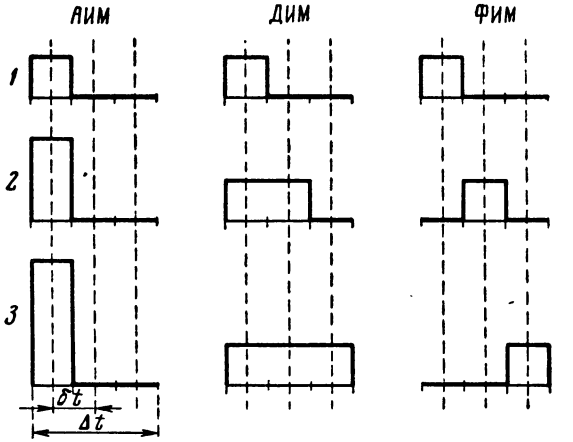


Рис. 46

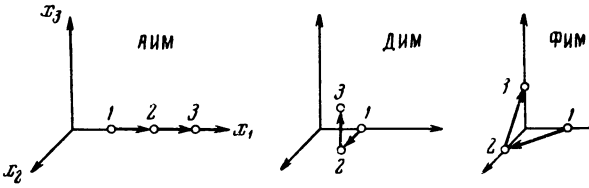


Рис. 47

говоря, $\delta t = 1/2F_f$. Но легко видеть, что в рассматриваемом случае следует взять интервал δt между соседними отсчетами, равный шагу шкалы квантования для ДИМ и ФИМ. Все описанные соотношения изображены на рис. 46 для трех сообщений, равных соответственно 1, 2 и 3 единицам. Моменты отсчетов обозначены вертикальными линиями.

Координаты изображенных на рис. 46 сигналов, т. е. мгновенные значения сигналов в моменты отсчетов, могут быть представлены следующими числами:

Сообщение	АИМ	ДИМ	ФИМ
1	100	100	100
2	200	110	010
3	300	111	001

Точки сигналов с такими координатами построены на рис. 47, дающем в отличие от рис. 45 уже не условное, а действительное изображение. Оно ясно показывает, в частности, что точки сигнала ФИМ располагаются на одном и том же расстоянии от начала координат, тогда как расстояния от начала точек сигнала АИМ возрастают пропорционально увеличению сообщения.

Однако три построения рис. 47 еще нельзя непосредственно сопоставлять друг с другом: нужно предварительно привести их к одному масштабу. Для этого следует уравнять средние энергии сигналов. Принимая за единицу энергию сигнала, соответствующего единичному сообщению, получим следующие значения энергии:

Сообщение	АИМ	ДИМ	ФИМ
1	1	1	1
2	4	2	1
3	9	3	1
Средняя энергия	4,67	2	1

Беря единицу линейного масштаба обратно пропорциональной корню квадратному из средней энергии, получим, что длины масштабных единиц на рис. 47 для АИМ, ДИМ и ФИМ должны относиться соответственно как 0,46 : 0,71 : 1,0.

Из всего сказанного следует, что совокупность точек сигнала, лежащих на линии сигнала, определяет *длину* линии сигнала в зависимости от способа модуляции. При одном и том же изменении сообщения ЧМ дает бóльшую длину линии сигнала, чем АМ, а для импульсной модуляции при ДИМ получается большая длина линии сигнала, чем при АИМ, но меньшая, чем при ФИМ.

Теперь мы можем обратиться непосредственно к помехоустойчивости. В § 44 было установлено, что помехоустойчивость определяется расстоянием d между двумя точками сигнала, соответствующими заданному приращению сообщения. Чем расстояние d больше, тем выше помехоустойчивость, так как для перехода точки сигнала из одного положения в другое вследствие наложения помехи требуется мгновенное значение помехи, равное d[1]. Очевидно, что чем d больше, тем меньше вероятность появления равного ему значения помехи.

Это положение сохраняет силу и в дальнейших рассуждениях. Но вспомним, как определялась в § 44 величина d. Мы полагали приращение сообщения малым и писали на этом основании

$$\Delta f_k = \frac{\partial \psi}{\partial u_k} \Delta u_k, \tag{1}$$

$$d^2 = \sum \Delta f_k^2 = \sum \left(\frac{\partial \psi}{\partial u_k} \Delta u_k \right)^2. \tag{2}$$

[1] Если речь идет об идеальном приемнике, для которого пространство сигналов разделено на граничащие друг с другом области, то соответствующее мгновенное значение помехи должно равняться $d/2$.

Но формула (2) определяет в сущности не расстояние между двумя точками, лежащими на линии сигнала, а приращение длины линии сигнала. Таким образом, в предыдущем оказывалось, что помехоустойчивость растет с увеличением длины линии сигнала и что, следовательно, большей помехоустойчивостью обладают те системы модуляции, которые при изменении сообщения в данных пределах дают большую длину линии сигнала.

Это верно в том единственном случае, когда линия сигнала есть прямая. Если же линия сигнала кривая, то ясно, что длина дуги этой кривой не равна хорде и всегда больше хорды. Иначе

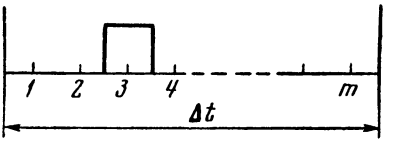


Рис. 48

говоря, расстояние между двумя точками должно измеряться по прямой, соединяющей эти точки, а не вдоль какой-либо проходящей через них кривой. Правильное выражение для расстояния между двумя точками сигнала, отвечающими двум значениям сообщения, можно записать в виде

$$d^2 = \sum \Delta f_k^2 = \sum [\psi (u_k + \Delta u_k) - \psi (u_k)]^2. \qquad (3)$$

Полагая по-прежнему, что ψ есть просто символ некоторой функции и пользуясь разложением разности в квадратных скобках в ряд Тэйлора, получим

$$d^2 = \sum \left[\psi' (u_k) \Delta u_k + \psi'' (u_k) \frac{\Delta u_k^2}{2!} + \psi''' (u_k) \frac{\Delta u_k^3}{3!} + \dots \right]^2. \qquad (4)$$

Последняя формула ясно обрисовывает положение: она показывает, что использованное ранее соотношение (2) есть приближение, справедливое при малых Δu_k; в то же время соотношение (2) является точным, если все производные порядка выше первого равны нулю, т. е. в случае, когда ψ есть линейная функция. Формула (4) показывает также, что в общем случае расстояние d растет не пропорционально Δu_k. Влияние членов высших порядков замедляет рост d. Для некоторых видов модуляции d оказывается постоянной величиной, не зависящей от изменений сообщения.

Поясним это на примере ФИМ. На рис. 48 изображен импульс, могущий занимать на протяжении интервала Δt m различных положений; иначе говоря, мы имеем квантованный сигнал ФИМ при коде с основанием m. Высота импульса пусть равна δ. Таким образом, на протяжении Δt сигнал определяется m координатами, из которых одна равна δ, а все остальные $m-1$ координат равны нулю. Любое изменение сообщения будет сопровождаться изме-

нением только двух координат; одна из них примет значение 0 вместо δ, вторая — наоборот.

Итак, при любом изменении сообщения изменяется по абсолютной величине только $2/m$ доля всех координат. Общее же число координат равно $mT/\Delta t$, где T — общая длительность сообщения. С другой стороны,

$$\Delta t = 1/2F_и,$$

где $F_и$ — ширина спектра сообщения. Таким образом, число претерпевших изменение координат равно

$$\frac{2}{m} 2mTF_и = 4TF_и.$$

Квадрат приращения каждой из изменившихся координат равен δ^2, и мы получим

$$d^2 = \sum \Delta f_k^2 = 4TF_и\delta^2, \qquad (5)$$

т. е. постоянную величину, не зависящую от приращения сообщения. С геометрической точки зрения это означает, что точки сигнала, лежащие на сферической поверхности, являются попарно равноотстоящими[1]. Таковы, например, три точки, представляющие сигнал ФИМ на рис. 47.

Теперь может быть поставлен вопрос о различии влияния малой и большой помех. Это различие проявляется для тех видов модуляции, которые характеризуются длинной линией сигнала, делающей много витков или петель. Пока помеха мала по сравнению с расстоянием между соседними витками линии сигнала, действуют соотношения линеаризованной теории § 44, и помехоустойчивость возрастает с увеличением длины линии сигнала, т. е. с увеличением $\overline{\psi'^2}$. Когда же помеха достигает по порядку величины расстояния между ближайшими витками, то возникают аномальные ошибки. На рис. 45 отмечены две точки сигнала, лежащие на соседних витках. Кружок означает область неопределенности. Радиус этой области равен средней энергии помехи. Аномальная ошибка произойдет при перебросе результирующего вектора (т. е. вектора суммы сигнала и помехи) с одного витка на другой. Надо заметить, что увеличение ширины полосы приведет лишь к более тесному расположению витков линии сигналов и аномальные ошибки возникнут при меньшей энергии помехи, хотя при малой помехе помехоустойчивость будет больше.

Из сказанного следует, что помехоустойчивость должна резко убывать после того, как превышение сигнала над помехой на входе приемника достигнет, уменьшаясь, некоторого предельного зна-

[1] Заметим, что наибольшее число попарно равноотстоящих точек, могущих быть размещенными в пространстве n измерений, равно $n+1$. Так, например, в трехмерном пространстве точки располагаются в вершинах равностороннего тетраэдра.

чения. Это предельное значение можно назвать порогом помехоустойчивости. На рис. 49 изображен примерный ход зависимости помехоустойчивости от превышения на входе приемника. Кривые *1*, *2* и *3* относятся соответственно к АМ (или АИМ), ЧМ (или ФИМ) и ЧМ с очень широкой полосой.

Существование порога для систем с постоянной средней мощностью (ЧМ, ФМ, ФИМ) может быть пояснено еще и следующим соображением. Для названных систем все точки сигналов лежат на сферической поверхности, причем радиус сферы равен средней

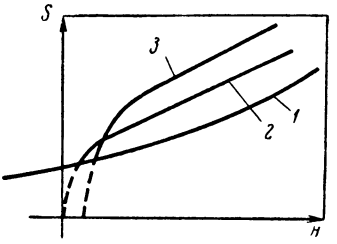


Рис. 49

энергии сигнала. Если теперь помеха будет возрастать, то она образует сферу, которая, увеличиваясь, может охватить всю сферу сигнала. Это произойдет, когда энергия помехи сравняется по порядку с энергией сигнала, т. е. когда превышение $H = \log \frac{P}{P_п}$ приблизится к нулю. При таких обстоятельствах правильный прием становится очень маловероятным, т. е. помехоустойчивость сделается весьма низкой. Для амплитудной же модуляции включение точек сигнала в сферу помехи происходит постепенно, так как точки сигнала при АМ расположены вдоль прямой. Поэтому резкого порога помехоустойчивости при АМ не обнаруживается.

Глава 4

РАЗДЕЛЕНИЕ СИГНАЛОВ

§ 47. Вводные замечания

При применении многоканальной связи задача состоит в передаче по одной линии нескольких независимых сообщений. На передающем конце линии сигналы нескольких каналов смешиваются и поступают в линию. На приемном конце линии полученную смесь нужно разделить. Задача разделения (селекции) составляет основную проблему многоканальной связи.

Нужно прежде всего заметить, что задача разделения сигналов имеет общие черты с задачей борьбы с помехами. Борьба

с помехами также имеет своей целью разделение некоторой смеси, а именно смеси сигнала с помехой; разделение должно быть в идеальном случае выполнено так, чтобы сигнал мог быть выделен в полностью очищенном от помех виде. Задача разделения сигналов при многоканальной связи сходна: она состоит в том, чтобы выделить из смеси многих сигналов сигнал данного канала, по возможности очистив его от влияния соседних каналов. Из этого сопоставления следует, между прочим, что влияние соседних каналов, т. е. остатки сигналов, принадлежащих другим каналам и наложенных на сигнал данного канала после разделения, естественно рассматривать как особого рода помеху, специфичную для многоканальной системы связи. Именно так вопрос и ставится в технике связи.

Однако между разделением сигналов и борьбой с помехами имеется и существенное различие. Дело в том, что помеха представляет собой явление, находящееся вне нашей власти; физические характеристики помехи, зачастую весьма неблагоприятные с точки зрения возможностей подавления помехи, не зависят от нашего усмотрения. В то же время выбор сигналов, применяемых при многоканальной связи, полностью зависит от нас; мы можем наделить эти сигналы вполне определенными признаками, по которым сигналы могут быть отличены друг от друга и разделены. Различие в способах разделения есть различие в том физическом признаке, по которому сигналы различных каналов отличаются друг от друга. Одна из проблем теории состоит в установлении общих требований, выполнение которых обеспечивает потенциальную разделимость сигналов. Далее, теория должна указать нам необходимые свойства разделяющих устройств; дать возможность обозреть все разнообразие теоретически мыслимых способов разделения; дать руководящие соображения для общей оценки качества разделения. Эти пожелания теория на современном своем уровне удовлетворяет лишь частично. Достаточно развита лишь теория линейного разделения (селекции), начало которой положил в 1935 г. своей глубокой работой Д. В. Агеев [2] [1].

Нелинейные же методы разделения не только не разработаны в сколько-нибудь общем виде, но и перспективы их применения пока далеко не ясны.

В практике многоканальной связи в настоящее время находят почти исключительное применение два метода разделения: *по частоте* и *по времени*. *Фазовое разделение*, представляющее собой частный случай временно́го, имеет ограниченное применение. Все три названных метода разделения относятся к числу линейных. Возможен еще один, в некотором смысле более общий вид линейного разделения — *разделение по форме сигнала* (по Агееву этот способ разделения называется компенсационной селекцией).

---

[1] К аналогичным положениям пришел лишь семнадцать лет спустя Задэ [29].

В дальнейшем будет показано, что разделением по частоте, по времени и по форме исчерпываются возможные виды линейного разделения.

§ 48. Частотное и временно́е разделение

Не вдаваясь в технические подробности, рассмотрим некоторые общие свойства двух способов разделения, имеющих основное значение в современной связи: частотного и временно́го разделения.

Идея частотного разделения весьма проста. Она состоит в том, что сигналы различных каналов размещаются в неперекрывающихся частотных полосах. Это осуществляется путем применения в каждом канале независимых модуляторов и соответственным образом разнесенных несущих частот. Разделение сигналов выполняется набором полосовых фильтров, каждый из которых пропускает полосу частот, относящуюся к данному каналу.

Не менее просто по идее и временно́е разделение. Передача в этом случае осуществляется так, что элементы сигнала, принадлежащего данному каналу, передаются в интервалы времени, свободные от сигналов других каналов. Таким образом, этот принцип применим лишь в импульсной связи. Для разделения сигналов на приемном конце необходим коммутатор, работающий синхронно с распределителем на передающем конце. Принцип временно́го разделения издавна применяется в телеграфии.

Но можно показать, опираясь на самые общие соображения, что идеальное разделение ни частотным, ни временны́м способами невозможно, т. е. что всегда должно существовать взаимное влияние каналов, состоящее в появлении сигналов соседних каналов на выходе данного канала.

Рассмотрим частотное разделение. Этот способ разделения основан на предположении, что сигнал данного канала занимает ограниченную полосу частот и что имеется полосовой фильтр, способный выделить эту полосу. Но всякий сигнал имеет конечную длительность, а следовательно, должен иметь бесконечно протяженный спектр. Таким образом, локализация спектра сигнала конечной длительности в конечной полосе частот принципиально невозможна. Если же мы искусственно ограничим спектр сигнала, то сигнал будет искажен. Следовательно, мы должны мириться либо с переходными помехами, либо с искажением сигнала. К этому нужно еще добавить, что идеальный фильтр, т. е. фильтр с бесконечно большим затуханием вне полосы прозрачности, можно себе представить лишь в виде цепочки из бесконечного числа звеньев; такой фильтр обладал бы бесконечным временем пробега, так что полная очистка сигнала от нежелательных частотных составляющих потребовала бы бесконечно большой задержки сигнала.

Аналогичные трудности возникают и при временно́м разделении. При применении этого способа предполагается, что каждый

элемент сигнала локализован во времени в пределах некоторого интервала и имеет вид импульса, равного нулю вне этого интервала. Но функция конечной длительности имеет бесконечно протяженный спектр, который не может быть передан по каналу связи. Результатом же ограничения спектра является расплывание импульса во времени на основе известного соотношения

$$\Delta f \Delta t = \mu,$$

где Δf — ширина спектра; Δt — длительность; μ — постоянная порядка единицы. При расплывании импульса он попадает в соседние интервалы, и таким образом возникают переходные помехи и при этом способе разделения [1].

Все это не означает, конечно, что частотная и временна́я многоканальные системы не могут работать; эти соображения указывают лишь на принципиальные ограничения, с которыми приходится считаться при практическом осуществлении этих систем. Как известно, для обеспечения хорошего качества передачи приходится оставлять запасные интервалы как по частоте, так и по времени, что позволяет удержать переходные помехи на достаточно низком уровне.

Что касается разделяющих устройств, то, говоря о них опять-таки в самом общем плане, можно сказать, что эти устройства должны быть чувствительны к тому признаку, по которому различаются подлежащие разделению сигналы. Иначе говоря, параметры разделяющих устройств должны зависеть от физической величины, положенной в основу разделения. Таким образом, параметры разделяющих устройств при частотном разделении должны зависеть от частоты, а при временно́м разделении — от времени. Так, например, действие разделяющих фильтров при частотном разделении определяется их частотными характеристиками. При временно́м же разделении характеристикой устройства являются, например, проводимости тех или иных цепей, изменяющиеся во времени по закону, задаваемому синхронным коммутатором.

Другими словами, системы, применяемые в качестве разделительных устройств при частотном и временном разделениях, будучи линейными, относятся к двум разным классам, а именно: частотное разделение осуществляется линейными системами с постоянными параметрами, а временно́е разделение — линейными системами с переменными параметрами. Соответственно различается и математическое описание этих систем. Некоторые черты временно́го способа разделения поясняются ниже на примере фазового разделения.

---

[1] Взаимное перекрытие импульсов не означает еще невозможности их разделения. Некоторые подробности по этому поводу см. в Добавлении 10.

§ 49. Фазовое разделение

Фазовое разделение есть частный случай временно́го разделения, когда разделяемые сигналы представляются синусоидальными колебаниями. Если нужно разделить два синусоидальных колебания, то разделение возможно, если частоты колебаний неодинаковы, — и это будет частотное разделение. Разделение возможно и в том случае, когда частоты колебаний одинаковы, но одно из них сдвинуто во времени относительно другого, т. е. если два сигнала записаны в виде

$$f_1(t) = a \sin \omega t,\ f_2(t) = b \sin \omega (t + \tau) = b \sin (\omega t + \varphi).$$

Таким образом, смещение во времени выражается через фазный сдвиг, и разделение по этому признаку и есть фазовое разделение (фазовая селекция).

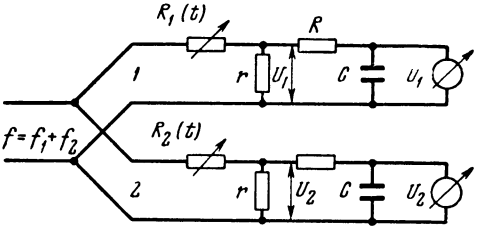


Рис. 50

Фазовое разделение простейшим образом осуществляется посредством синхронного детектирования. Пусть напряжение суммы сигналов

$$f(t) = f_1(t) + f_2(t)$$

подается на вход разделяющего устройства (рис. 50). В каждой ветви схемы имеется переменное сопротивление $R(t)$, которое включено последовательно с сопротивлением и изменяется таким образом, что проводимость меняется в первой ветви по закону

$$Y_1(t) = \frac{1}{r}(1 + \sin \omega t),$$

а во второй — по закону

$$Y_2(t) = \frac{1}{r}[1 + \sin (\omega t + \psi)].$$

Напряжения, снимаемые с сопротивлений r, усредняются интегрирующими звеньями RC и измеряются выходными вольтметрами. Мы имеем следующие выражения для выходных напряжений:

$$\bar{U}_1 = r(f_1 + f_2) Y_1 = \frac{1}{2}(a + b \cos \varphi),$$

$$\bar{U}_2 = r(f_1 + f_2)\, Y_2 = \frac{1}{2}(a\cos\psi +$$

$$+ b\cos\varphi\cos\psi + b\sin\varphi\sin\psi).$$

Нетрудно видеть, что возможно полное разделение, т. е. такое положение, когда $\bar{U}_1$ зависит только от f_1, $\bar{U}_2$ — только от f_2 [1].

Полное разделение получается в том единственном случае, когда

$$\varphi - \psi = \pi/2.$$

При этом

$$\bar{U}_1 = \frac{1}{2}a, \quad \bar{U}_2 = \frac{1}{2}b.$$

Таким образом, полностью разделимы сигналы, в которых переносчики представляют собой две синусоиды, сдвинутые по фазе на $\pi/2$, т. е. синусоиду и косинусоиду. Это значит, что при фазовом разделении возможно осуществить только два независимых канала; если же фазовое разделение применяется в комбинации с каким-либо другим способом разделения, то возможно, как известно, удвоение числа каналов, получаемого без фазового разделения.

При синхронном детектировании происходит собственно разделение не сигналов, а сообщений (как показывает и само название этого процесса); другими словами, процесс разделения сигналов неразрывно совмещен с отделением сообщения от сигнала, т. е. с детектированием.

§ 50. Разделение по форме

Для уяснения особенностей разделения сигналов, различающихся по форме, рассмотрим прежде всего два примера.

Пусть имеются два сигнала

$$f_1(t) = a_1\cos\omega_1 t + a_2\cos\omega_2 t, \qquad f_2(t) = b_1\cos\omega_1 t - b_2\cos\omega_2 t.$$

Частоты ω_1 и ω_2 произвольны, может быть, даже несоизмеримы, произвольны также амплитуды составляющих; наконец, могут быть произвольны и их фазы, но для простоты в нашем примере для составляющей частоты ω_1 взят фазный сдвиг, равный нулю, а для составляющей частоты ω_2 — сдвиг, равный π. При таких обстоятельствах различие между сигналами f_1 и f_2 уже не явля-

[1] Полное разделение оказывается возможным потому, что мы рассматриваем здесь для упрощения рассуждений абстрактные идеально периодические сигналы. При такой абстракции оказалось бы возможным полное разделение и частотным способом.

ется различием по частоте или различием по фазе; различие имеет более общий характер, и мы определяем его как различие *по форме*.

Если форма сигнала является его отличительным признаком, то и разделяющее устройство должно быть способно разделять сигналы, ориентируясь на этот признак. Задача разделяющего устройства состоит в том, чтобы дать на выходе канала *1* сигнал, зависящий только от f_1, а на выходе канала *2* — сигнал, зависящий только от f_2.

На вход разделяющего устройства поступает смесь обоих сигналов

$$f = f_1 + f_2 = (a_1 + b_1)\cos\omega_1 t + (a_2 - b_2)\cos\omega_2 t.$$

Разделим эту смесь прежде всего по частоте и после линейного детектирования получим два напряжения

$$U' = a_1 + b_1, \qquad U'' = a_2 - b_2.$$

Теперь составим суммы

$$U_1 = U' + \beta U'', \quad U_2 = \alpha U' - U''.$$

Если выбрать

$$\beta = b_1/b_2, \qquad \alpha = a_1/a_2,$$

то получится

$$U_1 = a_1 + \beta a_2, \qquad U_2 = \alpha b_1 + b_2.$$

Два выходных напряжения U_1 и U_2 зависят, таким образом, только от составляющих f_1 и f_2 соответственно, и сигналы, следовательно, разделены. Все описанные операции выполняются, например, схемой рис. 51 *.

В качестве второго примера возьмем рассмотренный в работе М. В. Назарова [13] случай, когда два сигнала на протяжении некоторого промежутка времени заданы в виде

$$f_1(t) = a_0, \quad f_2(t) = a_1 t.$$

Сигналы могут передаваться одновременно, их сплошные спектры перекрываются; таким образом, и в этом случае различие сигналов есть различие в форме. На вход разделительного устройства попадает сумма

$$f = f_1 + f_2 = a_0 + a_1 t.$$

Для разделения поступим так: продифференцируем сумму

$$\dot{f} = a_1,$$

---

* Несмотря на наличие детекторов, схема рис. 51 рассматривается как линейная. В детекторах нет необходимости; в работе Агеева [2], откуда заимствован разобранный пример, приведена несколько более сложная схема с преобразованием частоты, выполняющая те же операции.

а затем проинтегрируем от 0 до t

$$\int_0^t \dot{f}dt = a_1 t.$$

Таким образом выделен второй сигнал. Первый сигнал выделяется путем вычитания второго сигнала из суммы. Эти операции выполняются разделительным устройством, скелетная схема которого изображена на рис. 52. Через Д, И и $\ominus$ обозначены соответственно дифференцирующее, интегрирующее и вычитающее устройства. Все три операции, разумеется, линейны.

Аналогичным образом разделяется и большее количество сигналов. Так, например, если подлежит разделению сумма трех сигналов

$$f = f_1 + f_2 + f_3 = a_0 + a_1 t + a_2 t^2,$$

то для разделения требуются следующие операции:

$$\dot{f} = a_1 + 2a_2 t, \qquad \int_0^t \dot{f}(x)\,dx = f_2 + f_3,$$

$$\ddot{f} = 2a_2, \qquad \int_0^t dy \int_0^y \ddot{f}(x)\,dx = f_3,$$

$$f_1 = f - (f_2 + f_3), \qquad f_2 = (f_2 + f_3) - f_3.$$

Схема, выполняющая перечисленные операции, может быть составлена из двух дифференцирующих, трех интегрирующих и двух вычитающих устройств, соединенных между собой, как показано на скелетной схеме, рис. 53.

Сопоставляя оба рассмотренных примера, можно подметить одну характерную особенность. Особенность эта состоит в том, что разделение сводится к повторному выделению одного из сигналов с последующим вычитанием его из суммы. Эта операция должна повторяться до тех пор, пока все сигналы не окажутся разделенными. Применение вычитания является, возможно, общим признаком разделения по форме.

Мы можем теперь высказать некоторые общие суждения о рассмотренных способах разделения. Каковы бы ни были передаваемые сигналы, они могут быть представлены на конечном промежутке времени соответствующими рядами Фурье. Различие сигналов может быть выражено через различие параметров этих рядов. Если ряды различаются частотами своих составляющих, в частности, если частоты одного ряда укладываются в полосу частот, не перекрывающуюся с полосами, занимаемыми другими рядами (амплитуды и фазы при этом произвольны), то мы имеем дело с сигналами, разделимыми по частоте. Если частоты совпадают, но все составляющие одного ряда сдвинуты по фазе относительно соответствующих составляющих другого ряда, причем фазный

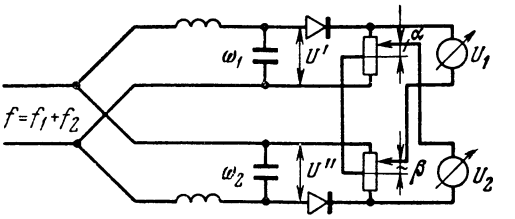


Рис. 51

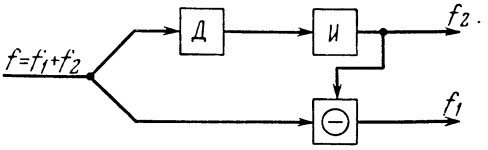


Рис. 52

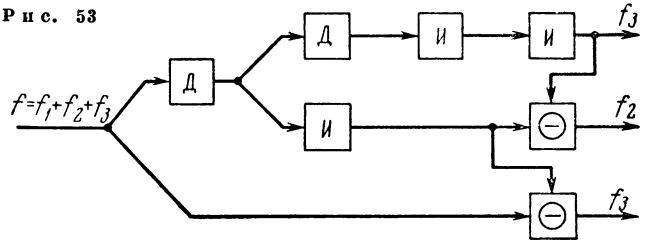


Рис. 53

сдвиг пропорционален частоте, то мы имеем дело с сигналами, раздвинутыми друг относительно друга во времени, и здесь применимо временно́е разделение. Если же частоты совпадают или полосы перекрываются, а амплитуды и фазы произвольны — и это, очевидно, наиболее общий случай, — то различие сигналов есть различие по форме и разделение сигналов может быть произведено только по этому признаку либо способами, зависящими от конкретного вида сигналов, либо на основе общего метода, описываемого в § 56.

Необходимо, однако, отметить еще один частный случай, а именно случай, когда частоты и фазы совпадают, а все амплитуды составляющих в одно и то же число раз больше соответствующих амплитуд другого ряда. В этом случае мы имеем дело с сигналами одинаковой формы, различающимися только величиной. И в этом случае разделение возможно, но только уже не линейными средствами. Разделение сигналов, различающихся только величиной, мы назовем разделением *по уровню*. Этому виду разделения посвящен следующий параграф.

§ 51. Разделение по уровню

Разделением по уровню назовем случай, когда сигналы различных каналов имеют одинаковую форму, посылаются одновременно и различаются только величиной. Пусть, например, имеются прямоугольные амплитудно-модулированные импульсы. Если вы-

соты импульсов могут принимать в каждом канале любое из значений $h_i = i\delta$, то разделение невозможно. В самом деле, если принимается, скажем, значение $h_i = 3$, то неизвестно, передается ли импульс высотой три единицы по одному каналу или импульс высотой 2 по другому каналу. К тому же неизвестны и не могут быть установлены номера каналов, которым принадлежат составляющие принятого сигнала. Разделение возможно лишь при соблюдении определенных условий, которые мы сейчас и установим.

Начнем с простейшего случая двух каналов, сигнал каждого из которых представляет собой произвольную временну́ю последовательность импульсов, имеющих высоту h_1 в первом канале и h_2 во втором. Сигналы обоих каналов могут быть принципиально всегда разделены при условии $h_1 \neq h_2$. Рассмотрим возможный процесс разделения.

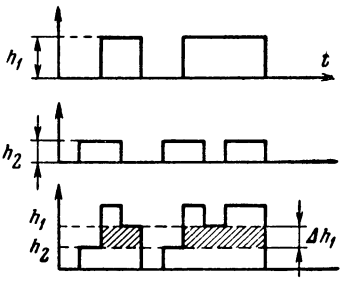


Рис. 54

Пусть $h_2 < h_1$. Тогда мы можем выделить первый сигнал при помощи ограничения смеси двух сигналов снизу на уровне h_2, а сверху на уровне h_1, т. е. путем вырезания из смеси полосы высотой

$$\Delta h = h_1 - h_2.$$

В результате такого ограничения сигнал № 1 выделяется в чистом виде (рис. 54). Для выделения сигнала № 2 нужно увеличить полученный сигнал № 1 в

$$h_1/\Delta h = h_1/(h_1 - h_2)$$

и вычесть его из смеси обоих сигналов. В результате этой второй операции будет выделен в чистом виде и сигнал № 2.

Перейдем теперь к случаю многих каналов. И в этом случае разделение возможно, если высоты импульсов всей совокупности сигналов образуют сходящийся ряд, причем сумма членов этого ряда, начиная с k-го, всегда меньше $(k-1)$-го члена, т. е.

$$\Sigma_{k-1} = \sum_{i=k}^{\infty} h_i < h_{k-1}. \qquad (1)$$

Положим, что требуется передать конечное число n сигналов. Для их разделения потребуется такая последовательность операций.

1. а). Ограничиваем смесь сигналов сверху на уровне h_1, а снизу на уровне $\Sigma_1 = \sum_{k=2}^{n} h_k$. б) Увеличиваем выделенный сигнал № 1 в отношении

$$\frac{h_1}{h_2 - \Sigma_1} = \frac{h_1}{\Delta_1}.$$

2. а) Ограничиваем сумму Σ_1, представляющую собой смесь всех сигналов, кроме № 1, сверху на уровне h_2, а снизу на уровне $\Sigma_2 = \sum_{k=3}^{n} h_3$. б) Увеличиваем выделенный сигнал № 2 в отношении

$$\frac{h_2}{h_2 - \Sigma_2} = \frac{h_2}{\Delta_2}$$

и вычитаем его из Σ_1; получаем, таким образом, сумму Σ_2.

3. а) Ограничиваем сумму Σ_2 сверху на уровне h_3, а снизу на уровне $\Sigma_3 = \sum_{k=4}^{n} h_k$, ..., и т. д.

Эта последовательность действий иллюстрируется рис. 55 для случая четырех каналов.

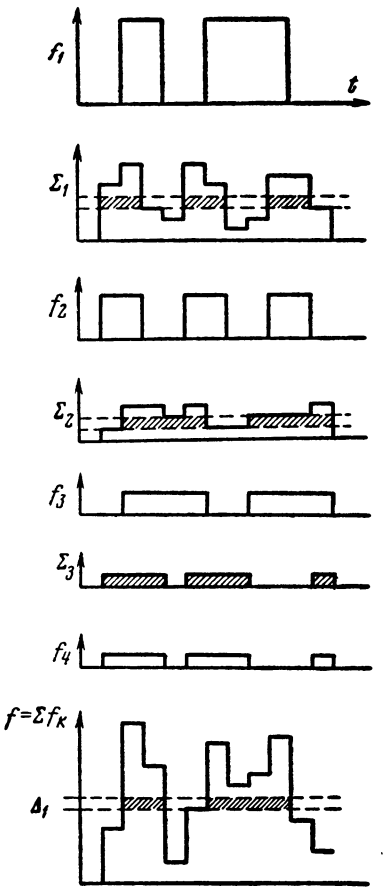


Рис. 55

В качестве простейшего ряда возьмем геометрическую прогрессию

$$h_1 = 1, \quad h_2 = q, \quad h_3 = q^2, \ldots, \quad h_k = q^{k-1}.$$

Нужно определить знаменатель прогрессии. Воспользуемся условием (1)

$$\Sigma_{k-1} = \sum_{i=k}^{\infty} h_i = \sum_{i=k}^{\infty} q^{i-1} < h_{k-1} = q^{k-2},$$

откуда

$$\frac{q^{k-1}}{1-q} < q^{k-2} \text{ или } q < 1/2.$$

Для конечного числа членов можно знак неравенства заменить знаком равенства и положить $q = 1/2$. К этому же значению мы придем, выбрав первую разность Δ_1 (высоту полосы, ограничиваемую при выделении сигнала № 1) равной уровню самого слабого сигнала, так как этот уровень предполагается еще различимым. Если принять это условие, то получится

$$\Delta_1 = h_1 - \Sigma_1 = h_1 - \sum_{k=2}^{n} h_k =$$

$$= 1 - \sum_{k=2}^{n} q^{k-1} = h_n = q^{n-1},$$

откуда $q = 1/2$.

При таком выборе знаменателя прогрессии все разности оказываются одинаковыми. Действительно,

$$\Delta_m = h_m - \Sigma_m = h_m - \sum_{k=m+1}^{n} h_k = q^{m-1} - \sum_{k=m+1}^{n} q^{k-1} =$$

$$= q^{m-1}\,\frac{1-2q+q^{n-m+1}}{1-q}.$$

Положив $q=1/2$, найдем

$$\Delta_m = \left(\frac{1}{2}\right)^{n-1},$$

т. е. высота ограничиваемой полосы уровней на всех ступенях разделения оказывается неизмененной. Это еще одно интересное свойство прогрессии со знаменателем, равным 1/2.

Итак, разделение сигналов, различающихся только по величине, возможно при соблюдении вышеописанных условий с применением нелинейных операций (ограничения).

§ 52. Комбинационное разделение

Возможен метод разделения, переводящий задачу построения многоканальной системы связи в несколько иную плоскость. Начнем с простейшего примера. Пусть имеются два канала и пусть оба канала работают двоичным кодом с элементами 0 и 1. Тогда возможны следующие комбинации сигналов в обоих каналах:

№ комбинации . . .	1	2	3	4
Канал № 1	0	1	0	1
Канал № 2	0	0	1	1
Сумма сигналов . .	0	1	1	2

Как видим, если сигналы обоих каналов будут смешаны, то разделить их будет невозможно, так как суммарный сигнал, равный единице, означает наличие импульса в одном канале и отсутствие в другом, но неизвестно, в каком именно. Но мы можем вместо суммарного сигнала передавать номер комбинации, так как этот номер однозначно определяет сигналы каждого из каналов в отдельности. Таким образом, дело сводится к передаче четырех чисел. Для понимания существа дела важно отметить, во-первых, что это могут быть какие угодно четыре различных числа, а во-вторых, что эти числа могут быть переданы любым способом, т. е. закодированы любым кодом и переданы посредством любого вида модуляции.

Построение многоканальной системы сводится теперь к созданию некоторого устройства, на n входов которого поступают сигналы n каналов (назовем их *канальными сигналами*) и которое вырабатывает посылаемый в линию результирующий сигнал (назовем его *линейным сигналом*) в форме кодовой комбинации,

отображающей совокупность мгновенных значений канальных сигналов в данный момент. Число таких комбинаций равно, очевидно,

$$N = m^n,$$

где n — число каналов, а m — основание кода в канале до преобразования. Так, при импульсной пятиканальной системе и при применении в каждом канале кода с основанием десять необходимо передавать в каждый тактовый момент пятизначное десятичное число. Число 20739 означает, например, что по первому каналу передается сигнал 2, по второму 0, по третьему 7 и т. д. Это

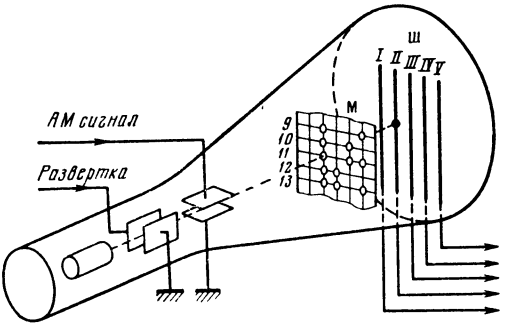


Рис. 56

число может быть далее закодировано как угодно, совершенно независимо от того, каков был код канала. Таким образом, линейный сигнал не есть просто сумма или смесь канальных сигналов; линейный сигнал представляет собой отображение определенной комбинации канальных сигналов, причем выбор способа отображения зависит от нашего усмотрения. Такой способ разделения сигналов мы и называем *комбинационным* разделением.

Что касается самого разделения, осуществляемого на приемном конце, то для этой цели может применяться любая система, пригодная для сигналов КИМ. Можно приспособить для разделения кодирующую трубку, как показано схематически на рис. 56. На этом рисунке для примера представлено устройство, способное разделить пять каналов, каждый из которых работает двоичным кодом, в предположении, что линейный сигнал представляет собой амплитудно-модулированные импульсы, закодированные по основанию 32. Электронный луч отклоняется линейным сигналом по вертикали, а пилообразным развертывающим напряжением с тактовой частотой по горизонтали. В результате этого луч пробегает по одной из 32 строк маски М, на которой имеются отверстия в точках, соответствующих наличию сигнала в данном канале. Пусть, например, сигналы имеются в данный момент в каналах № 2, 4 и 5 и отсутствуют в каналах № 1 и 3. Обозначая единицей наличие и нулем отсутствие сигналов, можем записать

состояние всех каналов двоичным числом 01011. В десятичной системе это соответствует числу 11. Линейный сигнал будет представлять собой импульс высотой 11 единиц. В приемном устройстве под действием этого импульса луч отклонится по вертикали и попадет на одиннадцатую строку маски. В этой строке отверстия пробиты на позициях 2, 4 и 5. Луч, пробегая по строке, проходит через отверстия и замыкает цепи соответствующих каналов через расположенные за маской вертикальные шины каналов Ш.

Широко известным примером применения принципа комбинационного разделения является система ДЧТ-двухканального частотного телеграфирования. Идея этой системы, предложенная А. Н. Щукиным еще в 1933 г. [16] и разработанная технически И. Ф. Агаповым [1], состоит в том, что работа двух телеграфных каналов (двоичных) отображается линейным сигналов с применением четверичного кода и частотной модуляции, т. е., проще говоря, для передачи применяются четыре различные частоты ($m=2$, $n=2$, $N=4$). В первоначальном варианте Щукина передача велась только тремя частотами, вместо четвертой давалась пауза («пассивная» пауза). Во внедренной же системе применяется «активная» пауза, т. е. передатчик все время излучает неизменную мощность, что имеет определенные технические преимущества.

§ 53. Общая теория линейного разделения

Пусть канальные сигналы записаны в виде

$$f_k = c_k \psi_k, \tag{1}$$

где $\psi_k(t)$ — переносчик, а c_k — коэффициент, который может представлять собой медленную функцию времени и выражать сообщение. Тогда произведение $c_k \psi_k$ представляет результат модуляции переносчика ψ_k сообщением c_k.

Линейный сигнал пусть представлен суммой канальных сигналов

$$f = \sum_{k=1}^{n} f_k = \sum_{k=1}^{n} c_k \psi_k. \tag{2}$$

Положим, что на приемном конце имеется разделительное устройство, состоящее из n избирательных систем. Каждая из этих систем обладает характеристикой Γ_k, имеющей смысл функционального оператора, связывающего входную и выходную функции избирательной системы. Мы будем называть Γ оператором разделения.

В этих обозначениях операция линейного разделения может быть выражена следующим соотношением, впервые установленным Д. В. Агеевым,

$$\Gamma_m f = \Gamma_m \sum c_k \psi_k = \sum c_k \Gamma_m \psi_k = c_m. \tag{3}$$

Остальные же члены должны выпасть. Иначе говоря, операторы разделения Γ_k по отношению к функциям ψ_k должны обладать следующими свойствами:

$$\Gamma_m \psi_k = \begin{matrix} 1\ [m = h], \\ 0\ [m \neq h]. \end{matrix} \tag{4}$$

Таким образом, каждая из избирательных систем должна реагировать на «свой» сигнал и не реагировать на все остальные. В этом и состоит сущность разделения.

Первый вопрос, который перед нами возникает, — это вопрос о том, какими свойствами должны обладать функции ψ_k для того, чтобы они могли быть разделены вышеописанной линейной операцией. Оказывается, что необходимым и достаточным условием разделимости функций ψ_k является их линейная независимость. Условие линейной независимости состоит, как известно, в том, что тождество

$$\sum c_k \psi_k \equiv 0 \tag{5}$$

удовлетворяется только при равенстве нулю всех коэффициентов c_k. Если же можно подобрать такие не равные нулю значения c_k, что тождество (5) удовлетворяется, то функции ψ_k линейно зависимы. Мы воспользуемся этим обстоятельством для доказательства от противного достаточности линейной независимости ψ_k для их разделимости.

Итак, положим, что ψ_k зависимы и что, следовательно, равенство (5) удовлетворяется хотя бы при одном $c_k = c_m \neq 0$. Применим к обеим частям (5) оператор Γ_m. Это даст

$$\Gamma_m \sum c_k \psi_k = \sum c_k \Gamma_m \psi_k = c_m = 0.$$

Но это невозможно, так как c_m нулю не равно. Мы должны, стало быть, предположить, что при $c_k \neq 0$

$$\sum c_k \psi_k \equiv 0,$$

а это и есть условие линейной независимости ψ_k.

Рассмотрим теперь более специальный случай, когда канальные сигналы образуют ортогональную систему функций φ_k, что является достаточным, но не необходимым условием для их разделимости. Условие ортогональности записывается в виде

$$\int_a^b \varphi_i(x)\, \varphi_k(x)\, dx = 0 \quad [l \neq k].$$

Пусть, кроме того, функции φ_k нормированы так, что

$$\int_a^b f_k^2(x)\, dx = 1.$$

Линейный сигнал пусть представляется суммой

$$f = \sum c_k \varphi_k. \tag{6}$$

При таких обстоятельствах разделение сводится к нахождению коэффициентов разложения f по ортогональным нормированным функциям φ_k. Умножая (6) на φ_m и интегрируя, получаем

$$\int_a^b f\varphi_m dx = \int_a^b \varphi_m \sum c_k \varphi_k dx = c_m \int_a^b \varphi_m^2 dx = c_m,$$

и, следовательно, оператор разделения Γ_m принимает в рассматриваемом случае вид

$$\Gamma_m f = \int_a^b f\varphi_m dx. \tag{7}$$

Результатом этой операции является выделение величины c_m, т. е. сообщения соответствующего канала.

Покажем эту операцию на примере разделения по фазе. В этом случае мы имеем только два переносчика, которые с учетом условий нормировки запишутся в виде

$$\varphi_1(t) = \sqrt{\frac{2}{T}} \sin \omega t,$$

$$\varphi_2(t) = \sqrt{\frac{2}{T}} \cos \omega t.$$

Линейный сигнал представляется суммой

$$f = a\varphi_1 + b\varphi_2.$$

Разделение происходит следующим образом:

$$\Gamma_1 f = \int_{-T/2}^{T/2} f\varphi_1 dt = \frac{2}{T} \int_{-T/2}^{T/2} (a \sin \omega t + b \cos \omega t) \sin \omega t dt = a,$$

$$\Gamma_2 f = \int_{-T/2}^{T/2} f\varphi_2 dt = \frac{2}{T} \int_{-T/2}^{T/2} (a \sin \omega t + b \cos \omega t) \cos \omega t dt = b.$$

При разделении по частоте и во времени мы также имеем дело с ортогональными функциями. При частотном разделении ортогональность сигналов непосредственно вытекает из ортогональности тригонометрических функций. При временно́м же разделении, например в импульсной связи, ортогональность сигналов следует из того, что функция, выражающая сигнал данного канала, тождественно равна нулю на протяжении интервала, отведенного для передачи сигналов других каналов. Что же касается разделения по форме, то здесь мы имеем как раз случай

линейно независимых, но не ортогональных сигналов, а поэтому операция вида (7) для разделения сигналов, различающихся по форме, непосредственно не применима. Однако известно, что система линейно независимых функций может быть ортогонализована, т. е. от данной системы можно путем линейного преобразования перейти (и притом не единственным образом) к другой системе, которая будет уже ортогональной. С этой точки зрения можно рассматривать процесс разделения по форме как операцию, включающую в себя ортогонализацию канальных сигналов, о чем подробнее говорится в § 55.

Мы рассматривали линейный сигнал вида

$$f(t) = \sum c_k \varphi_k(t)$$

и свели разделение к нахождению коэффициентов этой суммы по формуле

$$c_k = \int_a^b f(t)\,\varphi_k(t)\,dt. \tag{8}$$

Соотношения можно представить еще в несколько более общем виде, выразив линейный сигнал не суммой, а интегралом

$$f(t) = \int_{\lambda_1}^{\lambda_2} k(t,\ \lambda)\,S(\lambda)\,d\lambda, \tag{9}$$

где

$$S(\lambda) = \int_{-\infty}^{\infty} k^{-1}(t,\ \lambda)\,f(t)\,dt. \tag{10}$$

Формулы (9) и (10) образуют пару сопряженных преобразований. Смысл формулы (9) состоит в том, что линейный сигнал выражен непрерывной суммой сигналов, каждый из которых характеризуется сообщением $S(\lambda)$, модулирующим переносчик $k(t, \lambda)$. Различие канальных сигналов в данном случае определяется различием в значении параметра λ. Это различие и есть признак, по которому производится разделение. Операция разделения выражается формулой (10); оператор разделения есть

$$\Gamma f(t) = \int_{-\infty}^{\infty} k^{-1}(t,\ \lambda)\,f(t)\,dt. \tag{11}$$

В результате операции разделения находится несущее сведения сообщение $S(\lambda)$. С математической точки зрения $S(\lambda)$ есть обобщенный спектр разложения $f(t)$ по функциям $k(t, \lambda)$, и формула (10) представляет собой решение интегрального уравнения (9) относительно искомого спектра $S(\lambda)$. Сигналы разделяются, если обобщенные спектры $S(\lambda)$, принадлежащие двум различным

сообщениям, не перекрываются, т. е. если каждый из спектров занимает ограниченную полосу значений λ и границы каждой полосы находятся вне другой полосы. Функция $k(t, \lambda)$ носит название ядра интегрального уравнения. Для того чтобы записать решение интегрального уравнения (9) в конечной форме (10), нужно знать обращенное ядро $k^{-1}(t, \lambda)$, если оно вообще существует.

Поясним все эти соотношения на примере частотного и временно́го разделений. В случае частотного разделения

$$k(t, \lambda) = \frac{1}{2\pi} e^{j\omega t}, \quad k^{-1}(t, \lambda) = e^{-j\omega t},$$

и формулы (9) и (10) превращаются в пару трансформаций Фурье. Роль параметра разделения λ играет частота ω. Оператор разделения (11)

$$\Gamma_m f(t) = \int_{-\infty}^{\infty} f(t) e^{-j\omega_m t} dt$$

имеет физический смысл гетеродинирования линейного сигнала $f(t)$ с последующим усреднением. Это есть не что иное, как частотный анализ по методу вспомогательной частоты, или синхронное детектирование.

В случае временно́го разделения

$$k(t, \lambda) = \sigma_1(t - t_m), \quad h^{-1}(t, \lambda) = \sigma_1(t_m - t),$$

где σ_1 — единичный импульс (функция Дирака). Роль параметра разделения играет момент времени t_m. Оператор разделения имеет вид

$$\Gamma_m f(t) = \int_{-\infty}^{\infty} \sigma_1(t_m - t) f(t)\, dt.$$

Физический смысл этого выражения состоит в том, что некоторый ключ на короткое время замыкает цепь в момент $t = t_m$ и выделяет, таким образом, значение $f(t_m)$, которое и является канальным сигналом.

Итак, частотное и временно́е разделения являются частными случаями общего механизма линейного разделения, представляемого формулами (9) и (10).

Остается показать, что возможны иные способы разделения, кроме уже известных и применяемых в технике. Возьмем в качестве примера пару преобразований Меллина

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x^{-j\lambda} S(\lambda)\, d\lambda,$$

$$S(\lambda) = \int_{0}^{\infty} x^{-1+j\lambda} f(x)\, dx.$$

Мы имеем

$$x^{j\lambda} = e^{j\lambda \ln x} = \cos(\lambda \ln x) + j \sin(\lambda \ln x).$$

Введем безразмерное время

$$x = t/t_0$$

и пусть линейный сигнал представлен суммой

$$f(t) = \sum a_k \cos\left(\lambda_k \ln \frac{t}{t_0}\right).$$

Пусть далее оператор разделения действует на интервале ($1 < x < e^{2\pi}$)

$$\Gamma_m f(t) = \int_1^{e^{2\pi}} x^{-1+j\lambda_m} f(x)\, dx.$$

Если $\lambda_m = m\lambda_0$ и $\lambda_k = k\lambda_0$, то интеграл в правой части равен

$$\int_1^{e^{2\pi}} x^{-1+j\lambda_m} f(x)\, dx =$$

$$= \int_1^{e^{2\pi}} [\cos(\lambda_m \ln x) + j \sin(\lambda_m \ln x)] [\sum a_k \cos(\lambda_k \ln x)] \frac{dx}{x} =$$

$$= \int_0^{2\pi} (\cos \lambda_m y + j \sin \lambda_m y)(\sum a_k \cos \lambda_k y)\, dy = \frac{1}{\pi} a_m.$$

Итак, мы имеем здесь пример разделимых сигналов вида

$$f_k = a_k \cos\left(\lambda_k \ln \frac{t}{t_0}\right).$$

Процесс разделения состоит в умножении на аналогичное колебание

$$\frac{1}{t} \cos\left(\lambda_m \ln \frac{t}{t_0}\right)$$

с последующим усреднением произведения.

Этот пример не должен, разумеется, рассматриваться как техническое предложение; он служит лишь для иллюстрации общих принципов. Но несомненно стоит предпринять широкий обзор и оценку возможных новых видов разделимых сигналов; вовсе не исключено, что будут найдены новые технически целесообразные возможности построения систем многоканальной связи.

§ 54. Геометрическое представление разделения

Общие положения теории линейного разделения допускают геометрическое истолкование. Эта возможность интересна не только с точки зрения наглядного изложения теории: самая теория может строиться и развиваться как геометрическая теория.

Рассмотрим для начала фазовое разделение. Линейный сигнал представляется в этом случае суммой двух канальных сигналов:

$$f = f_1 + f_2 = a \sin \omega t + b \cos \omega t.$$

Эти сигналы могут быть представлены двумя векторами на обычной векторной диаграмме теории переменных токов; линейный сигнал представляется векторной суммой канальных сигналов. Из этого простейшего геометрического образа сразу вытекает представление о том, что операция разделения есть не что иное, как операция проектирования вектора линейного сигнала на ортогональные оси, каждая из которых принадлежит соответствующему каналу.

Такое представление естественно обобщается на случай произвольных функций времени, выражающих линейно разделимые сигналы. Всякая функция времени, заданная на промежутке $t_1 < t < t_2$, может быть представлена вектором в бесконечномерном пространстве — пространстве сигналов. Составляющие этого вектора суть мгновенные значения функции. Составляющие суммарного вектора, т. е. вектора линейного сигнала, будут представляться суммами мгновенных значений канальных сигналов в соответствующие моменты. Следовательно, линейный сигнал может быть представлен в пространстве сигналов как векторная сумма канальных сигналов

$$\boldsymbol{f} = \Sigma \boldsymbol{f}_k. \tag{1}$$

Оператор проектирования должен дать

$$\boldsymbol{\Gamma}_m \boldsymbol{f} = |f_m|. \tag{2}$$

Такой результат получится, если $\boldsymbol{\Gamma}_m$ есть орт (единичный вектор) соответствующей оси; операция (2), состоящая в скалярном умножении вектора на орт, есть не что иное, как операция проектирования вектора на соответствующую ось. Скалярное произведение двух канальных сигналов выражается соотношением

$$\boldsymbol{f}_k \boldsymbol{f}_e = \int_{t_1}^{t_2} f_k(t) f_e(t)\, dt.$$

Если канальные сигналы ортогональны, то это выражение при $k \neq l$ равно нулю; это означает, что координатные оси взаимно ортогональны и что проекция канального сигнала на ось любого другого равна нулю. При таких условиях и получается соотношение (2).

Обратимся теперь к канальным сигналам более общего вида

$$f_k = f(t, \lambda_k)$$

и учтем, что сигналы данного канала могут отвечать различным значениям λ_k; этот параметр может изменяться в известных пределах в соответствии с передаваемыми по данному каналу сообщениями, занимая, скажем, полосу значений от λ_{k1} до λ_{k2}. Поэтому сигналы данного канала будут представляться уже не одним вектором, а совокупностью векторов, отличающихся друг от друга по параметру λ. Концы векторов образуют совокупность точек, определяющих некоторое подпространство Λ_k пространства Λ. Подпространства Λ_k должны быть непересекающимися; иначе говоря, все (ненулевые) векторы f_k должны лежать в подпространстве Λ_k и ни один из них не должен попадать в подпространство Λ_l. Это есть условие разделимости сигналов, принадлежащих различным каналам. Но, с другой стороны, высказанное условие есть не что иное, как определение линейной независимости подпространства Λ_k. Таким образом, канальные сигналы разделимы, если подпространства, в которых они заключены, линейно независимы.

Теперь геометрическая картина такова: линейный сигнал размещается в пространстве Λ; составляющие его канальные сигналы локализованы в подпространствах Λ_k. Представим себе, что подпространства Λ_k взаимно ортогональны (т. е. что любой вектор подпространства Λ_k ортогонален любому вектору подпространства Λ_l). Тогда операция разделения каналов состоит в проектировании пространства Λ на соответствующие подпространства. Оператор проектирования будет при этом выражаться некоторой матрицей — так называемой матрицей проектирования.

§ 55. Разделение линейно независимых сигналов

Если канальные сигналы имеют вид

$$f = c_k \varphi_k,$$

а функции φ_k образуют ортогональную систему, то, как мы видели, оператор разделения, применяемый к линейному сигналу,

$$f = \sum f_k$$

имеет вид

$$\Gamma_m f = \int f \varphi_m dt = \overline{f\varphi_m} = c_m \qquad (1)$$

и дело сводится к разложению сигнала по ортогональным функциям φ_k, т. е. к нахождению коэффициентов этого разложения.

Однако ортогональность канальных сигналов не является необходимым условием их разделимости; как указывалось выше, необходимым условием разделимости является линейная незави-

симость канальных сигналов. Мы должны теперь рассмотреть способы разделения в этом более общем случае. Здесь перед нами открываются по меньшей мере две возможности. Первая из них состоит в том, чтобы еще на передающем конце подвергнуть систему линейно независимых канальных сигналов такой дополнительной обработке, в результате которой эта система превратится в ортогональную. После этого мы приведем задачу к ранее рассмотренной и можем сохранить на приемном конце способ разделения, выражаемый формулой (1).

Итак, мы рассмотрим прежде всего возможность ортогонализации канальных сигналов.

Если дана система векторов ψ_k и требуется заменить их системой ортогональных векторов φ_k, образующих то же подпространство, то построение новой системы производится по следующему правилу [1]:

$$\begin{aligned}
&\varphi_1 = \psi_1,\\
&\varphi_2 = \psi_2 - (\psi_2,\ \varphi_{1\text{н}})\,\varphi_{1\text{н}},\\
&\varphi_3 = \psi_3 - (\psi_3,\ \varphi_{1\text{н}})\,\varphi_{1\text{н}} - (\psi^3,\ \varphi_{2\text{н}})\,\varphi_{2\text{н}}\\
&\quad\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{2}$$

здесь

$$\varphi_{k\text{н}} = \varphi_k / |\varphi_k| \tag{3}$$

означает нормированный вектор, т. е. единичный вектор, направленный по φ_k.

Так как у нас под вектором понимается функция времени, а составляющими вектора являются мгновенные значения функции, то формулы (2) могут быть переписаны в виде

$$\begin{aligned}
&\varphi_1(t) = \psi_1(t),\\
&\varphi_2(t) = \psi_2(t) - \varphi_{1\text{н}}(t)\int_0^T \psi_2(t)\,\varphi_{1\text{н}}(t)\,dt,\\
&\varphi_3(t) = \psi_3(t) - \varphi_{1\text{н}}(t)\int_0^T \psi_3(t)\,\varphi_{1\text{н}}(t)\,dt -\\
&\qquad - \varphi_{2\text{н}}(t)\int_0^T \psi_3(t)\,\varphi_{2\text{н}}(t)\,dt,\\
&\quad\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{4}$$

---

[1] *В. И. Смирнов.* Курс высшей математики, т. I и III. Физматгиз, 1961.

или, короче,

$$\begin{aligned}
&\varphi_1 = \psi_1,\\
&\varphi_2 = \psi_2 - \varphi_{1\text{н}}\overline{\psi_2\varphi_{1\text{н}}},\\
&\varphi_3 = \psi_3 - \varphi_{1\text{н}}\overline{\psi_3\varphi_{1\text{н}}} - \varphi_{2\text{н}}\overline{\psi_3\varphi_{2\text{н}}}\\
&\cdots\cdots\cdots\cdots
\end{aligned} \tag{5}$$

Операция ортогонализации может выполняться соответствующей электрической схемой. На рис. 57 представлена схема для ортогонализации двух линейно независимых сигналов. Схема выполняет операции, соответствующие первым двум строкам (5). На схеме означают: Н — нормирующее устройство, выполняющее

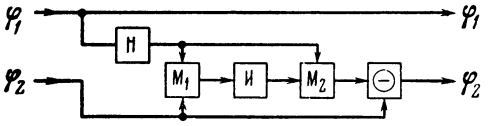


Рис. 57

операцию (3); M_1 и M_2 — модуляторы, осуществляющие перемножение двух функций, подаваемых на их входы, И — интегратор; ⊖ — вычитающее устройство.

В качестве примера рассмотрим ортогонализацию двух сигналов вида

$$f_1 = a\sin\omega t, \quad f_2 = b\sin(\omega t + \vartheta) \quad (0 < t < T),$$

где ϑ — произвольный угол $(0 < \vartheta < \pi/2)$.

Сигналы f_1 и f_2 линейно независимы, но не ортогональны. В результате обработки сигналов схемой рис. 59 получится следующее:

$$\psi_1 = \sin\omega t, \quad \psi_2 = \sin(\omega t + \vartheta),$$

$$\varphi_1 = \psi_1, \quad |\varphi_1| = \left(\int_0^T \sin^2\omega t\,dt\right)^{1/2} = \sqrt{T/2},$$

$$\varphi_{1\text{н}} = \frac{\varphi_1}{|\varphi_1|} = \sqrt{\frac{2}{T}}\sin\omega t,$$

$$\overline{\psi_2\varphi_{1\text{н}}} = \sqrt{\frac{2}{T}}\int_0^T \sin\omega t\sin(\omega t + \vartheta)\,dt = \sqrt{\frac{T}{2}}\cos\vartheta,$$

$$\varphi_2 = \sin(\omega t + \vartheta) - \sin\omega t\cos\vartheta = \sin\vartheta\cos\omega t.$$

Итак, мы получили два ортогональных сигнала

$$f_1 = a\varphi_1 = a\sin\omega t, \quad f_2 = b\varphi_2 = b\sin\vartheta\cos\omega t.$$

Легко представить себе геометрический смысл ортогонализации: мы выбираем систему ортогональных векторов φ_k, совмещаем один из них с одним из имеющихся независимых векторов ψ_k,

а остальные векторы ψ_k проектируем на соответствующие векторы φ_k. Для рассмотренного примера геометрическое представление изображено на рис. 58.

Перейдем к рассмотрению второй возможности разделения линейно независимых сигналов. Возможность эта состоит в разложении линейного сигнала

$$f = \sum f_k = \sum c_k \psi_k$$

на приемном конце по линейно независимым функциям ψ_k. Эта задача сложнее, чем разложение по ортогональным функциям, но всегда разрешима.

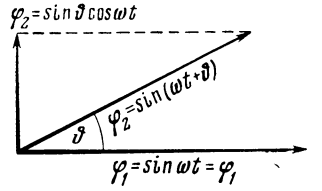


Рис. 58

Положим, что разделение осуществляется физически путем умножения линейного сигнала на некоторую функцию с последующим интегрированием этого произведения, т. е. аналогично тому, как происходит разделение ортогональных сигналов [1]. Тогда оператор разделения можно записать в виде

$$\Gamma_m f = \overline{f\eta_m}, \tag{6}$$

а функция η должна удовлетворять условиями

$$\overline{\eta_m \psi_k} = \begin{matrix} 1\,[k = m], \\ 0\,[k \neq m], \end{matrix} \tag{7}$$

т. е. функции η должны быть ортогональны канальным сигналам ψ. Функции η выражаются через ψ линейно посредством равенств

$$\begin{aligned} \eta_1 &= a_{11}\psi_1 + a_{12}\psi_2 + a_{13}\psi_3 + \ldots + a_{1n}\psi_n, \\ \eta_2 &= a_{21}\psi_1 + a_{22}\psi_2 + a_{23}\psi_3 + \ldots + a_{2n}\psi_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \eta_n &= a_{n1}\psi_1 + a_{n2}\psi_2 + a_{n3}\psi_3 + \ldots + a_{nn}\psi_n. \end{aligned} \tag{8}$$

Искомыми величинами являются n^2 коэффициентов a_{ik}. Для их определения нам требуется n^2 уравнений. Эти уравнения мы со-

[1] Оператор разделения в большинстве случаев будет иметь интегральный характер. Это положение следует из того, что для различения сигналов по тому или иному признаку следует наблюдать за ними в течение некоторого времени, достаточного для того, чтобы различие проявилось, т. е. чтобы характер сигнала определился. Степень общности этого положения еще неясна; в качестве исключения из правила можно привести разделение по уровню, которое относится, впрочем, к области нелинейного разделения.

ставим исходя из условий разделения. Для этого применим к (8) операцию (7) и получим систему уравнений

$$\begin{aligned}
&\overline{\eta_1\psi_1} = a_{11}\overline{\psi_1^2} + a_{12}\overline{\psi_1\psi_2} + a_{13}\overline{\psi_1\psi_3} + \ldots + a_{1n}\overline{\psi_1\psi_n} = 1,\\
&\overline{\eta_1\psi_2} = a_{11}\overline{\psi_2\psi_1} + a_{12}\overline{\psi_2^2} + a_{13}\overline{\psi_2\psi_3} + \ldots + a_{1n}\overline{\psi_2\psi_n} = 0,\\
&\quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\overline{\eta_1\psi_n} = a_{11}\overline{\psi_n\psi_1} + a_{12}\overline{\psi_n\psi_2} + a_{13}\overline{\psi_n\psi_3} + \ldots + a_{1n}\overline{\psi_n^2} = 0.
\end{aligned} \tag{9}$$

Из этой неоднородной системы находятся n коэффициентов a_{1k}. Для нахождения коэффициентов a_{2k} система уравнений составляется путем умножения η_2 на ψ_k и т. д. Всего мы располагаем n системами по n уравнений в каждой, а всего n^2 уравнениями, что и требуется. Канальные сигналы ψ_k известны; известны, следовательно, и величины $\overline{\psi_i\psi_k}$, являющиеся коэффициентами уравнений относительно искомых величин a_{ik}. Решение уравнений вида (9) на основании теоремы Крамера можно записать в виде

$$a_{ik} = \Delta_k^{(i)}/\Delta^{(i)}, \tag{10}$$

где $\Delta^{(i)}$ — определитель из величины $\overline{\psi_i\psi_k}$ для i-й системы уравнений, а $\Delta_k^{(i)}$ — тот же определитель, в котором k-й столбец заменен свободными членами. Формула (10) дает полное решение задачи, которое, конечно, оказывается довольно громоздким в случае большого числа каналов. Мы доведем решение до конца для случая двух каналов. В этом случае имеем

$$a_{11}\overline{\psi_1^2} + a_{12}\overline{\psi_1\psi_2} = 1, \quad a_{11}\overline{\psi_2\psi_1} + a_{12}\overline{\psi_2^2} = 0, \tag{11}$$

$$a_{21}\overline{\psi_1^2} + a_{22}\overline{\psi_1\psi_2} = 0, \quad a_{21}\overline{\psi_2\psi_1} + a_{22}\overline{\psi_2^2} = 1. \tag{12}$$

Для системы (11) мы имеем

$$\Delta^{(1)} = \begin{vmatrix} \overline{\psi_1^2} & \overline{\psi_1\psi_2} \\ \overline{\psi_2\psi_1} & \overline{\psi_2^2} \end{vmatrix} = \overline{\psi_1^2}\,\overline{\psi_2^2} - \overline{(\psi_1\psi_2)}^2,$$

$$\Delta_1^{(1)} = \begin{vmatrix} 1 & \overline{\psi_1\psi_2} \\ 0 & \overline{\psi_2^2} \end{vmatrix} = \overline{\psi_2^2},$$

$$\Delta_2^{(1)} = \begin{vmatrix} \overline{\psi_1^2} & 1 \\ \overline{\psi_2\psi_1} & 0 \end{vmatrix} = -\overline{\psi_1\psi_2},$$

откуда

$$a_{11} = \frac{\Delta_1^{(1)}}{\Delta^{(1)}} = \frac{\overline{\psi_2^2}}{\overline{\psi_1^2}\,\overline{\psi_2^2} - \overline{(\psi_1\psi_2)}^2},$$

$$a_{12} = \frac{\Delta_2^{(1)}}{\Delta^{(1)}} = -\frac{\overline{\psi_1\psi_2}}{\overline{\psi_1^2}\,\overline{\psi_2^2} - \overline{(\psi_1\psi_2)}^2}.$$

Действуя аналогично, получаем из системы (12)

$$a_{21} = \frac{\Delta_1^{(2)}}{\Delta^{(2)}} = a_{1,\,2},$$

$$a_{22} = \frac{\Delta_2^{(2)}}{\Delta^{(2)}} = \frac{\bar{\psi}_1^2}{\bar{\psi}_1^2\bar{\psi}_2^2 - (\overline{\psi_1\psi_2})^2}.$$

Когда коэффициенты a_{ik} найдены, то этим определены и функции η_k (см. (8)). Тогда операция разделения выражается следующим образом:

$$\Gamma_m f = \overline{\eta_m} f = \sum c_k \overline{\eta_m \psi_k} = c_m. \tag{13}$$

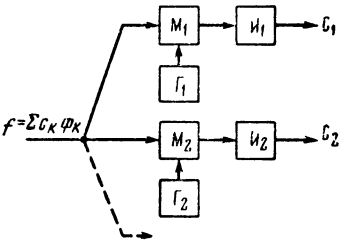


Рис. 59

Эта операция выполняется схемой рис. 59, на которой M_k — модуляторы (умножители); $И_k$ — интеграторы; Γ_k — генераторы функций η_k.

Рассмотрим пример. Пусть

$$f = f_1 + f_2 = a\psi_1 + b\psi_2 = at + bt^2 \quad (0 < t < 1).$$

Найдем $\overline{\psi_i\psi_k}$. Мы имеем

$$\bar{\psi}_1^2 = \int_0^1 t^2 dt = 1/3, \quad \overline{\psi_1\psi_2} = \int_0^1 t^3 dt = 1/4, \quad \bar{\psi}_2^2 = \int_0^1 t^4 dt = 1/5.$$

Определитель Δ равен

$$\Delta = \bar{\psi}_1^2\bar{\psi}_2^2 - (\overline{\psi_1\psi_2})^2 = 1/15 - 1/16 = 1/240.$$

Коэффициенты

$$a_{11} = 48, \; a_{12} = a_{21} = -60, \; a_{22} = 80.$$

Итак,

$$\eta_1 = 48t - 60t^2 = 48\left(t - \frac{5}{4}t^2\right),$$

$$\eta_2 = -60t + 80t^2 = 80\left(t^2 - \frac{3}{4}t\right).$$

Для проверки выделим из линейного сигнала $f = f_1 + f_2$ сигнал № 1, т. е. сделаем вычисление по формуле (13)

$$\Gamma_1 f = \overline{f\eta_1} = 48 \int_0^1 (at + bt^2)\left(t - \frac{5}{4} t^2\right) dt =$$

$$= 48\left(\frac{1}{3} a + \frac{1}{4} b - \frac{5}{16} a - \frac{1}{4} b\right) = a.$$

Требуемый сигнал, таким образом, выделен. По поводу рассмотренного примера следует еще заметить, что подобного же рода сигналы разделялись в § 50 другим способом, а именно путем дифференцирования и интегрирования. Таким образом, способ разделения, изложенный в данном параграфе, оказывается не единственным. Но дело в том, что сигналы в виде последовательности функций

$$1, \ t, \ t^2, \ \ldots$$

обладают тем специальным свойством, что каждый следующий член последовательности пропорционален интегралу от предыдущего. Конечно, возможно, используя те или иные специальные признаки выбранной для построения канальных сигналов системы функций, сконструировать специальные методы разделения, пригодные для данной системы. Способ же, изложенный в данном параграфе, является вполне общим и годится для любых функций, даже не объединяемых каким-либо родовым признаком в форме тех или иных рекуррентных формул или чего-либо подобного.

§ 56. Синтез разделяющих устройств

В предыдущих параграфах введен ряд обобщений общеизвестных понятий. Мы имели дело с обобщенным спектром $S(\lambda)$, появляющимся в качестве весовой функции при разложении сигнала по произвольным функциям $k(t, \lambda)$; отсюда возникло понятие об обобщенной полосе, как об интервале значений параметра λ. Операция разделения

$$\Gamma_m f = \int f\varphi_m dt = \overline{f\varphi}_m$$

может рассматриваться как обобщенное гетеродинирование с последующим усреднением. Об отделении сообщения от переносчика можно говорить, как об обобщенном детектировании и т. д.

Сейчас нам следует обратить внимание на то обстоятельство, что разделение по частоте возможно на основе двух различных принципов. Один принцип состоит в применении обычной частотной фильтрации, осуществляемой при помощи избирательных фильтров с требуемыми частотными характеристиками; эти системы являются системами с постоянными параметрами и описываются линейными дифференциальными уравнениями с постоянными коэффициентами. Второй принцип состоит в применении

гетеродинирования. Частотное разделение и выделение сообщения осуществляются по этому принципу посредством синхронного детектирования. Системы, выполняющие эти действия, являются системами с переменными параметрами и описываются линейными дифференциальными уравнениями с переменными коэффициентами.

В предыдущем рассмотрена в общем виде именно эта вторая возможность разделения; схема разделения рис. 59 представляет собой не что иное, как схему разделения посредством обобщенного гетеродинирования. Теперь нам остается рассмотреть вопрос об обобщенной фильтрации и, в частности, о синтезе обобщенных фильтров, т. е. систем с постоянными параметрами, способных разделить произвольного вида сигналы $k(t, \lambda)$, различающиеся между собой по параметру λ.

Свойства такого рода систем очень легко описать при помощи *обобщенных характеристик*. Обобщенной характеристикой $A(\lambda)$ мы назовем зависимость от λ отношения обобщенных спектров на выходе и на входе обобщенного фильтра. Очевидно, что для выделения сигнала, отвечающего значению $\lambda=\lambda_m$, обобщенная характеристика должна иметь вид

$$A(\lambda)=\sigma_1(\lambda-\lambda_m). \tag{1}$$

Если же обобщенный спектр сигнала лежит в полосе $\lambda_1 < \lambda < \lambda_2$, то обобщенная характеристика должна быть

$$A(\lambda)=\int_{\lambda_1}^{\lambda_2}\sigma_1(\lambda-\xi)\,d\xi=\sigma_0(\lambda-\lambda_1)-\sigma_0(\lambda-\lambda_2). \tag{2}$$

Знания обобщенной характеристики в принципе достаточно для синтеза схемы обобщенного фильтра. Но практически было бы удобнее задаваться не обобщенной характеристикой фильтра, а какой-либо из употребительных характеристик, например частотной или временной, так как эти характеристики нам более привычны; к тому же имеются разработанные приемы синтеза схем, основанные на применении именно этих характеристик. Таким образом, наша ближайшая задача состоит в нахождении частотной или временной характеристики обобщенного фильтра по заданной обобщенной характеристике.

Пусть назначение обобщенного фильтра состоит лишь в выделении сигнала $k(t, \lambda)$ (без детектирования); пусть, далее, отклик фильтра на сигнал $k(t, \lambda)$ есть $l(t, \lambda)$. Тогда требования к фильтру могут быть записаны в форме следующего условия [29]:

$$l(t, \lambda)=\begin{cases} k(t, \lambda) & \text{при } \lambda \text{ в пределе полосы } \Lambda_m, \\ 0 & \text{при } \lambda \text{ вне полосы } \Lambda_m. \end{cases} \tag{3}$$

Если вообще $y(t)$ представляет отклик фильтра на воздействие $x(t)$, то связь между x и y может быть выражена интегра-

лом Дюамеля

$$y(t)=\int_{-\infty}^{\infty} g(t-\tau)\,x(\tau)\,d\tau, \tag{4}$$

где $q(t)$ — импульсная реакция (временна́я характеристика), т. е. отклик фильтра на единичный импульс $\sigma_1(t)$. Если подставить $k(t, \lambda)$ вместо $x(t)$, то получится

$$l(t, \lambda)=\int_{-\infty}^{\infty} g(t-\tau)\,k(t, \lambda)\,d\tau, \tag{5}$$

что представляет собой разложение l по k со спектром g. Обратное преобразование, т. е. разложение g по k^{-1} со спектром l, имеет вид

$$g(t-\tau)=\int_{-\infty}^{\infty} k^{-1}(\lambda, \tau)\,l(\lambda, t)\,d\lambda \tag{6}$$

или, учитывая (3),

$$g(t-\tau)=\int_{\Lambda_m} k^{-1}(\lambda, \tau)\,k(\lambda, t)\,d\lambda. \tag{7}$$

Если требуется синтезировать фильтр, выделяющий дискретный сигнал $k(t, \lambda_n)$, то в качестве полосы Λ_m следует взять узкий интервал $\delta\lambda$, включающий в себя λ_m. Это дает

$$g(t-\tau)=\int_{\lambda_m-\delta\lambda/2}^{\lambda_m+\delta\lambda/2} k^{-1}(\lambda, \tau)\,k(\lambda, t)\,d\tau=$$

$$=\delta\lambda k^{-1}(\lambda_m, \tau)\,k(\lambda_m, t). \tag{8}$$

По этим формулам может быть вычислена временна́я характеристика (импульсная реакция) искомого фильтра [1].

---

[1] Предполагается, что мы исходим из определенной пары интегральных преобразований и что, следовательно, k и k^{-1} известны. Вообще же говоря, может быть указан следующий прием разыскания обращенного ядра. Пусть дано преобразование вида

$$f(t)=\int_{-\infty}^{\infty} k(\lambda, t)\,S(\lambda)\,d\lambda. \tag{a}$$

Запишем обратное преобразование в виде

$$S(\lambda)=\int_{-\infty}^{\infty} k^{-1}(t, \lambda)\,f(t)\,dt. \tag{b}$$

Положим

$$f(t)=\sigma_1(t-\tau). \tag{c}$$

Если предпочтительно синтезировать схему, основываясь на частотной характеристике, то можно воспользоваться тем известным обстоятельством, что частотная и временна́я характеристики связаны между собой парой преобразований Фурье

$$A(\omega) = \int_{-\infty}^{\infty} g(t)\, e^{-j\omega t} dt,$$

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A(\omega)\, e^{j\omega t} d\omega.$$

Само собой разумеется, что синтез фильтра не всегда возможен. Покажем это на примерах. В качестве первого примера возьмем частотное разделение. В этом случае

$$\lambda = j\omega, \quad k(t, \lambda) = e^{j\omega t}, \quad k^{-1}(\lambda, t) = \frac{1}{2\pi j} e^{-j\omega t}.$$

Пусть требуется синтезировать фильтр, пропускающий сигнал, лежащий в полосе $\omega_1 < \omega < \omega_2$, т. е. фильтр, обладающий характеристикой вида (2). Введя обозначения

$$\omega_0 = \frac{1}{2}(\omega_1 + \omega_2), \quad \Delta\omega = \frac{1}{2}(\omega_2 - \omega_1),$$

по формуле (7) получим

$$g(t-\tau) = \frac{1}{2\pi_j} \int_{j(\omega_0-\Delta\omega)}^{j(\omega_0+\Delta\omega)} e^{j\omega(t-\tau)} dj\omega =$$

$$= \frac{\Delta\omega}{\pi} \cdot \frac{\sin \Delta\omega (t-\tau)}{\Delta\omega (t-\tau)} e^{j\omega_0 t}. \qquad (9)$$

Легко узнать в полученном выражении импульсную реакцию идеального полосового фильтра.

Тогда

$$S(\lambda) = \int_{-\infty}^{\infty} k^{-1}(t, \lambda)\, \sigma_1(t-\tau)\, dt = k^{-1}(\tau, \lambda). \qquad (d)$$

Подставляя (c) и (d) в (a), получим интегральное уравнение, связывающее k и k^{-1},

$$\int_{-\infty}^{\infty} k(\lambda, t)\, k^{-1}(\tau, \lambda)\, d\lambda = \sigma_1(t-\tau).$$

Пусть теперь в основу разделения положено преобразование Лапласа

$$h(\lambda)=\int_0^{\infty} f(t)\,e^{-\lambda t}dt, \quad f(t)=\frac{1}{2\pi j}\int_{\varepsilon-j\infty}^{\varepsilon+j\infty} h(\lambda)\,e^{\lambda t}d\lambda$$

и пусть λ есть вещественное (положительное) число. Другими словами, семейство сигналов выражается функциями

$$k(t,\ \lambda)=e^{\lambda t} \quad [t>0,\ \lambda>0]. \tag{10}$$

Попытаемся построить фильтр, выделяющий один дискретный сигнал вида (10). Подставляя в (8)

$$k(t,\ \lambda_m)=e^{\lambda_m t}, \quad k^{-1}(\lambda_m,\ \tau)=\frac{1}{2\pi}e^{-\lambda_m \tau},$$

получим

$$g(t-\tau)=\frac{\delta\lambda}{2\pi}e^{\lambda_m (t-\tau)}.$$

Из этого выражения видно, что требуемый фильтр в виде пассивной системы с постоянными параметрами физически неосуществим. Однако посредством обобщенного гетеродинирования или применения специализированных приемов экспоненциальные сигналы могут быть разделены.

ДОБАВЛЕНИЕ

1. По поводу теоремы Котельникова (к § 9)

Теорема утверждает, что функция сообщения, имеющая конечную ширину спектра F, может быть передана на конечном интервале T при помощи сигнала, отображающего M дискретных чисел, где

$$m=T/\Delta t=2n=4\pi F/\omega_1.$$

Мы видели, что этими числами могут быть либо коэффициенты Фурье в разложении

$$f(t)=\sum_{k=1}^{n} c_k e^{jk\omega_1 t}, \tag{1}$$

либо коэффициенты разложения

$$f(t)=\sum_{k=1}^{m} f(k\Delta t)\,\frac{\sin\omega_c(t-k\Delta t)}{\omega_c(t-k\Delta t)}. \tag{2}$$

В (1) элементом разложения является синусоида, из которой взят ограниченный во времени отрезок. В (2) элементом разложения является функция, парная, по Фурье, синусоиде — единичный импульс, из которого взят отрезок, ограниченный по частоте. В (1) суммирование производится по элементам, равноотстоящим по частоте, а в (2) — по элементам, равноотстоящим во времени. В (1) элементы ограничены во времени, а суммирование по частоте, а в (2), наоборот, элементы ограничены по частоте, а суммирование во времени. Таким образом, мы имеем здесь пример двойственности частотно-временны́х представлений, двойственности, в основе которой лежат соотношения, определяемые парой сопряженных преобразований Фурье.

Если бы мы ограничили свойства функции сообщения не во времени и по частоте, а каким-либо другим способом, то получили бы возможность представить функцию другими конечными разложениями.

Но с точки зрения связи нас больше интересует другой вопрос, а именно вопрос о том, как осуществить технически передачу тех или иных m чисел за время T. С этой точки зрения то или иное разложение функции $f(t)$ отображает определенный способ передачи.

Что касается разложения (2), то соответствующий этому разложению способ передачи, состоящий в посылке коротких импульсов, величина которых пропорциональна мгновенным значениям функции, взятым через интервалы Δt, — способ этот подробно рассмотрен в § 9. Остается рассмотреть вопрос о том, как могла бы быть осуществлена передача, отвечающая разложению (1). В этом случае подлежащие передаче m чисел представляют собой амплитуды и фазы n гармоник разложения функции $\dot{f}(t)$ в ряд Фурье с периодом T. Положим для простоты, что фазовые соотношения не играют роли, как в телефонии, и что на приемном конце требуется лишь получить функцию с таким же амплитудным спектром, как $f(t)$. Тогда задача состоит в передаче за время T $n=m/2$ чисел, представляющих амплитуды гармоник $\dot{f}(t)$. Систему связи можно себе представить в следующем виде. На передающем конце имеется анализатор, производящий анализ отрезка функции $\dot{f}(t)$ длительностью T. Полученные значения амплитуд передаются по каналу связи со скоростью n/T значений в секунду, т. е. со средним интервалом $\Delta t=T/n$. В это время анализируется следующий отрезок длительностью T. На приемном конце значения амплитуд запоминаются, и по окончании передачи всей последовательности значений амплитуд, относящейся к данному периоду, в ход пускается генератор с основной частотой $\omega_1=2\pi/T$, амплитуды гармоник которого установлены в соответствии с принятыми значениями. Тем временем запоминаются значения амплитуд, относящиеся к следующему периоду, и т. д. Таким образом, на приемном конце с запозданием на $2T$ восстанавливается функция с тем же амплитудным спектром, что и $\dot{f}(t)$. Мы не

будем здесь обсуждать описанную систему с технической точки зрения, но очевидно, что она, существенно отличаясь от применяемых в настоящее время систем, дает в принципе тот же результат.

2. Код Морзе и статистика (к § 20)

Код Морзе, как неравномерный код, должен был бы строиться с учетом статистики букв русского языка. Для построения статистически оптимального неравномерного кода следует поступить

Таблица D

	1	2	3	4	5	6	7
№	l_k	буква	p_k (%)	$(p_k l_k)_{\text{опт}}$	буква	p_k (%)	$(p_k l_k)_{\text{факт}}$
1	4	О	11,0	44,0	Е	8,7	34,8
2	6	Е	8,7	52,3	И	7,5	45,0
3	6	А	7,5	45,0	Т	6,5	39,0
4	8	И	7,5	60,0	А	7,5	60,0
5	8	Т	6,5	52,0	Н	6,5	52,0
6	8	Н	6,5	52,0	С	5,5	44,0
7	10	С	5,5	55,0	М	3,1	31,0
8	10	Р	4,8	48,0	У	2,5	25,0
9	10	В	4,6	46,0	Р	4,8	48,0
10	10	Л	4,2	42,0	Д	3,0	30,0
11	10	К	3,4	34,0	Х	1,1	11,0
12	12	М	3,1	37,2	В	4,6	55,2
13	12	Д	3,0	36,0	К	3,4	40,8
14	12	П	2,8	33,6	Г	1,6	19,2
15	12	У	2,5	30,0	Ж	0,9	10,8
16	12	Я	2,2	26,4	Ф	0,2	2,4
17	12	Ы	1,9	22,8	Л	4,2	50,4
18	12	З	1,8	21,6	Б	1,7	20,4
19	14	Ь, Ъ	1,7	23,8	О	11,0	154,0
20	14	Б	1,7	23,8	Ю	0,7	9,8
21	14	Г	1,6	22,4	Я	2,2	30,8
22	14	Ч	1,5	21,0	П	2,8	29,2
23	14	Й	1,2	16,8	З	1,8	25,2
24	14	Х	1,1	15,4	Ц	0,5	7,0
25	14	Ж	0,9	12,6	Ь, Ъ	1,7	23,8
26	14	Ю	0,7	9,8	Э	0,3	4,2
27	16	Ш	0,7	11,2	Й	1,2	19,2
28	16	Ц	0,5	8,0	Ы	1,9	30,4
29	16	Щ	0,4	6,4	Щ	0,4	6,4
30	16	Э	0,3	4,8	Ч	1,5	24,0
31	18	Ф	0,2	3,6	Ш	0,7	12,6
			100,0	917,5		100,0	995,6

так: выписать все буквы в порядке убывающей вероятности, а кодовые обозначения в порядке возрастающей длины. Вероятность (как относительная частота) букв в русском языке дана в третьем столбце таблицы [1]. Длину кодовой комбинации мы будем выражать, приняв точку за единицу длительности и включая в комбинацию также и разделительную паузу между буквами. В таком случае самая короткая комбинация — одна точка — будет иметь длину четыре единицы, одно тире — шесть единиц, две точки — шесть единиц, точка — тире — восемь единиц и т. д. Эти цифры сведены в первом столбце таблицы. В четвертом столбце даны произведения $p_k l_k$. Сумма чисел этого столбца дает среднюю длину кодовой комбинации: она составляет 9,17 точек.

Этот результат относится к оптимальному в статистическом смысле коду. Но в принятом у нас коде Морзе оптимальный принцип не выдержан. Фактическое расположение букв дано в пятом столбце таблицы, а соответствующие вероятности — в шестом столбце. Как видим, расположение букв неправильно: ряд букв явно не на месте (например, Х, Ж, Ф и др.); в особенности же неудачно расположена буква О — вместо первого места она занимает девятнадцатое.

В седьмом столбце таблицы даны произведения $p_k l_k$, полученные путем перемножения чисел первого и шестого столбцов. Эти числа соответствуют фактическому положению; суммируя их, получим фактическую среднюю длину комбинации для принятого у нас кода Морзе. Она составляет 9,96 точек. При оптимальном построении мы получили бы среднюю комбинацию короче на

$$\frac{9{,}96-9{,}17}{9{,}96}=0{,}08\,(=8\%).$$

3. Функция корреляции кода Морзе (к § 23)

Для телеграфного сигнала, кодированного по Морзе, получаются своеобразные соотношения. Дело в том, что все элементы этого кода имеют определенную кратность. Элементы кода Морзе таковы: 1) точка-пауза — длительность две точки; 2) тире-пауза — длительность четыре точки; 3) пауза между буквами — длительность две точки. Пауза между словами в дальнейших рассуждениях не учитывается, но и она имеет длительность четыре точки, т. е. укладывается в те же кратные соотношения. Из наличия кратности следует, что в сигнале по коду Морзе содержится относительно большая периодическая составляющая, период которой равен продолжительности двух точек. Для пояснения этого положения на рис. 60, *а* изображен телеграфный сигнал по коду Морзе, на рис. 60, *б* — периодическая составляющая,

[1] Эти данные представлены П. В. Праховым.

а на рис. 60, *в* — остаток, получаемый в результате вычитания *б* из *а*. В этом остатке положительный импульс приходится на каждое тире, а отрицательный — на каждый пробел между буквами (пробел между словами дал бы два отрицательных импульса).

Периодическая составляющая сигнала имеет периодическую же функцию корреляции, изменяющуюся по треугольному закону, а импульсы остатка можно считать в первом приближении расположенными случайно. Приняв это предположение, мы получим, что функция корреляции остатка имеет треугольную форму,

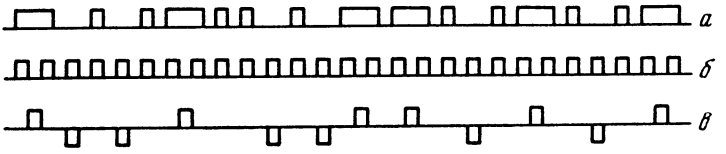


Рис. 60

как на рис. 15 для кода Бодо. Значение функции корреляции при $\tau=0$ определяется средним числом импульсов остатка. Это число можно подсчитать, зная среднее число знаков в кодовой комбинации. Пусть это число равно l_0. Точки и тире равновероятны. Тогда число тире на одну кодовую группу равно в среднем $^1/_2\, l_0$. Таково же и число точек. Длительность одной группы, выраженная через длительность t_0 одной точки, будет в среднем составлять (включая паузу между буквами)

$$\left(\frac{1}{2}\, l_0 2 + \frac{1}{2}\, l_0 4 + 2\right) t_0 = \left(3l_0 + 2\right) t_0.$$

Число тире в единицу времени, т. е. средняя частота тире, а следовательно, и средняя частота положительных импульсов остатка

$$f_+ = \frac{1/2l_0}{(3l_0+2)\, t_0} = f_0\, \frac{l_0}{3l_0+2},$$

где $f_0 = {}^1/_2 t_0$ — частота периодической составляющей сигнала. Далее, на каждую кодовую группу приходится одна пауза. Средняя частота пауз, а следовательно, и отрицательных импульсов составляет

$$f_- = \frac{1}{(3l_0+2)\, t_0} = f_0\, \frac{2}{3l_0+2}.$$

Средняя частота импульсов остатка (положительных и отрицательных вместе) равна

$$f = f_+ + f_- = f_0\, \frac{l_0+2}{3l_0+2}.$$

С учетом статистики букв в русском языке $l_0=2{,}65$. Беря эту цифру, получим

$$f/f_0=0{,}47.$$

Таким образом, если высота импульсов равна h, то наибольшее значение периодической составляющей функции корреляции будет $h^2/2$, а наибольшее значение функции корреляции остатка (при $\tau=0$) — 0,47 $h^2/2$, а всего при $\tau=0$ будем иметь 1,47 $h^2/2$. График функции корреляции для телеграфного сигнала по коду Морзе изображен на рис. 61.

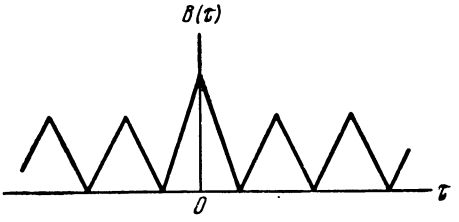


Рис. 61

Из сказанного следует, между прочим, что можно было бы декоррелировать сигнал Морзе, исключив из него периодическую составляющую рис. 60, *б*. Эту составляющую можно было бы при желании генерировать на приемном конце, по линии же достаточно передавать только остаточный сигнал рис. 60, *в*. При этом передача должна быть синхронной.

4. Отыскание экстремального распределения (к § 29)

Задача отыскания функции распределения, при которой получается наибольшее количество сведений, — вариационная задача. Она формулируется так: найти функцию $\varphi(x)$, дающую максимум интегралу

$$I'=-\int_{-\infty}^{\infty}\varphi(x)\log[\delta\varphi(x)]\,dx. \tag{1}$$

На искомую функцию $\varphi(x)$ накладываются дополнительные условия: условие нормировки

$$\int_{-\infty}^{\infty}\varphi(x)\,dx=1 \tag{2}$$

и условие неизменности среднего квадрата

$$\int_{-\infty}^{\infty}x^2\varphi(x)\,dx=\sigma^2. \tag{3}$$

Задача с дополнительными ограничивающими условиями называется задачей на условный экстремум. Если, как в нашем случае, условия заданы в интегральной форме, то задача относится к числу изопериметрических [1].

Общее правило для решения таких задач состоит в следующем: если дан функционал

$$v = \int_a^b F(x)\,dx$$

и дополнительные условия

$$\int_a^b F_i dx = l_i,$$

то нужно составить вспомогательный функционал

$$v^* = \int_a^b (F + \sum \lambda_i F_i)\,dx,$$

где λ_i — постоянные множители, и искать для этого нового функционала абсолютный экстремум обычными методами. Применяя это правило к нашему случаю, найдем

$$F = -\varphi \ln \delta\varphi, \quad F_1 = \varphi, \quad F_2 = x^2\varphi;$$

$$v^* = \int_{-\infty}^{\infty} (-\varphi \ln \delta\varphi + \lambda_1\varphi + \lambda_2 x^2\varphi)\,dx;$$

$$F^* = -\varphi \ln \delta\varphi + \lambda_1\varphi + \lambda_2 x^2\varphi.$$

Составим теперь уравнение Эйлера

$$\frac{\partial F^*}{\partial \varphi} = -\frac{d}{dx}\left(\frac{\partial F^*}{\partial \varphi'}\right) = 0.$$

Но в нашем случае дело упрощается тем, что F^* не зависит от $\varphi' = d\varphi/dx$ и, таким образом, уравнение Эйлера приводится к виду

$$\frac{\partial F^*}{\partial \varphi} = -\ln \delta\varphi - 1 + \lambda_1 + \lambda_2 x^2 = 0,$$

откуда искомая функция

$$\delta\varphi = e^{\lambda_1 - 1} e^{\lambda_2 x^2}. \tag{4}$$

Постоянные λ_1 и λ_2 определяются из условий (2) и (3). Условие нормировки (2) дает

---

[1] См., например, *Л. Э. Эльсгольц.* Вариационное исчисление. Гостехиздат, 1952 (гл. 4, § 3).

$$\int_{-\infty}^{\infty} \varphi(x)\,dx = \frac{1}{\delta} e^{\lambda_1 - 1} \int_{-\infty}^{\infty} e^{\lambda_2 x^2} dx = \frac{1}{\delta} e^{\lambda_1 - 1} \sqrt{\frac{\pi}{-\lambda_2}} = 1, \qquad (5)$$

а условие (3)

$$\int_{-\infty}^{\infty} x^2 \varphi dx = \frac{1}{\delta} e^{\lambda_1 - 1} \int_{-\infty}^{\infty} x^2 e^{\lambda_2 x^2} dx = \frac{1}{\delta} \frac{e^{\lambda_1 - 1}}{(-\lambda_2)^{3/2}} \frac{\sqrt{\pi}}{2} = \sigma^2. \qquad (6)$$

Из (5) и (6) находим

$$e^{\lambda_1 - 1} = \delta / \sqrt{2\pi\sigma}, \qquad \lambda_2 = -1/2\sigma^2.$$

Итак, окончательно

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}, \qquad (7)$$

что и представляет собой обычное выражение для симметричного нормального распределения (см. § 19).

Если ограничивающее условие (3), т. е. условие $\sigma = \text{const}$, не налагается, то результат получается иной. Мы имеем в этом случае

$$F^{**} = \varphi \ln \delta\varphi + \lambda\varphi, \quad \partial F^{**}/\partial\varphi = -\ln \delta\varphi - 1 + \lambda = 0,$$

откуда

$$\delta\varphi = e^{\lambda - 1} = \text{const},$$

т. е. оптимальным распределением в случае, когда средняя мощность не ограничивается, оказывается равномерное распределение.

Если бы мы искали распределение, дающее минимум мощности при заданном количестве сведений, т. е. взяли бы (3) за основной функционал, а (1) и (2) за дополнительные условия, то получили бы также нормальное распределение. Однако в этом случае постоянные были бы выражены уже не через σ, а через I', и мы получили бы

$$\varphi(x) = \frac{c}{\delta} e^{-\pi \frac{e^2}{\sigma^2} x^2}, \qquad (8)$$

где

$$c = e 2^{-I'}.$$

5. Распределение вероятностей для сигнала ошибки (к § 32)

Займемся вопросом о распределении вероятностей сигнала ошибки, мы имеем

$$\varepsilon = h_0 - h_{\text{п}}.$$

Стало быть, речь идет о распределении вероятностей для разности двух случайных величин. Рассмотрим разность

$$z = x - y,$$

где x и y — случайные величины с вероятностями соответственно $p(x)$ и $p(y)$. Но так как нас интересует случай коррелированных между собой величин, то должна существовать совместная вероятность $p(xy)$, не равная произведению простых вероятностей, как это было бы в случае, когда x и y независимы.

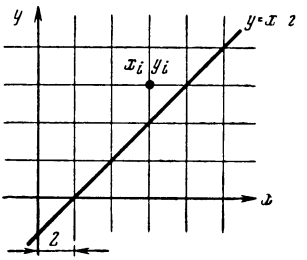


Рис. 62

Если x и y — два члена случайной последовательности, то вероятность разности равна сумме вероятностей величинам x и y принять такие значения x_i и y_j, что $x_i - y_j = z$. Таким образом,

$$p(z) = \sum_{i=1}^{n} \sum_{j=1}^{n} p(x_i y_j) = \sum_{i=1}^{n} p(x_i,\ x_i - z).$$

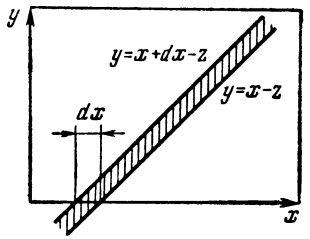


Рис. 63

Смысл этого соотношения поясняется рис. 62, на котором изображена координатная плоскость x, y. Узлы сетки имеют координаты x_i и y_j. Совместная вероятность $p(xy)$ характеризует вероятность точке попасть в узел с данными координатами. Вероятность $p(z)$ характеризует вероятность точке попасть в узлы, лежащие на прямой $y=x-z$.

Аналогично обстоит дело и в случае, когда x и y изменяются непрерывно. Только при этом связь между x и y характеризуется двумерной плотностью вероятностей $\varphi(x, y)$, и плотность вероятностей для $z=x-y$ выражается формулой

$$\varphi(z) = \int_{-\infty}^{\infty} \varphi(x,\ x - z)\, dx. \qquad (1)$$

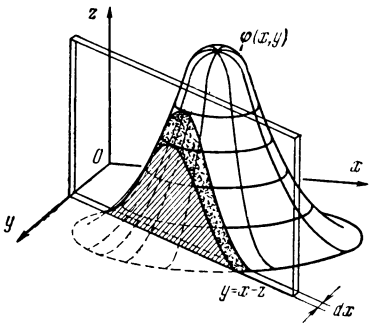


Рис. 64

Плотность $\varphi(z)$ есть вероятность точке попасть в полоску между прямыми $y=x-z$ и $y=x+dx-z$ (рис. 63).

Смысл интеграла (1) поясняется еще рис. 64; плотность вероятностей

$\varphi(z)$ равна объему, вырезаемому из тела $\varphi(xy)$ отмеченными на рисунке параллельными плоскостями.

Если совместная вероятность симметрична относительно x и y (т. е. не меняется от перестановки местами x и y) и если она убывает при увеличении разности x и y, то распределение вероятностей для z имеет максимум при $z=0$, каково бы ни было распределение вероятностей для x. Это объясняет с вероятностной точки зрения выигрыш в мощности при передаче сигнала ошибки.

6. Когерентность и корреляция (к § 37)

Когерентность определяют обычно ссылкой на закон суммирования. Именно два процесса некогерентны, если мощность их суммы равна сумме мощностей каждого из процессов в отдельности, или, говоря математическим языком, если средний квадрат суммы равен сумме средних квадратов. Понятие когерентности тесно связано также с ортогональностью функций.

С точки зрения приведенного определения гармоники периодического процесса некогерентны между собой; так, например, мощность несинусоидального тока равна сумме мощностей отдельных гармоник.

Рассмотрим среднее значение квадрата суммы двух процессов $x(t)$ и $y(t)$

$$P=\lim_{T\to\infty}\frac{1}{T}\int_0^T[x(t)+y(t)]^2\,dt.$$

Раскрывая подынтегральное выражение, получим

$$P=\lim_{T\to\infty}\frac{1}{T}\left[\int_0^T x^2(t)\,dt+\int_0^T y^2(t)\,dt+2\int_0^T x(t)\,y(t)\,dt\right].$$

Первый член дает мощность процесса x, второй — мощность процесса y. Третий член, который можно назвать взаимной мощностью,

$$P_{xy}=\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)\,y(t)\,dt$$

есть мера когерентности процессов x и y. Когда P_{xy} равно нулю, процессы некогерентны и имеет место закон энергетического суммирования. Но, как легко видеть,

$$P_{xy}=B_{xy}(0),$$

т. е. рассматриваемый член есть не что иное, как функция взаимной корреляции процессов x и y при $\tau=0$.

Это рассуждение можно распространить на более общий случай, когда один из процессов сдвинут относительно другого на произвольное время τ. Тогда

$$P=\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)\,y(t+\tau)\,dt=$$

$$=\lim_{T\to\infty}\frac{1}{T}\left[\int_0^T x^2(t)\,dt+\int_0^T y^2(t+\tau)\,dt+\right.$$

$$\left.+2\int_0^T x(t)\,y(t+\tau)\,dt\right]=P_x+P_y+B_{xy}(\tau)$$

и процессы некогерентны при условии, что функция взаимной корреляции $B_{xy}(\tau)$ равна нулю. Таким образом, некогерентность и отсутствие корреляции — одно и то же.

Нужно заметить, что при одних значениях τ процессы могут быть когерентны, а при других — нет. Так, например, если

$$x(t)=\sin\omega t,\quad y(t+\tau)=\sin\omega(t+\tau),$$

то эти процессы полностью когерентны при $\tau=nT_0$ и полностью некогерентны при $\tau=(n+{}^1/_2)T_0$.

7. Оптимальная фильтрация (к § 38)

Теория стационарных случайных процессов позволяет поставить и решить следующую общую задачу: найти частотную характеристику фильтра, назначение которого состоит в наилучшем отделении сигнала от помехи. Как помеха, так и сигнал представляются случайными процессами; предполагается, что известны их статистические спектры. Под наилучшим отделением понимается, что средний квадрат уклонения величины на выходе фильтра от сигнала получается наименьший.

Пусть $f(t)$ означает сигнал, а $\xi(t)$ — помеху. Обозначим через $G_f(\omega)$ и $G_\xi(\omega)$ соответственно спектры сигнала и помехи. Тогда можно показать [17], что частотная характеристика оптимального фильтра должна выражаться соотношением

$$A(\omega)=\frac{G_f(\omega)}{G_f(\omega)+G_\xi(\omega)}. \qquad (1)$$

При выводе предполагается, что сигнал и помеха не коррелированы между собой, так что функция корреляции и спектр смеси сигнала с помехой выражаются просто суммами соответствующих функций для сигнала и помехи, взятых в отдельности.

Средний квадрат ошибки, т. е. уклонения величины на выходе фильтра от сигнала, будет

$$\bar{\varepsilon}^2 = \int_{-\infty}^{\infty} \frac{G_f G_\xi}{G_f + G_\xi}\, d\omega. \tag{2}$$

Из этой формулы сразу видно, что ошибка может быть сделана равной нулю, т. е. что сигнал может быть полностью отделен от помехи лишь при условии $G_f G_\xi = 0$, что означает, что спектры сигнала и помехи не перекрываются ([17], стр. 106 и 136). В противном случае ошибка неизбежна.

Не приводя вывода формул (1) и (2), покажем их применение на примере. Пусть $f(t)$ есть обобщенный телеграфный сигнал, рассмотренный в § 23. Для такого сигнала функция корреляции убывает по экспоненциальному закону, а спектр выражается формулой

$$G_f(\omega) = \frac{4}{\pi} a^2 \frac{\mu}{4\mu^2 + \omega^2}. \tag{3}$$

Относительно помехи предположим, что она представляет собой белый шум с однородным спектром

$$G_\xi(\omega) = \rho = \text{const}. \tag{4}$$

Подставляя (3) и (4) в (1), получим для искомой частотной характеристики фильтра

$$A(\omega) = \frac{1}{1 + \frac{G_\xi}{G_f}} = \frac{1}{1 + \frac{\pi\mu\rho}{a^2} + \frac{\pi}{4} \cdot \frac{\rho}{a^2\mu} \omega^2}.$$

Величина μ означает среднее количество нулей в единицу времени. Следовательно, можно ввести среднюю круговую частоту манипуляции, выразив ее соотношением

$$\omega_0 = \pi\mu.$$

Введем, кроме того, обозначение

$$x = \pi\mu\rho/a^2 = \omega_0\rho/a^2.$$

Эта величина выражает отношение мощности помехи, приходящейся на полосу частот от нуля до ω_0, к мощности сигнала a^2. В этих обозначениях выражение для частотной характеристики фильтра запишется в виде

$$A(\omega) = \frac{1}{1 + x + \frac{\pi^2}{4} x \left(\frac{\omega}{\omega_0}\right)^2}.$$

Полученное выражение можно еще нормировать так, чтобы получить значение единица при $\omega=0$

$$a(\omega)=(1+x)A(\omega)=\frac{1}{1+\frac{\pi^2}{4}\cdot\frac{x}{1+x}\left(\frac{\omega}{\omega_0}\right)^2}. \qquad (5)$$

Таким образом, чем меньше x, тем более полого может идти характеристика фильтра. При $x=0$ (т. е. при отсутствии помехи) фильтрация вообще не нужна.

Обратимся к ошибке фильтрации. По формуле (2) находим

$$\bar{\varepsilon}^2=\int_{-\infty}^{\infty}\frac{G_f G_\xi}{G_f+G_\xi}\,d\omega=\rho\int_{-\infty}^{\infty}A(\omega)\,d\omega=$$

$$=\omega_0\rho\int_{-\infty}^{\infty}\frac{dy}{1+x+\frac{\pi^2}{4}xy^2}=\frac{2\omega_0\rho}{\sqrt{x(1+x)}}=2a^2\sqrt{\frac{x}{1+x}}.$$

Наименьшая, т. е. получаемая при оптимальной фильтрации, относительная ошибка полностью определяется относительной мощностью помехи. Если, например, допускается относительная среднеквадратичная ошибка 10%, то из уравнения

$$\frac{\bar{\varepsilon}^2}{a^2}=0{,}01=2\sqrt{\frac{x}{1+x}}\approx 2\sqrt{x}$$

находим, что x не должно превосходить $2{,}5\cdot10^{-5}$.

8. Асинхронное накопление (к § 39)

Метод накопления основан на использовании различия в свойствах сигнала и помехи: предполагается, что сигнал когерентен, а помеха — нет. В описанном выше синхронном варианте метода накопления когерентность сигнала достигается путем его периодического повторения. Но можно осуществить метод накопления и в другой форме, предложенной В. С. Воюцким [7].

Положим, что мы располагаем двумя каналами, по которым поступает смешанный с помехой сигнал, и пусть каналы устроены так, что помеха в обоих каналах некогерентна, а сигналы практически тождественны. Так, например, при внешней помехе можно применить прием на две разнесенные антенны, выбрав взаимное расположение антенн так, чтобы выставленное выше требование было удовлетворено.

Пусть сигнал в обоих каналах обозначен через $a(t)$, а помеха соответственно $\xi_1(t)$ и $\xi_2(t)$. На входе приемного устройства мы будем иметь от обоих каналов соответственно

$$x_1=a+\xi_1,\quad x_2=b+\xi_2.$$

Теперь составим сумму и разность сигналов, поступающих по обоим каналам,

$$y_1 = x_1 + x_2 = 2a + \xi_1 + \xi_2, \quad y_2 = x_1 - x_2 = \xi_1 - \xi_2.$$

Возведем в квадрат и усредним величины y

$$\bar{y}_1^2 = 4\bar{a}^2 + \overline{4a(\xi_1 + \xi_2)} + \overline{(\xi_1 + \xi_2)^2}, \quad \bar{y}_2^2 = \overline{(\xi_1 - \xi_2)^2}.$$

Если теперь вычесть друг из друга эти средние квадраты, т. е. составить разность

$$z = \bar{y}_1^2 - \bar{y}_2^2,$$

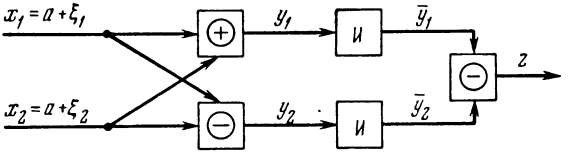


Рис. 65

то окажется, что z не зависит в среднем от помехи (рис. 65). В самом деле, в силу независимости сигнала и помехи произведение $a(\xi_1 + \xi_2)$ в среднем равно нулю. Что же касается квадратов суммы и разности помех в двух каналах, то для них мы имеем

$$(\xi_1 \pm \xi_2)^2 = \xi_1^2 + \xi_2^2 \pm 2\xi_1\xi_2.$$

Квадраты выпадают при вычитании; произведения же в среднем равны нулю в силу исходного предположения о некогерентности помех в обоих каналах. Таким образом,

$$\bar{z} = 4\bar{a}^2$$

и мы освобождаемся от помехи.

Нужно, однако, напомнить, что равенство справедливо в среднем и что величины, равные нулю в среднем, претерпевают флюктуации. Мы можем произвести усреднение лишь на конечном интервале. Заменим усреднение интегрированием в конечных пределах. Мы будем иметь

$$z = \int_0^T y_1^2 dt - \int_0^T y_2^2 dt = \int_0^T [4a^2 + 4a(\xi_1 + \xi_2) +$$

$$+ (\xi_1 + \xi_2)^2]\, dt - \int_0^T (\xi_1 - \xi_2)^2 dt = 4\left\{ \int_0^T a^2 dt + \right.$$

$$\left. + \int_0^T a(\xi_1 + \xi_2)\, dt + \int_0^T \xi_1\xi_2 dt \right\}. \qquad (1)$$

В такой форме результат показывает возможность накопления полезного сигнала. Первый член, подынтегральная функция которого положительна, неуклонно растет с течением времени. Два других члена представляют собой случайные величины, флюктуирующие около нуля. Величина флюктуаций, правда, растет со временем, но медленнее, чем энергия сигнала. Получаемые здесь соотношения отличаются от ранее рассмотренных (§ 39) лишь тем, что в формулу (1) входят не дискретные суммы, а интегралы. Флюктуации интегралов можно оценить следующим упрощенным способом: пусть дан интеграл

$$I = \int_0^T \zeta(t)\,dt,$$

где ζ — непрерывно изменяющаяся случайная величина.

Разобьем интервал интегрирования на участки Δt и применим теорему о среднем

$$I = \int_0^T \zeta(t)\,dt = \sum_{k=0}^{T/\Delta t} \int_{k\Delta t}^{(k+1)\Delta t} \zeta(t)\,dt = \Delta t \sum \zeta_k,$$

где ζ_k — некоторое значение ζ в интервале $k\Delta t < t < (k+1)\Delta t$. При малом Δt можно взять

$$\zeta_k = \zeta\left[\left(k+\frac{1}{2}\right)\Delta t\right].$$

Нас интересует дисперсия интеграла. Она равна

$$D(I) = D(\Delta t \sum \zeta_k) = \Delta t^2 D(\sum \zeta_k).$$

Пренебрегая корреляцией, т. е. считая, что ε_k независимы, получим

$$D(I) = \Delta t^2 \sum_{k=0}^{T/\Delta t} D(\zeta_k) = \Delta t T D(\zeta),$$

т. е. дисперсия интеграла при сделанных допущениях пропорциональна дисперсии случайной величины, стоящей под знаком интеграла, и интервалу интегрирования.

Среднеквадратичное значение флюктуирующих (второго и третьего) интегралов в формуле (1) пропорционально $\sqrt{T}$, тогда как первый интеграл пропорционален T (он выражает энергию сигнала и равен произведению T на среднюю мощность сигнала).

Если сигнал значительно слабее помехи, то вторым интегралом можно пренебречь по сравнению с третьим. Зная распределение, можно найти вероятность того, что флюктуация третьего интеграла превзойдет значение первого. Задавшись же вероятностью, можно найти требуемое время накопления.

9. Корреляционный прием двоичных сигналов (к § 44)

Пусть сигнал передается при помощи двух различных функций времени $a(t)$ и $b(t)$, чередующихся в известной последовательности. Такой сигнал кодирован по двоичной системе, а $a(t)$ и $b(t)$ — элементы двоичного кода. На сигнал налагается помеха $\xi(t)$, так что в результате получается функция $x(t)$, которая в зависимости от того, какой элемент кода передается в данный момент, равна либо $a+\xi$, либо $b+\xi$. Таким образом, мы имеем здесь уже разобранный случай двух различных сигналов при наличии помехи; разница состоит лишь в том, что два сигнала, которые мы теперь рассматриваем, — это не независимые сигналы, а элементы одного и того же сигнала.

Приемник должен отличать a от b. Если передается a, то идеальный (по В. А. Котельникову) приемник отметит различие и примет элемент a, т. е. даст правильный прием, если

$$\|x-a\|<\|x-b\|, \tag{1}$$

или

$$\int_0^T [x(t)-a(t)]^2\,dt < \int_0^T [x(t)-b(t)]^2\,dt, \tag{2}$$

или, выражая интегрирование как усреднение во времени,

$$\overline{(x-a)^2} < \overline{(x-b)^2}. \tag{3}$$

Раскрывая квадраты разностей, получим

$$\bar{a}^2 - 2\overline{ax} < \bar{b}^2 - 2\overline{bx}. \tag{4}$$

Естественно и вполне возможно выбрать элементарные сигналы $a(t)$ и $b(t)$ так, чтобы было

$$\bar{a}^2 = \bar{b}^2. \tag{5}$$

В таком случае (4) принимает вид

$$\overline{ax} > \overline{bx}. \tag{6}$$

К такому условию приходит в своей работе Е. А. Хмельницкий [15].

Но равенство (6) есть соотношение корреляционного типа. Как видим, условие приема сигнала a сводится к требованию, чтобы фактически принятый сигнал x имел бóльшую корреляцию с a, чем с b. Следовательно, идеальный приемник в рассматриваемом случае есть корреляционный приемник, который измеряет взаимные корреляции $\overline{ax}$ и $\overline{bx}$ сравнивает их между собой. Для того чтобы выполнение описанной операции было возможно, необходимо располагать на приемном конце системы связи всеми

тремя функциями a, b и x. Но $x\,(t)$ — это сигнал, который мы фактически принимаем. Что же касается a и b, то это — элементы сигнала, которые могут и должны быть наперед известны; в таком случае их можно генерировать на приемном конце, и скелетная схема приемника принимает вид, показанный на рис. 66. Входной сигнал x попадает на два модулятора М (умножающие схемы). Кроме того, на модуляторы подаются элементарные сигналы от местных генераторов Γ_a и Γ_b. На выходах модуляторов получаются соответственно произведения ax и bx. Произведения усредняются интегрирующими схемами И и подаются на вычи-

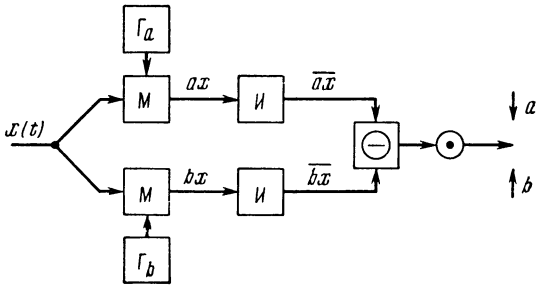


Рис. 66

тающее устройство. Выходной индикатор дает a или b в зависимости от знака разности.

Следует еще добавить, что мы имеем здесь дело не с обычным определением функции взаимной корреляции, предусматривающим усреднение на бесконечном интервале, а с определением так называемой кратковременной или текущей функции корреляции

$$B_T(\tau)=\frac{1}{T}\int_0^T x\,(t)\,y\,(t+\tau)\,dt.$$

Временем усреднения в рассматриваемом случае является, очевидно, длительность элементарной посылки.

10. О разделении перекрывающихся импульсов (к § 48)

Если при передаче бесконечно коротких амплитудулированных импульсов, образующих последовательность

$$f\,(t)=\sum q_k\sigma_1\,(t-k\Delta t), \qquad (1)$$

ограничить спектр сверху частотой ω_c, то импульсы расплываются и вместо бесконечно коротких импульсов мы получим возмущения, бесконечно протяженные во времени. Последовательность (1) заменится последовательностью

$$f(t)=\sum f_k \frac{\sin \omega_c (t-k\Delta t)}{\omega_c (t-k\Delta t)}, \qquad (2)$$

если отвлечься от общего сдвига во времени, обусловленного действием ограничивающего фильтра. Здесь f_k — значение k-го слагаемого суммы (2) при $t=k\Delta t$, равное $f(k\Delta t)$. Мы имеем

$$f_k=\frac{\omega_c}{\pi} q_k = \frac{q_k}{\Delta t}. \qquad (3)$$

Хотя возмущения вида $\sin x/x$ перекрываются, но, приняв последовательность (2), можно восстановить в точности последовательность (1). Эта возможность основана на том, что все слагаемые суммы (2) при $t=k\Delta t$ обращаются в нуль, кроме члена за номером k, который обращается в f_k. Таким образом, процесс восстановления последовательности (1) состоит в том, что берутся отсчеты мгновенных значений функции $f(t)$ в моменты $t=k\Delta t$; эти мгновенные значения равны f_k и пропорциональны величинам первоначальных импульсов q_k.

В описанной идеальной схеме возможно, таким образом, полное разделение импульсов, расплывшихся и наложившихся друг на друга вследствие ограничения спектра. Возникает естественный вопрос о том, возможно ли разделение в реальной схеме. Наша идеализация состоит в том, что, во-первых, исходные импульсы предположены бесконечно короткими, а во-вторых, фильтр предположен идеальным, т. е. бесконечной крутизной среза и с бесконечным затуханием вне полосы прозрачности. Мы рассмотрим здесь соотношения, получающиеся при отбрасывании первого из этих предположений, т. е. случай конечной длительности исходных импульсов.

Пусть исходные импульсы имеют прямоугольную форму; высота импульсов $h_k=q_k/\tau$, где τ — длительность импульсов. Рассмотрим прямоугольный импульс, середина которого приходится на момент $t=0$. Спектр такого импульса имеет, как известно, вид

$$S=q_0 \frac{\sin \omega \frac{\tau}{2}}{\omega \frac{\tau}{2}}. \qquad (4)$$

Положим, что мы ограничиваем полосу пропускания частотой ω_c. Искаженная вследствие ограничения спектра форма импульса будет

$$f(t)=\frac{q_0}{2\pi}\int_{-\omega_c}^{\omega_c} \frac{\sin \omega \frac{\tau}{2}}{\omega \frac{\tau}{2}} e^{j\omega t} d\omega = \frac{2q_0}{\pi\tau}\int_{}^{\omega_c} \sin \omega \frac{\tau}{2} \cos \omega t \frac{d\omega}{\omega} =$$

$$= \frac{h_0}{\pi}\left[Si\omega_c\left(t+\frac{\tau}{2}\right) - Si\omega_c\left(t-\frac{\tau}{2}\right)\right]. \qquad (5)$$

Нас интересуют значения этой функции в дискретные моменты $t = i\Delta t$. Полагая

$$\omega_c \Delta t = \pi,$$

получим

$$f(i\Delta t) = \frac{h_0}{\pi}\left[Si\left(i\pi + \frac{\pi}{2}\cdot\frac{\tau}{\Delta t}\right) - Si\left(i\pi - \frac{\pi}{2}\cdot\frac{\tau}{\Delta t}\right)\right]. \qquad (6)$$

Рассмотрим сперва случай большой скважности, т. е. $\tau \ll \Delta t$. В этом случае на основании формулы Тэйлора имеем

$$F(x_0 + \Delta x) - F(x_0 - \Delta x) = \\ = 2F'(x_0)\,\Delta x + 2F'''(x_0)\frac{\Delta x^3}{3!} + 2F^{V}(x_0)\frac{\Delta x^5}{5!} + \ldots$$

У нас

$$F(x) = Si(x), \quad F'(x) = \frac{d}{dx}Si(x) = \frac{\sin x}{x},$$

$$x_0 = i\pi, \quad \Delta x = \frac{\pi}{2}\cdot\frac{\tau}{\Delta t}.$$

Первый член выпадает, так как

$$F'(x_0) = \frac{\sin i\pi}{i\pi} = 0.$$

Поэтому первый не равный нулю член разложения есть член третьего порядка. Найдем третью производную

$$\frac{d^3}{dx^3}Si(x) = \frac{1}{x^3}[(2 - x^2)\sin x - 2x\cos x].$$

Подставляя

$$x = x_0 = i\pi,$$

получаем

$$Si'''(i\pi) = (-1)^{i+1}\frac{2}{i^2\pi^2}.$$

Итак,

$$f(i\Delta t) \approx \frac{q_0}{\pi\tau}Si'''(i\pi)\frac{\pi}{2}\cdot\frac{\tau}{\Delta t} = (-1)^{i+1}\frac{h_0}{12t^2}\cdot\frac{\tau^3}{\Delta t^3}.\ ^{*} \qquad (7)$$

Это — мгновенное значение данного возмущения в точках $t = i\Delta t$, в которых при идеальной схеме получаются нули. В мо-

* Эта приближенная формула дает тем более точный результат, чем меньше $\tau/\Delta t$, т. е. чем больше скважность. Однако она дает правильный порядок величин даже при $\tau/\Delta t = 1$. Для того чтобы убедиться в этом, сравним при $\tau/\Delta t = 1$ точную формулу (6), которая дает

$$f(i\Delta t) = \frac{h_0}{\pi}\left[Si\left(i + \frac{1}{2}\right)\pi - Si\left(i - \frac{1}{2}\right)\pi\right] = A_i h_0.$$

мент же $t=0$, т. е. посередине, мы получаем максимальное значение

$$f(0)=\frac{2h_0}{\pi}\,Si\left(\frac{\pi}{2}\cdot\frac{\tau}{\Delta t}\right)\approx h_0\frac{\tau}{\Delta t}. \tag{8}$$

При отсчете этого значения будет происходить ошибка вследствие наложения соседних возмущений. Ошибка эта, согласно (7), выразится суммой

$$\Delta f=\frac{1}{12}\cdot\frac{\tau^3}{\Delta t^3}\sum_{k=-\infty}^{\infty}(-1)^{k+1}\frac{h_k}{k^2}. \tag{9}$$

Ошибка Δf есть случайная величина. Найдем ее среднеквадратичное значение, предполагая, что h_k статистически независимы, т. е. пренебрегая корреляцией между ними. При этом средний квадрат суммы равен сумме средних квадратов

$$\overline{\Delta f^2}=\left(\frac{1}{12}\cdot\frac{\tau^3}{\Delta t^3}\right)^2\bar{h}_k^2 2\sum_{k=1}^{\infty}\frac{1}{k^4}=2{,}16\left(\frac{1}{12}\cdot\frac{\tau^3}{\Delta t^3}\right)^2\bar{h}_k^2,$$

откуда среднеквадратичное значение ошибки

$$\sigma=\sqrt{\overline{\Delta f^2}}=0{,}122\frac{\tau^3}{\Delta t^3}\sqrt{\bar{h}_k^2}. \tag{10}$$

Из этого соотношения видно, что σ быстро убывает с увеличением скважности $\Delta t/\tau$.

Найдем теперь $\sqrt{\bar{h}_k^2}$, предполагая, что шкала значений h_k симметрична относительно нуля, т. е. что имеются как положительные, так и отрицательные значения, что, кроме того, все значения h_k равновероятны. Обозначим шаг шкалы квантования через δ. Тогда

$$\bar{h}_k^2=2\frac{\delta^2}{m}\sum_{i=1}^{\frac{m-1}{2}}i^2=\frac{\delta^2}{12}(m^2-1), \tag{11}$$

с приближенной формулой (7), согласно которой

$$f(i\Delta t)=(-1)^{i+1}\frac{1}{12i^2}\,h_0.$$

Вычислив коэффициенты A_i (для этой цели хороши подробные таблицы В. В. Татаринова «Трехзначные таблицы интегральных синусов и косинусов». Связьтехиздат 1934), сопоставим их с коэффициентами $1/12i^2$

i	1	2	3
A_i	0,075	0,017	0,07
$1/12i^2$	0,083	0,021	0,09

В случае надобности можно было бы, конечно, построить лучшую приближенную формулу, чем (7); для наших целей это излишне.

где m — основание кода. Подставляя (11) в (10), получим

$$\frac{\Delta t^3}{\tau^3} = 0{,}0354\sqrt{m^2-1}\,\frac{\delta}{\sigma}. \qquad (12)$$

Это соотношение связывает, как видим, три величины: скважность $\Delta t/\tau$, основание кода m и коэффициент запаса δ/σ. Последняя величина определяет вероятность ошибки. Мы можем, основываясь на теореме Ляпунова, считать, что случайная величина Δf имеет нормальное распределение. Тогда можно использовать ранее полученные соотношения (см. § 22) и принять для коэффициента запаса δ/σ значение около десяти, что обеспечивает вероятность ошибки порядка 10^{-6}. Введя в (12) численное значение $\delta/\sigma=10$, получим связь между m и $\Delta t/\tau$

$$\Delta t^3/\tau^3 = 0{,}35\sqrt{m^2-1}. \qquad (13)$$

Оказывается, что достаточно малая ошибка при скважности единица обеспечивается уже при троичном коде, а при двоичном и подавно. При $m=128$ (число ступеней, принятое в импульсной телефонии) достаточно взять скважность, равную

$$\Delta t/\tau = \sqrt[3]{0{,}35\cdot 128} = 3{,}55.$$

Таким образом, возможность уверенного приема последовательности амплитудно-модулированных импульсов конечной длительности, расплывшихся вследствие ограничения полосы пропускания, сохраняется вплоть до очень малых значений скважности даже при большом основании кода.

Следует напомнить, что эти результаты относятся к идеализированным условиям: мы предполагаем, что исходные импульсы имеют прямоугольную форму и что ограничивающий фильтр идеален. Влияние отклонений от этих условий требует отдельного исследования.

Литература

1. *Агапов И. Ф.* Двухканальное частотное радиотелеграфирование с активной паузой (ДТЧ). Связьиздат, 1953.
2. *Агеев Д. В.* Основы теории линейной селекции. Научно-техн. сб. ЛИИС, 1935, № 10.
3. *Баев Н. А., Егоров К. П.* Основы дальней связи. Связьиздат, 1948.
4. *Боев Г. П.* Теория вероятностей. Гостехиздат, 1950.
5. *Вавилов В. С.* Опыты по радиолокации Луны. — УФН, 1949, т. 39, № 3.
6. *Владимирский К. В.* О синхронном фильтре. — ЖЭТФ, 1951, т. 21, № 1.
7. *Воюцкий В. С.* Обнаружение слабых сигналов способом асинхронного накопления. — Радиотехника, 1954, т. 7, № 6.
8. *Гнеденко Б. В.* Курс теории вероятностей. Гостехиздат, 1950.
9. *Железнов Н. А.* (ред.). Теория передачи электрических сигналов при наличии помех (сб. перев.). ИЛ, 1953.
10. *Котельников В. А.* О пропускной способности «эфира» и проволоки в электросвязи. — Всесоюзн. энерг. ком. Материалы к первому всесоюзн.

съезду по вопр. реконстр. дела связи и разв. слаботочн. пром-сти (изд. Ред. Упр. связи РККА, 1933).
11. *Котельников В. А.* Теория потенциальной помехоустойчивости при флюктуационных помехах. Докт. дисс. МЭИ, 1946.
12. *Котельников В. А.* Проблемы помехоустойчивой радиосвязи. Радиотехнич. сб. ЦБТИ МПСС. Госэнергоиздат, 1947.
13. *М. В. Назаров.* К теории разделения сигналов. Канд. дисс. МЭИС, 1953.
14. *Солодовников В. В.* Введение в статистическую динамику систем автоматического управления. М.—Л., Гостехиздат, 1952.
15. *Хмельницкий Е. А.* Исследование помехозащищенности при приеме на разнесенные антенны радиотелеграфных сигналов. Канд. дисс. МЭИС, 1954.
16. *Щукин А. Н.* Двукратная радиотелеграфная передача без потери мощности. — Техника связи, 1933, № 3.
17. *Яглом А. М.* Введение в теорию стационарных случайных функций. — УМН, 1952, т. 7, № 5 (51).
18. *Z. Bay.* Reflection of the microwave from the Moon. — Acta phys. Acad. scint. hung., 1947, v. 1, N 1.
19. *H. Dudley.* Remaking speech. — J. Acoust. Soc. America, 1939, v. 11.
20. *R. W. Hamming.* Error detecting and error correcting codes. — Bell System Techn. J., 1950, v. 29, N 2.
21. *R. V. L. Hartley.* Transmission of information. — Bell System Techn. J., 1928, v. 7.
22. *E. R. Kretzmer.* Statistics of television signals. — Bell System Techn. J., 1952, v. 31, N 4.
23. *Y. W. Lee, T. P. Cheatham, I. B. Wiesner.* Application of correlation analysis to the detection of periodic signals in noise. — Proc. IRE, 1950, v. 38, N 10.
24. *L. I. Lidois.* Un nouveau procedé de modulation codée: «la modulation end». — Onde electr., 1952, v. 32, N 298.
25. *B. M. Oliver, J. R. Pierce, C. E. Shannon.* The philosophy of PCM. Proc. — IRE, 1948, v. 36, N 11.
26. *B. M. Oliver.* Efficient coding. — Bell System Techn. J., 1952, v. 31, N 4.
27. *C. E. Shannon.* A mathematical theory of communication. — Bell System Techn. J., 1948, v. 27, N 3.
28. *C. E. Shannon.* Communication in the presence of noise. — Proc. IRE, 1949, v. 37, N 10.
29. *L. A. Zadeh, K. S. Miller.* Fundamental aspects of linear multiplexing. — Proc. IRE, 1952, v. 40, N 9.

О НАИЛУЧШЕМ КОДЕ

1. Для передачи сообщений можно применить самые различные коды. Естественно с самого начала поставить вопрос о наилучшем (оптимальном) коде. Предварительно нужно установить критерий качества кода.

Мы будем предполагать передаваемые сообщения равновероятными; наилучшим кодом мы назовем код, обеспечивающий передачу наибольшего количества сведений при заданной помехоустойчивости или наоборот.

Количество сведений тем больше, чем больше число возможных сообщений. Число возможных сообщений есть число кодовых комбинаций, которые можно составить при заданном числе знаков сигнала. Таким образом, тот код лучше, который при прочих равных условиях дает большее число кодовых комбинаций.

Помехоустойчивость кода тем больше, чем больше при прочих равных условиях различие между отдельными кодовыми комбинациями. Пусть каждая комбинация представлена последовательностью $f_1, f_2, \ldots, f_n$, где n — число знаков комбинации. Тогда мерой различия двух комбинаций a и b может служить величина

$$d = \sqrt{\sum_{k=1}^{n} (f_{ak} - f_{bk})^2}.$$

Эту величину мы будем называть расстоянием между сигналами a и b.

Ясно, что расстояние растет с увеличением энергии сигнала

$$E = \sum_{k=1}^{n} f_k^2,$$

поэтому сравнение следует производить для сигналов равной энергии или оперировать отношением $d/\sqrt{E}$.

Итак, наилучший код должен для сигналов равной энергии давать наибольшее число N возможных сигналов при наибольшем расстоянии d между ними. Это и есть критерий, которым мы воспользуемся в дальнейшем.

2. Дадим геометрическое истолкование основным соотношениям. Последовательность $f_1, f_2, \ldots, f_n$ можно рассматривать как координаты точки в n-мерном пространстве. Значит, каждый сиг-

нал (кодовая комбинация) отображается одной точкой в n-мерном пространстве. Сумма $\sum f_k^2$ выражает квадрат расстояния точки от начала координат. Таким образом, расстояние точки сигнала от начала координат выражается как $\sqrt{E}$, где E — энергия сигнала. Сигналы равной энергии находятся, следовательно, на равных расстояниях от начала координат, т. е. точки, представляющие эти сигналы, лежат на поверхности n-мерной сферы радиуса $\sqrt{E}$ с центром в начале координат. Величина d выражает действительное расстояние между сигналами, т. е. длину отрезка прямой, соединяющего две точки сигналов.

Построение наилучшего кода сводится с геометрической точки зрения к нахождению такого расположения точек на поверхности сферы, которое позволяет разместить наибольшее число точек при заданном наименьшем расстоянии между ними, или, наоборот, разместить заданное число точек так, чтобы наименьшее расстояние между ними было как можно больше. Это означает, что мы должны покрыть поверхность сферы шаровыми сегментами наиболее плотно и дело сводится к классической задаче о наиболее плотной укладке.

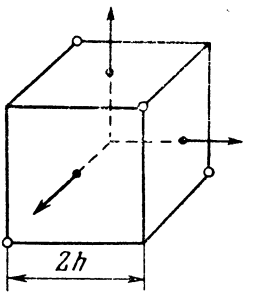


Рис. 1

Чтобы не потерять геометрической наглядности, мы ограничимся трехзначным сигналом; это позволит нам делать все построения в обычном трехмерном пространстве.

3. Начнем с двоичного сигнала вида $f_k = \pm h$. Возможно всего 8 трехзначных двоичных сигналов: $+h, +h, +h$; $+h, +h, -h$; $+h, -h, +h$; $-h, +h, +h$; $+h, -h, -h$; $-h, +h, -h$; $-h, -h, +h$; $-h, -h, -h$.

Каждая тройка чисел представляет координаты одной из восьми вершин куба, показанного на рис. 1. Наименьшее расстояние между вершинами составляет $2h$. Вершины куба лежат на сфере радиуса $\sqrt{3}h$. Отношение наименьшего расстояния к радиусу сферы составляет

$$d/\sqrt{E} = 2/\sqrt{3} = 1{,}15.$$

Если взять только вершины, отмеченные на рис. 1 кружками, то число сигналов сократится до четырех; зато наименьшее расстояние возрастет в $\sqrt{2}$ раз и будет $\frac{d}{\sqrt{E}} = 1{,}63$. Наконец, если взять только две вершины по концам диагонали куба (т. е. по концам диаметра описанной окружности), то будет $N = 2$, $d/\sqrt{E} = 2$.

4. Рассматривая расположение точек на сфере по вершинам вписанного в сферу куба (рис. 2), можно заметить, что это расположение не является плотнейшим, так как касающиеся друг друга

четыре сегмента оставляют большой промежуток. Наименьший промежуток получился бы при касании трех сегментов. Для наглядности эти два расположения показаны на рис. 3 в виде развертки на плоскость.

Из рис. 3, *б* можно заключить, что плотнейшая укладка сегментов на сфере получится при расположении центров сегментов (т. е. точек сигналов) по вершинам правильных многогранников с треугольными гранями. Такими многогранниками являются: тетраэдр (4 грани, 4 вершины), октаэдр (8 граней, 6 вершин), икосаэдр (20 граней, 12 вершин). Расположение точек на сфере по вершинам этих многогранников показано на рис. 4, 5 и 6.

Случай тетраэдра нами уже рассмотрен: отмеченные кружками на рис. 1 вершины куба лежат в вершинах тетраэдра. Таким образом, двоичный трехзначный сигнал при $N=4$ оптимален. Обратимся к октаэдру. Соотношения для него легко получить, принимая во внимание, что вершины октаэдра являются серединами граней некоторого куба; мы получаем для октаэдра $d/\sqrt{E}=\sqrt{2}=$ $=1{,}41$ при $N=6$. Для икосаэдра можно найти значение

$$\frac{d}{\sqrt{E}}=\frac{\sqrt{4\sin^2 36^\circ-1}}{\sin 36^\circ}=1{,}05.$$

Все эти результаты сведены в табл. 1. Столбцы 1 и 2 относятся как к двоичному, так и к наилучшему коду. Столбец 3 (куб) относится к двоичному коду. Последние два столбца отно-

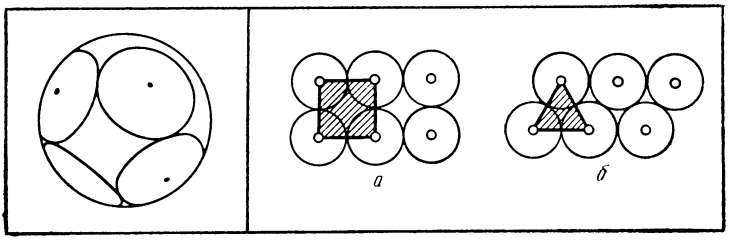


Рис. 2

Рис. 3

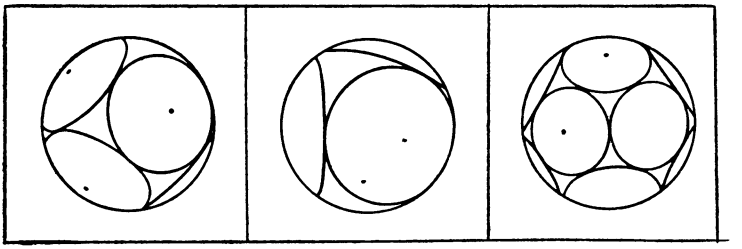


Рис. 4

Рис. 5

Рис. 6

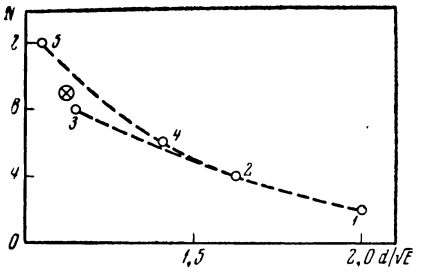


Рис. 7

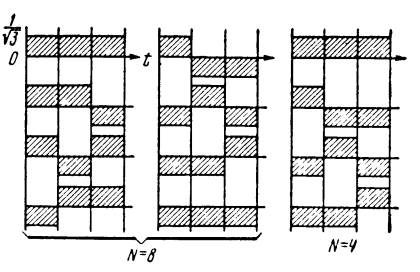


Рис. 8

сятся к наилучшему коду. Данные табл. 1 представлены графически на рис. 7. Нумерация точек соответствует нумерации столбцов табл. 1. Точки соединены линиями, показывающими общую тенденцию зависимости N от $d/\sqrt{E}$. Линия 1—2—3 относится к двоичному коду, линия 1—2—4—5 — к наилучшему коду. Преимущество наилучшего кода перед двоичным очевидно.

Таблица 1

№	1	2	3	4	5
Фигура	Диаметр	Тетраэдр	Куб	Октаэдр	Икосаэдр
N	2	4	8	6	12
$d/\sqrt{E}$	2	1,63	1,15	1,41	1,05

5. Можно получить хотя и не плотнейшие, но лучшие, чем при двоичном коде, расположения для любого $N > 4$. Так, рассматривая рис. 2, можно заметить, что на каждом полушарии располагаются венцом по четыре сегмента. Если поворотом на 45° сместить одно полушарие, то можно будет вдвинуть один венец в промежутки другого. При этом на полюсе освободится место для еще одного сегмента (при условии небольшого уменьшения d). Точки сигналов расположатся по вершинам неправильного многогранника, грани которого представляют собой 12 треугольников и 1 квадрат. Для этого случая мы имеем $N = 9$, $d\sqrt{E} \approx 1{,}12$. Соответствующая точка отмечена на рис. 7 крестиком. При всем несовершенстве данного расположения представляющая его точка все же лежит выше линии 1—2—3 двоичного сигнала.

6. Желательно составить представление о различных сигналах как функциях времени. Положим $E=1$; тогда двоичный сигнал будет представлен комбинациями импульсов высотой $1/\sqrt{3}$, пока-

занными на рис. 8. Для сигнала, соответствующего расположению по октаэдру, имеем шесть комбинаций импульсов высотой единица (рис. 9). Заметим, что характер комбинаций изменяется в зависимости от расположения координатных осей. Рис. 9 соответствует расположению, показанному на рис. 10, *а*. Если же, например, повернуть октаэдр на 45° вокруг вертикальной оси, как на рис. 10, *б*, то получатся кодовые комбинации, показанные на рис. 11. Нужно заметить, что изменения кода, обусловленные поворотом координатной системы, не оказывают влияния ни на E, ни на d. Так, например, изменяя расположение осей относительно куба, можно получить код, который уже не будет двоичным, но эквивалентен ему по качеству.

7. Для больших n координаты точек сигнала при наилучшем коде пока не могут быть найдены. Известно лишь, что эти точки не лежат в узлах какой-либо правильной решетки.

Однако известно (это установлено Шэнноном), что распределение вероятностей координат для наилучшего кода приближается к нормальному. С этой точки зрения интересно посмотреть распределения для вышеприведенных трехмерных моделей.

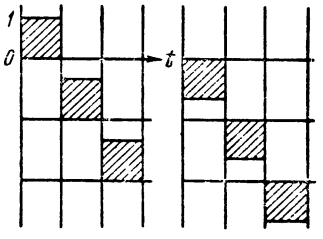


Рис. 9

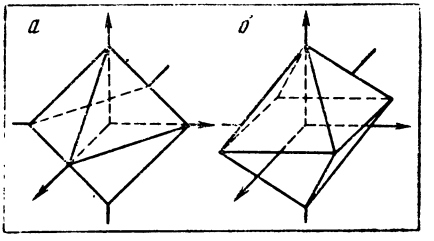


Рис. 10

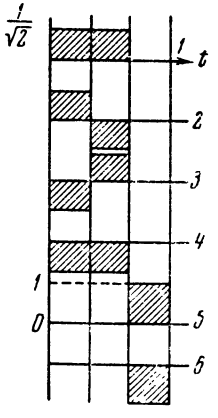


Рис. 11

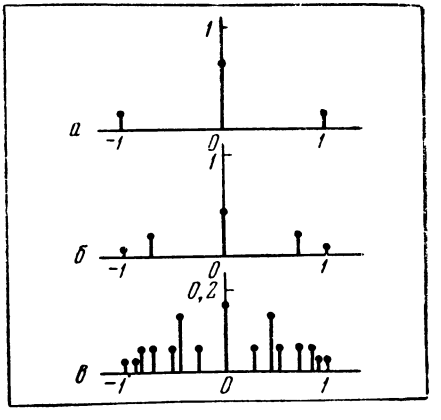


Рис. 12

Для расположения по октаэдру распределения показаны на рис. 12, *а* и *б*, соответствующих рис. 10, *а* и *б*. Для икосаэдра значения координат вершин даны в табл. 2 (одна из осей совпадает с диагональю многогранника).

Таблица 2

x_1	0	b	$b(1-2\beta^2)$	$b(1-2\beta^2)$	$-b\alpha$	$-b\alpha$	$b\alpha$	$b\alpha$	$-b(1-2\beta^2)$	$-b(1-2\beta^2)$	$-b$	0
x_2	0	0	$b2\alpha$	$-b2\alpha\beta$	$b\beta$	$-b\beta$	$b\beta$	$-b\beta$	$b2\alpha\beta$	$-b2\alpha\beta$	0	0
x_3	r	c	c	c	c	c	$-c$	$-c$	$-c$	$-c$	$-c$	r

В табл. 2 $\alpha = \cos 36°$; $\beta = \sin 36°$; $r = a\beta/\sqrt{4\beta^2 - 1}$ — радиус описанной сферы; a — длина ребра; $b = a/2\beta$; $c = a(1 - 2\beta^2)/2\beta\sqrt{4\beta^2 - 1}$.

Принимая $r=1$, получим распределение, изображенное на рис. 12, *в*.

Как видим, даже при $n=3$ начинает уже обозначаться тенденция перехода к нормальному распределению.

8. В заключение нужно сделать два замечания. Во-первых, разница между двоичным и оптимальным кодом в пользу последнего неограниченно растет с увеличением числа знаков n, или, что то же, с числом измерений пространства сигналов. Для двоичного кода $f_k = \pm h$ при $d=2h$, $N_{дв}=2^n$. Для наилучшего же кода при том же d и при той же энергии $E=nh^2$ имеем

$$N_{опт} > \left(\frac{1-\varepsilon}{2}\right)^n \left(\frac{n}{1-1/n}\right)^{n/2},$$

где ε убывает как n^{-1}. При очень больших n

$$N_{опт} > A(\sqrt{n}/2)^n$$

и отношение

$$N_{опт}/N_{дв} > A(\sqrt{n}/4)^n,$$

т. е. растет неограниченно с увеличением n.

Второе замечание относится к способу приема. Для того чтобы реализовать преимущества оптимального кода, приемник должен принимать n-значный сигнал целиком. Иначе говоря, приемник должен принять все значения f_k, построить по ним точку сигнала и определить, к какой из возможных точек принятый сигнал ближе.

О ПРИЕМЕ СЛАБЫХ СИГНАЛОВ

1. За последнее время ряд работ был посвящен исследованию различных методов приема слабых сигналов. Результаты всех этих работ сходятся на том, что все известные в настоящее время методы практически равноценны, т. е. дают одинаковый по порядку величины выигрыш в превышении сигнала над помехой на выходе по отношению к заданному превышению на входе.

Этот результат не случаен и может быть обоснован простыми обобщениями, изложенными ниже.

2. Все известные в настоящее время методы приема слабых сигналов могут быть сведены к интегральной операции вида

$$I = \int_0^T F(t)\,\varphi(t)\,dt, \tag{1}$$

где $F(t)=f(t)+\xi(t)$ — сумма сигнала и помехи; $\varphi(t)$ — весовая функция, определяющая способ приема. Так, например, имеем:

Метод накопления	$\varphi(t)=1$
Автокорреляционный прием	$\varphi(t)=F(t-\tau)$
Когерентный прием	$\varphi(t)=f(t)$
Фильтрация	$\varphi(t)=g(T-t)$

Здесь $g(t)$ — импульсная реакция фильтра.

3. Значение превышения на выходе приемника, выполняющего обобщенную операцию (1), может быть получено из следующих соображений. Представим (1) в виде

$$I = \int_0^T f(t)\varphi(t)\,\mathrm{d}t + \int_0^T \xi(t)\varphi(t)\,\mathrm{d}t = \eta + \zeta. \tag{2}$$

Первый член есть полезный сигнал: мы полагаем, что $\eta \neq 0$.

Второй член есть помеха: ζ есть случайная величина. Полагая $M\zeta=0$, будем считать, что помеха определяется дисперсией ζ. Тогда превышение сигнала над помехой на выходе приемника можно определить как

$$m = \gamma_1^2/D\zeta. \tag{3}$$

4. Найдем дисперсию ζ. Имеем

$$D\zeta = M\left(\int_0^T \xi(t)\,\varphi(t)\,dt\right)^2 = M\left(\int_0^T \xi(t)\,\varphi(t)\,dt \int_0^T \xi(s)\,\varphi(s)\,ds\right) =$$

$$= M\left(\int_0^T\int_0^T \varphi(t)\,\varphi(s)\,\xi(t)\,\xi(s)\,dtds\right) = \int_0^T\int_0^T \varphi(t)\,\varphi(s)\,M[\xi(t)\,\xi(s)]\,dtds,$$

или, вводя функцию корреляции,

$$D\zeta = \int_0^T \varphi(t)\,dt \int_0^T \varphi(s)\,B(t-s)\,ds. \qquad (4)$$

Определим интеграл корреляции для случайного процесса $\xi(t)$ как

$$\tau_\xi = \frac{1}{P_\xi}\int_0^T B_\xi(\tau)\,d\tau,$$

где $P_\xi = B_\xi(0)$ — мощность помехи на входе. Положим, что интервал корреляции τ_ξ мал. Тогда на основании теоремы о среднем вместо (4) можем записать приближенное соотношение

$$D\zeta \approx P_\xi \tau_\xi \int_0^T \varphi^2(t)\,dt = P_\xi \tau_\xi E_\varphi = \frac{\tau_\xi}{T} E_\varphi E_\xi, \qquad (5)$$

где E_φ и E_ξ — энергия функций $\varphi(t)$ и $\xi(t)$, т. е. интегралы от их квадратов на интервале T.

5. Найдем теперь полезный сигнал

$$\eta = \int_0^T f(t)\,\varphi(t)\,dt.$$

Ясно, что целесообразно выбрать весовую функцию $\varphi(t)$ так, чтобы получить наибольшее значение полезного сигнала. Это приводит к вариационной задаче о нахождении функции $\varphi(t)$, дающей максимум функционалу η. В качестве дополнительного условия возьмем

$$\int_0^T \varphi^2(t)\,dt = E_\varphi = \text{const}.$$

Решение этой задачи дает

$$\varphi(t) = \sqrt{\frac{E_\varphi}{E_f}}\,f(t)$$

и, таким образом,

$$\eta_{\max} = \sqrt{E_f E_\varphi}.$$

Как видим, наибольшее значение полезного сигнала получается при когерентном приеме[1].

6. Теперь определим превышение на выходе

$$m \leqslant \frac{\eta^2_{\max}}{D\zeta} = \frac{E_f}{E_\xi} \cdot \frac{T}{\tau_\xi}. \tag{6}$$

Этому соотношению можно придать иной вид, учитывая общую связь между интервалом корреляции и шириной спектра,

$$F\tau = \mu \approx 1.$$

Таким образом,

$$m \leqslant \frac{E_f}{E_\xi} F_\xi T, \tag{7}$$

или, деля на T,

$$m \leqslant \frac{P_f}{P_\xi} FT. \tag{8}$$

Если помеха имеет равномерный спектр с плотностью P_0, т. е. если

$$P_\xi = P_0 F_\xi,$$

то

$$m \leqslant \frac{P_f}{P_0} T = \frac{E_f}{P_0}. \tag{9}$$

Формула (6) или эквивалентные ей формулы (7)—(9) выражают (по порядку величины) наилучший возможный результат: единственное сделанное при выводе предположение состоит в том, что интервал корреляции τ_ξ мал, или, иначе, что ширина спектра помехи значительно больше, чем ширина спектра весовой функции.

Общий смысл полученных соотношений состоит в том, что для получения желаемого превышения на выходе приемника нужно увеличивать энергию сигнала; при заданной мощности сигнала на входе нужно увеличивать длительность сигнала T.

Как теперь известно, все применяемые методы приема слабых сигналов, описываемые формулой (1), дают результаты, очень близ-

---

[1] Заметим, что эти наиболее благоприятные соотношения достигаются в отдельных случаях и при других методах. Так, например, при приеме постоянного сигнала метод накопления оптимален. Другой пример: при приеме синусоидального сигнала на контур с нулевым затуханием и при $T = 2\pi n/\omega$ метод фильтрации также оптимален. Ясно, что в этих примерах $\varphi(t) = \text{const} \cdot f(t)$, что совпадает с условием когерентного приема.

кие к предельному соотношению. Это значит, что вопрос о выборе метода приема перемещается в настоящее время из плоскости теоретической в плоскость чисто технических или технико-экономических соображений.

ПРИЛОЖЕНИЕ

Вышеприведенный результат можно получить и геометрическим путем.

Представим сигнал, помеху и весовую функцию соответствующими векторами (рис. 1, *а*); длины векторов равны корням их энергий.

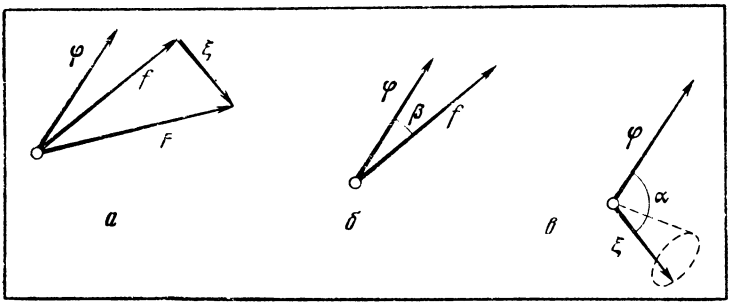


Рис. 1

Приемник составляет скалярное произведение векторов **E** и φ. Полезный сигнал есть $\eta = (\mathbf{f} \cdot \varphi) = \|\mathbf{f}\| \cdot \|\varphi\| \cos\beta$ (рис. 1, *б*). Очевидно, что η будет наибольшим (при фиксированных длинах векторов **f** и φ), когда векторы совпадают по направлению. При этом векторы **f** и φ представляют (с точностью до постоянного множителя) одну и ту же функцию; их скалярное произведение равно

$$\eta_{\max} = \|\mathbf{f}\| \cdot \|\varphi\| = \sqrt{E_f E_\varphi}.$$

Помеха зависит от скалярного произведения $\zeta = (\xi \cdot \varphi)$ (рис. 1, *в*). Вследствие статистической независимости ξ и φ угол α флюктуируется около $\pi/2$ и, таким образом, в среднем ζ равно нулю. Дисперсия ζ определяется дисперсией $\cos\alpha$ (влиянием флюктуаций длины вектора ζ пренебрегаем). Мы имеем

$$\cos\alpha = \zeta/\sqrt{E_\varphi E_\xi},$$

$$D(\cos\alpha) = D\zeta/E_\varphi E_\xi \approx 1/F_\xi T.$$

Таким образом, хотя длина вектора ξ растет с расширением полосы (так как $\|\xi\| = \sqrt{E_\xi} = \sqrt{P_\xi T} = \sqrt[4]{P_0 F_\xi T}$), но флюктуации

$\cos\alpha$ при этом убывают, и притом так, что оба эффекта компенсируют друг друга, и мы получаем

$$D\zeta = D(\|\varphi\| \cdot \|\xi\| \cdot \cos\alpha) = E_\varphi E_\xi D(\cos\alpha) = \\ = E_\varphi E_\xi / F_\xi T = E_\varphi P_0,$$

откуда

$$m_{\max} = \eta^2_{\max}/D\zeta = E_f/P_0.$$

Литература

1. *Y. W. Lee, T. P. Cheatham, I. B. Wiesner.* Detection of periodic signals in noise. — Proc. IRE, 1950, v. 38, N 10.
2. *D. E. Hampton.* Symposium on applications of communic. theory, London, 1953.
3. *R. M. Fano.* Symposium on applications of communic. theory, London, 1953.
4. *М. И. Карновский.* О подавлении флюктуационных помех при корреляционном методе приема. — Радиотехника, 1954, т. 9, № 3.
5. *D. G. Tucker, J. W. R. Griffits.* — Wireless engng., 1953, v. 30, N 11.
6. *P. Rudnich.* The detection of weak signals by correletion methods. — J. of appl. phys., 1953, v. 24, N 2.
7. *J. V. Harrington, T. F. Rogers.* Signal-to-noise improvement through integretion in a storage tube. — Proc. IRE, 1950, v. 38, N 10.
8. *S. F. George.* Effectiveness of crosscorrelation detectors. — Proc. Nat. Electronics Conf., 1954, v. 10.
9. *В. И. Чайковский.* Прием импульсных сигналов по методу взаимной корреляции. — Радиотехника, 1955, т. 10, № 6.

ОБ ОДНОЙ СХЕМЕ ПРИЕМА СИГНАЛОВ

Для реализации возможностей, предоставляемых помехоустойчивыми кодами, необходимо принимать сигнал (в виде некоторой кодовой комбинации) целиком, т. е. всю комбинацию в целом, а затем сличать его с множеством переданных сигналов и отождествлять с тем из них, от которого принятый сигнал наименее отличается. Это позволяет обнаруживать и исправлять ошибки.

С точки зрения геометрической теории всякий сигнал представляется точкой в пространстве соответствующего числа измерений. Наложение помехи смещает эту точку. Для того чтобы избежать ошибки, т. е. отождествления принятого сигнала не с тем, который фактически был передан, а с другим возможным, необходимо увеличивать расстояние между точками, представляющими возможные переданные сигналы.

Техника приема сигнала в целом должна состоять в общих чертах в следующем: 1) принятый сигнал запоминается; 2) он сличается поочередно со всеми возможными переданными сигналами, которые должны быть известны на приемной стороне системы связи и должны храниться в некотором запоминающем устройстве; 3) отмечается тот из возможных переданных сигналов, от которого принятый сигнал наименее отличается, — этот сигнал и считается переданным (истолкование выражения «наименее отличается» определяет способ действия приемника).

Общая схема приемника сигналов в целом представляется следующим образом: имеется память в виде, например, магнитной записи, на которой заранее записаны все возможные передаваемые сигналы; имеется возможность записать отдельно и принятый сигнал. Затем принятый сигнал многократно воспроизводится; одновременно воспроизводится каждый раз один из возможных передаваемых сигналов.

Сличение можно производить по-разному. Либо производится вычитание, возведение разности в квадрат и суммирование по формуле

$$d_k^2 = \int_0^T [y(t) - x_k(t)]^2 \, \mathrm{dt}, \tag{1}$$

где $y(t)$ — принятый сигнал; $x_k(t)$ — один из возможных переданных; переданный сигнал определяется по минимуму d_k; либо

сигналы перемножаются, а затем суммируются по формуле

$$R = \int_0^T y(t)\, x_k(t)\, \mathrm{dt}, \tag{2}$$

и переданный сигнал определяется по максимуму R*. Величина d_k по геометрическому смыслу есть расстояние между принятым сигналом и k-м возможным переданным. Величина R есть коэффициент взаимной корреляции между принятым сигналом и k-м возможным переданным. Формулы (1) и (2) относятся к непрерывным сигналам, T обозначает длительность сигналов. Для дискретных сигналов (1) и (2) заменяются соответствующими суммами:

$$d_k^2 = \sum_{i=1}^{n} (y_i - x_{ik})^2, \tag{1'}$$

$$R = \sum_{i=1}^{n} y_i x_{ik}, \tag{2'}$$

где y_i, x_{ik} — дискретные значения (знаки, символы) соответственно принятого и к-го возможного переданного сигналов, n — число знаков.

Вышеописанная схема имеет достаточно универсальный характер и применима для любых сигналов. Но техническое решение получается довольно громоздким. К тому же операция сличения и выбора требует времени; это время быстро возрастает с увеличением длительности сигналов, так как общее число возможных сигналов растет с увеличением длительности (или числа знаков в случае дискретных сигналов) по степенному закону. Соответственно растет и необходимая емкость памяти. Между тем преимущества помехоустойчивых кодов реализуются именно для больших отрезков сигнала. Наивыгоднейшие соотношения получаются в пределе при T (или n), стремящемся к бесконечности.

Поэтому представляет некоторый интерес не столь универсальная, но более простая схема для приема сигнала в целом. Мы опишем эту схему применительно к приему дискретных и, в частности, двоичных сигналов.

Идея состоит в перенесении многомерного образа множества передаваемых сигналов на плоскость, т. е. в пространство двух измерений. Это позволит воспользоваться в качестве оконечного звена приемного устройства обычной электронно-лучевой трубкой: множество передаваемых сигналов представится системой точек, расположенных по некоторой плоской решетке.

Любое n-значное двоичное число из общего количества

$$N = 2^n \tag{3}$$

может быть заменено двузначным числом по системе счисления с основанием a. Для полного отображения множества N нужно

* В обоих случаях предполагается, что сигналы равновероятны. Если к тому же они имеют равные энергии, то оба варианта просто совпадают.

только выполнить условие

$$M = a^2 \geqslant N, \tag{4}$$

откуда и определяется a. На основании (3) и (4) находим

n	3	4	5	6	7	8	9	...
a	3	4	6	8	12	16	23	...

Разберем для начала случай $n=3$, $a=3$. Трехзначный двоичный сигнал геометрически представляется вершинами куба; множество передаваемых сигналов содержит $N=2^3=8$ комбинаций.

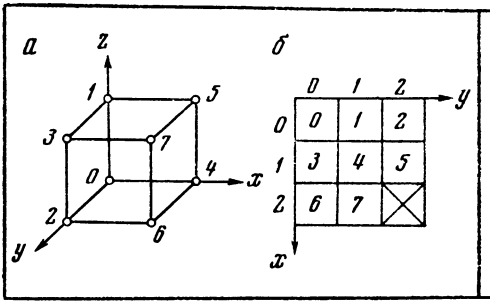


Рис. 1

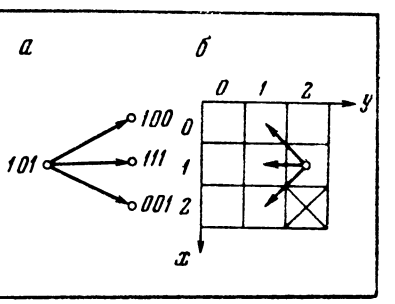


Рис. 2

Каждая комбинация выражается трехзначным двоичным числом. С другой стороны, эти двоичные числа могут быть выражены двузначными троичными числами; одно окажется лишним, так как

$$M=3^2=9.$$

Итак, мы имеем следующие соответствия:

№ комбинации (вершины куба)	0	1	2	3	4	5	6	7
Двоичная запись	000	001	010	011	100	101	110	111
Троичная запись	00	01	02	10	11	12	20	21

На рис. 1, *а* изображен куб с соответственно занумерованными вершинами, а на рис. 1, *б* — плоская таблица, отвечающая троичной записи.

Положим, что двоичное число 101 преобразовано в троичное число 12 (вопросы техники обсуждаются ниже). Мы рассматриваем это число как точку на плоскости с координатами $x=1$, $y=2$. Эта точка попадает в ячейку решетки за № 5 и отображает вершину № 5 трехмерного куба (с координатами $x=1$, $y=0$, $z=1$).

Ясно, что если при передаче двоичного сигнала произошла одиночная ошибка (т. е. ошибка в одном знаке), то принятая комбинация совпадет с одной из других возможных. Так, например, одиночная ошибка может превратить сигнал № 5 в один из трех других по схеме рис. 2, *а*, что соответствует положениям, отмеченным точками на рис. 2, *б*. Следовательно, ошибка при таких условиях не может быть ни исправлена, ни обнаружена.

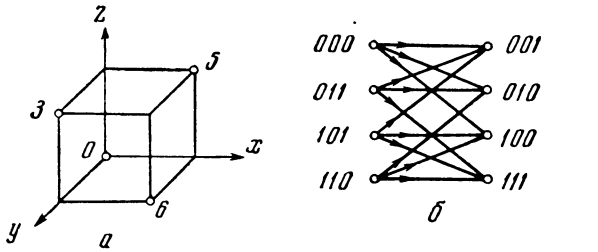


Рис. 3

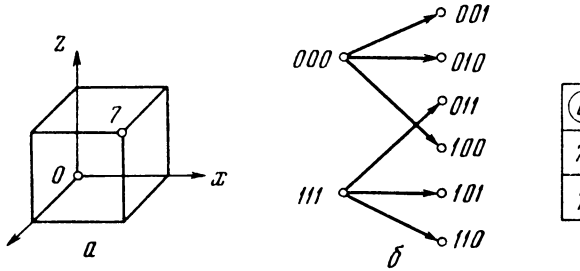


Рис. 4

Для того чтобы обнаружить одиночную ошибку, достаточно, как известно, выбрать двоичные комбинации так, чтобы они различались между собой не менее чем в двух знаках. На геометрической модели в виде трехмерного куба это соответствует выбору вершин, отстоящих друг от друга на два ребра (т. е. лежащих на диагоналях граней). Это означает, что мы используем только половину всех возможных комбинаций, остальные комбинации запрещены. Пусть разрешены комбинации 000, 011, 101, 110, соответствующие вершинам куба за № 0, 3, 5, 6 (рис. 3, *а*). Одиночная ошибка превращает разрешенные комбинации в запрещенные по схеме рис. 3, *б*. В двумерной таблице рис. 3, *в* места запрещенных комбинаций заштрихованы; попадание точки принятого сигнала в одну из заштрихованных клеток означает ошибку, которая таким образом и обнаруживается. Двойная ошибка обнаружена быть не может, так как она переводит любую разрешенную комбинацию в другую разрешенную.

Для того чтобы одиночная ошибка могла быть исправлена, необходимо, чтобы разрешенные комбинации различались не менее как в трех знаках. Мы можем построить только два трехзначных двоичных сигнала, удовлетворяющих этому условию. Соответствующие точки лежат на диагонали куба (например, 000 и 111, как показано на рис. 4, *а*). Схема перехода при одиночной ошибке показана на рис. 4, *б*. Таким образом, каждая из запрещенных комбинаций, образующихся в результате одиночной ошибки, связана только с одной из двух разрешенных комбинаций. Поэтому двумерная таблица принимает вид рис. 4, *в*. Основные положения (при отсутствии ошибки) сигналов № 0 и 7 отмечены кружками;

эти номера повторены в других клетках, куда сигналы могут попасть в результате одиночной ошибки. Таким образом, одиночная ошибка исправляется.

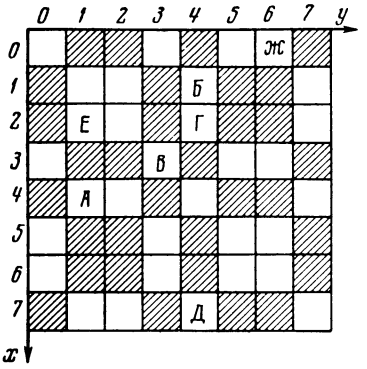


Рис. 5

Нас интересуют, однако, более обширные множества сигналов, требующие комбинаций с большим числом знаков при двоичной записи. Возьмем телеграфный код, служащий для передачи 32 букв. Обычный код является пятизначным (код Бодо), но мы сразу построим простейший код, обнаруживающий одиночную ошибку, приписав к пятизначным комбинациям еще 0 или 1 с таким расчетом, чтобы полученные шестизначные комбинации имели четное число единиц (или нулей). Для перевода шестимерного пространства на плоскость мы воспользуемся восьмеричной системой счисления. При этом

$$N=2^6=M=8^2=64.$$

Ниже даны буквы, их двоичная запись в виде шестизначных комбинаций, образованных из обычного кода Бодо, и соответствующая двузначная восьмеричная запись:

Буква	А	Б	В	Г	Д	Е	Ж
Двоичная запись	100001	001100	011011	010100	111100	010001	000110
Восьмеричная запись	41	14	33	24	74	21	06

Плоская таблица, соответствующая последней строке, имеет 8×8 клеток, как обычная шахматная доска. Половина клеток занята буквами, остальная половина соответствует запрещенным комбинациям (с нечетным числом нулей или единиц); эти клетки заштрихованы (см. рис. 5). Аналогичным образом может быть построен и двоичный код, исправляющий одиночную ошибку. Это будет девятизначный код [1]; его двумерным отображением послужит таблица из $23\times23=529$ клеток.

Для технического осуществления приемного устройства с отображением сигналов на плоскости может послужить скелетная схема, изображенная на рис. 6. Принимаемый двоичный сигнал вместе с наложенной на него помехой поступает в квантующее устройство Кв, которое относит принятый сигнал к одному из двух уровней и определяет, что именно принято: 0 или 1. Здесь и происходит возможная ошибка, вероятность которой подсчиты-

[1] *R. W. Hamming.* Error detecting and error correcting codes. Bell System Techn. J., 1950, v. 29, N 2.

вается обычным способом, если известно распределение вероятностей для помехи. Двоичный сигнал от квантующего устройства попадает в схему совпадений СС, туда же поступают импульсы от импульсного генератора ИГ, работающего синхронно с сигналом (он может запускаться стартовыми импульсами). На выходе СС получается двоичный сигнал стандартного уровня. Этот сигнал поступает на одно из двух декодирующих устройств $Д_1$ или $Д_2$ в зависимости от положения ключа К. Этот последний управляется от счетчика импульсов Сч.

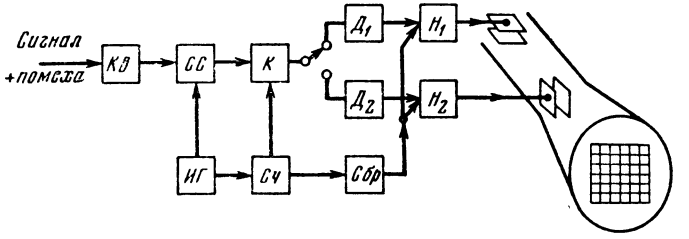


Рис. 6

Действие системы состоит в том, что, например, при шестизначном коде первая тройка двоичных знаков определяет первую цифру двузначного числа. Когда она определена, ключ перекладывается, и следующая тройка двоичных знаков, попадая на второе декодирующее устройство, определяет вторую цифру двузначного числа. Что касается декодирующих устройств, то это обычные декодирующие схемы КИМ, содержащие звено *RC* с постоянной времени, подобранной так, чтобы за тактовый период напряжение на конденсаторе убывало ровно вдвое. Этот принцип широко известен. На выходе декодирующих устройств стоят накопители Н, удерживающие окончательное значение напряжений, получаемых по окончании цикла работы декодирующих устройств. С накопителей напряжения подаются непосредственно на две пары отклоняющих пластин. Напряжение с накопителей снимается в нужный момент, задаваемый опять-таки счетчиком импульсов, при помощи сбрасывающей схемы Сбр. На экран трубки наложен транспарант в форме решетки (наподобие рис. 5).

Вышеприведенное описание представляет собой предварительный эскиз установки, задуманной в первую очередь в качестве демонстрационного устройства, а также для проведения некоторых исследований. Будущее покажет, смогут ли те или иные элементы подобного устройства найти применение в технике связи.

О ТЕОРЕТИЧЕСКИ-ОПТИМАЛЬНОЙ СИСТЕМЕ СВЯЗИ

В многочисленных работах последнего времени, посвященных общей теории связи, обсуждаются вопросы построения оптимальных кодов и вопросы идеального приема. По этому поводу следует заметить, что эти вопросы, вообще говоря, не должны были бы рассматриваться порознь. Выбор кода (т. е. способа передачи) и выбор способа приема — две стороны единой проблемы, состоящей в построении оптимальной системы связи. Эта проблема получает определенное решение, если заданы условия работы системы, в частности, характеристика воздействующей на систему помехи.

Условимся об определении оптимальной системы: назовем оптимальной такую систему, которая обладает наибольшей производительностью (эффективностью) при заданной помехоустойчивости, или, наоборот, обладает наибольшей помехоустойчивостью при заданной производительности.

Дадим количественное определение обоим названным показателям: будем выражать помехоустойчивость вероятностью правильного приема q_i; производительность будем характеризовать количеством информации на кодовую комбинацию, т. е. величиной $\log N$, где N — полное число кодовых комбинаций (полагаемых равновероятными).

Теперь попытаемся сформулировать задачу, пользуясь при этом геометрическим языком, уже достаточно привычным. Займемся сперва числом кодовых комбинаций. Если наложить обычное условие — равенство энергий всех кодовых комбинаций

$$E_i = \sum_{k=1}^{n} x_{ik}^2 = E = \text{const}, \tag{1}$$

где i—номер комбинации; n — число знаков (символов, координат), то задача сводится к размещению либо заданного, либо наибольшего числа кодовых точек на поверхности n-мерной сферы радиуса $\sqrt{E}$.

Что касается помехоустойчивости, то она зависит и от способа приема; задача может быть поставлена следующим образом. В результате наложения помехи точка принятого сигнала смещается относительно точки переданного сигнала. Действие приемника состоит в том, что он отождествляет принятый сигнал с i-м переданным в том случае, когда точка принятого сигнала оказывается

внутри некоторой n-мерной области Q_i-области правильного приема для i-го сигнала. При такой постановке вопроса вероятность q_i правильного приема i-го сигнала есть вероятность точке принятого сигнала при передаче i-го сигнала попасть в область Q_i. Конфигурация и расположение областей Q_i определяют способ действия приемника. Приемник, обеспечивающий наибольшую помехоустойчивость, можно назвать идеальным.

Вероятность точке принятого сигнала попасть в область Q_i зависит от многомерной плотности вероятностей помехи, представленной некоторой функцией $\varphi(x_1, x_2, \ldots, x_n)$. Теперь мы можем сформулировать задачу нахождения оптимальной системы для варианта, когда задано N, следующим образом: максимизировать величины

$$q_i = \int_{Q_i} \varphi(x_1 - x_{i1},\ x_2 - x_{i2},\ \ldots,\ x_n - x_{in})\, dx_1 dx_2 \ldots dx_n \qquad (2)$$

при дополнительном условии (1). В качестве второго дополнительного условия можно выставить равенство вероятностей правильного приема для всех сигналов

$$q_1 = q_2 = \ldots = q_N = q. \qquad (3)$$

Максимизировать q_i нужно путем одновременной вариации границы области Q_i (подбор способа приема) и вариации координат x_{ik} кодовых точек (подбор кода). Задача представляет собой, следовательно, совмещение вариационной задачи с подвижной границей с задачей на условный экстремум функции переменных. Ясно, что это вовсе не простая задача. Но мы и не ставим себе целью отыскание способов ее решения в общем виде. Общая формулировка задачи приводится здесь лишь для того, чтобы показать ее единство. Из выражения (2) непосредственно видно, что для данного кода (т. е. при заданных x_{ik}) можно найти наивыгоднейший приемник (т. е. наиболее благоприятную конфигурацию области Q_i), и наоборот. Распределение помехи (т. е. функция φ) предполагается заданным. Но можно представить себе случай, когда Q_i и x_{ik} заданы и ищется характеристика помехи, обеспечивающей при данных условиях наибольшую (а может быть, и наименьшую) помехоустойчивость. Короче говоря, код, способ приема и характеристика помехи взаимосвязаны и характер их взаимной зависимости выражается формулой (2).

Смысл приведенных общих соотношений мы постараемся пояснить на нескольких простейших примерах, рассмотрение которых не представляет математических трудностей.

Пример 1. Пусть имеется всего два сигнала ($N=2$) на расстоянии d друг от друга в n-мерном пространстве. Относительно плотности распределения помехи предположим, что она имеет сферическую симметрию, т. е. зависит только от

$$r = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Это совершенно естественное предположение, означающее, что все направления вектора помехи равновероятны. Пусть φ есть убывающая функция r (этими свойствами, в частности, обладает нормальное распределение). Для вероятностей правильного приема имеем

$$q_1 = \int_{Q_1} \varphi(x_1 - x_{11},\ x_2,\ \ldots,\ x_n)\, dx_1 dx_2 \ldots dx_n,$$

$$q_2 = \int_{Q_2} \varphi(x_1 - x_{21},\ x_2,\ \ldots,\ x_n)\, dx_1 dx_2 \ldots dx_n$$

(мы полагаем, что точки обоих сигналов лежат на оси X_1). Полагая

$$q_1 = q_2 = q,$$

выберем

$$x_{11} = d/2, \quad x_{21} = -d/2.$$

В силу симметрии, выражаемой равенством q_i, области Q_1 и Q_2 должны иметь зеркальную симметрию относительно гиперплоскости AA', нормальной к отрезку d и делящей его пополам, как показано на рис. 1, где область Q_2 заштрихована. Элементарным рассуждением можно показать, что для максимизации $q_1 = q_2$ нужно взять в качестве границы областей Q_1 и Q_2 плоскость симметрии AA'. В самом деле, рассмотрим пару элементов объема dQ_a и dQ_b, расположенных симметрично относительно AA', как показано на рис. 2. Пусть передается сигнал 1 и пусть dQ_b принад-

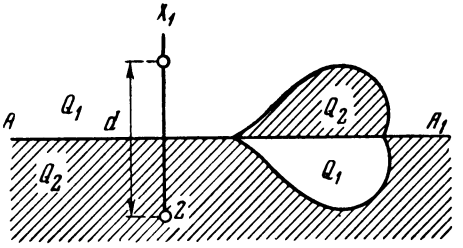


Рис. 1

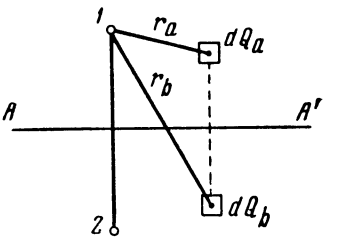


Рис. 2

лежит Q_1. Но так как $r_b > r_a$, то вероятность правильного приема q_1 возросла бы, если бы dQ_b был заменен dQ_a. Таким образом, dQ_b следует отнести к Q_2, а dQ_a к Q_1. Применяя такое рассуждение к любой паре симметричных элементов объема, приходим к заключению, что если плотность распределения помехи убывает с расстоянием (безразлично по какому закону), то идеальным будет приемник, относящий принятый сигнал к ближайшему возможному. Это есть приемник, идеальный в смысле Котельникова.

Остается найти оптимальный код. Но в данном простом случае едва ли нужно формально доказывать, что помехоустойчивость

растет с увеличением d и что, следовательно, принимая условие (1), следует расположить точки сигналов по концам диаметра сферы сигналов. Расстояние будет при этом $d=2\sqrt{E}$.

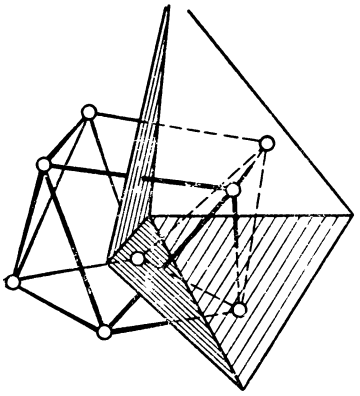


Рис. 3

Пример 2. Пусть по-прежнему плотность распределения помехи убывает с расстоянием, но возьмем теперь $N=8$, $n=3$. Точки сигналов располагаются на трехмерной сфере. Если выбрать двоичный код, то кодовые точки расположатся по вершинам куба. Области правильного приема будут представлять собой октанты, т. е. внутренности трехгранных углов, образованных плоскостями, нормальными к ребрам куба и делящими ребра пополам. Однако более выгодным оказывается недвоичный код, геометрический образ которого получается, если повернуть одну из граней куба в ее плоскости на 45°. Образованная таким путем фигура есть неправильный десятигранник (8 треугольников, 2 квадрата) с 16 равными ребрами, более длинными, чем ребро куба, вписанного в ту же сферу. Область правильного приема лежит внутри четырехгранного угла, как показано на рис. 3. Заметим, что рассматриваемый многогранник хотя и неправилен, но симметричен в том смысле, что при помещении в заданную точку любой вершины фигура совмещается сама с собой путем поворота относительно центра. Поэтому все области Q_i одинаковы, а следовательно, и вероятности q_i равны друг другу.

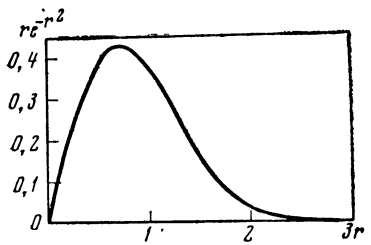


Рис. 4

Пример 3. Возьмем снова $N=2$ и пусть для простоты $n=2$ (что дает нам возможность изобразить всю картину на плоскости), но пусть теперь помеха имеет немонотонно изменяющуюся плотность вероятностей.

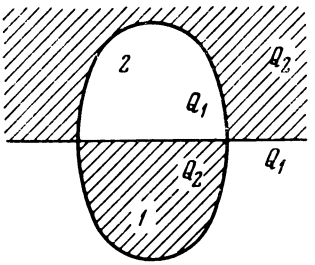


Рис. 5

Пусть, например,

$$\varphi(r) = Cre^{-r^2}$$

(рис. 4). Будем рассуждать о расположении границы области правильного приема в том же духе, как и в примере 1, но сформулируем вывод несколько иначе, а именно: каждый элемент объема dQ следует отнести к области правильного приема того сигнала, при приеме которого вероятность попадания в dQ больше. Иначе говоря, граница области правильного приема должна быть местом точек, где плотности вероятностей для двух сигналов становятся равными. В рассматриваемом случае это дает

$$[(x_1 - d/2)^2 + x_2^2]^{1/2} \exp - [(x_1 - d/2)^2 + x_2^2] =$$

$$= [(x_1 + d/2)^2 + x_2^2]^{1/2} \exp - [(x_1 + d/2)^2 + x_2^2]$$

или

$$x_1 d/\mathrm{th}\, 2x_1 d - r^2 + d^2/4 = 0,$$

где

$$r^2 = x_1^2 + x_2^2.$$

Граница области правильного приема показана на рис. 5. Очертание границы зависит, разумеется, от d.

КОДИРОВАНИЕ, УСТОЙЧИВОЕ ПО ОТНОШЕНИЮ К ЗАМИРАНИЮ[1] (АНТИФЭДИНГОВОЕ КОДИРОВАНИЕ)

Помехоустойчивые коды. Принцип построения современных помехоустойчивых кодов вкратце сводится к следующему: составляется набор кодовых комбинаций, различающихся между собой в достаточном числе знаков. Если при передаче одной из таких комбинаций некоторое число знаков заменится (в результате наложения помехи) неверными, то остающееся различие с другими возможными комбинациями должно все же позволить правильно идентифицировать переданную комбинацию. Число различающихся знаков называют обычно расстоянием (в метрике Хэмминга [1]).

Аддитивная и мультипликативная помехи. Вышеописанный принцип давно и с успехом применяется для борьбы с помехой, налагающейся на сигнал в том или ином звене тракта передачи. Эта помеха может быть импульсной или гладкой: большинство теоретических исследований относится к помехе в виде белого шума (гауссово распределение и равномерный спектр). Существенно лишь, что помеха предполагается аддитивной, т. е. на приемную сторону поступает сигнал плюс помеха. Другое важное предположение состоит в том, что помеха предполагается быстрой. Это значит, что если один знак в кодовой комбинации поражен помехой, то соседние могут остаться неповрежденными. Математически такое положение выражается тем, что поражение каждого знака рассматривается как независимое случайное событие.

Аддитивная помеха имеет большое значение. Но в ряде случаев большее значение может иметь мультипликативная помеха, связанная с изменениями условий передачи.

Этим термином обозначается явление, состоящее в том, что сигнал умножается на изменяющуюся со временем случайную величину (т. е. на случайный процесс). Говоря техническим языком, затухание тракта передачи флуктуирует. Явление носит общее название замирания (fading). Анализ его причин для наших целей излишен; существенно, однако, что замирание представляет собой относительно медленный процесс. В результате замирания на протяжении части времени передачи принимаемый сигнал оказывается ниже порога чувствительности приемной аппаратуры, и куски сигнала выпадают вовсе.

---

[1] Совместно с Э. Л. Блохом.

Способ передачи, ослабляющий влияние замирания. Известны и применяются методы борьбы с замиранием интерференционного происхождения. К числу таких методов относятся: прием на разнесенные антенны, применение широкополосных сигналов (предлагалось, в частности, применять для передачи сигналов модулированный шум [2]). Недавно появилось описание специальной системы Rake [3]. В настоящей заметке предлагается способ передачи в сочетании с применением помехоустойчивых кодов, состоящий в следующем.

Пусть передаваемое сообщение закодировано n-значными комбинациями некоторого равномерного кода. Возьмем группу из N таких комбинаций и запишем ее в табл. 1 таким образом, чтобы каждая кодовая комбинация занимала один столбец.

Таблица 1

	1	2	3	4	5	6	7	8	.	.	.	.	N
1	0	1	1	1	0	1	1	0	.	.	.	.	1
2	1	1	0	0	1	0	0	1	.	.	.	.	0
3	0	1	0	1	0	1	1	0	.	.	.	.	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
n	1	0	1	0	0	1	0	1					1

Выберем число N так, чтобы время передачи N двоичных знаков было достаточно велико по сравнению со средней продолжительностью замирания (точнее, со средней длительностью выпадения сигнала). Будем передавать табл. 1 не по столбцам (т. е. по исходным кодовым комбинациям), а по строкам. В результате замирания часть переданного сигнала выпадет. Обозначая выпавшие знаки символом *, получим на приеме примерно такую картину (табл. 2):

Таблица 2

	1	2	3	4	5	6	7	8	.	.	.	.	N
1	0	1	*	*	*	*	*	*	.	.	.	.	*
2	1	1	0	0	1	0	0	1	.	.	.	.	0
3	*	*	*	*	0	1	1	0	.	.	.	.	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
n	*	*	1	0	0	1	0	1	.	.	.	.	1

Если теперь сгруппировать принятые знаки по столбцам, то получим кодовые комбинации, из которых выпали отдельные знаки. При этом существенно, что если число N выбрано, сообразуясь со статистикой замирания, то выпадение отдельного знака в кодовой комбинации может считаться независимым событием. Суть дела сводится, таким образом, к тому, что выпавшие в результате замирания знаки разбрасываются случайно по кодовым комбинациям [1]. Заметим, что если в каждом столбце расположено по одной кодовой комбинации, то ошибки в одинаковых позициях соседних комбинаций сильно коррелированы. Однако, если разместить в одном столбце несколько кодовых комбинаций, представляющих некоторый отрезок сообщения, то с достаточно хорошим приближением ошибки можно уже считать независимыми не только внутри данной комбинации, но и в пределах отрезка сообщения. Именно это упрощающее предположение и лежит в основе формул раздела «Помехоустойчивость».

Применение помехоустойчивых кодов. Характер повреждения, наносимого сигналу аддитивной помехой и замиранием, существенно отличен. При аддитивной помехе некоторый знак заменяется неверным (например, нуль единицей или наоборот).

В случае же мультипликативной помехи типа замирания знак не заменяется другим, а вовсе выпадает; при синхронной передаче место выпавшего знака известно, и это создает значительно более благоприятные условия для построения надежной системы связи. Если требуется восстановить не более r выпавших знаков, то для этого достаточно применить код с расстоянием между кодовыми комбинациями не менее $r+1$*.

Сравнивая принятую комбинацию со всеми возможными, предварительно вычеркнув из них выпавшие позиции, получим, что принятая (и сокращенная таким образом) кодовая комбинация совпадает с той, которая была передана и отличается от всех остальных не менее чем в одном знаке. Таким образом, переданная комбинация может быть идентифицирована, а следовательно, все выпавшие знаки восстановлены. Если тот же код применяется в условиях действия аддитивной помехи, то он может лишь обнаружить ошибки, число которых не превосходит r; однако указать места этих ошибок, а следовательно, исправить их нельзя. (Для исправления не более чем r ошибок требуется код, расстояние между кодовыми комбинациями которого не менее $2r+1$.)

---

[1] Сущность описанного простого принципа можно наглядно пояснить аналогией с фототелеграфом. Известно, что фототелеграф относительно устойчив по отношению к замиранию. Это объясняется тем, что каждая буква передается не однократно, а несколькими своими «сечениями». Если одна или несколько строк выпали в результате замирания, то остающихся сечений может быть все-таки достаточно, чтобы букву можно было узнать.

Возможность декорреляции ошибок путем табличной записи сообщений использована (с другими целями) в [4].

* Этот результат приведен в [5], где рассматривается случай независимого выпадения отдельных символов передаваемой последовательности.

Коды с контрольными знаками. Общий метод приема, вытекающий из основной идеи помехоустойчивых кодов, состоит в сличении принятой кодовой комбинации со всеми возможными переданными. Однако практически более удобной оказывается программа проверок на четность, применяемая к кодам со специально добавленными так называемыми контрольными знаками (или позициями). Сущность и примеры такой программы приведены у Хэмминга [1] и во многих последующих работах. Здесь мы отметим лишь, что при совместном действии аддитивной и мультипликативной помех, вызывающем ошибки не более чем в r_1 знаках, и выпадении не более r_2 знаков следует применить код, обнаруживающий $r=r_1+r_2$ ошибок и исправляющий r_1 ошибок. (Расстояние между кодовыми комбинациями такого кода должно быть не менее чем $2r_1+r_2+1$.) Так, упомянутый у Хэмминга код, обнаруживающий две ошибки и исправляющий одну, способен в нашем случае исправить одну ошибку и восстановить один выпавший знак.

Помехоустойчивость. Будем, как обычно, характеризовать помехоустойчивость вероятностью безошибочного приема последовательности из L элементов сообщения. Параметром, определяющим действие замирания, может служить длительность выпадения сигнала, отнесенная к полному времени передачи. Это отношение мы обозначим через ε и будем считать малым.

При оценке помехоустойчивости предположим, что передаваемое сообщение предварительно закодировано двоичным статистическим кодом, так что последовательности из L элементов сообщения соответствует последовательность из M двоичных символов (0 и 1). Эта последняя разбивается на отрезки по m символов, для передачи которых используется упомянутый в начале статьи n-значный помехоустойчивый код, что приводит к уменьшению скорости передачи в n/m раз.

Величина n зависит от m и от минимального расстояния d между кодовыми комбинациями. Если код не восстанавливает и не исправляет ни одного кодового символа ($d=1$), то $n=m$.

Если код восстанавливает не более одного кодового символа ($d=2$), то $n=m+1$. (В общем случае для произвольного d величина n (m, d) неизвестна.)

Без применения исправляющих кодов $n=m$ вероятность безошибочного приема последовательности из M знаков равна

$$p_1=(1-\varepsilon)^M$$

или при $\varepsilon \ll 1$

$$p_1 \approx e^{-\varepsilon M}. \qquad (1)$$

При применении кода, восстанавливающего не более r знаков в каждой n-значной кодовой комбинации, вероятность правильного приема каждой комбинации равна

$$1-C_n^{r+1}\varepsilon^{r+1}(1-\varepsilon)^{n-r-1}-\ldots-\varepsilon^n,$$

и вероятность безошибочного приема отрезка из M знаков исходной последовательности будет

$$p_2^{(r)} = \left[1 - \sum_{k=r+1}^{n} C_n^k \varepsilon^k (1-\varepsilon)^{n-k}\right]^{M/m}$$

или при $\varepsilon \ll 1$

$$p_2^{(r)} \approx e^{-M/m} C_n^{r+1} \varepsilon^{r+1}. \tag{2}$$

В частности,

$$p_2^{(1)} \approx e^{-\frac{M}{m} \cdot \frac{n(n-1)}{2} \varepsilon^2} = e^{-M \frac{m+1}{2} \varepsilon^2}, \tag{3}$$

$$p_2^{(2)} \approx e^{-\frac{M}{m} \cdot \frac{n(n-1)(n-2)}{3!} \varepsilon^3}. \tag{4}$$

Из сопоставления (3) и (1) следует, что код, восстанавливающий один знак, целесообразно применять при условии

$$\frac{m+1}{2} \varepsilon < 1.$$

Аналогично, из (3) и (4) заключаем, что код, восстанавливающий два знака, следует применять только, если

$$\frac{n(n-1)(n-2)}{3m(m+1)} \varepsilon < 1.$$

Литература

1. *Р. В. Хэмминг.* Коды с обнаружением и исправлением ошибок. Сб.: Коды с обнаружением и исправлением ошибок. ИЛ, 1956.
2. *А. А. Харкевич.* Передача сигналов модулированным шумом. — Электросвязь, 1957, № 11.
3. *Price, Green.* Acommunication technique for multipath channels. — Proc. IRE, 1958, v. 46, N 3.
4. *П. Элиас.* Безошибочное кодирование. Сб.: Коды с обнаружением и исправлением ошибок. ИЛ, 1956.
5. *П. Элиас.* Кодирование для двух каналов с шумами. Сб.: Теория передачи сообщений. ИЛ, 1957.

НЕКОТОРЫЕ СВОЙСТВА СИСТЕМ СВЯЗИ С ЗАМИРАНИЕМ[1]

Введение. Теоретические построения общей теории связи образуют к настоящему времени стройную систему лишь применительно к случаю аддитивной флюктуационной помехи. Другими словами, предполагается, что принятый сигнал представляет собой сумму переданного сигнала и помехи. Конечно, этот случай имеет большое значение и его подробное исследование вполне оправдано. Но для радиосвязи не меньшее значение имеет и другой вид помехи, а именно мультипликативная помеха типа замирания (фэдинг). Действие этой помехи состоит в том, что сила принимаемого сигнала случайно изменяется, что можно себе представить как умножение передаваемого сигнала на некоторый случайный процесс. В результате замирания на протяжении отдельных промежутков времени сигнал лежит ниже порога чувствительности приемника, а поэтому некоторые отрезки сигнала вовсе выпадают. Более подробное исследование показывает, что ситуация при замирании существенно отличается от той, которая имеется в случае аддитивной шумовой помехи. Системам с замиранием были посвящены за последнее время многие работы. В данной статье сравниваются пропускные способности систем с аддитивной и мультипликативной помехами, а также рассматриваются некоторые вопросы, относящиеся к применению исправляющих кодов.

Характеристика исследуемых систем. О системах с аддитивной шумовой помехой говорить не приходится: они достаточно изучены и их общие свойства общеизвестны. Мы будем полагать, что в системе применяется двоичный сигнал и что действие помехи может заменить фактически переданный символ другим — неверным (т. е. 0 может замениться 1, и наоборот). Условия работы системы полностью определяются вероятностями перехода.

Что касается систем с мультипликативной помехой, то здесь обстановка иная: некоторый символ не заменяется другим, но вовсе выпадает. На приеме получается последовательность нулей, единиц и пропусков, образующихся на месте выпавших символов. Будем обозначать такой пропуск знаком $*$. При синхронной передаче места выпавших символов известны, что зна-

[1] Совместно с Э. Л. Блохом.

чительно облегчает задачу восстановления переданного сообщения.

Нужно различать два случая: система с активной паузой (ФМ, ЧМ) и система с пассивной паузой (АМ). К первой применимо все сказанное выше. По поводу второй надо пояснить, что выпадение нуля и единицы неравноценно. В самом деле: если 1 означает активный символ, а 0 — пассивную паузу, то выпадение единицы превращает ее в нуль, а выпадение нуля вообще ничего не меняет, и ясно, что этот случай существенно (и притом невыгодно) отличается от случая системы с активной паузой. В настоящее время преимущественно применяется система с активной паузой. Тем не менее ниже рассматриваются для сравнения все три случая, а именно: случай I — аддитивная помеха; случай II — замирание, система с активной паузой; случай III — замирание, система с пассивной паузой.

Случай I общеизвестен; случай II (в предположении независимости ошибок) рассмотрен Элиасом [1]. Случай III, по-видимому, не исследовался. Однако для полноты картины мы рассматриваем все три случая, применяя единую методику как при вычислении пропускной способности (п. 3), так и при обсуждении вопросов построения исправляющих кодов (п. 4).

Пропускная способность. При исследовании аддитивной помехи обычно предполагается, что ошибки в отдельных знаках независимы. В случае же мультипликативной помехи, когда выпадают целые отрезки сигнала, ошибки в соседних знаках сильно коррелированы. Однако, применяя специальный способ передачи [2], можно разрушить эту взаимосвязь и добиться практической независимости отдельных ошибок. При таких условиях можно получить общие выражения для пропускной способности системы связи для всех трех перечисленных выше случаев. Для вывода необходимо ввести условные вероятности переходов. Переданными символами могут быть 0 и 1. Принятыми символами могут быть 0, 1 и *. Запишем вероятности переходов в виде таблицы (матрицы переходов):

		Принятые символы		
		0	1	*
Переданные символы	0	ε_5	ε_1	ε_2
	1	ε_3	ε_6	ε_4

Из условий нормировки имеем

$$\varepsilon_5 = 1 - \varepsilon_1 - \varepsilon_2; \quad \varepsilon_6 = 1 - \varepsilon_3 - \varepsilon_4.$$

Введем, кроме того, априорные вероятности передачи нуля и единицы

$$p(0) = p; \quad p(1) = 1 - p$$

и вероятности приема символов 0, 1 и *

$$q_0 = p(1 - \varepsilon_1 - \varepsilon_2) + (1 - p)\varepsilon_3,$$

$$q_1 = p\varepsilon_1 + (1 - p)(1 - \varepsilon_3 - \varepsilon_4), \quad q_* = p\varepsilon_2 + (1 - p)\varepsilon_4.$$

Используя известное выражение для скорости передачи, имеем

$$R = H(q_0, q_1) - pH(\varepsilon_1, \varepsilon_2) - (1 - p)H(\varepsilon_3, \varepsilon_4),$$

где

$$H(x, y) = -x\log_2 x - y\log_2 y - \\ -(1 - x - y)\log_2(1 - x - y).$$

От этой общей формулы можно перейти к частным случаям, определяя пропускную способность, как

$$C = \max_{(p)} R. \tag{1}$$

Случай I

$$\varepsilon_2 = \varepsilon_4 = 0; \quad \varepsilon_1 = \varepsilon_3 = \varepsilon \neq 0^*,$$

и мы получаем известный результат

$$C = 1 - H(\varepsilon), \tag{2}$$

где

$$H(x) = -x\log_2 x - (1 - x)\log_2(1 - x).$$

Случай II

$$\varepsilon_1 = \varepsilon_3 = 0; \quad \varepsilon_2 = \varepsilon_4 = \varepsilon \neq 0.$$

При этом

$$C = 1 - \varepsilon. \tag{3}$$

Исключительная простота формулы (4) легко объясняется: ведь речь идет о случае, когда информация убывает просто пропорционально числу выпавших символов.

Случай III

$$\varepsilon_1 = \varepsilon_2 = \varepsilon_4 = 0, \quad \varepsilon_3 = \varepsilon \neq 0.$$

В этом случае для пропускной способности получается более сложное выражение

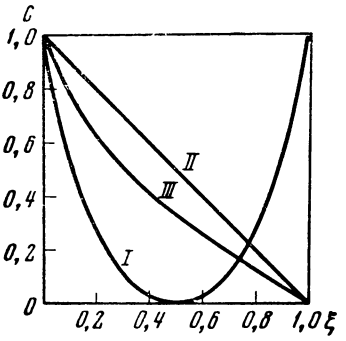
$$C = \log_2[1 + 2^{-H(\varepsilon)/(1-\varepsilon)}]. \tag{4}$$

Графики зависимости пропускной способности C от ε для всех трех рассмотренных случаев [по формулам (2)—(4)] даны на рисунке. Отметим, что в случае аддитивной помехи пропускная

* Мы ограничиваемся этим простым симметричным случаем. Не представляет затруднений исследование и более общего случая $\varepsilon_1 \neq \varepsilon_3$.

способность обращается в нуль при $\varepsilon=0{,}5$, после чего снова возрастает до единицы при $\varepsilon=1$, тогда как при замирании пропускная способность монотонно убывает, сохраняя отличное от нуля значение для всех $\varepsilon<1$.

Исправляющие коды. Общий принцип построения исправляющих равномерных кодов основан на способе приема, состоящем в том, что принятая (с ошибками) кодовая комбинация отождествляется с той из возможных переданных, от которой принятая комбинация наименее отличается. Отсюда следует, что в исправляющем коде должен быть достаточный запас различия между двумя комбинациями кода. Часть этого различия может утратиться в результате ошибок, но оставшееся различие должно быть достаточным, чтобы обеспечить возможность правильной идентификации принятой кодовой комбинации с фактически переданной.



Рисунок

Этот общий принцип сохраняет силу во всех трех рассматриваемых случаях; однако процесс сличения может выполняться по-разному.

Пользуясь общепринятыми геометрическими представлениями, будем называть число d несовпадающих знаков в двух кодовых комбинациях расстоянием. Будем рассматривать три кодовые комбинации, которые в дальнейшем для кратности обозначены буквами, а именно: A — фактически переданная комбинация, B — любая другая комбинация данного кода, C — принятая комбинация (при передаче комбинации A). A, B, C могут быть представлены точками в метрическом n-мерном пространстве (где n — число символов в кодовых комбинациях). Задача состоит в установлении наименьшего расстояния d между A и B, при котором C отождествляется с A, несмотря на наличие ошибок не более чем в r знаках. Для цельности изложения рассмотрим все три случая, хотя для первых двух результаты хорошо известны.

Случай 1. При наличии $r' \leqslant r$ ошибок расстояние между принятой и переданной комбинациями

$$d(AC)=r'.$$

Для того чтобы C отождествлялось с A, а не с B, необходимо и достаточно, чтобы расстояние до последней было больше, т. е. чтобы

$$d(CB)\geqslant r'+1 \qquad (5)$$

(r' — целое).

По определению расстояния (аксиома треугольника)

$$d(AB) \leqslant d(AC) + d(CB)$$

или

$$d(CB) \geqslant d(AB) - d(AC).$$

Отсюда видно, что условие (5) выполняется для всех $r' \leqslant r$, если

$$d(AB) \geqslant 2r + 1. \tag{6}$$

Итак, при числе неправильно принятых знаков, не превосходящем r, принятая кодовая комбинация будет правильно отождествлена с переданной при условии, что расстояние между любыми двумя комбинациями данного кода удовлетворяет неравенству (6).

Случай II. В этом случае действие помехи проявляется в том, что $r' \leqslant r$ знаков выпадают. Для нахождения переданной комбинации поступаем так: в принятой комбинации вычеркиваем все выпавшие позиции. Те же позиции вычеркиваем во всех комбинациях кода; таким образом, получаются укороченные комбинации (с геометрической точки зрения это означает проектирование n-мерного пространства сигналов на подпространство с числом измерений $n-r'$). Обозначим укороченные комбинации A', B' и C'. Очевидно, что A' и C' совпадают (так как в C' сохранены только неповрежденные знаки). Чтобы отличить C' от B', необходимо и достаточно, чтобы они отстояли друг от друга хотя бы на единицу

$$d(C'B') \geqslant 1.$$

Так как $d(A'C')=0$, то для $A'B'$ имеем

$$d(A'B') \geqslant 1. \tag{7}$$

Возвращаясь к полным (неукороченным) комбинациям, заметим, что в вычеркнутых позициях они могут различаться между собой не более чем в $r' \leqslant r$ знаках. Поэтому при выпадении $r' \leqslant r$ знаков условие (7) выполняется всегда, если

$$d(AB) \geqslant r + 1. \tag{8}$$

Случай III. В этом случае в результате действия помехи $r' \leqslant r$ единиц могут замениться нулями. Это можно толковать как обычные ошибки и, повторяя буквально относящиеся к случаю I рассуждения, прийти к формуле (6). Однако, используя особенности данного случая, можно указать более выгодный способ приема. Способ этот состоит в образовании укороченных комбинаций путем вычеркивания в принятой комбинации всех нулей и в вычеркивании во всех кодовых комбинациях тех же позиций.

Пусть A содержит a_1 единиц, B a_2 единиц, а C $a'=a_1-r'$ единиц. Прежде всего заметим, что выбирать для сравнения с C имеет смысл только такие кодовые комбинации, число единиц

в которых не более чем $a'+r$. Поэтому под B понимаются в дальнейшем лишь такие кодовые комбинации, для которых

$$a' \leqslant a_2 \leqslant a' + r^*. \tag{9}$$

Введем величину

$$\Delta a = a_2 - a' = (a_1 - a') + (a_2 - a_1) = r' \pm S,$$

где $S = |a_2 - a_1| \geqslant 0$.

Так как $a_2 \geqslant a'$, то по крайней мере Δa единиц в B приходится против нулей в C, что составляет расстояние $\Delta a = r' \pm S$. Но для правильного приема необходимо и достаточно (как и в предыдущем случае), чтобы $d(C'B') \geqslant 1$, т. е. против хотя бы одной из a' единиц в C должен стоять нуль в B. Но это значит, что необходимым и достаточным условием правильного приема является

$$d(CB) \geqslant r' \pm S + 1.$$

А так как $d(AC) = r'$, то для $d(AB)$ получаем условие

$$d(AB) \geqslant 2r' \pm S + 1. \tag{10}$$

Если $a_2 < a_1$ (т. е. $\Delta a = r' - S$), то r' ограничено лишь условием $r' \leqslant r$ и, следовательно, условие (10) выполняется для всех $r' \leqslant r$, только если

$$d(AB) \geqslant 2r - S + 1.$$

Если $a_2 > a_1$ (т. е. $\Delta a = r' + S$), то r' ограничено условием (9), которое означает, что $\Delta a \leqslant r$ и, следовательно,

$$r' + S \leqslant r \quad \text{или} \quad r' \leqslant r - S.$$

Но в этом случае условие (10)

$$d(AB) \geqslant 2r' + S + 1$$

будет выполняться для всех $r' \leqslant r - S$, только если

$$d(AB) \geqslant 2(r - S) + S + 1 = 2r - S + 1,$$

что совпадает со случаем $a_2 < a_1$.

Так как $d(AB)$ имеет ту же четность, что и S, равенства в последнем выражении быть не может и его следует записать либо в виде

$$d(AB) > 2r - S,$$

либо в виде

$$d(AB) \geqslant 2r - S + 2. \tag{11}$$

---

* Рассмотрение случая $a_2 < a'$ не представляет интереса, так как в этом случае B и C заведомо не могут совпадать.

Таким образом, для того чтобы принятый сигнал можно было правильно идентифицировать с переданным при условии, что вследствие помех не более чем r единиц могут быть заменены нулями, необходимо и достаточно, чтобы расстояние между любыми двумя кодовыми комбинациями, число единиц в которых отличается на величину S, удовлетворяло условию (11).

Сопоставляя формулы (6), (8) и (11), мы видим, что случай I требует наибольшего расстояния и, следовательно, наиболее длинных кодовых комбинаций. Случай II требует наименьшего расстояния, случай III занимает промежуточное положение. Это обусловлено тем, что в случае I нам неизвестно положение правильно принятых символов; в случае III места правильных символов частично известны (так как все принятые единицы заведомо правильны); в случае же II места всех правильных символов известны (так как все невыпавшие символы правильны). В соответствии с этим, применяя формулы (6), (8) и (11), не нужно забывать, что величина r имеет в них различный смысл, а именно:

Таблица 1

I	II	III
00000	000	0000
10011	011	0101
11100	101	1010
01111	110	1111

Таблица 2

I	II	III	I	II	III
000000	0000	00000	111000	1001	11100
010101	0011	01001	101101	1010	01111
100110	0101	00110	011110	1100	
110011	0110	10011	001011	1111	

Таблица 3

I	II	III
0000000	00000	000000
1101001	00011	100001
0101010	00101	010010
1000011	00110	001100
1001100	01001	010101
0100101	01010	100110
1100110	01100	111000
0001111	01111	001011
1110000	10001	110011
0011001	10010	101101
1011010	10100	011110
0110011	10111	111111
0111100	11000	
1010101	11011	
0010110	11101	
1111111	11110	

Таблица 4

I	II	III
00000000	00000000	00000000
11111111	11010010	00001111
	01010101	11110000
	10000111	11111111
	10011001	
	01001011	
	11001100	
	00011110	
	11100001	
	00110011	
	10111000	
	01100110	
	01111000	
	10101010	
	00101101	
	11111111	

в (6) r означает максимально допустимое число ошибок (замена нуля на единицу и обратно), в (8) — максимально допустимое число выпавших знаков, в (11) — максимально допустимое число единиц, замененных нулями.

Примеры исправляющих кодов. В заключение приведем примеры исправляющих кодов для всех трех рассматриваемых случаев.

В табл. 1, 2 и 3 приведены коды, исправляющие одиночную ошибку (причем в табл. 3 семизначный код, исправляющий ошибку для случая I, заимствован из работы [3]).

В табл. 4 приведены восьмизначные коды, исправляющие до трех ошибок (причем код для случая II заимствован из работы [3], где он приведен как пример кода, исправляющего одиночную и обнаруживающего двойную ошибку в случае I).

Литература

1. *П. Элиас.* Кодирование для двух каналов с шумами. Сб.: Теория передачи сообщений. ИЛ, 1957.
2. *Э. Л. Блох, А. А. Харкевич.* Кодирование, устойчивое по отношению к замиранию (антифэдинговое кодирование). — Электросвязь, 1960, № 4.
3. *Р. В. Хемминг.* Коды с обнаружением и исправлением ошибок. Сб.: Коды с обнаружением и исправлением ошибок. ИЛ, 1956.

АСИМПТОТИЧЕСКИЕ ВЫРАЖЕНИЯ СКОРОСТИ ПЕРЕДАЧИ ПРИ ВЫСОКОЙ НАДЕЖНОСТИ

При увеличении скорости передачи надежность, вообще говоря, убывает. В некоторых частных предположениях можно получить очень простые асимптотические формулы, связывающие скорость и надежность передачи.

Пусть двоичные символы передаются посредством двух функций $S_1(t)$ и $S_2(t)$ на интервале T. Наивыгоднейший случай есть $S_2 = -S_1$ [1]. Идеальный приемник дает при этом на выходе отношение сигнал/помеха

$$\rho = 8FT\rho_0, \tag{1}$$

где F — полоса частот; ρ_0 — отношение сигнал/помеха на входе.

Вероятность ошибки равна

$$p/\delta = \frac{1}{2}[1 - \Phi(z)], \tag{2}$$

где $z^2 = {}^1/_8\rho$.

Формула (1) выводится в предположении, что помеха имеет равномерный спектр, а формула (2) относится к случаю помехи с нормальным распределением.

В качестве меры надежности введем

$$S = \lg\frac{1}{P}. \tag{3}$$

При высокой надежности можно воспользоваться асимптотическим представлением (2)

$$p \sim \frac{1}{2}\cdot\frac{e^{-z^2}}{\sqrt{\pi}\,z}\left(1 - \frac{1}{2z^2} + \frac{1\cdot 3}{(2z^2)^2} - \dots\right)$$

и сохранить только первый член этого разложения. Логарифмируя, получим

$$\ln\frac{1}{p} = z^2 + \ln 2\sqrt{\pi}\,z.$$

Отбрасывая второй член и переходя к десятичным логарифмам, получим

$$S = 0{,}054\rho. \tag{4}$$

[1] *В. А. Котельников.* Теория потенциальной помехоустойчивости. Энергоиздат, 1956.

Если на передачу каждого символа затрачивается время T, то скорость передачи, т. е. число символов в секунду, равна

$$R = 1/T = 8F\rho_0/\rho.$$

Подставляя значения ρ_0 и ρ, находим

$$RS = 0{,}43F\rho_0 = 0{,}43P/N_0, \tag{5}$$

где P — мощность сигнала; N_0 — спектральная плотность помехи. Таким образом, RS=const, так как P и N_0 — заданные величины.

Интересно сравнить наш результат с пропускной способностью непрерывного канала. По Шеннону,

$$C = F\log_2(1 + P/N) = F\log(1 + \rho_0).$$

Перепишем эту формулу в виде

$$C = F\rho_0\log_2(1 + \rho_0)^{1/\rho_0} = kF\rho_0 = kP/N_0.$$

При ρ_0=1 (т. е. при P=N) k=1, а при $\rho \to 0$, $k \to \log_2 e$=1,44. Принимая это предельное значение, получим C=1,44P/N_0. Сопоставляя с ранее найденным выражением, находим

$$RS = 0{,}3C. \tag{6}$$

При передаче с пассивной паузой численные коэффициенты в формулах (4)—(6) в четыре раза меньше. Под P при этом понимается мощность в посылке.

Пример. При передаче с активной паузой пусть $P/N_0 = F\rho_0 = 10^3$ *сек*$^{-1}$ (например, F=1 *Мгц*, $\rho_0 = 10^{-3}$). Задана надежность S=8 ($p = 10^{-8}$). Тогда скорость передачи не должна превышать

$$R = 0{,}43P/N_0S = 0{,}43 \cdot 10^3/8 = 54 \text{ } дв.\ ед./сек,$$

т. е. длительность каждого символа должна быть не менее

$$T = 18{,}5 \text{ } м/сек.$$

При S=12 получим соответственно R=36 *дв. ед./сек* и T= =28 *мсек*. Ширина спектра сигнала имеет порядок R (*гц*), так что для передачи можно занимать полосу от $\approx$100 *гц* и более. Если отношение P/N_0 задано, то ни на скорость передачи, ни на надежность ширина полосы не влияет.

БОРЬБА С ПОМЕХАМИ

Предисловие

Передача информации посредством электрических сигналов играет очень большую и все возрастающую роль во всех видах человеческой деятельности. За последнее время резко повысились требования, предъявляемые к системам передачи информации. Необходимо вести передачу со все большими скоростями, на все большие расстояния, счет которых ведется уже не на тысячи, а на миллионы километров. Дело усложняется тем, что зачастую энергетические ресурсы передатчика жестко ограничены. И в то же время все более высокие требования предъявляются к верности передачи.

Верность зависит, с одной стороны, от исправности аппаратуры; этой стороны дела мы вовсе не будем касаться. С другой же стороны, верность зависит от помех, действующих в канале передачи.

Способность системы передачи противостоять вредному влиянию помех называется помехоустойчивостью. В современных условиях проблема помехоустойчивости выдвигается на передний план. Она останется важнейшей проблемой в области передачи информации и в предвидимом будущем.

Основы теории помехоустойчивости заложены В. А. Котельниковым в его выдающейся работе «Теория потенциальной помехоустойчивости». В этой работе впервые поставлены и решены многие основные задачи и введен ряд фундаментальных понятий. Она с полным основанием считается классическим исследованием проблемы помехоустойчивости.

Работа В. А. Котельникова появилась в 1946 г., намного опередив тогдашний уровень науки и техники. С тех пор немало сделано. Так, появились теория корректирующих кодов, теория оптимальных фильтров, вопросы приема сигналов стали рассматриваться с позиций теории статистических решений и т. д.

Существенно и то, что в связи с многообразными практическими применениями техники передачи информации чрезвычайно расширился круг специалистов, непосредственно соприкасающихся в своей повседневной деятельности с теми или иными сторонами проблемы помехоустойчивости. Это относится прежде всего к инженерам-исследователям, инженерам-разработчикам и инже-

нерам-эксплуатационникам, работающим в области техники передачи информации. Эта обширная категория работников нуждается в современном, но доступном и компактном изложении основных идей и фактов теории помехоустойчивости.

Предлагаемая монография и представляет собой попытку такого изложения. К сожалению, обстоятельства не позволили осуществить намеченную переработку книги, и во второе ее издание внесены лишь незначительные исправления.

Свои материалы (использованные в § 12, 20, 21, 27 и Д. VI) любезно предоставили мне Э. Л. Блох и О. В. Попов, которым приношу свою благодарность.

Я искренне благодарен также Б. Р. Левину за замечания, сделанные им по рукописи.

Автор

§ 1. Система передачи

Общая задача состоит в передаче некоторого сообщения. Сообщение создается источником сообщений. Источник, с которым имеет дело данная система передачи, выбирает сообщение из некоторого множества возможных сообщений. Во многих случаях это множество конечно.

Так, например, при передаче словесного текста каждая буква принадлежит конечному множеству, образующему алфавит, а каждое слово — конечному множеству, образующему словарь. Множество сообщений, заданное совместно с их априорными вероятностями, называют ансамблем сообщений.

Интересуясь технической стороной дела, мы будем представлять сообщение в виде некоторой функции $m(t)$. Это может быть как непрерывная функция непрерывного времени, так и последовательность чисел, т. е. функция дискретного времени. Об этом подробнее говорится в следующем параграфе.

Процесс передачи в общих чертах сводится к следующему. Прежде всего вырабатывается сигнал $s(t)$, отображающий сообщение. Формирование сигнала может включать операции кодирования и модуляции. Модуляция состоит в изменениях тех или иных параметров функции

$$f = f(a, b, c, \ldots, t),$$

называемой переносчиком. В качестве переносчика чаще всего применяются синусоидальные колебания высокой частоты или периодическая последовательность импульсов.

Операцию формирования сигнала можно кратко представить в виде

$$s = U(m, f), \tag{1.1}$$

где U — символ оператора, вообще говоря, нелинейного.

Далее сигнал передается по линии, в которой действует помеха $\xi(t)$*. Помеха представляет собой случайный процесс, задаваемый своими распределениями [1] или теми или иными моментами распределений. Обычно ограничиваются стационарными процессами, для которых перечисленные характеристики не зависят от начала отсчета времени.

Взаимодействие сигнала и помехи, происходящее в линии, можно выразить оператором

$$x = V(s, \xi). \tag{1.2}$$

Здесь x означает видоизмененный действием помехи сигнал на выходе линии.

Сигнал $x(t)$ поступает на вход приемника, который производит над ним некоторую операцию W, так что на выходе приемника имеем

$$y = W(x). \tag{1.3}$$

Объединяя выражения (1.1) — (1,3), запишем

$$y = W\{V[\xi, U(m, f)]\}. \tag{1.4}$$

Эта запись выражает действие системы передачи в целом.

Задача состоит в том, чтобы получить y, наименее уклоняющееся от m с точки зрения некоторого определенного критерия. Критерий не предуказан. Его выбор определяется в каждом данном случае обстоятельствами дела и требует широкой оценки всей ситуации. Заданными являются обычно: сообщение m, как член некоторого определенного ансамбля, помеха ξ, известная в результате статистического исследования условий передачи, и оператор V, характеризующий взаимодействие сигнала и помехи. Искомыми, таким образом, являются: переносчик f и операторы U и W, первый из которых определяет способ действия передатчика (т. е. правило образования сигнала), а второй — способ действия приемника.

Если представить выходной сигнал в виде

$$y = m + \varepsilon,$$

то задача состоит в подборе операторов U и W, минимизирующих ε (с точки зрения выбранного критерия, см. § 5).

Совершенно очевидно, что в столь общей постановке задача практически неразрешима (за исключением некоторых простейших вырожденных случаев). Но приведенная формулировка важна в том отношении, что она ясно показывает единство задачи.

---

* Предполагается, как это обычно делают, что помеха локализована в линии. В действительности источники различных помех могут действовать в разных звеньях системы передачи. В таком случае ξ есть эквивалентная суммарная помеха.

[1] Мы говорим для краткости «распределение» вместо «плотность распределения вероятностей».

Иначе говоря, выбор оптимальной системы сигналов и выбор оптимального способа приема — это две стороны одной и той же задачи — построения оптимальной системы передачи.

Однако для получения практических результатов задачу приходится расчленять. Суть дела сводится к тому, что прежде всего выбирают переносчик f. Затем, считая оператор U заданным, а следовательно, полагая заданным и сигнал x на входе приемника, стремятся построить оптимальный приемник. В такой постановке искомым является только оператор W. Так же раздельно решается и вопрос о построении сигналов, т. е. о выборе опера-

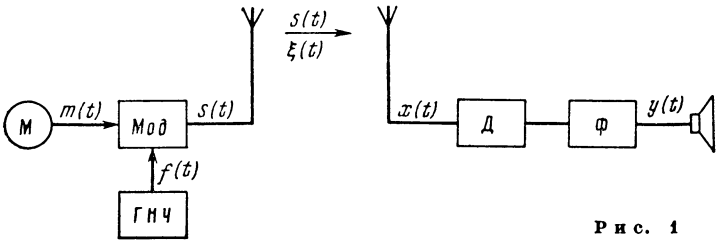


Рис. 1

тора U. Разумеется, при этом учитываются простейшие аспекты взаимосвязи этих операторов. Так, например, при выборе амплитудной модуляции (АМ) приемник снабжают амплитудным детектором, а при частотной модуляции (ЧМ) — частотным.

Можно в общих словах сформулировать принципы, которыми нужно руководствоваться при раздельном выборе операторов U и W. При построении системы сигналов нужно стремиться к тому, чтобы сигналы, соответствующие различным сообщениям, как можно более отличались друг от друга, чтобы действующая в системе передачи помеха как можно менее влияла на это отличие. При построении приемника нужно стремиться к тому, чтобы он по возможности подавлял помеху, т. е. очищал сигнал от искажений, вызванных помехой.

В этом смысле можно говорить о наилучших кодах, о наиболее помехоустойчивых видах модуляции, об оптимальных приемниках. Мы можем получить оптимальные решения для отдельных звеньев системы передачи, считая остальные заданными. Так пока что делают все, мы будем держаться той же методической линии. Это позволяет получить если не теоретически наилучшие, то по меньшей мере хорошие и работоспособные системы передачи, могущие удовлетворить весьма высоким требованиям по всем основным показателям.

В заключение поясним на простом примере обозначения, которыми мы воспользовались выше. В качестве примера возьмем обычный радиотелефон. На схеме рис. 1 изображены основные элементы системы передачи (усилители опущены): М — микрофон, Мод — модулятор, ГНЧ — генератор несущей частоты, Д —

детектор, Ф — фильтр. Операция U сводится здесь к простой модуляции сообщением $m(t)$ переносчика $f(t)$. Так, при обычной АМ имеем

$$s(t) = U(m, f) = [1 + \alpha m(t)] f(t),$$

где α — коэффициент (глубина) модуляции (предполагается, что $|m(t)| \leqslant 1$). Далее, если $\xi(t)$ есть аддитивный шум, то

$$x(t) = V(s, \xi) = s(t) + \xi(t).$$

И, наконец, если в приемнике, состоящем из детектора и фильтра, применен двухтактный линейный детектор, то действие приемника может быть представлено формулой

$$y(t) = W(x) = \int_{-\infty}^{t} |x(\tau)| g(t - \tau) d\tau,$$

где $g(t)$ — импульсная реакция фильтра.

Итак, в рассмотренном примере операция U сводится к умножению, операция V — к сложению. Выполняемая приемником операция W состоит в нелинейном преобразовании с последующим интегрированием с весом.

Если в операцию U входит кодирование, то ее представление в аналитической форме в общем случае затруднительно, так как код задается обычно не формулой, а таблицей.

§ 2. Сигналы

Рассмотрим некоторые подробности, относящиеся к формированию сигналов, и в первую очередь вопрос о дискретизации.

Сообщение может иметь дискретную природу, т. е. состоять из отдельных символов. В этом случае сигнал составляется из отдельных элементов и представляет собой дискретную последовательность. Примером может служить передача текста по телеграфу: символы сообщения — это буквы, соответствующие им элементы сигнала — кодовые комбинации телеграфного кода.

Таким образом, операция образования сигнала начинается в рассматриваемом случае с кодирования. Правило кодирования выражается кодовой таблицей, в которой каждому символу сообщения сопоставляется определенная кодовая комбинация. Следуя этому принципу, можно, разумеется, кодировать не только отдельные буквы, но и целые слова или фразы. При этом кодовая таблица разрастается до размеров кодовой книги (примеры: коммерческие коды, морской сигнальный код и т. п.).

Сообщение может представлять собой и непрерывную функцию времени. В простейшем случае эта функция непосредственно используется в качестве сигнала. Так обстоит, например, дело

при обычной городской телефонной связи. Для передачи на большие расстояния прибегают к модуляции, к которой и сводится образование сигнала.

Если мы желаем воспользоваться для передачи непрерывной функции импульсными или кодовыми методами, то нужно произвести дискретизацию функции по времени, т. е. перейти от функции непрерывного аргумента к функции дискретного аргумента. Эта операция выполняется путем взятия отсчетов функции в определенные дискретные моменты t_k. В результате функция $m(t)$ заменяется совокупностью мгновенных значений

$$\{m_k\} = \{m(t_k)\}.$$

Обычно моменты отсчетов располагаются по оси времени равномерно, т. е.

$$t_k = k\Delta t.$$

Выбор интервала Δt производится на основании теоремы Котельникова, которая гласит:

Функция с ограниченным спектром полностью определяется своими значениями, отсчитанными через интервалы

$$\Delta t = {}^1/_2 F,$$

где F — ширина спектра.

Это положение может применяться и к функциям с неограниченным, но быстро убывающим за пределами интервала F спектром. В таком случае функция восстанавливается по своим отсчетам не точно, но с легко оцениваемым приближением [1].

Исходное сообщение может представлять собой функцию не одного, а многих аргументов. В этом случае она превращается в функцию $m(t)$, зависящую от одного аргумента. Это осуществляется посредством операции, называемой вообще разверткой. При этом может произойти дискретизация по одному или нескольким, или всем аргументам. Примером может послужить образование телевизионного сигнала. Изображение может быть представлено как $B(x, y, t)$, где x и y — пространственные координаты (координаты плоскости изображения); B — яркость. Время дискретизируется в результате покадровой передачи ($\Delta t = {}^1/_{25}$ *сек*). При обычной строчной развертке координата x (вдоль строки) остается непрерывной, а координата y дискретизируется. Шаг Δy определяется числом строк развертки. Таким образом, получается функция

$$m(t) = m(i\Delta y,\ k\Delta t,\ vt),$$

где v — скорость развертки вдоль строки; i — номер строки; k — номер кадра.

[1] Теорема Котельникова в своем первоначальном виде относилась к детерминированным функциям. В дальнейшем она была распространена на случайные функции с практически ограниченным спектром (т. е. на функции, спектральная плотность которых вне конечной полосы достаточно мала).

До сих пор речь шла о дискретизации по аргументам. Но возможна (а иногда необходима) дискретизация по значениям функции. Предполагается, что функция ограничена, т. е. ее значения лежат в конечном интервале. В таком случае дискретизация состоит в замене несчетного множества возможных значений функции конечным множеством. Обычно дискретные значения располагаются по шкале функции равномерно, так что $m_i = i\Delta m$. Дискретизация значений функции носит название квантования. Если при квантовании любое значение m заменяется ближайшим значением m_i, то операция квантования может быть выражена формулой

$$m_i = [m/\Delta m + {}^1/_2]\,\Delta m,$$

где скобки означают «целая часть», Δm — шаг квантования.

Само собой квантование, заменяющее истинное значение округленным квантованным значением m_i, вносит погрешность

$$\varepsilon = m - m_i.$$

Однако существенно, что эта погрешность не превосходит половины шага квантования и, следовательно, находится под нашим контролем.

Квантование принципиально необходимо при применении кодовых методов передачи, так как кодовая таблица должна быть конечна.

Итак, при импульсной передаче необходима дискретизация по времени, а при кодовой передаче, кроме того, и дискретизация по значениям функции, т. е. квантование.

Обратимся к вопросам модуляции. Берется некоторая функция

$$f = (a,\ b,\ c,\ \ldots,\ t),$$

называемая переносчиком. Величины a, b, c ... представляют собой в отсутствие модуляции постоянные параметры. Сущность модуляции состоит в том, что один из параметров получает приращение, пропорциональное передаваемому сообщению, например

$$a = a_0 + \delta a = a_0 + \Delta a m\,(t) = a_0\left[1 + \frac{\Delta a}{a_0}\, m\,(t)\right],$$

где δa — переменное приращение; Δa — постоянная величина, выражающая степень изменения параметра. Если $|m\,(t)| \leqslant 1$, то отношение $\Delta a/a_0$ есть наибольшее относительное изменение параметра a, или глубина модуляции.

Таким же образом может изменяться и любой параметр. Если изменяется (модулируется) параметр a, то мы имеем a-модуляцию, если изменяется параметр b, то имеем b-модуляцию и т. д. Число возможных видов модуляции при данном переносчике равно числу его параметров. Так, например, если в качестве переносчика выбрано синусоидальное колебание

$$f\,(t) = A \sin\,(\omega t + \psi),$$

то параметрами являются амплитуда A, частота ω и начальная фаза φ. Каждый из этих параметров можно модулировать, и мы получаем соответственно АМ, ЧМ и ФМ.

Если переносчиком является периодическая последовательность импульсов определенной формы, например прямоугольной, то параметрами являются: высота (амплитуда), длительность, частота следования и фаза (т. е. положение импульса относительно тактовой точки). Это дает четыре основных вида импульсной модуляции: АИМ, ДИМ, ЧИМ и ФИМ. Переход от видеоимпульсов к радиоимпульсам позволяет получить еще два вида модуляции: по частоте и по фазе высокочастотного заполнения.

Возможны в принципе многочисленные виды модуляции по параметрам, определяющим форму видеоимпульсов; однако в практике такие виды модуляции применения пока не имеют.

В качестве переносчика можно воспользоваться не только периодической функцией, но и стационарным случайным процессом. В этом случае в качестве модулируемого параметра можно взять любую числовую характеристику, которая в силу стационарности является, по определению, постоянной (т. е. не зависящей от начала отсчета времени) величиной. Таковы, к примеру, моменты распределения или их Фурье-преобразования. Первый момент, т. е. среднее значение, обычно равен нулю. Второй момент есть функция корреляции, зависящая от временного сдвига τ. Фурье-преобразование функции корреляции есть спектр мощности. Второй момент при $\tau=0$ есть просто мощность. Модуляция по мощности представляет собой аналогию амплитудной модуляции. Модуляция по положению спектра на шкале частот есть нечто вроде частотной модуляции. Аналога фазовой модуляции для случайного процесса не существует.

Следует иметь в виду, что мощность, определенная для конечного отрезка реализации случайного процесса, есть случайная величина, флюктуирующая около среднего значения. То же относится и к любым другим моментам или их преобразованиям. Поэтому при использовании случайного процесса в качестве переносчика в сигнал с самого начала примешивается специфическая помеха хотя и не устранимая, но с известными статистическими характеристиками.

Переходя в заключение к вопросу о кодах, заметим, что в рамках нашей темы нас интересуют только помехоустойчивые коды, о которых дальше придется говорить со всеми необходимыми подробностями. Поэтому ограничимся здесь лишь пояснением некоторых общих терминов.

Кодом будем называть совокупность условных сигналов, обозначающих дискретные сообщения. Код представляют обычно таблицей, в которой приведен список сообщений, каждому из которых сопоставляется условное кодовое обозначение.

Код строится из элементов. Число различных элементов называется основанием кода. Код с основанием два называется

двоичным, код с основанием три — троичным и. т. д. Условные сигналы, составляющие код, называются кодовыми комбинациями. Число элементов или знаков, образующих комбинацию, называется значностью кода. Коды, все комбинации которых имеют одинаковое число знаков, называются равномерными. Пример: телеграфный код Бодо является равномерным пятизначным двоичным кодом.

Для равномерного кода число всех возможных кодовых комбинаций выражается формулой

$$N_0 = m^n,$$

где m — основание кода; n — значность кода. Код Бодо содержит $N_0=2^5=32$ кодовые комбинации, чего как раз достаточно для обозначения всех букв алфавита.

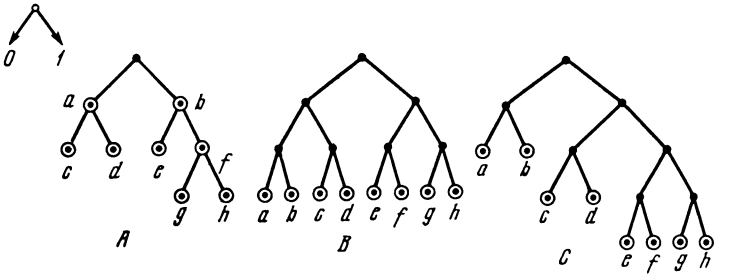


Рис. 2

Неравномерные коды требуют либо специальных разделительных знаков, указывающих конец одной и начало другой кодовой комбинации, либо же должны строиться так, чтобы никакая кодовая комбинация не являлась началом другой. Коды, удовлетворяющие этому последнему условию, называются неприводимыми. Кодовые комбинации представляются обычно в виде чисел, записанных по системе счисления с основанием, равным основанию кода. Так, элементы двоичного кода могут быть представлены двоичными цифрами 0 и 1, и кодовая комбинация представляется n-значным двоичным числом, т. е. последовательностью из нулей и единиц, содержащей n позиций. Кодовые комбинации троичного кода могут быть представлены троичными числами, составленными из цифр 0, 1, 2 и т. д. Так, например, телеграфный код Морзе, будучи неравномерным, требует разделительного знака (паузы между буквами). Обозначая точку нулем, тире единицей и паузу двойкой, получим следующую цифровую запись по троичной системе закодированного кодом Морзе слова «Москва»

11211120002101201120122

(последние две двойки означают двойную паузу, отделяющую одно слово от другого).

Нужно отметить и подчеркнуть, что строение кода никак не связано с физическим осуществлением различных элементов.

Двоичный код характеризуется наличием двух разных элементов; физический характер этого различия не играет никакой роли. Два элемента двоичного кода, 0 и 1, могут представлять собой два сигнала, различающиеся между собой либо амплитудой, либо частотой, либо фазой, либо длительностью, либо, наконец, формой в самом общем смысле. Физическое различие элементов кода — это вопрос вида модуляции, а не строения кода.

Строение кода удобно представлять в виде кодового дерева. Из каждого узла исходит число ветвей, равное основанию кода.

Для примера на рис. 2 показаны кодовые деревья для трех разных двоичных кодов. Шаг влево означает 0, шаг вправо — 1. Общее число сообщений равно 8, сообщения обозначены буквами от *a* до *h*. Кодовые обозначения даны ниже:

0	1	00	01	10	11	110	111
000	001	010	011	100	101	110	111
00	01	100	101	1100	1101	1110	1111

Код *A* — неравномерный. Он требует разделительных знаков между комбинациями. Код *B* является равномерным. Код *C* есть неравномерный неприводимый код, не требующий разделительных знаков. Любая последовательность комбинаций кода *C* легко расшифровывается при помощи кодового дерева. Например, последовательность

$$11011011111000011100$$

расшифровывается как *fdhaabe*.

Интересно отметить, что равномерный код также является неприводимым. Поэтому при работе с равномерным кодом (вопреки распространенному мнению) в принципе не требуется ни синхронного, ни стартстопного режима. Требуется лишь знать начало последовательности кодовых комбинаций. Однако, кроме того, требуется, чтобы ни один знак последовательности не выпал и не был принят ошибочно. С этой точки зрения стартстопная работа может рассматриваться как средство повышения надежности. Сказанное относится к любому неприводимому коду.

§ 3. Помехи

Помехой называется стороннее возмущение, действующее в системе передачи и препятствующее правильному приемнику сигналов.

Источники помех могут находиться как вне, так и внутри самой системы передачи.

Если помеха регулярна и известна, то борьба с ней не представляет затруднений. Так, фон переменного тока может быть устранен простейшей компенсацией; помеха от определенной радиостанции с модуляционным спектром нормальной ширины устраняется соответствующим фильтром.

Такого рода помехи нас не будут интересовать. Мы будем заниматься только непредсказуемыми случайными помехами. Очевидно, что борьба со случайными помехами представляет наибольшие трудности.

В общем виде влияние помехи ξ на передаваемый сигнал может быть выражено оператором

$$x = V(s, \xi). \tag{3.1}$$

В том частном случае, когда этот оператор вырождается в сумму

$$x = s + \xi, \tag{3.2}$$

помеха ξ называется аддитивной. Аддитивную помеху часто называют шумом. Если же оператор V может быть представлен в виде

$$x = \nu s, \tag{3.3}$$

где случайный процесс $\nu(t)$ неотрицателен, то помеху ν называют мультипликативной. Если ν — медленный (по сравнению с s) процесс, то явление, вызываемое мультипликативной помехой, носит название замирания (фединг).

В более общем случае оператор V не может быть приведен к основным формам (3.2) и (3.3). При одновременном наличии шума и мультипликативной помехи удобно ввести два случайных процесса, выражающих оба вида помехи, т. е. записать

$$x = \nu s + \xi. \tag{3.4}$$

С физической точки зрения случайные помехи порождаются различного рода флюктуациями. Флюктуациями в физике называются случайные отклонения тех или иных физических величин от их средних значений. Так, источником шума в электрических цепях постоянного тока могут являться флюктуации тока около среднего значения, обусловленные дискретной природой носителей заряда (ионов и электронов). Это явление носит название дробового эффекта.

Наиболее универсальной причиной шума являются флюктуации, обусловленные тепловым движением. Случайное тепловое движение носителей заряда в любом проводнике вызывает случайную разность потенциалов на его концах. Эта разность потенциалов флюктуирует около среднего значения, равного нулю; ее средний квадрат пропорционален абсолютной температуре. Возникающая помеха называется тепловым шумом.

Из сказанного видно, что флюктуации и обусловленные ими помехи заложены глубоко в природе вещей. Флюктуации есть результат дискретного строения вещества и статистической природы ряда физических величин. Действительно, многие физические величины представляют результат усреднения по большому числу индивидуальных частиц, поведение и действие которых подчиняется законам случая. Поэтому флюктуация этих физических

величин принципиально неустранима, и можно лишь ставить вопрос о том, какова относительная величина флюктуаций и каким образом мы можем на нее повлиять находящимися в нашем распоряжении средствами.

Имеется еще один источник принципиально неустранимого шума. Речь идет о дискретной природе электромагнитного излучения. Как известно, излучение совершается дискретными порциями — квантами, энергия которых равна $h\vartheta$, где h — постоянная Планка, ϑ — частота. Квант электромагнитного излучения называется фотоном. В настоящее время в технике имеются две ясные тенденции: к увеличению расстояний и к повышению частоты. Увеличение расстояний означает уменьшение потока энергии, а повышение частоты — укрупнение фотонов. Таким образом, при определенных условиях не только начинает ощущаться дискретная фотонная структура излучения, но обусловленный этой причиной шум может превзойти все остальные помехи. Канал, работающий при таких условиях, получил название фотонного канала.

Мы говорили до сих пор о шуме (аддитивной помехе) и основных его источниках. Обратимся теперь к мультипликативной помехе.

Природа этой помехи состоит в случайном изменении параметров канала передачи. Здесь уместно заметить, что при передаче сигнал подвергается искажениям вследствие того, что коэффициент передачи канала не есть постоянное число; свойство канала описывается частотными или временны́ми характеристиками, определяющими так называемые линейные искажения. Кроме того, канал может вносить и нелинейные искажения, обусловленные нелинейностью тех или иных звеньев канала [1].

Как линейные, так и нелинейные искажения обусловлены известными характеристиками канала, а потому, по крайней мере в принципе, могут быть устранены надлежащей коррекцией. Поэтому искажения следует четко отделить от действия помехи случайного характера, которая заранее не может быть известна (об этом подробнее говорится в § 17).

Если же коэффициент передачи канала претерпевает случайные изменения, то влияние этих изменений следует уже рассматривать как действие случайной помехи, которая и является, как легко сообразить, помехой мультипликативной.

Примером медленной мультипликативной помехи является изменение силы принимаемого сигнала, обусловленное интерференцией при многолучевом распространении (замирание). Быстрая мультипликативная помеха возникает при использовании шума в качестве переносчика.

---

[1] Здесь и ниже мы пользуемся термином канал не в математическом, а в физическом смысле. Под каналом понимается совокупность линии и оконечных устройств. Канал характеризуется коэффициентом передачи и действующими в канале помехами.

Перейдем к вопросу о математическом описании помех. Помеха представляется случайной функцией времени. Случайную функцию дискретного времени называют обычно случайной последовательностью, случайную функцию непрерывного времени — случайным процессом. Случайные функции характеризуются своими распределениями. Применяются также числовые характеристики в виде моментов распределения. Обычно рассматриваются стационарные случайные процессы.

Среди всех случайных процессов особое место занимает процесс с нормальным распределением (гауссов процесс). Большое число действительных случайных процессов являются гауссовыми. Это обстоятельство находит себе объяснение в известной теореме Ляпунова, согласно которой распределение суммы независимых случайных величин (при некоторых достаточно широких условиях) сходится к нормальному вне зависимости от характера распределения слагаемых. А многие процессы, наблюдаемые нами, представляют собой как раз совокупность множества отдельных независимых случайных актов. Так обстоит дело во всех явлениях, где мы имеем дело со множеством частиц или квантов. Гауссов процесс обладает также многими замечательными с математической точки зрения свойствами. Эти свойства подробно описаны в руководствах по теории вероятностей и теории случайных процессов. Отметим здесь лишь то очень важное для расчетов обстоятельство, что гауссов процесс полностью определяется своим смешанным моментом второго порядка.

Момент первого порядка (первый момент)

$$M\xi = \int_{-\infty}^{\infty} xw(x)\,dx = a \tag{3.5}$$

выражает среднее значение, или, говоря техническим языком, постоянную составляющую процесса.

Центральный момент второго порядка (второй момент) называется дисперсией. Он равен

$$D\xi = \int_{-\infty}^{\infty} (x-a)^2 w(x)\,dx = M\xi^2 - a^2 = \sigma^2. \tag{3.6}$$

Дисперсия выражает мощность переменной составляющей, а средний квадрат $M\,\xi^2$ — общую мощность. В большинстве случаев $M\,\xi=0$, так что дисперсия совпадает со средним квадратом.

Смешанный второй момент

$$M[\xi(t)\,\xi(t+\tau)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 w(x_1,\ x_2)\,dx_1 dx_2 = B(\tau) \tag{3.7}$$

называется функцией автокорреляции процесса $\xi(t)$. Величина $B(0)$ есть мощность процесса, т. е.

$$B(0) = M\xi^2 = P. \tag{3.8}$$

Многие случайные процессы, встречающиеся в практике, обладают свойством эргодичности. Свойство это состоит в том, что средние по множеству (т. е. математические ожидания, вычисляемые по распределениям) с вероятностью единица совпадают со средними по времени, найденными по одной реализации процесса. Для эргодических процессов имеем

$$a = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \xi(t)\, dt, \quad P = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \xi^2(t)\, dt,$$

$$B(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \xi(t)\,\xi(t+\tau)\, dt.$$

Спектральная плотность мощности $G(\omega)$ (называемая в дальнейшем просто спектром) связана с функцией автокорреляции парой преобразований Фурье

$$\begin{aligned} G(\omega) &= \frac{2}{\pi} \int_0^\infty B(\tau) \cos \omega\tau d\tau, \\ B(\tau) &= \int_0^\infty G(\omega) \cos \omega\tau d\omega. \end{aligned} \tag{3.9}$$

Положив во второй формуле $\tau=0$, получим соотношение, поясняющее смысл функции $G(\omega)$:

$$B(0) = \int_0^\infty G(\omega)\, d\omega = P.$$

Наряду с $G(\omega)$ часто пользуются функцией $A(\omega) = 2\pi G(\omega)$. Как G, так и A представляют спектральную плотность мощности, но функция A есть мощность, приходящаяся на полосу 1 *гц*. Иногда удобнее записать преобразование (3. 9) в виде

$$\begin{aligned} A(f) &= 4 \int_0^\infty B(\tau) \cos 2\pi f\tau d\tau, \\ B(\tau) &= \int_0^\infty A(f) \cos 2\pi f\tau df. \end{aligned} \tag{3.10}$$

Задание $B(\tau)$ (или $G(\omega)$) исчерпывающим образом характеризует гауссов процесс; во многих случаях этого достаточно и для описания свойств процесса с распределением, отличным от нормального.

Помеху, представляющую собой случайный процесс с равномерным спектром, т. е.

$$A(f) = A_0 = \text{const},$$

называют белым шумом. Мощность белого шума в полосе равна

$$P_F = \int_0^F A(f)\,df = A_0 F,$$

т. е. просто пропорциональна ширине полосы; например, тепловой шум является белым. В известной формуле Найквиста

$$P_F = \bar{E}^2/R = 4kT_A F,$$

где k — постоянная Больцмана; T_A — абсолютная температура; величина $4kT_A$ непосредственно выражает постоянную спектральную плотность A_0 теплового шума, возникающего в сопротивлении R.

При белом шуме полосу нужно всегда предполагать конечной. В противном случае получится либо бесконечная мощность, либо нулевая спектральная плотность.

Заметим, что канал с аддитивной помехой характеризуют обычно не абсолютной мощностью помехи, а отношением средних мощностей сигнала и помехи

$$\rho = P_s/P_\xi.$$

Это отношение, называемое кратко отношение сигнал/помеха, играет большую роль в теории помехоустойчивости.

Физическое ограничение полосы вносит корреляцию. Значения случайного процесса являются некоррелированными только при неограниченной полосе. При ограниченной же полосе некоррелированными можно считать только значения случайного процесса, отстоящие друг от друга не менее чем на интервал корреляции. Интервал корреляции можно определить (см. (3. 10)) как

$$\tau_0 = \frac{1}{B(0)} \int_{-\infty}^{\infty} B(\tau)\,d\tau = \frac{1}{2} \cdot \frac{A(0)}{B(0)}.$$

Если $A(0) = A_0$, то, учитывая, что $B(0)$ есть мощность, равная $A_0 F$, находим

$$\tau_0 = {}^1/_2 F.$$

Это соотношение поясняет, между прочим, смысл теоремы Котельникова: интервал Δt есть не что иное, как интервал корре-

ляции, так что отсчеты функции, по Котельникову, представляют собой ближайшие некоррелированные значения функции.

Многообразие случайных процессов, с которыми нам приходится иметь дело, не исчерпывается гауссовыми процессами. Всякое нелинейное преобразование изменяет распределение. Таким образом, гауссов процесс на входе нелинейного устройства дает негауссов процесс на выходе. Из числа часто встречающихся распределений упомянем о рэлеевом распределении

$$w(x)=\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}\ [x>0].$$

Такому закону подчинено, в частности, одномерное распределение огибающей гауссова процесса, т. е. процесса на выходе линейного детектора, на вход которого подан гауссов процесс. Рэлеево распределение встречается также при исследовании замирания.

Мы говорили до сих пор только об одномерном распределении, которое характеризует значение случайной функции в данный момент, и о двумерном распределении для двух значений, относящихся к двум различным моментам времени. Если помеха представлена последовательностью

$$\xi_1,\ \xi_2,\ \ldots,\ \xi_n,$$

то ее исчерпывающим описанием будет n-мерное распределение

$$w(x_1,\ x_2,\ \ldots,\ x_n).$$

Если значения ξ_i статистически независимы, то

$$w(x_1,\ x_2,\ \ldots,\ x)_n=\prod_{i=1}^{n}w(x_i). \tag{3.11}$$

В частности, для нормального многомерного распределения, если все ξ_i имеют одинаковое распределение, т. е. $\sigma_i=\sigma$, полагая $a_i=0$, получим

$$w(x_1,\ x_2,\ \ldots,\ x_n)=\frac{1}{(\sqrt{2\pi}\,\sigma)^n}e^{-\frac{1}{2\sigma^2}\sum x_i^2}. \tag{3.12}$$

Если же значения ξ_i коррелированы, т. е.

$$M(\xi_i\xi_j)=\sigma^2k_{ij}\neq 0,$$

то распределение принимает вид

$$w(x_1, x_2, \ldots, x_n)=\frac{1}{(\sqrt{2\pi}\,\sigma)^n}e^{-\frac{1}{2\sigma^2D}\sum_i\sum_j D_{ij}x_ix_j}, \tag{3.13}$$

где D — определитель корреляционной матрицы (k_{ij}), а D_{ij} — алгебраическое дополнение элемента k_{ij}.

Для двумерного случая имеем из общей формулы

$$w(x_1,\ x_2)=\frac{1}{2\pi\sigma^2\sqrt{1-k^2}}\,e^{-\frac{1}{2\sigma^2(1-k^2)}\left(x_1^2+x_2^2-2kx_1x_2\right)}, \qquad (3.14)$$

где k — коэффициент корреляции.

Заметим, что для нормального процесса условие статистической независимости сводится к более мягкому условию некоррелированности. Это связано с тем, что нормальный процесс полностью определяется первыми двумя моментами распределения.

Мы не обсуждаем здесь вопроса о природе и статистике так называемой импульсной помехи, действие которой проявляется в резком ухудшении условий передачи (вплоть до полного нарушения связи) на протяжении отдельных промежутков времени. Меры борьбы с этой помехой описаны в параграфах, посвященных корректирующим кодам.

§ 4. Геометрические представления

В современной теории передачи сигналов широко используют геометрические представления. Приступая к обсуждению этих представлений, нужно прежде всего заметить, что геометрия нашего времени существенно отличается от геометрии классической древности.

Если первоначальное понятие пространства складывалось на основе непосредственного чувственного восприятия окружающего мира, то в современной математике пространство определяется как абстрактное множество, свойства которого аксиоматически выражаются некоторыми соотношениями между элементами этого множества.

В современной теории передачи сигналов применяются такие термины, как вектор, пространство, расстояние, проекция и т. п. Однако эти геометрические термины подразумевают более широкие понятия, относящиеся к области функционального анализа. В рамках технической монографии не может быть и речи о систематическом изложении этих понятий. В этом нет и надобности, так как существуют руководства, специально приспособленные к возможностям и потребностям инженера. Мы ограничимся в этом параграфе, имеющем, как и предыдущие, вспомогательный характер, самыми краткими справочными данными о терминах и понятиях, которыми нам предстоит пользоваться в дальнейшем.

Начнем с перечисления терминов и определений, относящихся к n-мерному евклидову пространству.

Вектор определяется как совокупность чисел, называемых координатами вектора

$$x=(x_1,\ x_2,\ \ldots,\ x_n). \qquad (4.1)$$

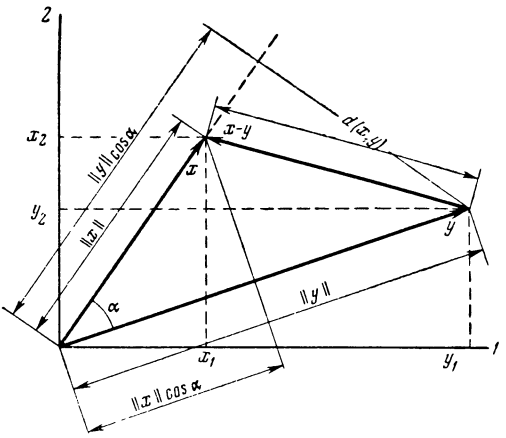


Рис. 3

Совокупность всех векторов x — это и есть n-мерное векторное евклидово пространство, обозначаемое R_n. Можно также сказать, что пространство R_n есть множество точек, представляемых концами векторов.

Норма вектора есть

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{4.2}$$

Как видим, норма есть обобщение длины вектора.

Расстояние между двумя векторами x и y определяется как норма разности векторов

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{4.3}$$

Скалярное произведение двух векторов x и y есть

$$xy = \sum_{i=1}^{n} x_i y_i. \tag{4.4}$$

Вводя угол α между двумя векторами, имеем выражения для

$$\cos\alpha = \frac{xy}{\|x\|\|y\|} \tag{4.5}$$

и для проекций x на y и обратно y на x

$$\|x\|\cos\alpha = xy/\|y\|, \quad \|y\|\cos\alpha = xy/\|x\|. \tag{4.6}$$

Координаты вектора представляют собой проекции вектора на оси. Иначе говоря, координаты вектора выражаются скалярными произведениями данного вектора на орты, т. е. на векторы с единичной нормой, все координаты которых равны нулю, кроме одной, соответствующей номеру орта.

Все эти соотношения можно пояснить на двумерной модели, изображенной на рис. 3.

Покажем сразу же, какое отношение имеют эти определения к нашей теме.

Пусть сообщение m выражено отрезком функции дискретного времени, т. е. конечной последовательностью

$$m=(m_1,\ m_2,\ \ldots,\ m_n).$$

Так обстоит дело при передаче любого сообщения импульсным методом. Но в таком случае сообщение m может быть представлено вектором в n-мерном пространстве, называемом пространством сообщений. Пространство сообщений — это множество всех возможных сообщений, составленных из n каких угодно элементов m_i. Различие между двумя какими-либо сообщениями выражается расстоянием между векторами, изображающими их. Расстояние зависит от норм (длин) векторов и от угла между ними.

Перейдем к обобщениям геометрических понятий. Можно определить так называемое метрическое пространство, введя расстояние как неотрицательную величину, удовлетворяющую следующим аксиомам:

$$\begin{aligned}&1.\ d(x,\ x)=0,\\&2.\ d(x,\ y)=d(y,\ x),\\&3.\ d(x,\ y)\leqslant d(x,\ z)+d(z,\ y),\end{aligned}\qquad(4.7)$$

где x, y, z — элементы (точки) пространства.

Первая аксиома устанавливает, что расстояние между некоторым элементом и им самим равно тождественно нулю. Вторая аксиома носит название аксиомы симметрии и устанавливает равноправие элементов. Третья аксиома называется аксиомой треугольника, так как в простейшем случае двумерного евклидова пространства она выражает тот факт, что сторона треугольника меньше суммы двух других (знак равенства относится к вырожденному случаю, когда все стороны лежат на одной прямой).

Расстояние может выражаться любой функцией координат, удовлетворяющей перечисленным выше аксиомам. Эту функцию называют обычно метрикой пространства. Легко убедиться, что метрика евклидова пространства R_n, выражаемая формулой (4. 3), удовлетворяет аксиомам (4. 7).

Возможно, однако, построить сколько угодно пространств с метрикой, отличной от евклидовой. Нам встретится, в частности, n-мерное пространство, которое мы обозначим l_n со следующей метрикой:

$$d(x,\ y)=\sum_{i=1}^{n}|x_i-y_i|.\qquad(4.8)$$

Если для множества, образующего данное пространство, определены операции сложения элементов множества и умножения элемента на число, причем обе операции удовлетворяют условиям коммутативности, ассоциативности, дистрибутивности, то пространство называется линейным.

К линейным метрическим пространствам относятся, в частности, нормированные пространства. Этот вид пространств определяется заданием нормы, удовлетворяющей следующим аксиомам:

$$\begin{aligned} &1.\ \|x\| \geqslant 0, \\ &2.\ \|\lambda x\| = |\lambda|\,\|x\|, \\ &3.\ \|x+y\| \leqslant \|x\| + \|y\|. \end{aligned} \tag{4.9}$$

Первая аксиома устанавливает, что норма есть положительное вещественное число, равное нулю только для нулевого вектора (т. е. вектора, все координаты которого равны нулю). Во второй аксиоме λ есть любое число. Третья аксиома есть аксиома треугольника.

Расстояние определяется как норма разности

$$d(x,\ y) = \|x - y\|.$$

Все пространства, которые мы будем применять в дальнейшем, относятся к числу нормированных.

В линейном пространстве можно аксиоматически ввести скалярное произведение, приписав ему следующие свойства:

$$\begin{aligned} &1.\ xy = yx, \\ &2.\ (x+y)\,z = xz + yz, \\ &3.\ (\lambda x)\,y = \lambda\,(xy), \\ &4.\ xx \neq 0. \end{aligned}$$

Если определить норму через скалярное произведение, т. е. положить

$$\|x\|^2 = xx,$$

то говорят, что норма порождена скалярным произведением, и пространство, отвечающее такому определению, называется гильбертовым.

Таково, в частности, пространство всех непрерывных функций аргумента t, заданных на интервале $a < t < b$, в котором скалярное произведение определено соотношением

$$xy = \int_a^b x(t)\,y(t)\,dt, \tag{4.10}$$

а норма есть

$$\|x\| = \sqrt{\int_a^b x^2(t)\,dt}. \tag{4.11}$$

Это пространство обозначается C_{L^2}. Оно имеет бесконечное число измерений. Расстояние между двумя векторами в пространстве C_{L^2} определяется соотношением

$$d(x,y) = \|x - y\| = \sqrt{\int_a^b [x(t) - y(t)]^2\,dt}. \tag{4.12}$$

Пространство C_{L^2} представляет собой естественное обобщение пространства R_n, получаемое путем перехода от последовательности к функции непрерывного аргумента. Для наших целей пространство C_{L^2} имеет особое значение, так как позволяет применить общие геометрические представления к сообщениям и сигналам, являющимся функциями непрерывного времени.

Теперь нужно ввести важное понятие о случайном векторе. Этим термином мы обозначаем вектор, координаты которого случайные величины.

Случайный вектор $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ в пространстве R_n не занимает какого-либо определенного положения. Его конец может оказаться в той или иной области пространства с известной вероятностью, которую можно подсчитать, зная распределение величин ξ_i. Таким образом, конец случайного вектора можно представить себе не как определенную точку, а как облако, переменная плотность которого выражает вероятность концу вектора оказаться в данном элементе объема.

В качестве примера подсчитаем вероятность концу случайного вектора попасть в элементарный объем.

$$dV = dx_1 dx_2 \ldots dx_n,$$

если координаты вектора $\xi_1, \xi_2, \ldots, \xi_n$ распределены по одному и тому же нормальному закону и статистически независимы. Мы имеем в этом случае

$$dp\{x_1 < \xi_1 < x_1 + dx_1,\ x_2 < \xi_2 < x_2 + dx_2,\ \ldots,\ x_n < \xi_n < x_n + dx_n\} = w(x_1, x_2, \ldots, x_n)\,dx_1 dx_2 \ldots dx_n =$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \prod_{i=1}^{n} e^{-x_i^2/2\sigma^2} dV = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-r^2/2\sigma^2} dV.$$

Таким образом, искомая вероятность оказывается зависящей только от

$$r = \sqrt{\sum_{i=1}^{n} x_i^2},$$

т. е. от расстояния объема dV от начала координат.

Иначе говоря, упомянутое выше облако, плотность которого можно выразить объемной плотностью вероятностей

$$\frac{dp}{dV} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-r^2/2\sigma^2},$$

имеет сферическую симметрию. Все направления рассматриваемого случайного вектора равновероятны.

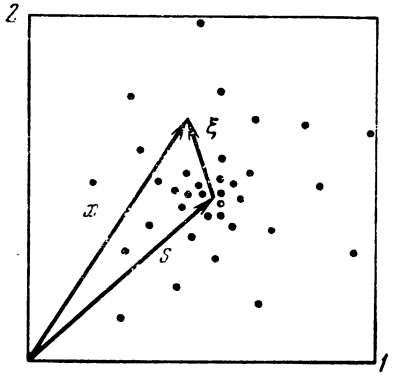


Рис. 4

Сходным образом можно рассматривать и случайные процессы, как случайные векторы в пространстве C_{L^2}. При этом норма

$$\|\xi\| = \sqrt{\int_a^b \xi^2(t)\,dt}$$

представляет собой случайную величину, распределение которой зависит не только от распределения ξ, но и от величины интервала (a, b).

Отметим, что если даны два случайных процесса, ξ и ζ, то косинус угла между векторами, представляющими эти процессы, равный (см. (4,5))

$$\cos\alpha = \frac{\xi\zeta}{\|\xi\| \cdot \|\zeta\|},$$

есть не что иное, как нормированный коэффициент взаимной корреляции. Полное отсутствие корреляции (т. е. равенство нулю коэффициента корреляции) выражается ортогональностью векторов.

Рассмотрим с геометрической точки зрения влияние аддитивной помехи. Пусть сигнал $s(t)$ представлен вектором в пространстве S, которое является пространством сигналов. На сигнал накладывается помеха в виде случайного процесса $\xi(t)$. В результате получается сигнал

$$x = s + \xi.$$

Положение представлено на двумерной модели рис. 4. Изображенный на рисунке вектор ξ — одна из реализаций случайной

помехи. Соответственно и вектор x представляет только одну из возможных реализаций принятого сигнала. Точки, разбросанные по полю рисунка, соответствуют разным реализациям, т. е. представляют положения конца вектора ξ (или x), которые можно было бы зафиксировать в ряде испытаний. Эти точки (аналогично пробоинам на стрелковой мишени) отображают статистику помехи. При возрастании числа испытаний плотность точек сходится к упомянутой выше объемной плотности вероятностей. Так как распределение помехи обычно не ограничено, то конец вектора может с той или иной вероятностью оказаться где угодно, в том числе и в тех областях пространства, где расположены векторы других возможных сигналов s. Это и может повлечь за собой ошибку при приеме. Вероятность такого события, зависящая от свойств сигнала и помехи, а также от способа приема, будет подробно исследована ниже.

Заметим, что нормы векторов равны корням из энергий, т. е.

$$\|s\|^2 = \int_a^b s^2(t)\,dt = E_s, \qquad \|\xi\|^2 = \int_a^b \xi^2(t)\,dt = E_\xi.$$

Для нормы вектора x имеем

$$\|x\|^2 = \int_a^b (s+\xi)^2\,dt = E_s + F_\xi + 2E_{s\xi},$$

где

$$E_{s\xi} = \int_a^b s(t)\,\xi(t)\,dt -$$

скалярное произведение s и ξ. Она имеет размерность энергии, и мы будем в дальнейшем называть ее иногда взаимной энергией [1] двух функций времени.

Выше говорилось о том, что нам понадобится конечномерное пространство с неевклидовой метрикой, а именно, пространство l_n, в котором расстояние определено как

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{4.8}$$

Сейчас мы рассмотрим одно из основных применений этого пространства. Речь пойдет о геометрических представлениях двоичных кодов. Конечный отрезок двоичного сигнала представляется конечной последовательностью нулей и единиц. Под от-

---

[1] Величина $\frac{1}{b-a}\int_a^b x(t)\,y(t)\,dt = \frac{1}{b-a}Exy$ называется иногда в литературе функцией кратковременной (short-time) корреляции функций x и y.

резком мы будем далее понимать отдельную кодовую комбинацию. Для n-значного кода последовательность содержит n двоичных цифр. Координаты сигнала могут быть только нулем или единицей, поэтому геометрической моделью n-значного кода является n-мерный куб с ребром, равным единице, каждая вершина которого представляет одну из возможных кодовых комбинаций.

Число вершин n-мерного куба равно

$$N_0 = 2^n.$$

Модель трехзначного двоичного кода, содержащего восемь возможных кодовых комбинаций, а именно:

1	2	3	4	5	6	7	8
000	001	010	011	100	101	110	111

представлена в виде трехмерного куба на рис. 5.

Различие между двумя двоичными числами принято выражать числом знаков, в которых они различаются. Любые две комбинации (из восьми приведенных выше) различаются между собой не менее чем в одном знаке. Комбинации 010 и 001 различаются в двух знаках; комбинации 101 и 010 — в трех знаках. Отождествляя различие, выражаемое числом различающихся знаков, с расстоянием, мы и приходим к метрике (4. 8), так как

$$|0-0|=0,\ |0-1|=1,\ |1-0|=1,\ |1-1|=0.$$

Заметим, что к тем же результатам приводит сложение по модулю два, так что в двоичном случае метрику можно записать в виде

$$d(x,\ y)=\sum_{i=1}^{n}(x_i + y_i),$$

где x_i и y_i принимают значения нуль или единица, $+$ есть знак сложения по модулю два.

С геометрической точки зрения это означает, что в рассматриваемом пространстве l_n расстояние измеряется числом ребер куба, которые надо пройти на пути от одной точки к другой. Занумеровав кодовые комбинации так же, как в приведенной выше таблице, можно составить матрицу расстояний:

	1	2	3	4	5	6	7	8
1	0	1	1	2	1	2	2	3
2		0	2	1	2	1	3	2
3			0	1	2	3	1	2
4				0	3	2	2	1
5					0	1	1	2
6						0	2	1
7							0	1
8								0

Матрица симметрична относительно диагонали в силу аксиомы симметрии. Для каждой вершины имеются три другие на расстоя-

нии единица, еще три на расстоянии два и одна на расстоянии три. Нетрудно заметить, что между метрикой (4. 8) и евклидовой метрикой существует в рассматриваемом двоичном случае простая связь. Если обозначить d_E евклидово расстояние, то имеем просто

$$d_E = \sqrt{d}.$$

В ряде случаев желательно построить код, все комбинации которого находились бы друг от друга на равных расстояниях (эквидистантный код). Простейшим кодом такого рода является ортогональный код. Его кодовые комбинации содержат по одной единице, все остальные знаки — нули. Геометрическим образом

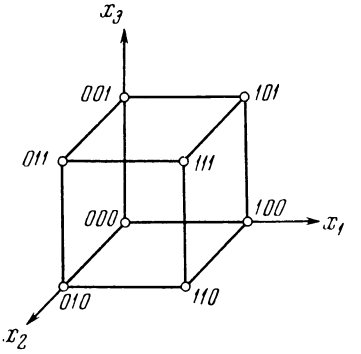


Рис. 5

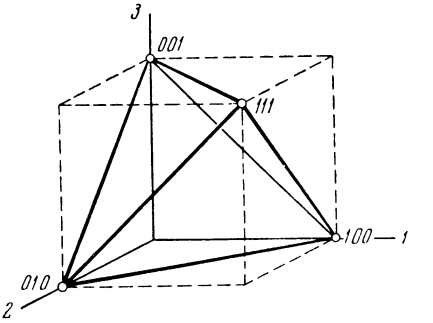


Рис. 6

такого кода являются точки на координатных осях (или, иначе говоря, орты). В трехмерном случае комбинациями ортогонального кода являются следующие три: 100, 010, 001 (см. рис. 5). Независимо от значности кода расстояние между любой парой кодовых комбинаций равно двум. Число кодовых комбинаций равно, очевидно, $N = n$.

Комбинации ортогонального кода обладают равной энергией. Геометрически это выражается в том, что все точки, представляющие кодовые комбинации, находятся от начала координат на одинаковом расстоянии, равном единице.

Заметим, что, добавляя кодовые комбинации с заменой 1 на —1, мы получим так называемый биортогональный троичный код с символами 0, 1, —1, содержащий $2n$ кодовых комбинаций при $d = -2$. Число кодовых комбинаций двоичного эквидистантного кода можно увеличить на единицу, применяя симплексный код. Симплексом называется правильный многогранник, все вершины которого находятся на равных расстояниях друг от друга. В трехмерном пространстве таким многогранником является тетраэдр (рис. 6), вершины которого соответствуют комбинациям 100,

010, 001, 111. Расстояние между любой парой вершин равно двум. Число вершин симплекса в n-мерном пространстве равно

$$N=n+1.$$

Кодовые комбинации симплексного кода не обладают равной энергией (например, на рис. 6, вершины 100, 010 и 001 находятся от начала на расстоянии $d=1$, а вершина 111 — на расстоянии $d=3$). Можно уравнять энергии, поместив начало координат в центр симплекса, но тогда код уже не будет двоичным [1].

Переходим теперь к преобразованиям сообщений и сигналов. И здесь нам потребуются некоторые понятия, относящиеся к функциональному анализу.

Напомним прежде всего определение функции из учебника математики. Величина y называется функцией независимой переменной x, если каждому значению x (из множества его возможных значений) соответствует определенное значение y.

Таким образом, если значения x выражены некоторым множеством чисел, а значения y — другим множеством чисел, то функциональная зависимость устанавливает соответствие между элементами обоих множеств. Для наших целей удобно выразиться так: функция устанавливает зависимость одного числа от другого.

Более общим понятием является понятие функционала, который устанавливает соответствие между множеством чисел, с одной стороны, и некоторым множеством функций — с другой. Можно сказать, что функционал устанавливает зависимость числа от функции.

Функционал есть величина, зависящая от выбора функции из некоторого определенного множества. Примером функционала может служить определенный интеграл, величина которого (при неизменных пределах) зависит от вида подынтегральной функции [2].

И, наконец, еще более общим понятием является понятие функционального оператора, или, для краткости, просто оператора. Оператор устанавливает соответствие между двумя множествами функций, так что всякой функции из одного множества соответствует определенная функция из другого множества. Можно сказать, что оператор устанавливает зависимость функции от функции.

Если обозначить функцию символом f, функционал — символом Φ, а оператор — символом ψ, то приведенные определения

---

[1] Исключение составляет случай $n=3$. Для этого случая преобразованием координат (переводящим начало в центр куба) можно привести симплексный код к двоичному.

[2] Именно такие функционалы рассматриваются в вариационном исчислении, где и возникла впервые надобность во введении понятия функционала.

можно представить в виде

$$y = f(x), \quad y = \Phi[f(x)], \quad f_2(x) = \psi[f_1(x)],$$

где x и y — переменные величины, выражаемые числами.

Так как функции могут быть представлены векторами, и множество функций определяется как векторное пространство, то действие оператора может быть описано в геометрических терминах как преобразование пространства X векторов x в пространство Y векторов y или как отображение пространства X на пространство Y. Если оператор, выполняющий преобразование X в Y, обозначен через Ψ, то обратный оператор, т. е. оператор, преобразующий Y в X, обозначается Ψ^{-1}.

В теории операторов существенную роль играют линейные операторы. Линейный оператор обладает свойствами

1) аддитивности, т. е. $\psi(x_1 + x_2) = \psi(x_1) + \psi(x_2)$;

2) однородности, т. е. $\psi(\lambda x) = \lambda\psi(x)$.

В n-мерном евклидовом пространстве R_n линейное преобразование пространства X в пространство V определяется формулами

$$\begin{aligned}
y_1 &= \psi_{11}x_1 + \psi_{12}x_2 + \ldots + \psi_{1n}x_n,\\
y_2 &= \psi_{21}x_1 + \psi_{22}x_2 + \ldots + \psi_{2n}x_n\\
&\cdots\cdots\cdots\cdots\\
y_n &= \psi_{n1}x_1 + \psi_{n2}x_2 + \ldots + \psi_{nn}x_n,
\end{aligned}$$

или сокращенно

$$y_i = \sum_{R=1}^{n} \psi_{iR}a_R \quad (i = 1, 2, \ldots, n).$$

Таким образом, линейный оператор ψ можно рассматривать как тензор 2—го ранга, а ψ_{iR} — как компоненты (составляющие) этого тензора.

Итак, линейный оператор ψ полностью определяется коэффициентами ψ_{ik}, образующими квадратную матрицу

$$\psi = \begin{pmatrix}
\psi_{11} & \psi_{12} & \ldots & \psi_{1n}\\
\psi_{12} & \psi_{22} & \ldots & \psi_{2n}\\
\cdots & \cdots & \cdots & \cdots\\
\psi_{n1} & \psi_{n2} & \ldots & \psi_{nn}
\end{pmatrix}.$$

Мы не вдаемся в подробности хорошо разработанной теории линейных операторов, так как, к сожалению, большинство операторов, с которыми мы встречаемся в технике передачи сигналов, являются нелинейными или параметрическими. Такие операторы тоже можно представить матрицей коэффициентов, но эти коэффициенты будут уже не постоянными величинами, в параметрическом случае они являются функциями времени, а в нелинейном — функциями координат x_k. Поэтому такого рода операторы будем

рассматривать в индивидуальном порядке, не прибегая к какой-либо общей теории, из которой мы заимствуем только общее понятие о функциональном операторе.

Линейный функционал удовлетворяет тем же условиям аддитивности и однородности, что и линейный оператор. В пространстве R_n линейный функционал может быть представлен формулой

$$y = \Phi(x) = \Phi_1 x_1 + \Phi_2 x_2 + \ldots + \Phi_n x_n = \sum_{i=1}^{n} \Phi_i x_i.$$

Таким образом, линейный функционал может рассматриваться как скалярное произведение вектора x на вектор

$$\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)$$

и, следовательно,

$$y = \Phi(x) = x\Phi.$$

Это положение распространяется на весьма общие бесконечномерные пространства, в том числе и на интересующее нас пространство C_{L^2} непрерывных функций, для которого общий вид линейного функционала можно представить в виде

$$y = \Phi(x) = \int_a^b x(t)\,\Phi(t)\,dt.$$

Это важное соотношение широко используется в дальнейшем. Теперь мы можем вернуться к общему описанию системы передачи, данному в § 1, и изложить его заново в геометрических терминах.

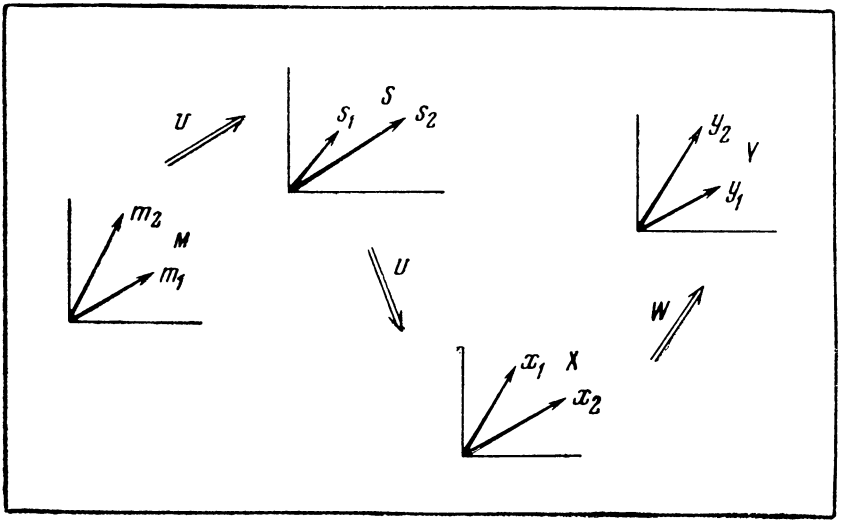


Рис. 7

Сообщения представляются векторами в пространстве сообщений M. На рис. 7 представлена двумерная модель этого пространства с двумя разными сообщениями m_1 и m_2. Оператор U преобразует пространство сообщений M в пространство сигналов S, вектора m_1 и m_2 преобразуются в векторы s_1 и s_2. Оператор V преобразует пространство сигналов S в пространство принятых сигналов X; векторы s_1 и s_2 переходят в x_1 и x_2. Наконец, оператор W преобразует пространство принятых сигналов X (на входе приемника) в пространство выходных сигналов V, векторы x_1 и x_2 переходят в y_1 и y_2.

Если бы помеха отсутствовала, то оператор V тождественно преобразовывал бы пространство S само в себя. Мы получили бы в этом случае идеальную передачу, выбрав $W=U^{-1}$, т. е. проделав на приемной стороне операцию, обратную той, которая применена для формирования сигнала.

К сожалению, это не так. Помеха всегда присутствует, и для получения удовлетворительного соответствия выходных сигналов переданным сообщениям необходимы специальные меры, обсуждению которых и посвящено все дальнейшее изложение.

§ 5. Общие соображения о приеме сигналов

Начнем со следующего утверждения; принимая любой сигнал, мы обязательно что-то о нем знаем и чего-то не знаем. Предварительные сведения о сигнале используются для рационального выбора метода приема в соответствии с поставленными целями. Если бы о сигнале заранее не было известно абсолютно ничего, то его нельзя было бы принять, так как не известно было бы, чем сигнал отличается от несигнала, в частности от любой помехи. То, чего мы заранее не знаем, может являться носителем полезной информации. Если бы о сигнале нам абсолютно все было известно наперед, то такой сигнал не нужно было бы и передавать: он не принес бы нам никакой информации. Говоря конкретно, объектом априорного знания или незнания могут являться те или иные физические параметры сигнала, например, интенсивность, несущая частота, время появления, длительность и т. д.

Некоторые из этих параметров остаются неизменными. В таком случае их следует рассматривать как постоянные признаки сигнала, и наличие этих признаков, естественно, нужно наилучшим образом использовать для отличения сигнала от помехи.

Другие же параметры можно так или иначе модулировать. Эти параметры мы назовем информационными. В их изменениях, неизвестных заранее на приемной стороне, заложена переносимая сигналом информация. Обычно известен лишь диапазон возможных изменений информационных параметров.

Рассмотрим теперь задачи, возникающие при приеме сигналов. В зависимости от назначения сигналов эти задачи сводятся к трем

основным: 1) обнаружение сигнала, 2) различение сигналов, 3) восстановление сообщения.

Разберем все три случая подробно.

Под обнаружением сигнала понимается установление его наличия. При наличии шума (аддитивной помехи) задача сводится к получению ответа на вопрос, имеется ли на входе приемника сигнал плюс шум или только шум? Констатация наличия сигнала — это и есть обнаружение сигнала.

Часто полагают, что обнаружение сигнала не требует измерения каких-либо параметров сигнала. Это не так. Всякое обнаружение неразрывно связано с измерением тех или иных параметров. Более того, обнаружение это в сущности и есть измерение, хотя, может быть, и грубое. В самом деле, когда мы обнаруживаем сигнал, то обнаруживаем его не вообще, а в определенной полосе частот, в определенном интервале времени, в определенном телесном угле. Интервалы, в которых заключены значения этих параметров, определяют точность, с которой они измеряются. В дальнейшем, если требуется, точность может быть повышена. Так, например, поисковый локатор имеет малую направленность; когда цель обнаружена, т. е. когда грубо определено направление на цель, ее перехватывает локатор сопровождения или наводки, имеющий значительно более острую направленность.

Эти соображения можно пояснить следующим примером. Пусть мы собираемся измерить длину бруска при помощи линейки с делениями. Мы совмещаем нуль шкалы с одним концом бруска; измерение же сводится к тому, что мы обнаруживаем второй конец бруска между n-м и $(n+1)$-м делениями шкалы. При самом грубом измерении мы имеем в своем распоряжении мерную линейку определенной длины l. Иначе говоря, мы имеем мерную линейку, на которой имеется всего два деления: 0 и l. В этом случае мы можем лишь отвечать на вопрос о том, больше или меньше длина измеряемого бруска по сравнению с длиной l. Такая ситуация ближе всего к тому, что понимается обычно под обнаружением сигнала.

Если мы в состоянии обнаружить сигнал, т. е. отличить наличие сигнала от его отсутствия, то это открывает возможность передачи любой информации при помощи двоичного кода. Наличие сигнала (посылка) будет соответствовать символу 1, отсутствие сигнала (пауза) — символу 0. Такая система носит название передачи с пассивной паузой, так как в паузе передатчик бездействует.

При передаче двух различных сигналов s_1 и s_2 положение несколько иное. Здесь речь идет уже не об обнаружении, а о *различении* двух сигналов. Дело сводится к ответу на вопрос, имеется ли на входе приемника сигнал s_1 плюс шум или сигнал s_2 плюс шум? Ответ на этот вопрос определяется уже не свойствами каждого сигнала в отдельности, эти свойства в принципе могли бы оставаться неизвестными, а физическим различием между сигналами.

Сигналы могут различаться между собой значениями тех или иных параметров. При выборе двух сигналов нужно стремиться к тому, чтобы различие между ними было по возможности более стойким по отношению к действию помехи. Это значит прежде всего, что различие должно быть по возможности велико: если оно уменьшится в процессе передачи под действием помехи, то оставшееся различие должно быть все же достаточно для уверенного различения сигналов. Но, кроме того, нужно выбирать различие по тому параметру (из всех параметров данного сигнала), который в наименьшей степени подвержен влиянию помехи данного типа.

Очевидно, случай обнаружения может рассматриваться как вырожденный случай различения двух сигналов, когда один из них есть тождественный нуль.

Передача двоичным кодом, в котором символу 1 соответствует сигнал s_1, а символу 0 — сигнал s_2, называется передачей с активной паузой.

Случай различения многих сигналов в принципиальном отношении мало отличается от случая различения двух сигналов. Все сказанное выше по поводу различения двух сигналов должно быть отнесено в случае набора из нескольких сигналов к любой паре сигналов, входящих в этот набор. Однако техника различения многих сигналов может оказаться существенно отличной от техники различения двух сигналов.

Задача восстановления сообщения значительно отличается от задач обнаружения и различения сигналов. Она состоит в том, чтобы получить выходной сигнал $y(t)$, наименее отличающийся от передаваемого сообщения $m(t)$. При этом существенно, что сообщение m заранее неизвестно; известно лишь, что оно принадлежит к некоторому множеству. При таких условиях можно рассматривать данное сообщение как одну из реализаций некоторого случайного процесса. Следовательно, заранее известными могут быть только распределения или моменты распределения этого процесса. В частности, на приемной стороне могут быть заранее известны мощность и спектр случайного процесса, реализацией которого является передаваемое сообщение. Борьба с помехами при такой постановке задачи, конечно, более трудна.

При восстановлении сообщения нужно опираться на некоторый критерий верности, на основе которого оценивается уклонение выходного сигнала y от передаваемого сообщения m. Этот критерий должен, вообще говоря, выводиться из требований, предъявляемых к передаче данного вида сообщений.

Часто применяется критерий квадратичного уклонения. Для непрерывных функций, заданных на интервале (a, b), квадратичное уклонение выражается соотношением

$$\varepsilon^2_{\text{кв}} = \int_a^b [m(t) - y(t)]^2 \, dt.$$

Но можно применить и критерий абсолютного уклонения

$$\varepsilon_{\text{абс}} = \int_a^b |m(t) - y(t)|\, dt.$$

Можно также воспользоваться критерием наибольшего уклонения

$$\varepsilon_{\max} = \max |m(t) - y(t)|, \quad [a < t < b].$$

С точки зрения геометрических представлений (§ 4) уклонение ε есть не что иное, как расстояние $d(m, y)$, а выбор критерия — это выбор метрики пространства сообщений.

Нет решительно никаких общих оснований для предпочтения одного критерия другому. Критерий квадратичного уклонения (евклидова метрика для пространства M) применяется особенно часто только потому, что при пользовании этим критерием получаются, как правило, сравнительно простые выкладки.

Если получателем сообщения является человек, то критерий верности должен, естественно, выбираться на основе свойств восприятия человеком данного вида сообщений. Так, например, при передаче звука критерий верности должен строиться на основе психофизиологических свойств слуха, а при передаче изображений — на основе тех же свойств зрения.

Как ни странно, несмотря на давность задачи, научно обоснованные критерии верности для обоих названных видов передачи до сих пор не выработаны. Если для передачи речи и существуют кое-какие экспериментальные данные, дающие ответ на некоторые практические вопросы построения телефонных систем, то в области телевидения такие данные отсутствуют. Это положение следует, по-видимому, объяснить тем, что в области передачи речи существует такой полуобъективный показатель, как разборчивость речи (этот показатель выражается числом, получаемым в результате артикуляционных испытаний), тогда как в области передачи изображений не существует никакого показателя качества изображения (не считая таких, как четкость и контрастность). Вместе с тем техника передачи речи оказывается совершенно беспомощной, когда мы выходим за пределы оценок по разборчивости. Так обстоит, например, дело в области вокодеров, дающих высокую разборчивость, но неприятное звучание, причем никому пока не известно, как объективно оценить и измерить степень «неприятности». В области критериев, связанных с особенностями восприятия, нужно еще очень много сделать.

Гораздо проще обстоит дело, когда элементы человеческого восприятия могут не приниматься в расчет. Если, например, непрерывное сообщение $m(t)$ вырабатывается некоторой телеметрической системой и отображает изменение во времени наблюдаемой физической величины, то критерий верности сводится к точности измерения, требования к которой сравнительно легко

вывести из общего назначения данного измерения и условий применения его результатов.

Все эти рассуждения приведены здесь для обоснования ранее высказанного тезиса: критерий верности не универсален и не задан заранее; он должен выбираться в каждом данном случае, исходя из задач и обстановки.

§ 6. Понятие помехоустойчивости

Помехоустойчивость определена во введении, как способность системы передачи противостоять вредному действию помех. Эта общая формулировка должна быть уточнена применительно к различным условиям передачи; должна быть установлена количественная мера помехоустойчивости.

Прежде всего заметим, что по смыслу определения помехоустойчивость понимается как свойство системы в целом. Однако, как об этом говорилось в § 1, мы вынуждены отказаться от исследования, по крайней мере от синтеза, системы в целом из-за непомерной сложности этой задачи. При таких условиях имеет смысл говорить о помехоустойчивости отдельных звеньев системы. Так, можно говорить о помехоустойчивости кодов, видов модуляции, приемников. При этом достаточно оперировать сравнительной или относительной помехоустойчивостью, что позволяет сравнить между собой различные варианты технических решений.

Предельно достижимая помехоустойчивость называется, по В. А. Котельникову, потенциальной помехоустойчивостью. Сравнение фактической помехоустойчивости каждого конкретного устройства с его потенциальной помехоустойчивостью дает оценку качества устройства и показывает наличие еще не использованных резервов.

Действие помехи проявляется в том, что принятый сигнал (а следовательно, и сообщение) отличается от переданного. Поэтому помехоустойчивость можно характеризовать как степень соответствия принятого сигнала (или сообщения) переданному при заданной помехе. Таким образом, при сравнении нескольких систем та из них будет более помехоустойчивой, в которой при одинаковой помехе различие между принятым и переданным сигналами (или сообщениями) будет меньше.

Ввести единое количественное определение помехоустойчивости затруднительно, так как и критерий соответствия принятого сигнала переданному и характеристики действующей в системе помехи могут в зависимости от условий передачи существенно различаться.

Вместе с тем оказывается полезным ввести определения для меры соответствия принятого сигнала переданному. С одной стороны, эта мера зависит, очевидно, от помехи и характеризует, таким образом, помехоустойчивость с необходимой полнотой. С другой стороны, эта мера, взятая не как функция помехи, а как

численный показатель системы, работающей в заданных условиях, представляет собой показатель, вполне характеризующий систему с точки зрения потребителя. В самом деле, эта мера характеризует качество системы как совокупности средств для передачи информации. Потребителя интересует лишь получение сообщения, соответствующего переданному; его не интересуют ни условия, в которых работает система, ни меры, принятые для обеспечения гарантированной потребителю степени соответствия.

Для обозначения степени соответствия принятого сообщения переданному введем общий термин верность. Количественную меру соответствия приходится выбирать по-разному, в зависимости от характера сообщения. Именно, выбор меры зависит от того, передаются ли дискретные символы или непрерывная функция непрерывного аргумента, которую нужно восстановить при приеме.

Рассмотрим первый случай. Пусть сообщение представляет собой последовательность символов из некоторого ансамбля (т. е. конечного множества, заданного вместе с априорными вероятностями его элементов, см. § 1). В этом случае влияние помехи проявляется в том, что вместо фактически переданного символа принимается какой-либо другой. Такое событие мы называем ошибкой. Так как ошибка есть случайное событие, то верностью естественно характеризовать вероятность отсутствия ошибки, т. е. вероятностью правильного приема. Если вероятность ошибки обозначена $p_{\text{ош}}$, то вероятность правильного приема

$$p_{\text{пр}} = 1 - p_{\text{ош}}$$

(так как ошибка и правильный прием образуют полную группу событий). В качестве количественной меры верности можно взять либо самую вероятность $p_{\text{пр}}$, либо любую возрастающую функцию этой вероятности. Но во всякой хорошей системе передачи вероятность ошибки $p_{\text{ош}}$ очень мала и выражается обычно десятичной дробью вида

$$p_{\text{ош}} = 10^{-s},$$

где s имеет значение от 5 и выше. Поэтому за количественную меру верности принятого сигнала переданному в рассматриваемом случае удобно взять следующую убывающую функцию вероятности ошибки (или, что то же, возрастающую функцию вероятности правильного приема)

$$S = \lg \frac{1}{p_{\text{ош}}} \left(= \lg \frac{1}{1 - p_{\text{пр}}} \right). \tag{6.1}$$

Определенная таким образом мера верности, или просто верность, выражается положительным числом, обычно целым и в пределах первого десятка. Таким образом, вопрос о количественной мере верности в случае передачи дискретных символов мы будем считать решенным.

Можно пояснить, что к этому случаю относятся все виды передачи при помощи импульсов с квантовыми параметрами, а также все виды передачи с применением тех или иных кодов. Таким образом, речь идет о весьма обширной категории видов передачи, практическое применение которой будет, как можно предвидеть, со временем еще больше расширяться.

В случае передачи непрерывных функций мы не можем столь же просто ввести количественную меру верности. При передаче дискретных символов мы имеем дело с простой системой событий (ошибка есть — ошибки нет), и распределение вероятностей для этой системы исчерпывает вопрос. При передаче же неправильного сообщения отличие принятого сообщения от переданного имеет также непрерывный характер. И в этом случае применяют термин верность передачи, под которым понимается по-прежнему степень соответствия принятого сообщения переданному. Однако количественная мера соответствия зависит от выбранного критерия верности. Вообще говоря, мерой соответствия или уклонения может служить некоторая величина ε, представляющая собой расстояние между принятым и переданным сообщениями. Критерий верности определяет метрику пространства сообщений.

Теперь можно ввести количественную меру верности, определив ее возрастающей функцией вероятности,

$$p = p\{\varepsilon \leqslant \varepsilon_0\}, \tag{6.2}$$

т. е. вероятности того, что уклонение ε не превзойдет некоторой заранее назначенной величины ε_0. Характер этой функции не играет роли; все определения произвольны и оцениваются только с точки зрения внутренней непротиворечивости и удобства применения.

Таким образом, центр тяжести вопроса переносится на проблему выбора критерия верности.

Заметим, что выражение (6. 2) имеет очень наглядную геометрическую интерпретацию. Уклонение ε, как говорилось в § 5, геометрически представляется расстоянием между концами векторов m и y, т. е.

$$\varepsilon = d(m, y) = \|m - y\|.$$

В таком случае условие $\varepsilon < \varepsilon_0$ означает, что вектор y остается в пределах сферы радиуса ε_0 с центром в конце вектора m_1 (рис. 8), если под сферой независимо от метрики пространства понимать множество точек, расстояния которых от точки, называемой центром, не превосходят постоянной величины, называемой радиусом. Множество точек, находящихся на равных расстояниях от центра, соответственно будет называться сферической поверхностью.

Вероятность (6. 2) есть вероятность концу вектора y попасть в сферу $\varepsilon < \varepsilon_0$. Для другого сообщения (вектор m_2 на том же рис. 8) имеется своя сфера с центром в конце вектора. Таким обра-

зом, количественная мера верности определяется и в случае непрерывного сообщения. Вся трудность переносится на вопрос о критерии верности, т. е., говоря более определенно, на выбор, во-первых, метрики пространства сообщений и, во-вторых, порогового значения ε_0.

Заметим, что геометрическая картина рис. 8 имеет весьма общий характер и может быть применена к любым сообщениям (или сигналам). Особенно плодотворным является представление об определенной области пространства, принадлежащей данному сообщению (или сигналу).

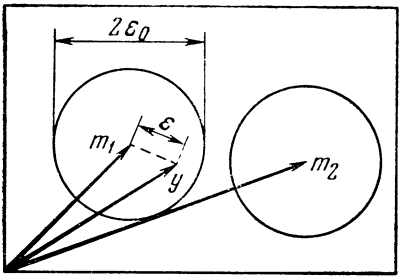


Рис. 8

Точно такую же картину мы могли бы построить и для случая передачи дискретных символов. Разница заключалась бы лишь в том, что множество непрерывных функций несчетно, и пространство сообщений представляет собой многомерный континуум, тогда как в случае дискретных символов пространство сообщений (но не принятых сигналов!) представляет собой конечное множество изолированных точек, находящихся на конечных расстояниях друг от друга. Область, принадлежащую данному сообщению, мы назовем собственной областью. Ее называют также областью правильного приема.

В дискретном случае вероятность правильного приема есть вероятность концу вектора сигнала попасть в собственную область данного сообщения. Вероятность же ошибки есть вероятность концу вектора сигнала оказаться вне этой области. Собственные области сообщений чаще всего не имеют сферической формы, как будет видно из дальнейшего. Но, так или иначе, интересующие нас вероятности сведены к геометрическим вероятностям.

Следует сделать несколько замечаний по поводу связи между верностью и отношением сигнал/помеха.

Верность тем ниже, чем больше относительная интенсивность помехи, т. е. чем меньше отношение сигнал/помеха. Иначе говоря, верность (как бы мы ее ни определили), есть возрастающая функция отношения сигнал/помеха. Но дело в том, что отношение сигнал/помеха — не единственный фактор, влияющий на верность. Во-первых, при неизменной мощности сигнала можно выбирать разные системы сигналов, т. е. строить множество сигналов с различными расстояниями между элементами этого множества. Выбор расстояний должен быть таков, чтобы помеха с заданными свойствами в наименьшей степени влияла на различимость принятых сигналов. В простейшем случае дело сводится к построению

системы сигналов с наибольшими расстояниями. Во-вторых, верность зависит от способа приема, так что она может быть различной при одном и том же отношении сигнал/помеха. Часто дело сводится к тому, что отношение сигнал/помеха в некотором звене приемника больше, чем то же отношение на входе приемника, а между тем канал характеризуется именно отношением сигнал/помеха на выходе линии, т. е. на входе приемника. Правильно сконструированный приемник может увеличить отношение сигнал/помеха, и притом весьма значительно.

Итак, верность зависит от отношения сигнал/помеха и обычно возрастает с его увеличением. Но верность зависит также от рационального построения системы в целом и, в частности, от выбора системы сигналов и способа действия приемника.

Остается пояснить, что понимается под помехоустойчивостью отдельных звеньев системы передачи.

Помехоустойчивость кода можно характеризовать верностью при заданном отношении сигнал/помеха и при определенном способе приема.

Помехоустойчивость системы модуляции непосредственно характеризуется относительным изменением модулируемого параметра под действием данной помехи. Можно сравнивать различные системы модуляции по относительному увеличению отношения сигнал/помеха, даваемому некоторым идеализированным приемником.

Помехоустойчивость приемника также удобно выражать относительным увеличением отношения сигнал/помеха, сравнивая значение этого отношения на выходе звена приемника, производящего основную обработку сигнала (из дальнейшего будет ясно, какое именно звено имеется в виду), со значением того же отношения на входе приемника.

Из этих пояснений видно, что тесно переплетены функции отдельных звеньев системы передачи. Полностью изолированное рассмотрение этих звеньев было бы бессмысленно, единое слишком сложно. Компромисс состоит в том, что мы рассматриваем данное звено, задавая по возможности простым образом показатели, характеризующие влияние остальных звеньев.

§ 7. Влияние вида модуляции

В этом параграфе мы рассматриваем сравнительную помехоустойчивость различных видов модуляции.

Модуляция состоит, как известно, в том, что те или иные параметры функции

$$f = f(a_1, a_2, a_3, \ldots, a_n, t), \tag{7.1}$$

называемой переносчиком, изменяются в соответствии с передаваемым сигналом. Именно параметр a_k получает приращение, пропорциональное передаваемому сигналу (модулирующей функ-

ции). Для данного переносчика возможно столько различных видов модуляции, сколько имеется независимых параметров.

Будем полагать, что приемник снабжен соответствующим детектором, устройство которого таково, что выходной сигнал пропорционален изменениям модулируемого параметра. Поэтому при обсуждении вопроса о помехоустойчивости различных видов модуляции можно не рассматривать действие приемника и свести дело к изменениям параметров переносчика, обусловленным действием помехи. Такого рода изменения мы будем называть паразитной модуляцией.

Итак, задача сводится к нахождению глубины паразитной модуляции, вызванной помехой. Помеху мы будем считать аддитивной и малой [1]. Последнее условие значительно упрощает исследование, хотя и лишает нас возможности рассмотреть ряд интересных вопросов, в частности, вопрос о пороговом эффекте, наблюдаемом при отношении сигнал/помеха порядка единицы при широкополосных видах модуляции.

В результате наложения помехи $\xi(t)$ на немодулированный переносчик (7. 1) получаем функцию

$$F(t) = f(a_1,\ a_2, \ldots, t) + \xi(t). \tag{7.2}$$

Заменим эту функцию другой функцией

$$F_1(t) = f(a_1 + \delta a_1,\ a_2 + \delta a_2, \ldots; t), \tag{7.3}$$

потребовав, чтобы F_1 в определенном смысле наименее уклонялась от F. Сущность этой замены состоит в том, что мы представляем результат наложения помехи как изменение параметров переносчика. Приращение δa_k и представляет собой паразитную модуляцию каждого из параметров a_k [2].

Если $\xi(t)$ имеет среднее значение, равное нулю, то и δa_k в среднем равно нулю. Поэтому эффект помехи можно выразить через дисперсию или средний квадрат δa_k.

Наибольшие допустимые изменения параметров при полезной модуляции обозначим через Δa_k. Тогда отношение сигнал/помеха для модуляции по a_k можно представить в виде [3]

$$\rho_k = \frac{(\Delta a_k)^2}{D\delta a_k}. \tag{7.4}$$

---

[1] Точный смысл условия малости состоит в применимости линеаризованного соотношения (7. 7), см. ниже.

[2] С геометрической точки зрения переход от (7. 2) к (7. 3) означает отображение пространства сигналов на пространство параметров p, о котором говорится в § 14. Приращение δa_k представляют собой проекции вектора разности $f - F_1$ на координатные оси пространства параметров.

[3] При таком определении в числителе стоит величина, равная наибольшей мгновенной мощности. Средняя мощность зависит от вида модулирующей функции. При синусоидальной модуляции средняя мощность равна половине пиковой.

Напомним, что Δa_k — постоянные, а δa_k — случайные величины. Оказывается, что ρ_k различны, т. е. отношения сигнал/помеха неодинаковы для различных видов модуляции. Поэтому интересно выяснить, по какому параметру выгоднее модулировать переносчик, т. е. какой вид модуляции обладает большей помехоустойчивостью. Для ответа на этот вопрос мы выведем общее выражение для ρ_k, а затем рассмотрим конкретные примеры.

Прежде всего выберем квадратичный критерий, т. е. введем в качестве меры уклонения F_1 от F его средний квадрат

$$d^2 = \overline{[F_1(t) - F(t)]^2} \tag{7.5}$$

и будем искать значения δa_k, минимизирующие эту величину. Перепишем (7. 5) в виде

$$d^2 = \frac{1}{T}\int_0^T [F_1(t) - F(t)]^2\, dt = \frac{1}{T}\int_0^T [\delta f(t) - \xi(t)]^2\, dt. \tag{7.6}$$

Здесь

$$\delta f(t) = F_1(t) - f(t).$$

Интервал усреднения T берется равным интервалу корреляции для модулирующей функции; на протяжении этого интервала изменениями параметров при полезной модуляции пренебрегается. Интервал T связан с шириной спектра F модулирующей функции соотношением $T = {}^1/_2 F$.

Если помеха мала, то справедливо следующее линеаризованное соотношение:

$$\delta f(t) = \sum_{k=1}^{n} \frac{\partial f}{\partial a_k}\, \delta a_k. \tag{7.7}$$

Подставив (7. 7) в выражение (7. 6), будем минимизировать d^2, для чего найдем частные производные $\partial d^2/\partial\, \delta a_k$ и приравняем их к нулю

$$\frac{\partial d^2}{\partial \delta a_k} = \frac{1}{T}\int_0^T \frac{\partial}{\partial \delta a_k}\left(\sum \frac{df}{da_i}\, \delta a_i - \xi\right)^2 dt =$$

$$= \frac{2}{T}\int_0^T \left(\sum \frac{\partial f}{\partial a_i} \cdot \frac{\partial f}{\partial a_k}\, \delta a_i - \frac{\partial f}{\partial a_k}\, \xi\right) dt = 0.$$

Отсюда

$$\sum_{i=1}^{n} b_{ik}\delta a_i = \frac{1}{T}\int_0^T \frac{\partial f}{\partial a_k}\, \xi dt, \tag{7.8}$$

где

$$b_{ik} = \frac{1}{T}\int_0^T \frac{\partial f}{\partial a_i} \cdot \frac{\partial f}{\partial a_k}\, dt. \tag{7.9}$$

Разрешая систему (7.8) относительно искомых δa_i, находим

$$\delta a_i = \frac{1}{T} \int_0^T y_i(t)\, \xi(t)\, dt, \tag{7.10}$$

где

$$y_i(t) = \frac{1}{D} \sum_{k=1}^{n} D_{ik} \frac{\partial f}{\partial a_k}. \tag{7.11}$$

Теперь нужно вычислить дисперсию этих величин. Мы имеем

$$D\delta a_i = \frac{1}{T^2} \int_0^T \int_0^T y_1(t)\, y_1(t_1)\, B_\xi(t - t_1)\, dt dt_1, \tag{7.12}$$

где $B_\xi(\tau) = M[\xi(t)\,\xi(t+\tau)]$ — функция корреляции помехи; D — определитель матрицы; (b_{ik}), D_{ik} — соответствующие алгебраические дополнения. Помеха задается обычно не функцией корреляции, а спектром. Поэтому вместо (7.12) можно воспользоваться спектральным выражением

$$D\delta a_i = \frac{1}{2T^2} \int_{-\infty}^{\infty} G_\xi(\omega)\, \Phi_{y_i}^2(\omega)\, d\omega. \tag{7.13}$$

В этой формуле использованы следующие определения спектральных функций:

$$G_\xi(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} B_\xi(\tau)\, e^{-j\omega\tau} d\tau -$$

спектр мощности помехи,

$$\Phi_y^2(\omega) = S_y(\omega)\, S_y^*(\omega) -$$

спектр энергии функции $y(t)$. Здесь

$$S_y(\omega) = \int_{-\infty}^{\infty} y(t)\, e^{-j\omega t} dt.$$

Звездочка означает комплексно-сопряженную величину.

Мы ограничимся в дальнейшем частным случаем, когда $G(\omega) = C_0 = \text{const}$ в полосе $\pm\Omega$. Тогда вместо (7.13) будем иметь

$$D\delta a_i = \frac{G_0}{T^2} \int_0^{\Omega} \Phi_{y_i}^2(\omega)\, d\omega. \tag{7.14}$$

Но по теореме Рэлея интеграл в правой части выражает энергию функции $y_1(t)$

$$E_i = \frac{1}{\pi}\int_0^{\Omega} \Phi_{y_i}^2(\omega)\,d\omega = \int_0^T y_i^2(t)\,dt,$$

так что

$$D\delta a_i = \frac{\pi G_0}{T^2} E_i = \frac{\frac{1}{2}A_0}{T^2} E_i. \qquad (7.15)$$

Перейдем к примерам. Рассмотрим сначала синусоидальный переносчик, т. е. возьмем

$$f = a_0 \sin \omega t,$$

и сравним между собой обычные амплитудную и частотную модуляции. В нашем примере

$$a_1 = a_0,\ a_2 = \omega,\quad \partial f/\partial a_1 = \sin \omega t,\quad \partial f/\partial a_2 = a_0 t \cos \omega t.$$

Положим для упрощения $\omega T = n\pi$ и будем учитывать, что $\omega T \gg 1$ (т. е. что несущая частота ω много больше наивысшей частоты в спектре модулирующей функции). Тогда для коэффициентов b_{ik} по формуле (7.9) получаем

$$b_{ik} = \begin{pmatrix} \frac{1}{2} & -\frac{a_0}{4\omega} \\ \frac{a_0}{4\omega} & \frac{a_0^2 T^2}{6} \end{pmatrix}.$$

Далее, по формуле (7.11) находим

$$y_1(t) = 2\sin \omega t + 3\frac{1}{\omega T}\cos \omega t \simeq 2 \sin \omega t,$$

$$y_2(t) = \frac{1}{a_0 T}\left(3\frac{1}{\omega T}\sin \omega t + 6\frac{t}{T}\cos \omega t\right).$$

Для энергий имеем $E_1 = T$, $E_2 = 3/a_0^2 T$ и по формуле (7.15)

$$D\delta a_1 = A_0/2T,\quad D\delta a_2 = 3A_0/2a_0^2 T^3.$$

Теперь найдем отношение сигнал/помеха. При амплитудной модуляции наибольшее приращение амплитуды равно ей самой (100%-ная модуляция), так что $\Delta a_1 = a_0$. Подставляя в (7.4), находим для АМ

$$\rho_a = 2a_0^2 T/A_0 = 2FT\rho_0,$$

где ρ_0 — отношение сигнал/помеха на входе приемника[1]. Таким образом, выигрыш в отношении сигнал/помеха при АМ получается только за счет накопления, т. е. за счет интегрирования на интервале T.

[1] См. сноску [3] на стр. 270.

Совершенно иначе обстоит дело с частотной модуляцией. При ЧМ задается частотное отклонение (девиация), так что $\Delta a_2 = \Delta\omega$ и для отношения сигнал/помеха имеем

$$\rho_\omega = \frac{2(\Delta\omega)^2 a_0^2 T^3}{3A_0} = \frac{(\Delta\omega)^2 a_0^2 T}{6F^2 A_0} = \frac{2\pi^2}{3}\left(\frac{\Delta\omega}{\Omega}\right)^2 \frac{a_0^2 T}{A_0} = \frac{\pi^2}{3}\beta^2 2FT\rho_0 = \frac{\pi^2}{3}\beta^2\rho_a.$$

Таким образом, отношение сигнал/помеха для ЧМ в $\frac{\pi^2}{3}\beta^2 \simeq 3{,}3\beta^2$ раз больше, чем для АМ. Здесь

$$\beta = \Delta\omega/\Omega = \Delta\omega/2\pi F —$$

индекс ЧМ. Этот выигрыш в помехоустойчивости получается, конечно, не даром; мы расплачиваемся за него расширением спектра модулированного сигнала. При больших индексах ширина спектра ЧМ прямо пропорциональна индексу (ширина спектра равна удвоенному частотному отклонению).

В качестве второго примера возьмем переносчик в виде периодической последовательности трапецеидальных импульсов. Импульсы будут определяться тремя параметрами: высотой («амплитудой») h, моментом начала («фазой») t и длительностью τ. Модуляция по каждому из этих трех параметров есть соответственно АИМ, ФИМ и ДИМ. Длительность фронтов обозначена через μ и остается постоянной. График функции

$$f = f(h,\ t_1,\ \tau,\ t)$$

на протяжении одного периода T показан на рис. 9. Аналитическая запись функции f может быть представлена в виде

$$f = \frac{h}{\mu}[(t-t_1)\sigma(t-t_1) - (t-t_1-\mu)\sigma(t-t_1-\mu) - (t-t_1-\tau- \\ -\mu)\sigma(t-t_1-\tau+\mu) + (t-t_1-\tau)\sigma(t-t_1-\tau)].$$

Дифференцируя это выражение, находим все три частные производные по параметрам

$$\frac{df}{da_2} = \frac{df}{dt_1} = \frac{h}{\mu}[-\sigma(t-t_1) + \sigma(t-t_1-\mu) + \\ + \sigma(t-t_1-\tau+\mu) - \sigma(t-t_1-\tau)],$$

$$\frac{df}{da_3} = \frac{df}{d\tau} = \frac{h}{\mu}[\sigma(t-t_1-\tau+\mu) - \sigma(t-t_1-\tau)].$$

Графики двух последних производных изображены на рис. 10. Матрица коэффициентов b_{ik} имеет вид

$$b_{ik} = \begin{pmatrix} \frac{\mu}{T}\left(\frac{\tau}{\mu} - \frac{4}{3}\right) & 0 & \frac{h}{2T} \\ 0 & \frac{2h^2}{T\mu} & \frac{h^2}{T\mu} \\ \frac{h}{2T} & \frac{h^2}{T\mu} & \frac{h^2}{T\mu} \end{pmatrix}.$$

Для дальнейших вычислений примем следующие упрощения:

1. $\alpha = \tau/\mu \gg 1$, т. е. будем полагать относительную крутизну фронтов большой или, иначе говоря, длительность фронтов малой по сравнению с длительностью импульса.

2. $\Omega\tau \gg 1$, что следует из того, что обычно полоса пропускания Ω выбирается так, что $\Omega\mu \simeq 0{,}4$ и, следовательно, $\Omega\tau = \Omega\mu\tau/\mu \simeq 0{,}4\tau/\mu \gg 1$.

Опуская выкладки, приведем сразу значения дисперсий паразитных приращений параметров $D\delta h = A_0/2\tau_0$, $D\delta t_1 = 0{,}064A_0\mu/h_0^2$, $D\delta\tau = 0{,}128A_0\mu/h_0^2$ (нуликами обозначены начальные, т. е. немодулированные значения параметров).

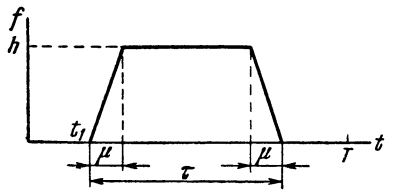


Рис. 9

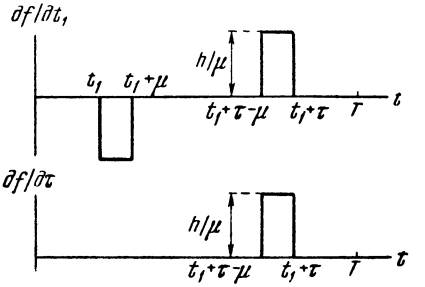


Рис. 10

Остается ввести полезные приращения. Для АИМ имеем $\Delta h = h_0$ (т. е. амплитуда может меняться от нуля до удвоенного значения). Для ФИМ t_1 может получать приращение от нуля до $T/2$ в обе стороны, если $t_{10} = T/2$ и если скважность велика, т. е. $\nu = T/\tau \gg 1$.

В этих условиях $\Delta t_1 = T/2$. Наконец, в случае ДИМ, выбирая $\tau_0 = T/2$ (т. е. $\nu_0 = 2$), имеем $\Delta\tau = T/2$. Пользуясь этими значениями, находим отношение сигнал/помеха.

Для АИМ

$$\rho_A = \frac{(\Delta h)^2}{D\delta h} = \frac{2h_0^2\tau_0}{A_0} = \frac{2E_0}{A_0},$$

где $E_0 = h_0^2\tau_0$ — энергия импульса.

Для ФИМ

$$\rho_\Phi \simeq 3{,}76T^2h_0^2/A_0\mu = 3{,}76\nu_0^2\alpha_0E_0/A_0,$$

где $\nu_0 = T/\tau_0$ — начальная скважность, $\alpha_0 = \tau_0/\mu$ — относительная крутизна фронтов.

Наконец, для ДИМ

$$\rho_D \simeq 1{,}88\ T^2h_0^2/A_0\mu = 1{,}88\nu_0^2\alpha_0E_0/A_0 = 7{,}52\alpha_0E_0/A_0.$$

Итак, для отношений сигнал/помеха для трех рассматриваемых видов импульсной модуляции можно записать

$$\rho A : \rho\Phi : \rho D = 1 : 1{,}88\nu_0^2\alpha_0 : 3{,}76\alpha_0,$$

откуда легко сделать заключение о значительных преимуществах ФИМ. Помехоустойчивость ФИМ и ДИМ возрастает с увеличением относительной крутизны $\alpha = \tau/\mu$, т. е. с увеличением полосы пропускания.

По поводу паразитной модуляции, обусловленной аддитивной помехой, можно еще заметить следующее.

Величины δa_k оказываются коррелированными между собой. Из этого следует, что если для полезной модуляции используется только часть параметров, в простейшем случае один, то остальные можно использовать для компенсации паразитной модуляции. Компенсация может быть осуществлена путем вычитания из выходного сигнала линейной комбинации напряжений, пропорциональных паразитным приращениям нерабочих параметров. Оптимальные коэффициенты этой комбинации определяются коэффициентами корреляции между величинами δa_k, взятыми попарно.

С другой стороны, если величины δa_k коррелированы слабо, то возможен другой способ улучшения отношения сигнал/помеха. Способ этот состоит в том, что одним и тем же сигналом модулируется несколько параметров переносчика. Это равносильно образованию нескольких каналов. Если в отношении действия помехи эти каналы статистически независимы, то, сложив выходные сигналы всех каналов, мы получим увеличение отношения сигнал/помеха в n раз, где n — число каналов, т. е. число модулируемых параметров. Речь идет, таким образом, о методе накопления, о котором подробно говорится в § 9.

Итак, для ослабления влияния помехи целесообразно использовать все параметры переносчика. При большой корреляции можно воспользоваться возможностью компенсации, при малой корреляции — методом накопления, а в промежуточных случаях, возможно, комбинацией обоих методов.

§ 8. Обнаружение при однократном отсчете

Рассмотрим для начала простейший вариант задачи обнаружения сигнала. Условия задачи пусть будут таковы:

1. На сигнал s наложена аддитивная помеха ξ.

2. Способ приема состоит в том, что в некоторый момент $t = t_0$ берется отсчет мгновенного значения сигнала $s(t_0) = a > 0$; это значение считается заранее известным.

Если сигнал налицо, то отсчет равен $a + \xi$; если сигнал отсутствует, то отсчет равен ξ. Приемник содержит решающее устройство, дающее ответ «да» или «нет» на вопрос о наличии сигнала на входе.

С геометрической точки зрения задача является одномерной, так как сигнал определяется одной-единственной координатой. Следовательно, собственная область сигнала представляется отрезком числовой оси. Отметим на оси x порогового значения

$x_0 < a$, и пусть собственная область сигнала есть отрезок $x_0 < x < \infty$.

Действие решающего устройства состоит в том, что если входное напряжение больше x_0, устройство выдает решение «да» (сигнал есть), в противном случае — решение «нет» (сигнала нет).

Составим выражение для верности ошибки. Пусть $w_0(x)$ означает распределение помехи, а $w_a(x)$ — распределение суммы сигнала и помехи. Тогда условная вероятность получить решение «нет», когда сигнал фактически передается, есть

$$p_a(0) = p\{a + \xi < x_0\} = \int_{-\infty}^{x_0} w_a(x)\,dx, \tag{8.1}$$

а условная вероятность получить решение «да» при отсутствии сигнала равна

$$p_0(a) = p\{\xi > x_0\} = \int_{x_0}^{\infty} w_0(x)\,dx. \tag{8.2}$$

Если обозначить $p(a)$ и $p(0)$ соответственно априорные вероятности передачи сигнала (посылки) и его отсутствия (паузы), то полная вероятность ошибки запишется в виде

$$p_{\text{ош}} = p(a)\,p_a(0) + p(0)\,p_0(a) = \\ = p(a)\int_{-\infty}^{x_0} w_a(x)\,dx + p(0)\int_{x_0}^{\infty} w_0(x)\,dx. \tag{8.3}$$

Так как a — постоянная, то имеем

$$w_a(x) = w_0(x - a).$$

Положим для упрощения

$$p(a) = p(0) = {}^1\!/_2,$$

т. е. что посылка и пауза равновероятны. Тогда вместо (8.3) имеем

$$p_{\text{ош}} = \frac{1}{2}\left(\int_{-\infty}^{x_0-a} w_0(x)\,dx + \int_{x_0}^{\infty} w_0(x)\,dx = \right. \\ \left. = \frac{1}{2}\left(1 - \int_{x_0-a}^{x_0} w_0(x)\,dx\right) = \frac{1}{2}\left(1 - \int_0^a w_0(x_0 - x)\,dx\right)\right). \tag{8.3а}$$

Здесь использовано условие нормировки вероятностей

$$\int_{-\infty}^{\infty} w_0(x)\,dx = 1.$$

Плотность распределения вероятностей $w(x)$ имеет размерность x^{-1}. Вводя новую переменную $y = x/\sigma$, перейдем к безразмерной функции $v(y)$, определяемой соотношением

$$w(x)\,dx = v(y)\,dy,$$

т. е.

$$v(y) = \sigma w(\sigma y),$$

где $\sigma = \sqrt{D\xi}$ — среднеквадратичное значение помехи. Получим

$$p_{\text{ош}} = \frac{1}{2}\left(1 - \int_0^{a/\sigma} v(y_0 - y)\,dy\right), \tag{8.4}$$

где $y = x_0/\sigma$. Интеграл есть функция верхнего предела, которую мы обозначим через $f_0(a/\sigma)$. Но $w(x)$ и $v(y)$ — положительны; следовательно, эта функция есть возрастающая функция своего аргумента. Таким образом, вероятность ошибки с возрастанием a/σ.

Введя обозначение отношения сигнал/помеха $\rho = a^2/\sigma^2$, можем записать (8. 4) в виде

$$p_{\text{ош}} = \frac{1}{2}(1 - f_0(\sqrt{\rho})). \tag{8.5}$$

Итак, верность возрастает с увеличением отношения сигнал/помеха. Никаких ограничений на характер распределения помехи при выводе этого заключения не накладывалось.

Определим теперь собственную область сигнала так, чтобы вероятность ошибки была наименьшей. Приемник, удовлетворяющий этому требованию, называется, по В. А. Котельникову, идеальным. Параметром, определяющим границу собственной области, является пороговое значение x_0. Поэтому для минимизации вероятности ошибки достаточно продифференцировать (8. 3) по x_0

$$\frac{dp_{\text{ош}}}{dx_0} = p(a)\,w_a(x_0) - p(0)\,w_0(x_0).$$

Приравняв эту производную нулю, получим уравнение для значения x_0, соответствующего экстремальным значениям $p_{\text{ош}}$. Но предварительно нужно установить характер экстремума. Для этого найдем вторую производную

$$\frac{d^2p_{\text{ош}}}{dx_0^2} = p(a)\,w_a'(x_0) - p(0)\,w_0'(x_0).$$

Если эта величина положительна, то в точке, абсцисса которой определяется из уравнения

$$\frac{dp_{\text{ош}}}{dx_0} = 0,$$

имеется минимум $p_{\text{ош}}(x_0)$; если же отрицательна, то максимум. Полагая $p(a)=p(0)$, запишем условие минимума

$$w_a(x_0) > w_0(x_0). \tag{8.6}$$

Если это условие выполнено, то оптимальное значение находится из уравнения

$$w_a(x_0) = w_0(x_0), \tag{8.7}$$

т. е. x_0 есть абсцисса точки пересечения кривых $\omega_a(x)$ и $\omega_0(x)$. На рис. 11 изображены эти кривые. Рисунок наглядно поясняет,

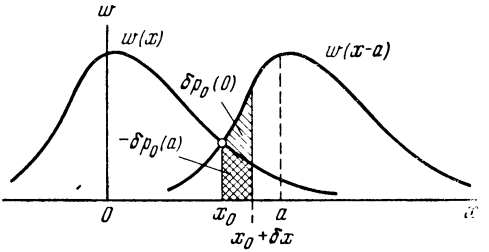


Рис. 11

почему найденное значение x_0 минимизирует вероятность ошибки. Пусть пороговое значение x_0 получило приращение δx. Тогда условная вероятность $p_a(0)$ возрастет, получив приращение $\delta p_a(0)$ (заштрихованная площадь); условная вероятность $p_0(a)$ уменьшается, получив отрицательное приращение $\delta p_0(a)$ (дважды заштрихованная площадь). Сумма этих приращений положительная, если выполнено условие (8. 6). Легко видеть, что для этого достаточно, чтобы $w_0(x)$ была монотонно убывающей функцией $|x|$. При этом условии всякое отклонение от значения x_0, удовлетворяющего уравнению (8. 7), увеличивает вероятность ошибки [1].

Мы имеем

$$w_a(x) = w_0(x-a),$$

так что (8. 7) можно записать в виде

$$w_0(x_0) = w_0(x_0-a). \tag{8.8}$$

Если $w_0(x)$ — четная функция, то, меняя знак аргумента в правой части (8. 8), получим

$$w_0(x_0) = w_0(a-x_0).$$

Здесь оба аргумента положительны. Приравняв их, находим

$$x_0 = \frac{1}{2}a.$$

[1] Случай немонотонного распределения рассмотрен в Добавлении III.

Для этого случая, полагая по-прежнему $p(a) = p(0)$, получаем из (8.3а)

$$p_{\text{ош}} = \frac{1}{2}\left(\int\limits_{-\infty}^{1/2a} w_0(x)\,dx + \int\limits_{1/2}^{\infty} w_0(x)\,dx\right) = \int\limits_{a/2}^{\infty} w_0(x)\,dx, \qquad (8.9)$$

или

$$p_{\text{ош}} = \frac{1}{2} - \int\limits_{0}^{a/2} w_0(x)\,dx. \qquad (8.10)$$

Можно также на основании (8.9) представить вероятность ошибки в виде

$$p_{\text{ош}} = p\left\{\xi > \frac{1}{2}a\right\}. \qquad (8.11)$$

Эта форма будет нам встречаться и в последующих параграфах.

Переходя к безразмерным функциям, представим (8.10) в виде

$$p_{\text{ош}} = \frac{1}{2} - \int\limits_{0}^{\frac{1}{2}\cdot\frac{a}{\sigma}} v_0(y)\,dy = \frac{1}{2} - f\left(\frac{1}{2}\sqrt{\rho}\right), \qquad (8.12)$$

где

$$f(z) = \int\limits_{0}^{z} v_0(y)\,dy.$$

Из формулы (8.12) следует, что вероятность ошибки зависит только от распределения помехи (функция f) и отношения сигнал/помеха (аргумент $\sqrt{\rho}/2$). Но не нужно забывать, что этот результат получен для заданного сигнала (постоянная величина a) и для заданного способа приема (однократный отсчет). Поэтому влияние выбора вида сигнала и способа приема и не нашло в этой формуле отражения.

Рассмотрим в качестве примера случай нормального распределения помехи:

$$w_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x/2\sigma^2}.$$

Безразмерная функция распределения

$$v_0(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Вероятность ошибки по формуле (8.12) равна

$$p_{\text{ош}} = \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{a/2\sigma} e^{-y^2/2}\,dy$$

или

$$p_{\text{ош}} = \frac{1}{2} - \frac{1}{\sqrt{\pi}} \int_0^{a/2\sqrt{2}\sigma} e^{-t^2} dt = \frac{1}{2}[1 - \Phi(z)], \qquad (8.13)$$

где Φ — интеграл вероятностей [1], называемый также функцией Лапласа или функцией Крампа,

$$z = \frac{1}{2\sqrt{2}} \cdot \frac{a}{\sigma} = \sqrt{\frac{1}{8}\rho}. \qquad (8.14)$$

Для вычисления при малых вероятностях ошибки может оказаться полезным асимптотическое разложение

$$1 - \Phi(z) \sim \frac{e^{-z^2}}{\sqrt{\pi}z}\left(1 - \frac{1}{2z^2} + \frac{1 \cdot 3}{(2z^2)^2} + \cdots\right).$$

Верность по принятому нами определению равна

$$S = \lg \frac{1}{p_{\text{ош}}} = -\lg \frac{1}{2}[1 - \Phi(z)]. \qquad (8.15)$$

Асимптотическое выражение для верности имеет вид

$$S \sim 0{,}55 + \frac{1}{2}\lg z^2 + 0{,}434z^2, \qquad (8.16)$$

где $z^2 = \rho/8$. График зависимости (8.16) представлен на рис. 12. Напомним еще раз, что этот график относится к вполне определенному частному случаю; он выражает верность обнаружения при гауссовой аддитивной помехе методом однократного отсчета.

Рассмотренный пример с технической точки зрения можно трактовать как прием телеграфного сигнала методом пробы при телеграфировании постоянным током. Обратимся теперь к радиотелеграфному сигналу, выражаемому на протяжении посылки синусоидальным колебанием высокой частоты. И в этом случае мы будем предполагать аддитивную гауссову помеху, но отсчет будет браться не на входе приемника, а на выходе линейного детектора. Иначе говоря, будут сравниваться отсчеты огибающих при наличии и отсутствии сигнала.

[1] Заметим во избежание недоразумения, что мы пользуемся обозначением

$$\Phi(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt.$$

В литературе применяются и другие обозначения. Часто встречается

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx.$$

Очевидно,

$$\Phi(z) = 2F(\sqrt{2}z) - 1.$$

Особенность рассматриваемого случая состоит в том, что распределения для огибающих при отсутствии и наличии сигнала выражаются разными функциями, а именно

$$w_0(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, \tag{8.17}$$

$$w_a(x) = \frac{x}{\sigma^2} e^{-\frac{x^2+a_0^2}{2\sigma^2}} I_0\left(\frac{a_0 x}{\sigma^2}\right). \tag{8.18}$$

Распределение (8. 17) есть рэлеево распределение, уже упоминавшееся в § 3, а распределение (8. 18) называется обобщенным

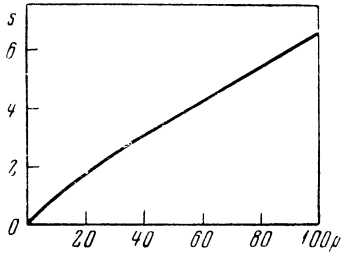


Рис. 12

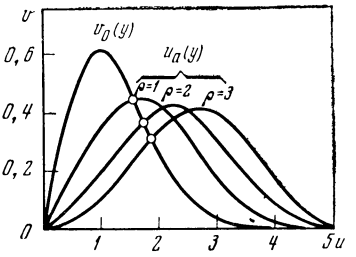


Рис. 13

рэлеевым распределением. Входящая в (8. 18) величина a_0 означает амплитуду синусоидального сигнала на входе.

Так как огибающая, по определению, положительна, то распределения (8. 17) и (8. 18) относятся к значениям $x \geqslant 0$. Распределения (8. 17) и (8. 18) можно заменить соответствующими безразмерными функциями

$$v_0(y) = ye^{-\frac{1}{2}y^2}, \tag{8.19}$$

$$v_a(y) = ye^{-\left(\frac{1}{2}y^2+\rho\right)} I_0(y\sqrt{2\rho}), \tag{8.20}$$

где $y=x/\sigma$. Величина $\rho=a_0^2/2\sigma^2$ означает отношение сигнал/помеха на входе приемника. Полагая, как и раньше,

$$p(0) = p(a) = 1/2,$$

т. е. считая посылки и паузы равновероятными, будем иметь уравнение порога x_0 в виде (8. 7),

$$w_a(x_0) = w_0(x_0),$$

или в безразмерных функциях,

$$v_a(y_0) = v_0(y_0). \tag{8.21}$$

Подставляя сюда (8. 19) и (8. 20), получаем для порога трансцендентное уравнение

$$I_0(y_0\sqrt{2\rho}) = e^{\rho},$$

где $y_0 = x_0/\sigma$. Решения этого уравнения представлены абсциссами точек пересечения кривых (8. 19) и (8.20), изображенных на рис. 13. Из этого рисунка видно, что при увеличении отношения сигнал/помеха распределения v_0 и v_a все более раздвигаются, а потому верность возрастает.

Вероятность ошибки вычисляется по общей формуле (8. 3), для практического применения которой в данном случае нужно располагать таблицами интегральных рэлеевых распределений.

§ 9. Обнаружение методом накопления

Одним из сильных методов борьбы с помехами, известным в многочисленных вариантах и играющим в технике передачи сигналов видную роль, является метод накопления. Мы рассмотрим сейчас этот метод в применении к простейшей задаче обнаружения постоянного сигнала.

Итак, пусть передаются посылки, на протяжении которых сигнал имеет постоянное значение a. На сигнал наложена помеха ξ. В посылке имеем $a+\xi$, в паузе только помеху ξ. Требуется отличить посылку от паузы, т. е. констатировать наличие сигнала. Это в точности та же задача, которая была поставлена в начале предыдущего параграфа. Но способ приема мы теперь применим иной. Он состоит в том, что на протяжении посылки берется не один отсчет (однократный отсчет по терминологии предыдущего параграфа), а несколько. Число отсчетов мы обозначим через n. Затем все отсчеты складываются в некотором суммирующем устройстве — накопителе. Решающее устройство, дающее ответ «да» или «нет» на вопрос о наличии сигнала, подключено к выходу накопителя и выносит свое суждение не по отдельному отсчету, а по сумме n отсчетов. Как сейчас будет показано, такой способ приема позволяет увеличить верность при заданном отношении сигнал/помеха на входе приемника.

Отдельные отсчеты можно представить в виде

$$x_1 = a + \xi_1, \quad x_2 = a + \xi_2, \ \ldots, \quad x_n = a + \xi_n,$$

где ξ_k — значение помехи в момент k-го отсчета.

Сумма отсчетов равна

$$x = \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} (a + \xi_k) = na + \sum_{k=1}^{n} \xi_k = b + \eta.$$

Величина $b = na$ представляет собой полезный сигнал на входе решающего устройства. Случайная величина

$$\eta = \sum_{k=1}^{n} \xi_k$$

представляет собой помеху. Отношение сигнал/помеха на входе решающего устройства выражается так:

$$\rho = \frac{b^2}{D\eta} = \frac{n^2a^2}{D\left(\sum \xi_k\right)},$$

где D — дисперсия (мы полагаем, что $M\left(\sum \xi_k\right) = 0$). Предположим вначале, что значения ξ не коррелированы. Тогда дисперсия суммы равна сумме дисперсий, и мы получаем

$$\rho = \frac{n^2a^2}{\sum D\xi_k},$$

а так как все ξ_k имеют одинаковое распределение (потому что они являются мгновенными значениями одного и того же случайного процесса $\xi(t)$), то

$$\rho = \frac{n^2a^2}{nD\xi} = \frac{na^2}{\sigma^2}.$$

Вводя отношение сигнал/помеха на входе приемника

$$\rho_0 = P_a/P_\xi = a^2/\sigma^2,$$

получаем окончательно

$$\rho = n\rho_0. \tag{9.1}$$

Таким образом, при описанных условиях накопление отсчетов приводит к увеличению отношения сигнал/помеха на входе решающего устройства ровно в n раз (по сравнению с отношением сигнал/помеха на входе приемника, являющимся заданной величиной). Соответственно возрастает и верность; для вероятности ошибки остается в силе формула (8. 9). Заметим, что для величины η распределение нормально, если ξ_k имеют нормальное распределение; по теореме Ляпунова для η можно принять нормальное распределение и в том случае, когда ξ_k распределены по какому-либо другому закону (если их число достаточно велико). Если так, то вероятность ошибки и верность можно определять по формулам (8. 10) и (8. 12). Теоретически возможно обнаружить описанным способом сколь угодно слабый сигнал. При этом, разумеется, время наблюдения возрастает пропорционально n, если отсчеты берутся с некоторым постоянным интервалом во времени.

Мы предполагали, что сигнал есть постоянная величина, но тот же результат получится, если сигнал представляет собой произвольную периодическую функцию, а отсчеты берутся синхронно с интервалом, равным периоду (в моменты, задаваемые синхронной последовательностью стробирующих импульсов [1]).

[1] Накопление применимо и в случае произвольного сигнала при условии, что среднее значение отсчетов не равно нулю. Мы имеем в этом случае

$$x = \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} (s_k + \xi_k) = n\frac{1}{n}\sum_{k=1}^{n} s_k + \sum_{k=1}^{n} \xi_k = b + \eta,$$

Отбросим теперь предположение о независимости величин. Найдем дисперсию суммы зависимых случайных величин. Мы имеем $\left(\text{полагая } M\left(\sum_k \xi_k\right) = 0\right)$

$$D\left(\sum_k \xi_k\right) = M\left(\sum_k \xi_k\right)^2 = M\left(\sum_k \xi_k^2 + \sum_i \sum_k \xi_i \xi_k\right).$$

Займемся отдельно второй суммой. Она берется по всем $i \neq k$. Введем $l = |k - i|$. Тогда

$$\sum_i \sum_k \xi_i \xi_k = 2 \sum_i \sum_l \xi_i \xi_{i+l} + l = 2 \sum_{l=1}^{n-1} (n - l)\, \xi\xi_l,$$

где $\xi\xi_l$ означает любую пару значений, порядковые номера которых различаются на l. Среднее значение

$$M(\xi\xi_l) = R(l)$$

называется коэффициентом корреляции. Итак,

$$D\left(\sum_k \xi_k\right) = M\left(\sum_{k=1}^{n} \xi_k^2 + 2 \sum_{l=1}^{n-1} (n - l)\, \xi\xi_l\right) =$$

$$= \sum_{k=1}^{n} M\xi_k^2 + 2 \sum_{l=1}^{n-1} (n - l)\, M(\xi\xi_l) = nD\xi + 2 \sum_{l=1}^{n-1} (n - l)\, R(l)$$

или, окончательно,

$$D\left(\sum \xi_k\right) = nD\xi \left[1 + \frac{2}{n} \sum_{l=1}^{n-1} (n - l)\, k(l)\right], \tag{9.2}$$

где $k(l) = R(l)/D\xi$ — нормированный коэффициент корреляции. При $k(l) \equiv 0$ второе слагаемое в правой части (9.2) пропадает, и мы имеем $D\left(\sum \xi_k\right) = \sum D\xi_k$, т. е. дисперсия суммы равна сумме дисперсий.

Составляя теперь выражение для отношения сигнал/помеха, находим

$$\rho = \frac{n^2 a^2}{D\left(\sum \xi_k\right)} = \frac{n^2 a^2}{nD\xi\,(1 + \lambda)} = \frac{na^2}{\sigma^2 (1 + \lambda)}$$

или

$$\rho = \frac{n}{1 + \lambda}\, \rho_0, \tag{9.3}$$

где $b = \overline{ns}$ — полезный сигнал на входе накопителя, а $s = \frac{1}{n} \sum_{k=1}^{n} s_k$ — среднее значение отсчетов. Отношение сигнал/помеха вычислено в Добавлении X.

где обозначено для краткости

$$\lambda = \frac{2}{n} \sum_{l=1}^{n-1} (n - l)\, k\,(l). \qquad (9.4)$$

Итак, при наличии корреляции между значениями помехи в моменты отсчета выигрыш в отношении сигнал/помеха всегда меньше, чем в случае независимых значений помехи. Рассмотрим несколько примеров.

1. Пусть $k\,(1)=k$, $k\,(2)=k\,(3)= \ldots =k\,(n-1)=0$, т. е. пусть корреляционные связи существуют только между соседними значениями ξ. Тогда

$$\lambda = 2\frac{n-1}{n}\, k \simeq 2k$$

(если n достаточно велико) и

$$\rho = \frac{n}{1+2k}\, \rho_0.$$

2. Пусть $k\,(1)=k\,(2)=\ldots=k\,(s)=k$; $k\,(s+1)=\ldots=k\,(n-1)=0$, т. е. пусть корреляционные связи распространяются на значения, отстоящие друг от друга на s номеров. Тогда

$$\lambda = \frac{2}{n} \sum_{l=1}^{s} (n - l)\, k = 2ks\left(1 - \frac{s+1}{2n}\right).$$

Если $s=n-1$, т. е. любая пара значений имеет один и тот же коэффициент корреляции k, то

$$\lambda = (n-1)\, k.$$

3. Пусть $k\,(l)=e^{-\alpha l}$. Тогда

$$\lambda = \frac{2}{n} \sum_{l=1}^{n-1} (n - l)\, e^{-\alpha l}.$$

Это есть арифметико-геометрическая прогрессия. Пользуясь общей формулой для суммы членов такой прогрессии, получаем

$$\lambda = \frac{2y}{1-y}\left(1 - \frac{1-y^n}{n\,(1-y)}\right),$$

где $y=e^{-\alpha}$. При $n \gg 1$

$$\lambda \simeq \frac{2y}{1-y} = \frac{2}{e^{\alpha}-1}.$$

Чем больше α, т. е. чем быстрее затухают корреляционные связи, тем меньше λ, следовательно, тем больше выигрыш, выражаемый отношением ρ/ρ_0.

Теперь заметим, что, осуществляя метод накопления, можно взять не сумму дискретных отсчетов x_k, а интеграл непрерывно изменяющейся функции

$$x(t) = a + \xi(t)$$

за время T, равное длительности посылки. Это соответствует переходу от пространства R_n к пространству C_{L^2}. Таким образом, мы будем рассматривать величину

$$\int_0^T x(t)\,dt = aT + \int_0^T \xi(t)\,dt = b + \eta,$$

где b — постоянная, выражающая полезный сигнал на выходе накопителя (интегратора), а η — случайная величина, выражающая помеху на выходе накопителя. Для определения отношения сигнал/помеха найдем дисперсию случайной величины η.

Мы имеем

$$D\eta = M\eta^2 = M\left(\int_0^T \xi(t)\,dt\right)^2 = M\int_0^T\int_0^T \xi(t)\xi(t_1)\,dt\,dt_1 =$$

$$= \int_0^T dt \int_0^T M[\xi(t)\xi(t_1)]\,dt_1$$

или

$$D\eta = \int_0^T dt \int_0^T B_\xi(t - t_1)\,dt_1, \tag{9.5}$$

где $B_\xi(\tau)$ — функция корреляции помехи $\xi(t)$. Для вычисления D_η эта функция должна быть известна [1]. Но кое-что можно сделать и в общем виде.

Рассмотрим отдельно внутренний интеграл в (9.5)

$$C = \int_0^T B_\xi(t - t_1)\,dt_1.$$

Вводя переменную

$$\tau = t - t_1,$$

запишем

$$C = \int_{t-T}^{t} B_\xi(\tau)\,d\tau = C(t).$$

Так как $B_\xi(\tau)$ — четная функция, будем, кроме того, считать ее убывающей так, что интеграл достигает максимума при сим-

[1] Пример вычисления для случая ограниченного по полосе белого шума дан в Добавлении I.

метричном расположении интервала T относительно начала, т. е. при $t=T/2$. Таким образом,

$$C \leqslant \max C = \int_{-T/2}^{T/2} B(\tau)\, d\tau.$$

Если интервал T достаточно велик в том смысле, что на краях этого интервала функция корреляции мала (по сравнению с $B(0)$), то пределы можно заменить бесконечными, и мы получим

$$C \leqslant \max C < \int_{-\infty}^{\infty} B(\tau)\, d\tau.$$

Введем определение интервала корреляции

$$\tau_0 = \frac{1}{B(0)} \int_{-\infty}^{\infty} B(\tau)\, d\tau_1,$$

т. е. определим τ_0 как основание прямоугольника, высота которого есть $B(0)$, а площадь равна площади под кривой (рис. 14) [1]. Тогда

$$C < B(0)\,\tau_0 = P_\xi \tau_0,$$

и мы получаем оценку для D_η

$$D\eta = \int_0^T dt \int_{t-T}^{t} B(\tau)\, d\tau < P_\xi \tau_0 T.$$

Отношение сигнал/помеха на выходе накопителя определим как

$$\rho = b^2/D\eta.$$

Подставляя найденные значения, получаем

$$\rho > \frac{a^2 T}{P_\xi \tau_0} = \frac{T}{\tau_0}\rho_0. \qquad (9.6)$$

Итак, выигрыш, получаемый в результате интегрирования, тем больше, чем больше отношение T/τ_0 *.

Описанный метод известен под названием интегрального приема. Как мы увидим в дальнейшем, этот метод является частным случаем некоторого общего оптимального метода приема.

Заметим, что $n=T/\tau_0$ — это не что иное, как число некоррелированных значений помехи на интервале T.

Таким образом, и по смыслу и по порядку величины формула (9. 6) совпадает с (9. 1). Другими словами, добавление зависимых

---

[1] Это простое определение, к сожалению, не универсально. Его удобно применять только, если $B(\tau)$ нигде не отрицательна.

* См. по этому поводу Добавление II.

значений (отстоящих во времени на интервал, меньший τ_0) не влияет существенно на результат, и замена суммирования независимых значений непрерывным интегрированием дополнительного выигрыша не дает. Однако нужно учесть, что технически интегрирование осуществляется гораздо проще суммирования дискретных значений.

По поводу метода накопления следует сделать несколько общих замечаний. Прежде всего поясним, что суть дела заключается в том, что при суммировании нескольких отсчетов суммируются как полезные сигналы, так и мгновенные значения помехи.

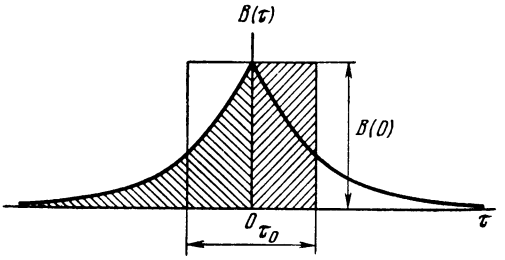


Рис. 14

Но полезный сигнал во всех слагаемых один и тот же, поэтому сумма растет как n, а ее квадрат как n^2. Сумма же случайных величин растет по другому закону. Если эти величины независимы, то суммируются средние квадраты, так что средний квадрат суммы пропорционален первой степени n. Отсюда и следует выигрыш в n раз в отношении сигнал/помеха в случае, когда значения помехи независимы.

Теперь заметим, что получение нескольких значений x_k вовсе не обязательно связано с отсчетами, которые берутся в разные моменты времени. Мы могли бы получить n значений x_k, если бы один и тот же сигнал передавался по n независимым каналам. Под независимыми каналами понимаются такие каналы, в которых действуют независимые помехи. Способ образования этих каналов не играет никакой роли с точки зрения существа метода накопления. Так, можно применить каналы, разделенные по частоте, или в пространстве, или по направлению поляризации электромагнитной волны и т. д. Описанные выше отсчеты можно представить как образование нескольких каналов, разделенных во времени.

Таким образом, техническое осуществление метода накопления связано с построением многоканальной системы передачи одного и того же сигнала.

Можно высказать следующее положение: существует столько вариантов технического выполнения метода накопления, сколько имеется различных способов разделения каналов.

Общие теоретические соображения, относящиеся к эффективности метода накопления, при любом варианте его осуществления остаются неизменными.

Метод накопления позволяет увеличить отношение сигнал/помеха без увеличения мощности сигнала. Однако за этот выигрыш приходится расплачиваться. При временном разделении каналов возрастает время передачи, при частотном — занимаемая полоса (причем и то и другое в n раз). При любом другом способе разделения, не затрагивающем времени и частоты, цена выигрыша — усложнение аппаратуры и увеличение затрат.

§ 10. Оптимальный линейный приемник

Обсуждая задачу обнаружения сигнала, мы рассматривали до сих пор лишь частные случаи, когда задан какой-либо определенный простейший вид сигнала. Теперь мы поставим задачу в значительно более общем виде, имея в виду пространство C_{L^2}. Пусть на интервале $0 < t < T$ сигнал задан произвольной (но известной) функцией времени $s(t)$. Требуется обнаружить этот сигнал при наличии аддитивной помехи $\xi(t)$. Иначе говоря, нужно установить, представляет ли входной сигнал $x(t)$ сумму $s(t)$ и $\xi(t)$ или он определяется только помехой $\xi(t)$. С этой целью мы подвергнем входной сигнал специальной обработке, результатом которой будет некоторая величина, поступающая на вход решающего устройства. Если указанная обработка может быть представлена линейной операцией, то величина на входе решающего устройства распадается на два слагаемых, из которых одно зависит только от сигнала, а второе — только от помехи. При этих условиях можно говорить об отношении сигнал/помеха после обработки входного сигнала, и задача теперь сводится к отысканию такой оптимальной обработки, которая дает наибольшее отношение сигнал/помеха на входе решающего устройства при заданном отношении сигнал/помеха на входе приемника. Единственное ограничение, которое мы наложим на операцию обработки входного сигнала, будет состоять в том, что операция эта линейна, т. е. выражается линейным функционалом [1].

Как сказано в § 4, общим видом линейного функционала является скалярное произведение

$$y = \Phi(x) = x\varphi = \int_0^T x(t)\,\varphi(t)\,dt, \tag{10.1}$$

[1] Мы ограничиваемся рассмотрением линейной операции, для которой доказано, что она является оптимальной только для гауссовых процессов. Теория нелинейных операций в настоящее время еще недостаточно развита. Наглядный пример нелинейной оптимальной операции можно найти в Добавлении III.

где $\varphi(t)$ — некоторая весовая функция. Подставляя в (10. 1) значение $x(t)=s(t)+\xi(t)$, получим

$$y=\int_0^T s(t)\varphi(t)\,dt+\int_0^T \xi(t)\varphi(t)\,dt=b+\eta,$$

где b — постоянная величина, выражающая полезный сигнал, а случайная величина η характеризует помеху.

Схема приемника показана на рис. 15, на котором Ф означает преобразователь, выполняющий операцию (10. 1), Р — решающее устройство, Г — генератор весовой функции.

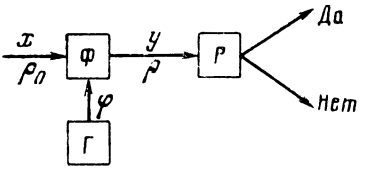


Рис. 15

Повторяя рассуждения § 8, получим для случая четного монотонного распределения помехи η следующую формулу для вероятности ошибки [1]:

$$P_{\text{ош}}=p\left\{\eta>\frac{1}{2}b\right\} \tag{10.2}$$

(ср. формулу (8. 11)) или

$$p_{\text{ош}}\doteq\frac{1}{2}-f\left(\frac{1}{2}\sqrt{\rho}\right), \tag{10.2а}$$

(формула (8. 12)), где

$$\rho=\frac{b^2}{D\eta}.$$

Поставим себе задачу подобрать весовую функцию φ так, чтобы максимизировать полезный сигнал b, т. е. функционал

$$b=\int_0^T s(t)\varphi(t)\,dt.$$

Это, вообще говоря, вариационная задача, но она решается очень просто. Воспользуемся неравенством Буняковского, согласно которому

$$b^2=\left(\int_0^T s(t)\varphi(t)\,dt\right)^2\leqslant\int_0^T s^2(t)\,dt\int_0^T\varphi^2(t)\,dt=E_sE_\varphi,$$

[1] Нужно заметить, что помеха η на выходе преобразователя Ф выражается интегралом от входной помехи ξ. Иногда можно ожидать, что распределение η приближается к нормальному, так что формулы (10. 2) и (10. 2а) могут

где E_s и E_φ означают энергии соответственного сигнала и весовой функции. Знак равенства достигается только при условии

$$\varphi(t) = ks(t)$$

и оптимальная весовая функция, таким образом, определена. Постоянный множитель k мы в дальнейшем опустим, так как он все равно сократится при образовании отношения сигнал/помеха. Заметим, что с геометрической точки зрения полученный результат совершенно очевиден, так как скалярное произведение пропорционально проекции одного вектора на другой, а проекция имеет наибольшее значение, когда оба вектора совпадают по направлению. Итак, при оптимальном выборе весовой функции имеем

$$b^2 = E_s E_\varphi = E_s^2. \tag{10.3}$$

Займемся теперь помехой, выражаемой интегралом

$$\eta = \int_0^T \xi(t)\varphi(t)\,dt.$$

Составим квадрат этой величины

$$\eta^2 = \int_0^T \int_0^T \varphi(t)\varphi(t_1)\xi(t)\xi(t_1)\,dt\,dt_1.$$

Усредняя, находим

$$D\eta = M\eta^2 = \int_0^T \varphi(t)\,dt \int_0^T \varphi(t_1) B_\xi(t - t_1)\,dt_1, \tag{10.4}$$

где $B(\tau)$ — функция автокорреляции помехи. Итак, если заданы весовая функция $\varphi(t)$ и корреляционная функция $B(\tau)$ помехи, то дисперсия помехи на выходе преобразователя находится по формуле (10.4).

При определенном допущении можно получить из общей формулы (10.4) один важный результат в весьма компактной форме. Именно, предположим, что интервал корреляции помехи настолько мал, что на протяжении этого интервала весовая функция заметно не изменяется. Тогда

$$D\eta \simeq \int_0^T \varphi^2(t)\,dt \int_0^T B_\xi(t - t_1)\,dt_1 \simeq P_\xi \tau_0 E_\varphi \tag{10.5}$$

оказаться применимыми и в том случае, когда распределение ξ не удовлетворяет указанным условиям.

(рассуждения по поводу интеграла от функции корреляции см. в § 9). Используя (10. 3) и (10. 5), находим отношение сигнал/помеха

$$\rho = \frac{b^2}{D\eta} \simeq \frac{E_s}{P_\xi \tau_0}, \tag{10.6}$$

т. е. при данной помехе отношение сигнал/помеха зависит только от энергии сигнала [1].

Если мощность сигнала задана, то $E_s = P_s T$, и мы получаем

$$\rho \simeq \frac{T}{\tau_0} \cdot \frac{P_s}{P_\xi} = \frac{T}{\tau_0} \rho_0, \tag{10.7}$$

и видим, что рассматриваемый оптимальный способ приема может трактоваться как обобщение метода накопления на сигналы произвольной формы (см. формулу (9. 6)).

Дисперсия помехи на выходе преобразователя может быть выражена через спектры. Для этого нужно воспользоваться связью между функцией корреляции и спектром мощности случайного процесса

$$B(\tau) = \frac{1}{2} \int_{-\infty}^{\infty} G(\omega) e^{j\omega\tau} d\omega. \tag{10.8}$$

Подставив выражение (10. 8) в (10. 4), получаем

$$D\eta = \frac{1}{2} \int_0^T \varphi(t)\, dt \int_0^T \varphi(t_1)\, dt_1 \int_{-\infty}^{\infty} G_\xi(\omega) e^{j\omega(t-t_1)} d\omega =$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} G_\xi(\omega)\, d\omega \int_0^T \varphi(t) e^{j\omega t} dt \int_0^T \varphi(t_1) e^{-j\omega t_1} dt_1 =$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} G_\xi(\omega) S_\varphi(\omega) S_\varphi^*(\omega) d\omega,$$

где

$$S_\varphi(\omega) = \int_0^T \varphi(t) e^{-j\omega t} dt -$$

---

[1] Теперь видно, что выбор весовой функции $\varphi(t) = s(t)$, максимизирующей полезный сигнал b, максимизирует также и отношение сигнал/помеха $\rho = b^2/D\eta$, так как $D\eta$ зависит только от

$$E_\varphi = \int_0^T \varphi^2(t)\, dt.$$

текущий спектр функции $\varphi(t)$, $S_\varphi^*(\omega)$ — комплексно-сопряженная функция. Итак,

$$D\eta = \int_0^\infty G_\xi(\omega) \mid S_\varphi(\omega) \mid^2 d\omega. \qquad (10.9)$$

Эта формула является общей, как и формула (10. 4), из которой она непосредственно получена. Предположим, что функция φ имеет весьма узкий спектр, сосредоточенный около $\omega=0$. Тогда

$$D\eta \simeq G_\xi(0) \int_0^\infty \mid S_\varphi(\omega) \mid^2 d\omega = \pi G_\xi(0) E_\varphi, \qquad (10.10)$$

так как $|S|^2$ представляет собой спектр энергии [1]. Формула (10. 10) совпадает с (10. 5). В самом деле, из соотношения

$$G(\omega) = \frac{1}{\pi} \int_{-\infty}^\infty B(\tau) e^{-j\omega\tau} d\tau \qquad (10.11)$$

находим

$$G(0) = \frac{1}{\pi} \int_{-\infty}^\infty B(\tau)\, d\tau = \frac{1}{\pi} P\tau_0 —$$

по определению интервала корреляции.

Таким образом, допущение, сделанное при выводе формулы (10. 5), состоящее в том, что функция φ мало изменяется на интервале τ_0 в переводе на спектральный язык, означает, что ширина спектра функции φ мала по сравнению с шириной спектра помехи.

Можно теперь составить общее спектральное выражение для отношения сигнал/помеха. Мы имеем

$$b = E_s = \int_0^T s^2(t)\, dt = \frac{1}{\pi} \int_0^\infty \mid S_s(\omega) \mid^2 d\omega.$$

---

[1] Здесь использована теорема Рэлея, согласно которой

$$\int_{-\infty}^\infty x^2(t)\, dt = \frac{1}{\pi} \int_0^\infty \mid S_x(\omega) \mid^2 d\omega.$$

Практическое значение этой теоремы состоит в том, что она позволяет вычислить энергию не только путем интегрирования мгновенной мощности по времени (левая часть равенства), но и путем интегрирования спектра энергии по частоте.

и, используя (10.9), получаем

$$\rho = \frac{b^2}{D\eta} = \frac{\left(\frac{1}{\pi}\int\limits_0^\infty |S_s(\omega)|^2\, d\omega\right)^2}{\int\limits_0^\infty G_\xi(\omega)\,|S_s(\omega)|^2\, d\omega}. \qquad (10.12)$$

При помощи неравенства Буняковского можно получить из (10.12) следующую оценку[1]:

$$\rho \leqslant \frac{1}{\pi^2}\int\limits_0^\infty \frac{|S_s(\omega)|^2}{G_\xi(\omega)}\, d\omega. \qquad (10.13)$$

В частном случае помехи с равномерным спектром, т. е.

$$G_\xi(\omega) = \begin{cases} G_0 = \text{const} & [0 < \omega < \omega_c], \\ 0 & [\omega_c < \omega < \infty], \end{cases}$$

при условии, что спектр сигнала укладывается в ту же полосу, находим

$$\rho = \frac{1}{\pi^2 G_0}\int\limits_0^{\omega_c} |S_s(\omega)|^2\, d\omega = \frac{E_s}{\pi G_0}. \qquad (10.14)$$

Тот же результат мы получили бы и из (10.13), так что наибольшее значение отношения сигнал/помеха, соответствующее знаку равенства в (10.13), получается при помехе, имеющей равномерный спектр во всей полосе частот, занимаемой сигналом.

Результат (10.14) можно переписать в виде

$$\rho = \frac{2E_s}{A_0}. \qquad (10.15)$$

Здесь A_0 — также спектральная плотность, но отнесенная к 1 *гц*, т. е.

$$A = 2\pi G,$$

так что

$$P_\xi = G_0\omega_c = A_0 F,$$

и соотношение (10.15) можно еще представить в виде

$$\rho = 2FT\,\frac{P_s}{P_\xi} = 2FT\rho_0. \qquad (10.16)$$

[1] Для получения этого результата нужно положить в числителе (10.12)

$$|S_s(\omega)|^2 = \frac{|S_s(\omega)|}{\sqrt{G_\xi(\omega)}}\,|S_s(\omega)|\sqrt{G_\xi(\omega)}.$$

Заметим, что при выводе формул (10. 14)—(10. 16) предположение о том, что спектр ξ много шире спектра φ, уже снято. Предполагается лишь, что спектр сигнала вмещается в полосу F.

Важные формулы (10. 15) и (10. 16) показывают, что если задана спектральная плотность белого шума, то отношение сигнал/помеха вовсе не зависит от полосы. Если же задано отношение сигнал/помеха на входе, то оно увеличивается на выходе преобразователя пропорционально полосе.

Для понимания этих соотношений нужно учесть, что отношение сигнал/помеха на входе убывает с расширением полосы, так как

$$\rho_0 = \frac{P_s}{P_\xi} = \frac{P_s}{A_0 F} = \frac{E_s}{A_0} \cdot \frac{1}{FT}. \qquad (10.17)$$

С другой стороны, при заданном ρ_0 отношение сигнал/помеха на выходе преобразователя возрастает с расширением полосы вследствие того, что чем шире полоса, тем лучше усредняется помеха при интегральной операции, выполняемой преобразователем. Оба эффекта взаимно компенсируются, что наглядно видно, если получить (10. 15) подстановкой (10. 17) в (10. 16).

Как уже отмечалось, формула (10. 1) дает наиболее общий вид линейного функционала. Многие известные методы приема являются частными случаями этой линейной операции. Пока что мы можем отметить два таких частных случая, а именно:

1) если $s(t) = a = \text{const}$, то мы имеем обычно накопление в форме так называемого интегрального приема;

2) если $s(t)$ представляет собой синусоидальное колебание с известной частотой и фазой, то это есть случай так называемого когерентного приема, а операция (10. 1) есть не что иное, как синхронное детектирование.

В дальнейшем нам встретятся и другие частные случаи линейной операции (10. 1).

По поводу схемы приемного устройства рис. 15 нужно заметить, что в состав этого устройства входит еще не показанный на схеме синхронизатор. Назначение его состоит в том, чтобы включить преобразователь в момент $t=0$, т. е. в начале посылки, подключить на выход преобразователя в момент $t=T$ решающее устройство и вернуть систему в исходное состояние путем сброса накопленных значений сигнала и помехи после того, как решающее устройство вынесет свое суждение о наличии или отсутствии сигнала.

§ 11. Активные и пассивные фильтры

Общий вид линейной операции, выполняемой преобразователем в схеме описанного выше приемника, представляется функционалом

$$y = \Phi(x) = \int_0^T x(t)\,\varphi(t)\,dt, \qquad (11.1)$$

где $x(t)=s(t)+\xi(t)$ — принятый сигнал; $s(t)$ — переданный сигнал; $\xi(t)$ — помеха. Преобразователь, выполняющий операцию (11. 1), состоит из перемножителя, образующего произведение $x(t)\varphi(t)$, и интегратора, интегрирующего это произведение. Кроме того, имеется генератор, вырабатывающий весовую функцию $\varphi(t)$, и синхронизатор, управляющий ритмом работы всего приемного устройства. Схема приемного устройства, более подробная, чем схема предыдущего параграфа, показана на рис. 16. Итак, преобразователь Ф (обведен рамкой на рис. 16) состоит из перемножителя П, интегратора И и генератора весовой функции Г.

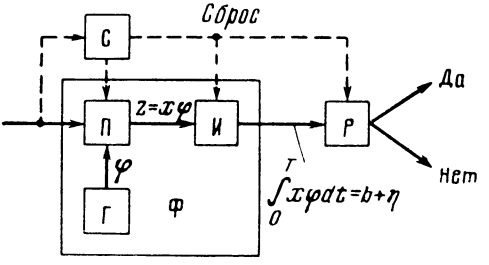


Рис. 16

С физико-математической точки зрения преобразователь Ф представляет собой параметрическую систему, описываемую линейным дифференциальным уравнением с переменными коэффициентами. В самом деле, умножение на заданную функцию есть линейная операция по определению (см. § 4), так как она удовлетворяет условию аддитивности

$$\varphi(t)[x_1(t)+x_2(t)]=\varphi(t)x_1(t)+\varphi(t)x_2(t)$$

и условию однородности

$$\varphi(t)[\lambda x(t)]=\lambda\varphi(t)x(t).$$

На выходе перемножителя, являющегося линейным устройством без реактивных элементов, мы имеем

$$z(t)=\varkappa x(t).$$

Коэффициент $\varkappa$ может рассматриваться как параметр. Но в нашем случае $\varkappa=\psi(t)$, т. е. параметр является заданной функцией времени. Это и означает, что перемножитель, а следовательно, и преобразователь в целом есть параметрическая система.

Теперь мы покажем, что операция (11. 1) может быть выполнена также устройством, представляющим собой линейный пассивный четырехполюсник, т. е. систему, описываемую линейным дифференциальным уравнением с постоянными коэффициентами.

Пусть $x(t)$ означает входное воздействие, а $y(t)$ — отклик четырехполюсника. Отклик $y(t)$ может быть выражен интегралом Дюамеля

$$y(t) = \int_{-\infty}^{t} x(\tau)\, g(t-\tau)\, d\tau_1, \qquad (11.2)$$

где $g(t)$ — импульсная реакция четырехполюсника. Действие такого четырехполюсника в системе нашего приемника определяется тем, что в момент $t=0$ он включается, а в момент $t=T$ берется отсчет напряжения на выходе четырехполюсника. Этот отсчет равен

$$y = y(T) = \int_{0}^{T} x(t)\, g(T-t)\, dt, \qquad (11.3)$$

где переменная интегрирования обозначена буквой t вместо буквы τ. Теперь (11. 3) в точности совпадает с (11. 1), если приравнять весовую функцию в обеих формулах, т. е. положить

$$\varphi(t) = g(T-t)$$

или

$$g(t) = \varphi(T-t),$$

или, так как при оптимальной операции $\varphi(t) = s(t)$ *,

$$g(t) = s(T-t). \qquad (11.4)$$

Таким образом, мы определили, какой должна быть импульсная реакция пассивного четырехполюсника, выполняющего в точности такое же линейное преобразование, как и описанный выше линейный преобразователь Ф. Дело сводится по существу к фильтрации; рассмотренные преобразователи можно назвать фильтрами. Однако, как видим, имеется вполне определенная принципиальная разница между параметрическим преобразователем с перемножителем и пассивным четырехполюсником.

Пользуясь термином «фильтр» как общим термином, мы будем, чтобы подчеркнуть различие, называть параметрический преобразователь активным фильтром, а пассивный четырехполюсник — пассивным фильтром.

Рассмотрим теперь подробнее свойства оптимального пассивного фильтра, удовлетворяющего условию (11. 4). Прежде всего поясним смысл этого условия. На рис. 17, *а* изображен сигнал $s(t)$ в виде произвольной (но известной) функции, заданной на

---

* Вообще $[\varphi(t)]_{\text{опт}} = ks(t)$, но мы условимся игнорировать постоянные множители, не играющие в наших рассуждениях никакой роли.

интервале $0 < t < T$. Как следует из (11. 4), импульсная реакция $g(t)$ воспроизводит значения $s(t)$ в обратной последовательности, начиная с момента $t=T$ и кончая моментом $t=0$ (рис. 17, *б*). Таким образом, график $g(t)$ получается как зеркальное отображение графика $s(t)$ относительно вертикали, делящей интервал $(0, T)$ пополам (штрих на рис. 17). Заметим, что по физическому смыслу $g(t) \equiv 0$ при $t < 0$ (так как отклик не может предшествовать воздействию в силу принципа причинности). С другой стороны, существенно отметить, что условие (11. 4) должно выполняться только на интервале $(0, T)$. За пределами этого интервала импульсная реакция может быть какой угодно, как показано штрихом на рис. 17, *б*. Дело в том, что в момент $t=T$ берется отсчет, решающее устройство срабатывает, производится сброс всех накопленных значений, и система, возвратясь в исходное состояние, начинает новый цикл работы.

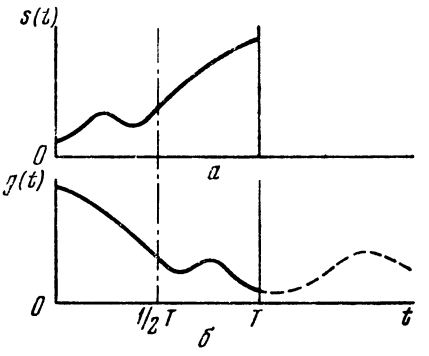


Рис. 17

Итак, вместо (11. 4) можно записать следующую более подробную характеристику оптимального фильтра:

$$g(t) = \begin{cases} 0 & [-\infty < t < 0], \\ s(T-t) & [0 < t < T], \\ \text{произвольная функция} & [T < t < \infty]. \end{cases} \qquad (11.5)$$

Рассмотрим несколько примеров.

1. Пусть $s(t)=a=\text{const}$. Тогда имеем согласно (11. 5)

$$g(t) = \begin{matrix} 0\,[t<0], \\ a\,[0<t<T], \end{matrix}$$

а в качестве произвольного продолжения за пределами интервала $(0, T)$ сохраним для $g(t)$ постоянное значение a.

В результате получаем

$$g(t) = \begin{matrix} 0\,[t<0], \\ a\,[0<t<\infty] \end{matrix} = a\sigma(t),$$

т. е. импульсная реакция должна выражаться единичной функцией. Напомним, что импульсной реакцией называется отклик системы на единичный импульс, выражаемый дельта-функ-

цией $\delta(t)$. Но, по определению,

$$\sigma(t) = \int_{-\infty}^{t} \delta(t)\, dt.$$

Из этого следует, что искомый оптимальный фильтр, обладающий импульсной реакцией $g(t) = \sigma(t)$, есть попросту идеальный интегратор. Этот результат подтверждает, что интегральный прием есть оптимальный способ приема постоянного сигнала.

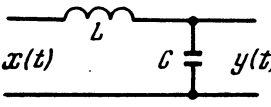


Рис. 18

2. Пусть $s(t) = a_0 \sin \omega_0 (t)$, причем положим для простоты, что $\omega_0 T = (2n+1)\pi$, т. е. на интервале $(0, T)$ укладывается нечетное число полупериодов. Тогда на интервале $(0, T)$

$$g(t) = a_0 \sin \omega_0 (T - t) = a_0 \sin [(2n+1)\pi - \omega_0 t] = a_0 \sin \omega_0 t,$$

так что на всей оси времени

$$g(t) = \begin{matrix} 0 & [t < 0], \\ a_0 \sin \omega_0 (t) & [0 < t < \infty]. \end{matrix}$$

Такой импульсной реакцией (с точностью до постоянного множителя) обладает контур без потерь, схема которого изображена на рис. 18. В операционной форме его уравнение имеет вид

$$\bar{y} = \frac{1}{p^2 LC + 1} \bar{x} = \frac{\omega_0^2}{p^2 + \omega_0^2} \bar{x}.$$

Подставляя $x(t) = \delta(t)$, $\bar{x} = p$, $y(t) = g(t)$, имеем

$$\bar{g} = \omega_0 \frac{\omega_0 p}{p^2 + \omega_0^2},$$

откуда

$$g(t) = \omega_0 \sin \omega_0 t \quad [t > 0].$$

Поучительно проследить, как протекает процесс преобразования. Мы имеем для полезного сигнала

$$y_0(t) = \int_0^t s(\tau)\, \varphi(\tau)\, d\tau = \int_0^t s(\tau)\, g(t - \tau)\, d\tau = \omega_0 a_0 \int_0^t \sin^2 \omega_0 \tau d\tau =$$

$$= \frac{1}{2} \omega_0 a_0 \int_0^t (1 - \cos 2\omega_0 \tau)\, d\tau = \frac{1}{2} \omega_0 a_0 \left(t - \frac{\sin 2\omega_0 t}{2\omega_0}\right),$$

или

$$y_0(t) = \frac{1}{2} a_0 \left(\omega_0 t - \frac{1}{2} \sin 2\omega_0 t \right).$$

Отсчет берется в момент $t = T$. При этом, как мы условились, $\omega_0 T = (2n + 1)\pi$ и, следовательно, $\sin 2\omega_0 T = 0$.

Таким образом,

$$b = y_0(T) = \frac{1}{2} a_0 \omega_0 T,$$

как показано на рис. 19. Далее, помеха на выходе фильтра

$$\eta = \int_0^T \xi(t)\varphi(t)\,dt = \int_0^T \xi(t)\, g(T - t)\,dt,$$

при белом шуме

$$D\eta = \frac{1}{2} A_0 E_\varphi = \frac{1}{2} \omega_0^2 A_0 \int_0^T \sin^2 \omega_0 t dt = \frac{1}{4} \omega_0^2 A_0$$

и отношение сигнал/помеха

$$\rho = b^2/D\eta = a_0^2 T/A_0 = 2E_s/A_0,$$

в соответствии с формулой (10. 15).

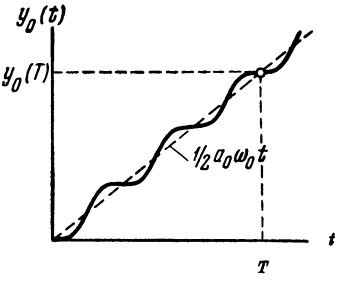


Рис. 19

Теперь займемся частотными характеристиками оптимального фильтра. Воспользуемся известным соотношением, связывающим коэффициент передачи и импульсную реакцию

$$K(\omega) = \int_{-\infty}^{\infty} g(t)\, e^{-j\omega t} dt.$$

Подставляя сюда

$$g(t) = \varphi(T - t) \quad [t > 0],$$

получим

$$K(\omega) = \int_0^{\infty} \varphi(T - t)\, e^{-j\omega t} dt = e^{-j\omega T} \int_{-\infty}^{T} \varphi(t_1)\, e^{j\omega t_1} dt_1 = e^{-j\omega T} S_\varphi(-\omega, T),$$

или

$$K(\omega) = e^{-j\omega T} S_{\varphi}^{*}(\omega, T). \tag{11.6}$$

Здесь

$$S_{\varphi}(\omega, T) = \int_{-\infty}^{T} \varphi(t)\, e^{-j\omega t} dt$$

означает текущий спектр функции $\varphi(t)$, зависящий от T и являющийся, таким образом, функцией двух аргументов. Символ S_{φ}^{*} означает комплексно-сопряженную функцию. При оптимальном выборе весовой функции, т. е. при

$$\varphi(t) = s(t),$$

имеем вместо (11.6)

$$K(\omega) = e^{-j\omega T} S_{s}^{*}(\omega, T). \tag{11.7}$$

Эта формула позволяет найти коэффициент передачи оптимального фильтра для обнаружения полностью известного сигнала.

Рассмотрим на основе этой формулы те же два примера, что и выше.

1. Пусть $s(t) = a = \text{const}$. В этом случае [1]

$$S_{s}^{*}(\omega, T) = a \int_{-\infty}^{T} e^{j\omega t} dt = \frac{a}{j\omega} e^{j\omega T}$$

и коэффициент передачи

$$K(\omega) = e^{-j\omega T} S_{s}^{*}(\omega, T) = \frac{a}{j\omega},$$

что и представляет собой выражение для коэффициента передачи идеального интегратора.

2. Пусть $s(t) = a_0 \sin \omega_0 t$. В этом случае

$$S_{s}^{*}(\omega, T) = a_0 \int_{-\infty}^{T} \sin \omega_0 t e^{j\omega t} dt =$$

$$= \frac{a_0 e^{j\omega T}}{\omega^2 - \omega_0^2} (\omega_0 \cos \omega_0 T - j\omega \sin \omega_0 T).$$

Но мы условились, что $\omega_0 T = (2n + 1)\pi$. Поэтому

$$S_{s}^{*}(\omega, T) = e^{j\omega T} \frac{a_0 \omega_0}{\omega^2 - \omega_0^2}$$

и

$$K(\omega) = \frac{a_0 \omega_0}{\omega^2 - \omega_0^2}.$$

[1] Полагают, что колебание затухает на бесконечности, т. е. $\lim\limits_{t \to \infty} e^{\pm j\omega t} = 0$. Это обосновывают обычно, добавляя в показатель отрицательную постоянную a и переходя в окончательном результате к пределу при $a \to 0$.

Это выражение (с точностью до постоянного множителя $-a/\omega_0$) представляет коэффициент передачи контура без потерь (см. рис. 18).

Рассмотрим еще один пример. Пусть сигнал $s(t)$ представляет собой периодическую последовательность весьма коротких импульсов, которые можно выразить дельта-функциями, т. е. положим

$$s(t) = \sum_{i=-\infty}^{\infty} \delta(t - it_0),$$

где t_0 — период следования. Тогда

$$S_s^*(\omega, T) = \int_{-\infty}^{T} \left(\sum_{i=-\infty}^{\infty} \delta(t - it_0) \right) e^{j\omega t} dt$$

или, меняя порядок операций,

$$S_s^* = \sum_{i=-\infty}^{n-1} \int_{-\infty}^{\infty} \delta(t - it_0)\, e^{j\omega t} dt = \sum_{i=-\infty}^{n-1} e^{ji\omega t_0},$$

где $n = T/\tau_0$ — число импульсов на интервале $(0, T)$. Разобьем сумму на две

$$\sum_{i=-\infty}^{n-1} e^{ji\omega t_0} = \sum_{i=0}^{\infty} e^{-ji\omega t_0} + \sum_{i=1}^{n-1} e^{ji\omega t_0}.$$

Первая сумма равна

$$\sum_{i=0}^{\infty} e^{-ji\omega t_0} = 1/(1 - e^{-j\omega t_0}).$$

Вторая сумма есть сумма геометрической прогрессии

$$\sum_{i=1}^{n-1} e^{ji\omega t_0} = (e^{j(n-1)\omega t_0} - 1)/(1 - e^{-j\omega t_0}).$$

Итак,

$$S_s^* = e^{j(n-1)\omega t_0}/(1 - e^{-j\omega t_0}) = e^{j(n-1/2)\omega t_0} \Big/ 2j \sin \frac{1}{2} \omega t_0.$$

и искомый коэффициент передачи

$$K = e^{-j\omega T} S_s^* = e^{-j1/2\omega t_0} \Big/ 2j \sin \frac{1}{2} \omega t_0.$$

Модуль коэффициента передачи, т. е. амплитудно-частотная характеристика фильтра, выражается формулой

$$|K| = 1/2 \left| \sin \frac{1}{2} \omega t_0 \right|.$$

Таким образом, характеристика фильтра есть периодическая функция частоты, показанная на рис. 20. Подобного рода фильтры носят название гребенчатых.

Физический смысл полученного результата очень прост: коэффициент передачи бесконечно велик на частотах всех гармоник

периодической функции, представляющей сигнал. Очевидно, что при этом наилучшим образом используется полная мощность сигнала, распределенная между гармониками. Если бы имелась только одна гармоника, т. е. если бы сигнал был синусоидален, то требовался бы только один пик характеристики на соответствующей частоте. Этот случай нами рассмотрен выше.

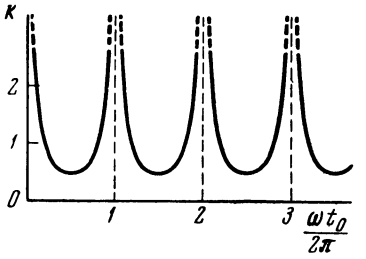


Рис. 20

Если импульсы имеют конечную длительность, то спектр последовательности практически ограничен частотой $F \approx 1/\tau$, где τ — длительность импульса. Соответственно ограничена и частотная характеристика фильтра.

§ 12. Различение двух сигналов

Поставим себе задачу построения идеального приемника для различения нескольких сигналов.

Способ действия приемника исчерпывающе характеризуется собственными областями сигналов. Данный сигнал s_k принимается правильно, если вектор принятого сигнала

$$x = s_k + \xi$$

попадает в собственную область V_k. В противном случае совершается ошибка.

Идеальным мы будем называть приемник, дающий наименьшую вероятность ошибки.

Рассмотрим сначала идеальный приемник, различающий два сигнала s_1 и s_2. Определим собственные области. Вероятность ошибки можно выразить следующим соотношением:

$$p_{\text{ош}} = p(1)\, p_1(2) + p(2)\, p_2(1), \qquad (12.1)$$

где $p(1)$ и $p(2)$ — априорные вероятности передачи сигналов s_1 и s_2; $p_1(2)$ — условная вероятность принять сигнал s_2, когда в действительности передавался сигнал s_1; $p_2(1)$ — вероятность принять s_1 при передаче s_2.

При наличии двух сигналов все пространство V делится на два полупространства V_1 и V_2, которые и являются собственными областями сигналов s_1 и s_2. Задача состоит в нахождении границы между собственными областями.

Приведенные выше условные вероятности можно также представить в виде [1]

$$p_1(2) = p_1\{x \in V_2\}, \quad p_2(1) = p_2\{x \in V_1\}.$$

Мы ограничимся пока случаем равновероятных сигналов s_1 и s_2. При этом

$$p(1) = p(2) = 1/2, \quad p_{\text{ош}} = \frac{1}{2}[p_1(2) + p_2(1)].$$

Выразим условные вероятности через объемные плотности вероятностей, а именно:

$$p_1(2) = \int_{V_2} q_1 dV, \quad p_2(1) = \int_{V_1} q_2 dV,$$

где $q_1 = dp_1(2)/dV$; $q_2 = dp_2(1)/dV$ выражают объемные плотности условных вероятностей, т. е. вероятности вектору x попасть в данный единичный объем, а $q_1 dV$ и $q_2 dV$ выражают вероятности попасть в объем dV.

Формула для вероятности ошибки принимает вид

$$p_{\text{ош}} = \frac{1}{2}\left(\int_{V_2} q_1 dV + \int_{V_1} q_2 dV\right). \tag{12.2}$$

Учитывая, что по условию нормировки вероятностей

$$\int_V q_1 dV = \int_V q_2 dV = 1,$$

можем переписать (12.2) в следующем виде:

$$p_{\text{ош}} = \frac{1}{2}\left(1 - \int_{V_1} (q_1 - q_2)\, dV\right) = \frac{1}{2}\left(1 - \int_{V_2} (q_2 - q_1)\, dV\right)$$

и для получения наименьшей вероятности ошибки нужно максимизировать интегралы

$$\int_{V_1} (q_1 - q_2)\, dV = \int_{V_2} (q_2 - q_1)\, dV.$$

Для этого нужно выбрать границы областей V_1 и V_2 так, чтобы подынтегральные выражения были положительны.

Это значит, что в области V_1 должно быть $q_1 > q_2$, а в области $V_2 - q_2 > q_1$. Отсюда следует, что граница собственных областей определяется условием

$$q_1 = q_2. \tag{12.3}$$

Заметим, что для случая двух равновероятных сигналов это заключение имеет совершенно общий характер. Мы не делали ни-

[1] Запись $x \in V_k$ означает: элемент x принадлежит множеству V_k. У нас вектор x попадает в область V_k.

каких предположений ни о виде сигналов, ни о свойствах помехи; более того, не вводили никаких характеристик пространства V, не определяли его размерности и даже не требовали, чтобы оно было метрическим.

Можно теперь применить общее условие (12. 3) к частному случаю метрического пространства, введя расстояния r_1 и r_2 от сигналов s_1 и s_2 до вектора принятого сигнала x, как показано на рис. 21, на котором 00′ означает границу между собственными областями V_1 и V_2. Положим, что помеха имеет сферическую сим-

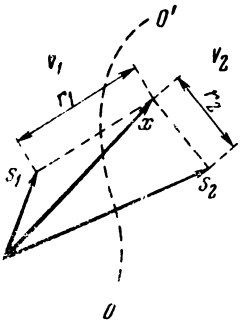


Рис. 21

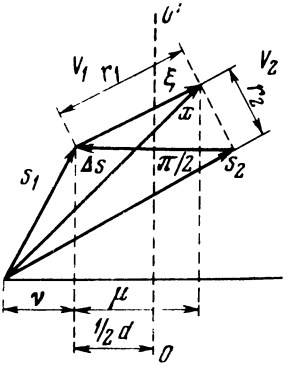


Рис. 22

метрию (т. е. что все направления вектора помехи равновероятны). Тогда объемные плотности вероятностей зависят только от расстояний, и мы можем записать

$$q_1 = q(r_1), \quad q_2 = q(r_2).$$

В том частном случае, когда $q(r)$ есть монотонная функция (и только в этом случае), из (12. 3) следует

$$r_1 = r_2. \tag{12.4}$$

Это означает, что граница 00′ представляет собой геометрическое место точек, равноотстоящих от s_1 и s_2. Это заключение относится к любому метрическому пространству. В R_n граница 00′ представляет собой гиперплоскость, перпендикулярную к вектору разности $\overline{\Delta s} = s_1 - s_2$ и делящую его пополам.

Перейдем теперь к нормированному пространству. Вероятность ошибки равна $p_{\text{ош}} = p_1(2)$ (если оба сигнала равноправны и $p_1(2) = p_2(1)$). Во всех нижеследующих формулах предполагается, что передается сигнал s_1. Можно составить два выражения для вероятности ошибки

$$p_{\text{ош}} = p\{r_1 > r_2\}, \tag{12.5}$$

$$p_{\text{ош}} = p\left\{\mu > \frac{1}{2}d\right\}, \tag{12.6}$$

где $r_1 = \|x - s_1\|$, $r_2 = \|x - s_2\|$, μ — проекция ξ на вектор, коллинеарный $\overline{\Delta s}$, т. е. $\mu = \xi\overline{\Delta s}/\|\overline{\Delta s}\|$, $d = \|\overline{\Delta s}\|$ (рис. 22). Оба выражения совершенно эквивалентны. Однако, как мы увидим, они приводят к различным схемам приемников.

Преобразуем неравенство из (12.6), прибавив к обеим частям проекцию s_1 на $\overline{\Delta s}$,

$$V = s_1\overline{\Delta s}/\|\overline{\Delta s}\|.$$

Имеем

$$\mu + \nu = \frac{x\overline{\Delta s}}{\|\overline{\Delta s}\|} > \frac{1}{2}d + \nu = \frac{1}{2}\|\overline{\Delta s}\| + \frac{s_1\overline{\Delta s}}{\|\overline{\Delta s}\|} = \frac{\|s_2\|^2 - \|s_1\|^2}{2\|\Delta s\|}.$$

Итак, вместо (12.5) и (12.6) можно записать

$$p_{\text{ош}} = p\{\|x - s_1\|^2 > \|x - s_2\|^2\}, \tag{12.7}$$

$$p_{\text{ош}} = p\left\{x\overline{\Delta s} > \frac{1}{2}(\|s\|^2 - \|s_1\|^2)\right\}. \tag{12.8}$$

В бесконечномерном пространстве непрерывных функций, заданных на интервале, при евклидовой метрике (пространство C_{L^2}), имеем

$$p_{\text{ош}} = p\left\{\int_0^T [x(t) - s_1(t)]^2\,dt > \int_0^T [x(t) - s_2(t)]^2\,dt\right\}, \tag{12.9}$$

$$p_{\text{ош}} = p\left\{\int_0^T x(t)\,[s_2(t) - s_1(t)]\,dt > \frac{1}{2}(E_2 - E_1)\right\}. \tag{12.10}$$

Заметим, что формула (12.9) выражает обычно даваемое определение идеального приемника Котельникова.

Умножая обе части неравенства в (12.6) на $\|\overline{\Delta s}\|$, получим

$$p_{\text{ош}} = p\left\{\eta > \frac{1}{2}b\right\} \tag{12.6а}$$

и

$$p_{\text{ош}} = \frac{1}{2} - f\left(\frac{1}{2}\sqrt{\rho}\right),$$

где $\rho = b^2/D\eta$, $\eta = \xi\overline{\Delta s}$, $b = \|\overline{\Delta s}\|^2$ $(= d^2)$.

Таким образом, повторяются все соотношения, выведенные ранее (§ 10); к приемнику, различающему два сигнала, можно применить результаты, относящиеся к обнаруживающему приемнику, если в качестве весовой функции взять Δs вместо s. Это и естественно, потому что, как уже отмечалось, обнаружение можно рассматривать как различение двух сигналов, из которых один тождественно равен нулю.

Заметим, что при фиксированных $\|s_1\|$ и $\|s_2\|$ (т. е. при фиксированных энергиях обоих сигналов) наибольшее значение $\|\overline{\Delta s}\|$, а следовательно, и наибольшее значение полезного сигнала (см. (12.6) или (12.6а)) достигаются в том случае, когда векторы

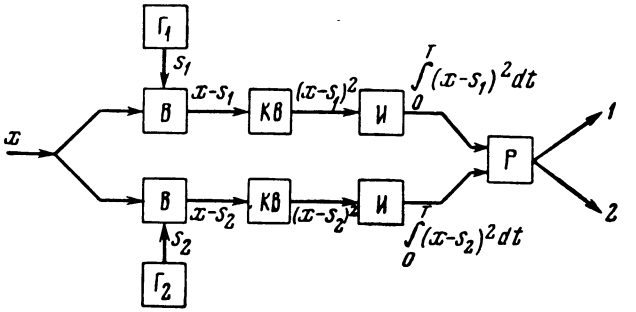


Рис. 23

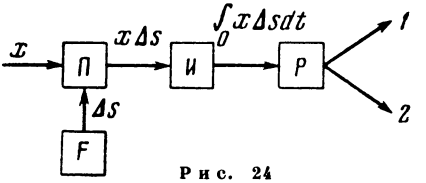


Рис. 24

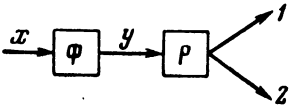


Рис. 25

s_1 и s_2 коллинеарны и направлены в противоположные стороны. При этом

$$\|\overline{\Delta s}\| = \|s_1\| + \|s_2\|,$$

тогда как в общем случае

$$\|\overline{\Delta s}\|^2 = \|s_1\|^2 + \|s_2\|^2 - 2s_1 s_2.$$

Теперь мы можем составить блок-схемы приемников, выполняющих действия (12. 9) и (12. 10), учитывая, что решающие устройства действуют по-разному, а именно: в случае (12. 9) решающее устройство сравнивает между собой две случайные величины, а в случае (12. 10) сравнивает случайную величину с постоянным порогом. Схемы показаны на рис. 23 и 24.

На схеме рис. 23 имеются два идентичных канала, каждый из которых состоит из вычитающего устройства В, квадратора Кв и интегратора И. Кроме того, имеются генераторы Г обеих сигнальных функций s_1 и s_2 и решающее устройство Р.

Схема рис. 24 содержит преобразователь, состоящий из перемножителя П, генератора весовой функции Г и интегратора И. Эта схема в точности совпадает со схемой рис. 15. Заметим лишь, что в качестве весовой функции берется

$$\Delta s(t) = s_2(t) - s_1(t).$$

Наконец, можно воспользоваться преобразователем в виде пассивного фильтра (рис. 25), характеристика которого должна быть

$$K(\omega) = e^{-j\omega T} \int_0^T [s_2(t) - s_1(t)]\, e^{j\omega t} dt$$

(см. § 11).

Схема рис. 24 проще схемы рис. 23. Однако она имеет тот недостаток, что постоянный порог зависит от энергий сигналов. При наличии замираний система будет работать неуверенно, если не применить автоматическую регулировку порога. Но этот недостаток устраняется, если сигналы имеют равную энергию. Тогда пороговое значение равно нулю, и решающее устройство определяет только знак случайной величины на выходе преобразователя, что, конечно, очень удобно. При $E_1=E_2$ упрощаются соотношения для обоих приемников. Раскрывая квадраты разностей в (12. 9), получаем

$$p_{\text{ош}} = p\left\{\int_0^T x(t)\,s_1(t)\,dt < \int_0^T x(t)\,s_2(t)\,dt\right\}, \qquad (12.11)$$

а вместо (12. 10) имеем

$$p_{\text{ош}} = p\left\{\int_0^T x(t)\,\Delta s(t)\,dt > 0\right\}. \qquad (12.12)$$

Итак, при $E_1=E_2$ приемник Котельникова превращается в корреляционный приемник. Он измеряет взаимную корреляцию [1] принятого сигнала x с каждым из двух возможных сигна-

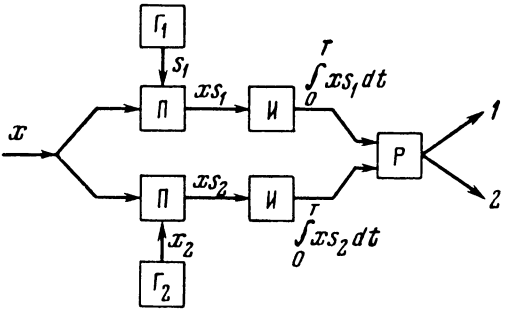


Рис. 26

лов s_1 и s_2; решающее устройство признает за переданный сигнал тот, с которым принятый сигнал имеет большую корреляцию. Таким образом, возникает еще один вариант схемы идеального приемника (рис. 26).

Интересно отметить, что нелинейный преобразователь схемы рис. 23 заменяется на схеме рис. 26 линейным преобразователем. Схемы рис. 24 и 25, разумеется, также линейны.

Рассмотрим теперь пример, показывающий, как зависит форма границы между собственными областями сигналов от распределения помехи. Как уже говорилось, при помехе, объемная плотность которой выражается монотонной функцией расстояния, граница

[1] Точнее, взаимную энергию, которая равна функции кратковременной корреляции, умноженной на интервал интегрирования (см. § 4).

есть гиперплоскость, расположенная симметрично относительно двух сигнальных точек.

В качестве примера мы имели нормальное распределение

$$q(r) = (2\pi)^{-n/2} e^{-r^2}.$$

(Здесь и в последующих формулах все величины, имеющие размерность длины, нормированы делением на $\sqrt{2}\sigma$.) Подставляя это распределение в общее уравнение границы

$$q(r_1) = q(r_2),$$

сразу получаем единственное решение

$$r_1 = r_2.$$

Возьмем теперь немонотонную функцию, а именно:

$$q(r) = \frac{2}{n} (2\pi)^{-n/2} r^2 e^{-r^2}.$$

Уравнение границы имеет вид

$$r_1^2 e^{-r_1^2} = r_2^2 e^{-r_2^2}.$$

Тождественным решением является по-прежнему $r_1 = r_2$. Однако выбор этой гиперплоскости в качестве границы между собственными областями не минимизирует вероятность ошибки. Следует взять другое решение трансцендентного уравнения границы.

На рис. 27 представлена двумерная модель. Пользуясь обозначениями чертежа, получаем

$$\left[\left(x_1 - \tfrac{1}{2} d\right)^2 + x_2^2\right] e^{-\left[\left(x_1 - \frac{1}{2} d\right)^2 + x_2^2\right]} =$$

$$= \left[\left(x_1 + \tfrac{1}{2} d\right)^2 + x_2^2\right] e^{-\left[\left(x_1 + \frac{1}{2} d\right)^2 + x_2^2\right]}$$

или

$$\frac{x_1 d}{\operatorname{th} x_1 d} - x_1^2 - x_2^2 - \frac{1}{4} d^2 = 0.$$

График этой функции (с соблюдением масштаба) нанесен на рис. 27. Собственная область сигнала s_1, обозначенная V_1, заштрихована. Любопытно, что точка каждого сигнала и ее ближайшие окрестности принадлежат собственной области другого сигнала. Это является результатом того, что $q(0) = 0$, т. е. вероятность вектора принятого сигнала попасть в элементарный объем, охватывающий точку переданного сигнала, равна нулю, а для ближайшей окрестности сигнальной точки эта вероятность мала.

Выясним теперь, к чему приведет отказ от двух принятых ранее допущений, а именно: равновероятности обоих сигналов s_1 и s_2 и наличия сферической симметрии помехи.

Если помеха симметрична, т. е. объемная плотность зависит только от расстояния, но не от направления, и если $q=q(r)$ есть монотонно убывающая функция, то при равновероятных сигналах имеем (см. (12. 3))

$$q_1/q_2=1, \tag{12.13}$$

откуда получается $r_1=r_2$ (см. (12. 4)), и граница между собственными областями есть гиперплоскость, перпендикулярная к вектору $\overline{\Delta s}$ и делящая его пополам.

Если же априорные вероятности сигналов s_1 и s_2 не равны, то вместо (12. 13) получаем

$$\frac{q_1}{q}=\frac{p(1)}{p(2)}. \tag{12.14}$$

Так, например, для нормального распределения

$$\frac{q_1}{q}=e^{r_2^2-r_1^2}, \quad r_2^2-r_1^2=\ln\frac{p(1)}{p(2)},$$

а это есть уравнение гиперплоскости, перпендикулярной к вектору $\overrightarrow{\Delta s}$, но пересекающейся с ним в точке, отстоящей от середины (в сторону менее вероятного сигнала) на величину

$$\Delta l=\frac{1}{2d}\ln\frac{p(1)}{p(2)}.$$

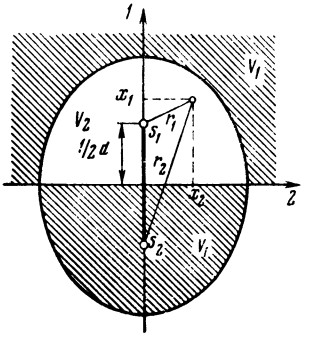


Рис. 27

Здесь r — безразмерное расстояние; d — безразмерная длина отрезка между концами векторов s_1 и s_2.

Обратимся к предположению о симметрии помехи. Отбросив это предположение, мы введем зависимость от координат, и вероятность ошибки будет определяться не только расстоянием между сигналами, но и ориентацией векторов сигналов относительно координатной системы. Естественно, что все соотношения при этом усложнятся. Для того чтобы разъяснить существо дела, мы начнем наше рассмотрение с двумерной модели.

Пусть даны два сигнала s_1 (s_{11}, s_{12}), s_2 (s_{21}, s_{22}) и помеха ξ (ξ_1, ξ_2), которую мы предположим нормально распределенной. При этом будем полагать

$$\sigma_1=\sigma_2=\sigma, \quad M\xi_1=M\xi_2=0, \quad \frac{M(\xi_1\xi_2)}{\sigma^2}=k,$$

где σ — среднеквадратичное значение; k — коэффициент корре-

ляции. Двумерная плотность вероятностей для величины ξ есть

$$w(x_1, x_2) = \frac{1}{2\pi\sigma^2\sqrt{1-k^2}}\, e^{-\frac{x_1^2+x_2^2-2kx_1x_2}{2\sigma^2(1-k^2)}}. \qquad (12.15)$$

Как видим, плотность вероятностей постоянна не на окружности

$$x_1^2 + x_1^2 = \text{const},$$

как это было бы при отсутствии корреляции (т. е. при $k=0$), а на эллипсе

$$x_1^2 + x_2^2 - 2kx_1x_2 = \text{const}.$$

Эксцентриситет этого эллипса равен

$$e = \sqrt{\frac{2k}{1+k}},$$

а отношение полуосей

$$\frac{b}{a} = \frac{1-k}{1+k}.$$

Большая часть направлена по биссектрисе координатного угла.

Итак, нарушение сферической симметрии помехи означает наличие корреляции. Наоборот, предположение о симметрии, принятое во всех предыдущих рассуждениях, равносильно предположению о том, что помеха не коррелирована, т. е. имеет неограниченно широкий спектр [1].

Построим геометрическую картину в пространстве сигналов. На рис. 28 s_1 и s_2 изображают два сигнала, $\overline{\Delta s} = s_2 - s_1$ — вектор разности сигналов. Эллипсы представляют места точек, на которых плотность вероятности постоянна. (Мы рассматриваем случай равновероятных сигналов, поэтому эллипсы одинаковы.) Граница $00'$ собственных областей проходит через точки пересечения эллипсов. Она представляет собой, как мы видим, прямую, проходящую через середину вектора $\overline{\Delta s}$, но не перпендикулярную к нему, а составляющую с вектором $\overline{\Delta s}$ угол ψ. Для определения этого угла нужно решить совместно уравнения обоих эллипсов, т. е. составить равенство

$$(x_1 - s_{11})^2 + (x_2 - s_{12})^2 - 2k(x_1 - s_{11})(x_2 - s_{12}) =$$
$$= (x_1 - s_{21})^2 + (x_2 - s_{22})^2 - 2k(x_1 - s_{21})(x_2 - s_{22}),$$

где s_{ik} — координаты сигнальных точек, т. е. концов векторов s_i (рис. 28).

[1] Это условие в той или иной форме фигурировало при выводе приближенных выражений для отношения сигнал/помеха в § 9. Отметим также, что при сферической симметрии составляющие вектора помехи независимы только в случае нормального распределения (напомним, что независимость и отсутствие корреляции, вообще говоря, не одно и то же).

Из этого равенства получаем уравнение границы

$$x_1[2(s_{21}-s_{11})-2k(s_{22}-s_{12})]+x_2[2(s_{22}-s_{12})-$$
$$-2k(s_{21}-s_{11})]+s_{11}^2+s_{12}^2-s_{21}^2-s_{22}^2-2k(s_{11}s_{12}-s_{21}s_{22})=0.$$

Отсюда для угла φ, образуемого границей с осью 1, находим

$$\operatorname{tg}\varphi=-\frac{s_{21}-s_{11}-k(s_{22}-s_{12})}{s_{22}-s_{12}-k(s_{21}-s_{11})}=\frac{1-k\operatorname{tg}\vartheta}{k-\operatorname{tg}\vartheta},$$

а для искомого угла $\psi=\varphi-\vartheta$

$$\operatorname{tg}\psi=\frac{1-k\sin 2\vartheta}{k\cos 2\vartheta}, \qquad (12.16)$$

где ϑ — угол между $\overline{\Delta s}$ и осью 1.

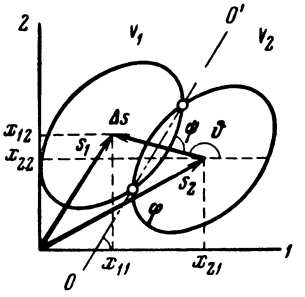


Рис. 28

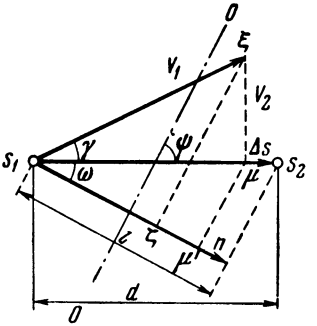


Рис. 29

Найдем вероятность ошибки, основываясь на геометрическом построении рис. 29.

Способ приема пусть состоит в проектировании вектора принятого сигнала на вектор нормали n к границе $00'$. Обозначая проекцию через ξ, можем записать для вероятности ошибки

$$p_{\text{ош}}=p\left\{\zeta>\frac{1}{2}l\right\}.$$

Для вычисления этой величины ограничимся двумерной моделью при нормальном распределении помехи. Мы имеем при передаче сигнала s_1

$$\omega(x_1, x_2)=$$
$$=\frac{1}{2n\sigma^2\sqrt{1-k^2}}e^{-\frac{(x_1-s_{11})^2+(x_2-s_{12})^2-2k(x_1-s_{11})(x_2-s_{12})}{2\sigma^2(1-k^2)}},$$

и вероятность ошибки

$$p_{\text{ош}} = \iint\limits_{V_2} w(x_1, x_2)\, dx_1 dx_2.$$

Для вычисления удобно перенести начало координат в точку s_1 (s_{11}, s_{12}) и повернуть координатную систему так, чтобы ось 1 совпала с вектором нормали n (рис. 30). Таким образом, новые переменные вводятся по формулам

$$x_1 - s_{11} = x_1' \cos\alpha - x_2' \sin\alpha, \qquad x_2 - s_{12} = x_1' \sin\alpha + x_2' \cos\alpha.$$

В этих переменных интегрирование производится в пределах

$$\frac{1}{2} l < x_1' < \infty, \quad -\infty < x_2' < \infty.$$

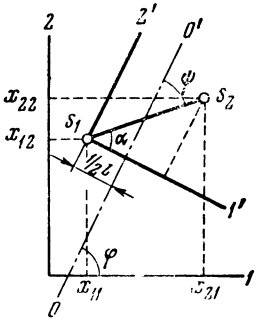


Рис. 30

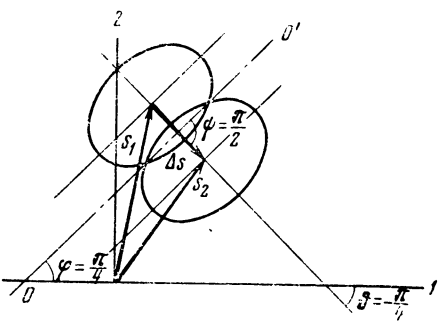


Рис. 31

Вычисление дает

$$p_{\text{ош}} = \frac{1}{2}[1 - \Phi(z)], \tag{12.17}$$

где

$$z = \frac{d}{2\sqrt{2\sigma}} \sqrt{\frac{1 - k\sin 2\vartheta}{1 - k^2}}, \tag{12.18}$$

тогда как при некоррелированной помехе мы имели бы (при $k=0$):

$$z = \frac{d}{2\sqrt{2\sigma}} = z_0. \tag{12.19}$$

Полученный результат интересен в том отношении, что, как показывает формула (12.18), при наличии корреляции вероятность ошибки зависит от расположения сигнальных точек относительно координатной системы. В зависимости от угла ϑ между вектором $\overline{\Delta s}$ и осью 1 аргумент z изменяется в пределах

$$z_{\min} = \frac{z_0}{\sqrt{1+k}} < z < z_{\max} = \frac{z_0}{\sqrt{1-k}}.$$

Наибольшее значение $z=z_{\max}$ (а следовательно, и наименьшее значение вероятности ошибки) достигается при $\vartheta=-\pi/4$. При этом $\psi=\pi/2$, $\varphi=\pi/4$, $\alpha=0$ (рис. 31). Для построения системы сигналов служит соотношение

$$\operatorname{tg}\vartheta=\frac{s_{22}-s_{12}}{s_{21}-s_{11}}=-1.$$

Если $E_1=E_2$, т. е. если сигналы имеют равные энергии и представлены векторами равной длины, то картина рис. 31 становится симметричной; граница $00'$ проходит через начало координат. При этом должно быть

$$s_{21}=s_{12},\quad s_{22}=s_{11}.$$

Заметим, что при надлежащем подборе сигналов, обеспечивающем получение наибольшего значения

$$z=z_{\max}=\frac{z_0}{\sqrt{1-k}},$$

увеличение коэффициента корреляции ведет к неограниченному увеличению верности. В то же время неоптимальный выбор сигналов влечет за собой лишь ограниченное уменьшение верности, так как в наихудшем случае

$$z_{\min}=\frac{z_0}{\sqrt{1+k}}\xrightarrow[k\to 1]{}\frac{z_0}{\sqrt{2}}.$$

Мы рассматривали до сих пор двумерный случай. Приведенные выше соотношения без особых затруднений обобщаются на n-мерное пространство сигналов. В этом случае граница между собственными областями представляет собой гиперплоскость, рассекающую отрезок $d=\|\overline{\Delta s}\|$ пополам (при равных априорных вероятностях обоих сигналов). Вектор нормали к граничной гиперплоскости образует с вектором $\overline{\Delta s}$ угол α, для которого получается

$$\sin\alpha=\frac{\sum_{i=1}^{n}A_i\cos\vartheta_i}{\sqrt{\sum_{i=1}^{n}A_i^2}},$$

где

$$A_i=\frac{d}{\sigma^2}\sum_{j=1}^{n}D_{ij}\cos\vartheta_j,$$

D_{ij} — алгебраические дополнения элементов k_{ij} корреляционной матрицы; $\cos\vartheta_j$ — направляющие косинусы (т. е. косинусы углов ϑ_j между вектором $\overline{\Delta s}$ и j-й осью).

Мы не будем больше углубляться в этот вопрос. Ограничимся общим заключением: наличие корреляции помехи позволяет по-

высить верность; для реализации этой возможности следует строить оптимальную систему сигналов. При оптимальной системе сигналов верность беспредельно растет с увеличением корреляции.

§ 13. Различение многих сигналов

Перейдем к проблеме различения многих сигналов. Простейший случай многих сигналов — это квантованный набор m равновероятных постоянных значений

$$0,\ a,\ 2a,\ \ldots,\ (m-1)\,a.$$

Эти m сигналов наилучшим образом различаются приемником с интегратором И и полосовым ограничителем О, схема которого показана на рис. 32. При передаче i-го сигнала имеем на выходе интегратора

$$b_i = \int_0^T iadt = iaT = ib, \quad \eta = \int_0^T \xi dt,$$

и вероятность ошибки

$$p_{\text{ош}} = p\left\{|\eta| > \tfrac{1}{2}b\right\}. \tag{13.1}$$

С геометрической точки зрения задача является одномерной. Числовая шкала показана на рис. 33. Там же отмечены собственные области. Действие полосового ограничителя состоит в том, что он выдает напряжение на i-й выход, если величина

$$y = \int_0^T x(t)\,dt$$

оказывается в интервале V_i.

Итак, верность определяется так же, как и в случае обнаружения постоянного сигнала при интегральном приеме. Разница состоит лишь в том, что при обнаружении собственные области простираются от порога вниз до нуля и вверх до бесконечности. Здесь же у каждого сигнала (кроме крайних) имеется по два соседних. Поэтому собственная область V_i ограничена сверху и снизу значениями $(i \pm {}^1/_2)\,b$. В формуле (13.1) это обстоятельство нашло свое отражение в том, что в неравенство введена не величина η, а ее абсолютное значение $|\eta|$, что соответственно понижает верность различения многих сигналов по сравнению с верностью обнаружения или различения двух сигналов.

Столь же прост и случай, когда имеется набор из m функций, различающихся только постоянным множителем

$$0,\ s(t),\ \ 2s(t),\ \ldots,\ \ (m-1)\ s(t).$$

Оптимальным будет в данном случае приемник с весовой функцией $s(t)$, и мы будем иметь

$$b_i = i\int_0^T s^2(t)\,dt = iE_s, \quad \eta = \int_0^T \xi(t)\,s(t)\,dt.$$

Формула (13. 1) и рис. 33 сохраняют силу и для этого случая, если положить $b=E_s$.

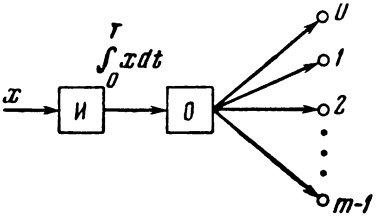


Рис. 32

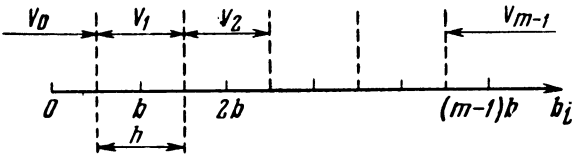


Рис. 33

Обратимся теперь к случаю, когда m сигналов выражаются разными функциями

$$s_1(t), \quad s_2(t), \ldots, s_m(t).$$

Начнем с простейшей ситуации, когда сигналы образуют ортогональную систему функций, т. е.

$$\boldsymbol{s}_i\boldsymbol{s}_j = \int_0^T s_i(t)\,s_j(t)\,dt = 0 \quad [i \neq j], \tag{13. 2}$$

Будем, кроме того, полагать, что все сигналы имеют одинаковую энергию, т. е.

$$\|\boldsymbol{s}_i\|^2 = \int_0^T s_i^2(t)\,dt = E_s \text{ const.} \tag{13. 3}$$

Формула (13. 2), представляющая собой определение свойства ортогональности, подсказывает оптимальный метод приема ортогональных сигналов. Он состоит в образовании скалярных произведений

$$y_i = \int_0^T x(t)\,s_i(t)\,dt = b_i + \eta_i$$

и в сравнении их между собой. Схема приемника (рис. 34) должна содержать m преобразователей Φ_i. Решающее устройство Р должно сравнить между собой b_i и η_i для каждого значения i. Дело сводится к обнаружению сигнала в i-й собственной области, т. е. к m-кратному повторению операции обнаружения.

Вероятность ошибки

$$p_{\text{ош}} = 1 - (1 - p_0)^m \simeq mp_0$$

(если $p_0 \ll 1$). Здесь p_0 означает вероятность ошибки по отношению к каждой из m собственных областей.

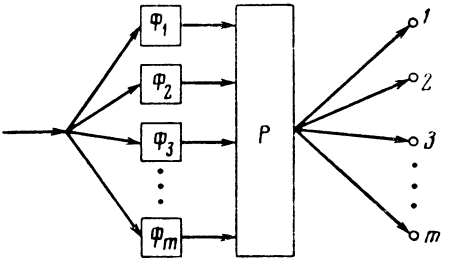


Рис. 34

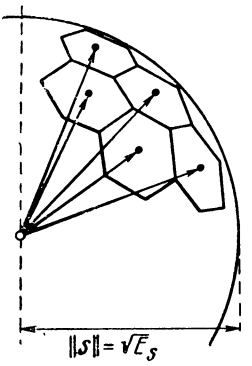


Рис. 35

Заметим, что вариант схемы идеального приемника с проектированием на вектор разности в данном случае нецелесообразен, так как число разностей равно m $(m-1)/2$. При $m=2$ имеется одна разность, при $m=3$ уже три, а дальше число разностей растет примерно пропорционально квадрату числа сигналов, так что схема приемника по этому варианту при большом m получается громоздкой, а принципиальных преимуществ такая схема не имеет.

В заключение рассмотрим в общих чертах значительно более общий случай набора из m произвольных функций, подчиненных лишь условию (13. 3).

Геометрической моделью такой системы сигналов является пучок векторов, концы которых лежат на сферической поверхности в n-мерном пространстве сигналов. Ясно, что при равновероятных сигналах сигнальные точки должны располагаться на поверхности сферы равномерно. При заданном числе сигналов нужно стремиться разместить их так, чтобы наименьшее расстояние между любой парой сигналов было наибольшим. Мы рассмотрим некоторое правильное размещение сигналов, характеризующееся тем, что каждый сигнал окружен m_0 ближайшими (соседними) сигналами, находящимися от данного сигнала и между собой на одинаковых расстояниях d_0. Точки сигналов расположены, таким образом, в центрах правильных сферических m-угольников, покрывающих поверхность сферы без промежутков.

На рис. 35 схематически показана трехмерная модель такой системы сигналов при $m_0=6$. Такая правильная система удобна в том отношении, что единственным параметром, определяющим ее свойства с точки зрения помехоустойчивости, является наименьшее расстояние d_0 между любой парой сигналов (т. е. расстояние между данным сигналом и соседним). В качестве параметра можно также взять угол 2ϑ между векторами соседних сигналов.

Собственные области сигналов представляют собой m_0-гранные пирамиды с вершинами в центре сферы. Одна такая пирамида

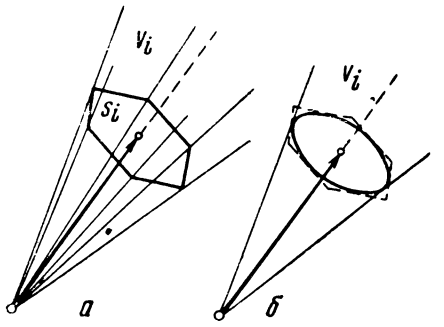


Рис. 36

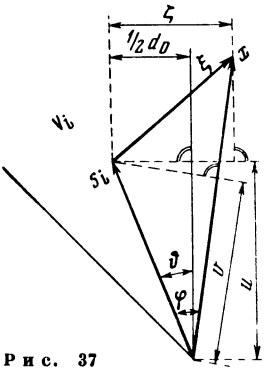


Рис. 37

показана на рис. 36, *а*. Для упрощения можно заменить пирамиду вписанным в нее конусом, как показано на рис. 36, *б*. При такой замене собственная область меньше действительной, так что верность определяется с запасом.

Рассмотрим теперь вероятность ошибки различения. Ошибка произойдет, если при передаче сигнала $\boldsymbol{s}_1$ вектор принятого сигнала $\boldsymbol{x}$ окажется за пределами собственной области V_i. Вероятность этого события можно записать в различных равносильных формах, а именно (рис. 37):

$$p_{\text{ош}} = p\{\varphi > \vartheta\}, \tag{13.4a}$$

$$p_{\text{ош}} = p\{\cos\varphi < \cos\vartheta\}, \tag{13.4b}$$

$$p_{\text{ош}} = p\{v < u\}, \tag{13.4c}$$

$$p_{\text{ош}} = p\left\{\zeta > \frac{1}{2}d_0\right\} \tag{13.4d}$$

и т. п. Здесь φ — угол между $\boldsymbol{x}$ и $\boldsymbol{s}_i$; ϑ — угол при вершине конуса, определяющего собственную область V_i; v — проекция $\boldsymbol{s}_i$ на $\boldsymbol{x}$; u — проекция $\boldsymbol{s}_i$ на граничную поверхность (т. е. на образующую конуса); ζ — проекция ξ на нормаль к граничной поверх-

ности (или, иначе, на вектор разности $s_i - s_j$, где s_j — соседний сигнал).

Все формулы (13. 4) равносильны; одна может быть выведена из другой на основании определений нормы и скалярного произведения. Выбор той или иной формы может нас интересовать с двух точек зрения. Во-первых, та или иная запись (13. 4) определяет способ действия приемника. Во-вторых, от выбора одной из формул (13. 4) зависит техника вычисления вероятности ошибки.

Займемся прежде построением схемы приемника. Заметим, что для приемника данным является принимаемый сигнал x, над которым могут производиться те или иные операции. Поэтому неравенствам в формулах (13. 4) должен быть придан такой вид, чтобы характер выполняемой приемником операции был ясно виден.

Возьмем формулу (13. 4b) и преобразуем ее, имея в виду, что

$$\cos\varphi = \frac{s_i x}{\|s\|\|x\|}$$

и выражение для вероятности ошибки можно представить в виде

$$p_{\text{ош}} = p\{xs_i \lessgtr \|x\|\|s\|\cos\vartheta\}. \qquad (13.4e)$$

Приемник, построенный на основании этой формулы, представляет собою корреляционный приемник. Он производит скалярное умножение принятого сигнала x на каждый из сигналов s_i. Затем все скалярные произведения $E_i = xs_i$ сравниваются с пороговым значением $E_0 = \|x\|\|s\|\cos\vartheta$, стоящим в правой части неравенства. Все значения E_i лежат ниже порога, за исключением одного, которое и указывает номер фактически переданного сигнала. Формула (13. 4) дает вероятность ошибки при этом способе приема.

Заметим, что в выражении для порогового значения величины $\|s\| = \sqrt{E_s}$ и $\cos\vartheta$ — постоянные, задаваемые строением системы сигналов, а $\|x\|$ — случайная величина, которая флюктуирует из-за того, что $x = s_i + \xi$. Кроме того, величина $\|x\|$ может изменяться еще и из-за замираний. Поэтому нужно сделать одно из двух: либо нормировать скалярные произведения перед подачей их на решающее устройство, чему соответствует запись (13. 4c), которую можно представить в следующем развернутом виде:

$$p_{\text{ош}} = p\left\{\frac{xs_i}{\|x\|} \lessgtr \|s\|\cos\vartheta\right\}, \qquad (13.4f)$$

либо, что, вероятно, технически удобнее, сделать порог регулируемым, введя в решающее устройство легко измеряемое значение $\|x\|^2 = E_x$. В этом варианте схема приемника принимает вид, показанный на рис. 38. На этом рисунке Φ_i означают преобразователи, т. е. пассивные или активные фильтры, вырабатывающие скаляр-

ные произведения. Дополнительная цепь, вырабатывающая значения $\|x\|^2$, состоит из квадратора Кв и интегратора (сумматора) И.

Другой вариант схемы приемника получим, взяв за исходное неравенство из (13. 4d). Введем в рассмотрение один из ближайших сигналов s_j, рсстояние до которого

$$d_0 = \|s_i - s_j\|$$

(рис. 39). Формулу (13. 4d) можно переписать в виде

$$p_{\text{ош}} = p\{\zeta > d_0 - \zeta\}. \qquad (13.4g)$$

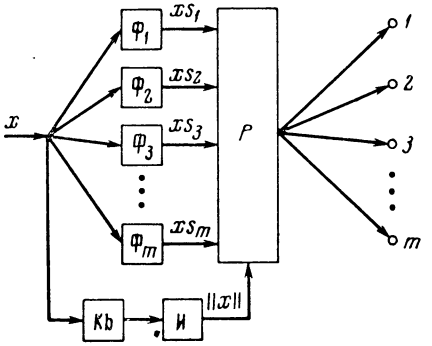


Рис. 38

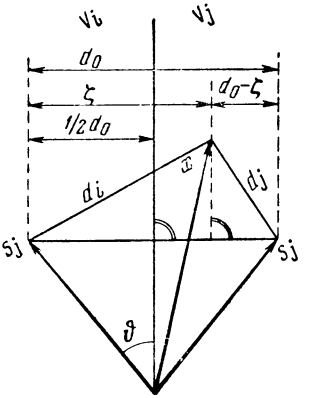


Рис. 39

Но, как это видно из рис. 39, эту формулу можно заменить равносильной [1]

$$p_{\text{ош}} = p\{d_i > d_j\}, \qquad (13.4h)$$

где d_i и d_j — расстояния от принятого сигнала x до фактически переданного сигнала s_i и до ближайшего сигнала s_j. Так как s_j — ближайший сигнал, то формула (13. 4g) и подавно справедлива для любого сигнала s_k, входящего в систему сигналов. Итак, дело сводится к измерению расстояний и к отождествлению принятого сигнала с тем из возможных переданных сигналов, к которому принятый сигнал ближе.

Таким образом, приемник должен найти расстояния и сравнить

$$d_k = \|x - s_k\|$$

[1] Равносильность (13.4g) и (13.4h) доказывается следующим образом:

$$d_i^2 = h^2 + \zeta^2, \quad d_j^2 = h^2 + (d_0 - \zeta)^2 = d_i^2 - 2d_0\left(\zeta - \frac{1}{2}d_0\right).$$

Но если $\xi > d_0/2$, то $d_i > d_j$.

их между собой. Переданным сигналом считается тот, для которого получается наименьшее d_k. Заметим, что для переданного сигнала

$$d_i = \|x - s_i\| = \|\xi\|.$$

Постоянного порога сравнения при этом способе приема не существует, так что действие решающего устройства усложняется. Зато не требуется никакой нормировки. Схема приемного устройства, действующего описанным образом, показана на рис. 40. Схема состоит из m идентичных каналов, в каждом из которых

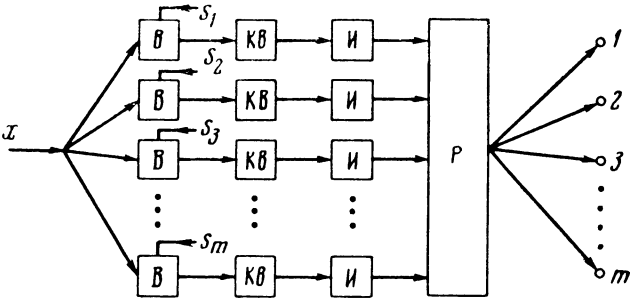


Рис. 40

имеется вычитающее устройство В. На вычитающее устройство, кроме принятого сигнала x, поступает один из сигналов s_i. Затем идут квадратор Кв и интегратор И. Эта схема повторяет схему рис. 23 с соответственным увеличением числа каналов.

Обратимся к вопросу о вычислении вероятности ошибки. Выбирая ту или иную из формул (13. 4), придем к различной постановке математической задачи. Нужно интегрировать объемную плотность вероятностей по собственной области. Беря ту или иную случайную величину, получим для нее соответствующее выражение для плотности и соответствующее выражение для пределов интегрирования. Стремясь упростить первое, мы усложняем второе, и наоборот. Так, если поместить начало координат в конец вектора s_i, то получим простое многомерное распределение для вектора ξ, но пределы интегрирования при этом будут заданы весьма сложными выражениями.

Поясним возникающие трудности на примере.

Будем исходить из формулы (13. 4а) и найдем распределение вектора $x = s + \xi$, считая для простоты, что ξ — некоррелированная гауссова помеха. Тогда (в пространстве R_n)

$$w(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum (x_i - s_i)^2}.$$

Сумма в показателе может быть представлена в виде

$$\sum (x_i - s_i)^2 = r^2 + E_s - 2r\sqrt{E_s}\cos d\varphi,$$

где $r^2 = \sum x_i^2$, $E_s = \sum s_i^2$, $\cos\varphi = \sum x_i s_i / r\sqrt{E_s}$.

Теперь нужно выразить элемент объема dV в полярных координатах r и φ. В n-мерном пространстве (рис. 41)

$$dV = \Omega_{n-1} r^{n-1} \sin^{n-2}\varphi dr d\varphi, \tag{13.5}$$

где

$$\Omega_{n-1} = \frac{(n-1)\,\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \tag{13.6}$$

— угловая мера сферы $n-1$ измерений (т. е. поверхность единичной сферы). Таким образом, для вероятности правильного приема можем записать

$$p_{\text{пр}} = 1 - p_{\text{от}} =$$

$$= \frac{\Omega_{n-1}}{(\sqrt{2\pi}\sigma)^n} e^{-\sqrt{E_s}/2\sigma^2} \int_0^\infty r^{n-1} e^{-r^2/2\sigma^2} dr \int_0^\vartheta \sin^{n-2}\varphi e^{\frac{\sqrt{E_s}\, r\cos\varphi}{\sigma^2}} d\varphi. \tag{13.7}$$

Мы выбрали в качестве случайного вектора вектор x. При этом, очевидно, простейшим образом выражаются пределы интегрирования. Но подынтегральные функции таковы, что вычисление интегралов в конечной форме невозможно. Напомним, что рассматривается собственная область в виде конуса, а не пирамиды, т. е. уже введено довольно существенное упрощение задачи. К сожалению, никаким образом не удается пока свести задачу к вычислимым интегралам, и вероятность ошибки для рассматриваемого случая до сих пор не найдена. Были предложены лишь различные оценки.

Не останавливаясь на вопросе об оценках, укажем лишь на одну сравнительно простую возможность получения асиптотического выражения для вероятности ошибки. Эта возможность основана на некоторых особенностях асимптотического при $n \to \infty$ поведения многомерной сферы.

К таким особенностям относится, например, тот факт, что с увеличением n объем сферы сосредоточивается у ее поверхности, что непосредственно следует из того, что объем n-мерной сферы пропорционален n-й степени ее радиуса.

Но нас интересует другое, менее известное свойство, состоящее в том, что с ростом n площадь поверхности сферы сосредоточивается у ее экватора.

Поясним это обстоятельство подробнее. Площадь n-мерного сферического сегмента, т. е. части поверхности сферы, вырезае-

мой круговым конусом с углом φ при вершине, находящейся в центре сферы (рис. 42), выражается формулой

$$S(n, \varphi, r) = \omega(n, \varphi) r^{n-1}. \tag{13.8}$$

Здесь $\omega(n, \varphi)$ — угловая мера сегмента, т. е. площадь поверхности сегмента единичного радиуса. Для этой величины имеем

$$\omega(n, \varphi) = \Omega_{n-1} \int_0^{\varphi} \sin^{n-2} u du, \tag{13.9}$$

где Ω_{n-1} — угловая мера сферы $n-1$ измерений (см. (13.6)). Легко видеть, что, так как $|\sin u| \leqslant 1$, то при возрастании n зна-

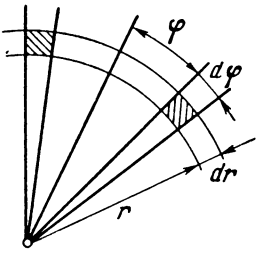


Рис. 41

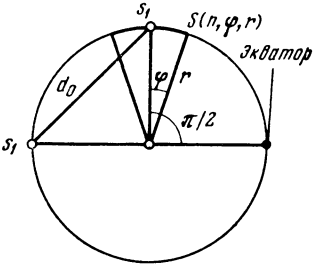


Рис. 42

чения подынтегральной функции будут уменьшаться, стремясь к нулю при всех значениях u, кроме $u = \pi/2$, где $\sin u$ обращается в единицу.

На рис. 43 показаны для примера графики $\omega(n, \varphi)$ при нескольких значениях n. Эти графики и поясняют смысл высказанного положения об асимптотическом сосредоточении площади сферы на экваторе, т. е. при $\varphi = \pi/2$.

Предположим теперь, что все сигналы распределены по поверхности с равномерной плотностью, т. е. что на единицу поверхности приходится везде одинаковое число сигналов. Если выбрать один сигнал s, поместив его в полюсе сферы (рис. 42), то в пределе при $n \to \infty$ все остальные сигналы сосредоточатся на экваторе, т. е. на расстоянии

$$d = \sqrt{2E_s} \tag{13.10}$$

от полюса. Таким образом, задача нахождения вероятности ошибки, т. е. события, состоящего в том, что данный сигнал s_1 будет спутан с каким-либо другим, сводится к одномерной задаче о вероятности ошибки при различении двух сигналов, находящихся друг

от друга на расстоянии d. Эта вероятность равна (см. выражение (11. 6))

$$p_{\text{ош}} = p\left\{\mu > \frac{1}{2}d\right\},$$

или при нормальном распределении

$$p_{\text{ош}} = \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{8}\rho}\right)\right], \tag{13.11}$$

где

$$\rho = \frac{d^2}{D\mu}, \tag{13.12}$$

μ — проекция вектора помехи на направление d. Остается найти дисперсию этой величины.

Мы имеем вообще

$$\mu = \frac{\xi\overline{\Delta s}}{\|\overline{\Delta s}\|}.$$

В пространстве R_n

$$\mu = \frac{1}{d}\sum_{k=1}^{n}\xi_k\Delta s_k.$$

Для дисперсии получаем

$$D\mu = M\mu^2 = \frac{1}{d^2}\sum_{k}\sum_{l}M(\xi_k\xi_l)\,\Delta s_k\Delta s_l.$$

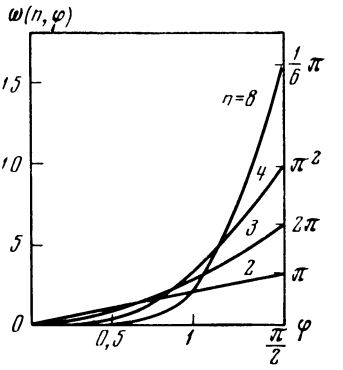


Рис. 43

Но помеха предполагается некоррелированной, т. е.

$$M(\xi_k\xi_l) = \begin{cases} 0 & [l \neq k], \\ P_\xi & [l = k]. \end{cases}$$

Таким образом,

$$M\mu^2 = \frac{P_\xi}{d^2}\sum_{k=1}^{n}\Delta s_k^2 = P_\xi. \tag{13.13}$$

С помощью (13. 10) и (13. 13) находим искомое асимптотическое выражение

$$p_{\text{ош}} \sim \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{E_s}{4P_\xi}}\right)\right],$$

или

$$p_{\text{ош}} \sim \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{4}n\rho_0}\right)\right],$$

где $\rho_0 = P_s/P_\xi$.

§ 14. Обнаружение неполностью известного сигнала

В § 10 обсуждался вопрос о приеме сигнала, представляющего собой полностью известную функцию, определенную на интервале (О, Т); обычно речь идет о функции вида

$$s = s(a,\ b,\ c,\ \ldots,\ t), \tag{14.1}$$

где a, b, c — некоторые параметры, и, говоря о полностью известной функции, имеют в виду случай, когда все эти параметры известны. Функция же, не вполне известная, будет обладать одним или несколькими неизвестными параметрами, и именно такую ситуацию мы в дальнейшем и имеем в виду.

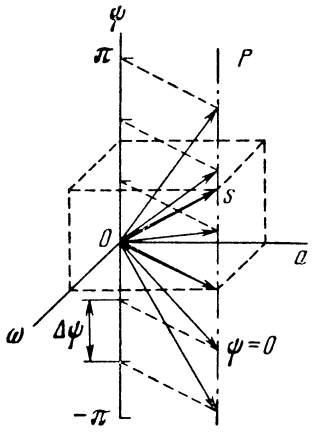


Рис. 44

Для геометрического представления положения удобно ввести специальное пространство, которое мы будем называть пространством параметров и обозначать буквой P. Это пространство представляет собой множество всех функций данного вида (14. 1). За координаты принимаются значения параметров, так что размерность пространства P равна числу параметров, полностью определяющих функцию вида (14. 1). Поясним это сразу же примером. Пусть функции s имеют вид

$$s = a \sin(\omega t + \psi) = s(a,\ \omega,\ \psi,\ t), \tag{14.2}$$

где a, ω, ψ, т. е. амплитуда, частота и начальная фаза, — три параметра, полностью определяющие любую функцию (14. 2), т. е. любую синусоиду. Пространство P будет в этом случае трехмерным (рис. 44), и любая синусоида представляется в этом пространстве одним вектором (т. е. одной точкой).

Если один из параметров, например начальная фаза ψ, неизвестен, то это значит, что нам задан не вектор s, а лишь его проекция на плоскость a, ω (вектор $\psi=0$ на рис. 44). Ясно, что оптимальный метод приема, основанный на полном знании функции s, при таких условиях неосуществим, и мы должны найти методы, оптимальные при ограниченном знании представляющей сигнал функции.

Совершенно очевидно, что чем меньше мы знаем о сигнале, тем менее совершенны возможные методы его приема. Мы ограничимся здесь лишь рассмотрением задачи обнаружения. Наша цель состоит в том, чтобы, во-первых, указать целесообразные методы приема и, во-вторых, оценить потерю верности, обусловленную неполным знанием сигнала.

Естественно попытаться сохранить и в рассматриваемом случае общую процедуру, описанную в § 9, состоящую в образовании

линейного функционала, т. е. скалярного произведения. Но мы не можем взять в качестве весовой функции синусоиду с произвольной фазой ψ_0, так как сдвиг фаз между сигналом и весовой функцией может оказаться каким угодно; в частности, при $\delta\psi=\psi-\psi_0=\pi/2$ полезный сигнал на выходе преобразователя будет равен нулю. Поэтому альтернатива такова: либо стремиться тем или иным способом измерить неизвестную фазу и вернуться, таким образом, к случаю полностью известного сигнала, либо прибегнуть к какому-либо подходящему усреднению результатов, получаемых при различных фазовых соотношениях.

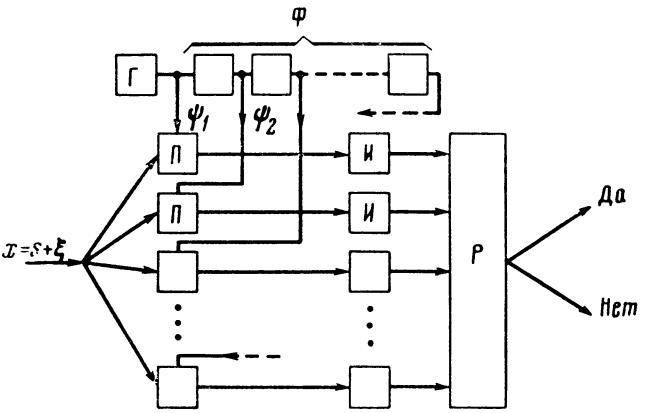


Рис. 45

Первый путь в принципе позволяет, разумеется, сохранить наибольшую верность, достигаемую при полном знании сигнала, но требует усложнения процедуры приема. Если бы мы располагали возможностью длительного наблюдения за сигналом, то его фаза могла бы быть измерена. Точнее говоря, можно подобрать такую фазу весовой функции ψ_0, чтобы $\delta\psi=0$, что, собственно, и требуется для осуществления оптимального приема. Однако эта возможность нам не дана, поэтому для получения того же результата придется применить многоканальную схему приемника, показанную на рис. 45. На этой схеме имеется m преобразователей, состоящих каждый из перемножителя *П* и интегратора *И*. На вторые входы перемножителей подаются напряжения одной и той же частоты, но различной фазы. На схеме рис. 45 эти напряжения берутся с различных звеньев фазосдвигающей цепочки Ф, питаемой одним генератором Г. Решающее устройство Р совмещает две функции. Во-первых, оно выбирает наибольшее из поступающих на его вход напряжений. Во-вторых, оно сравнивает это наибольшее напряжение с пороговым значением. Вторая операция в точности такова же, как и в случае полностью известного сигнала. Первая же операция есть не что иное, как измерение неизвестной фазы сигнала. Применение многоканальной схемы

приема позволяет избежать дополнительной затраты времени. Наша задача сведена теперь к различению многих сигналов с квантованной начальной фазой.

Не нужно думать, что требуемое число каналов непомерно велико. Оно определяется допустимой потерей верности, которую естественно выразить через потерю в отношении сигнал/помеха. Это отношение пропорционально $\cos^2 \delta\psi$.

Если мы допустим наибольшее уменьшение ρ на 25%, то

$$\cos^2 \delta\psi_{max} = {}^3/_4, \quad \cos \delta\psi_{max} = \sqrt{{}^3/_2}, \quad \delta\psi_{max} = \pi/6 = 30°.$$

Шаг фазы в фазосдвигающей цепочке должен быть не более

$$\Delta\psi = 2\delta\psi_{max} = \pi/3 = 65°,$$

а число каналов равно

$$m = \frac{2\pi}{\Delta\psi} = \frac{\pi}{\delta\psi_{max}} = 6$$

(см. рис. 44). При этом в среднем отношение сигнал/помеха будет уменьшено не на 25%, а меньше. Если, как обычно, принять для фазы равномерное распределение в интервале $(0{,}2\ \pi)$, то среднее уменьшение ρ составляет

$$\frac{\rho_{ср}}{\rho_{max}} = M \cos^2 \Delta\psi = \frac{1}{\Delta\psi} \int\limits_{-\frac{1}{2}\Delta\psi}^{\frac{1}{2}\Delta\psi} \cos^2 \alpha d\alpha = \frac{1}{2}\left(1 + \frac{\sin \Delta\psi}{\Delta\psi}\right).$$

При $\Delta\psi = \pi/3 = 60°$ имеем $\rho_{ср}/\rho_{max} = 0{,}92$. Итак, потеря в отношении сигнал/помеха составляет в среднем при шести каналах всего 8%.

Обратимся теперь к другой возможности — к усреднению по фазе. Один из способов осуществления этой возможности состоит в том, что в качестве весовой функции берется синусоидальное колебание частоты $\omega + \Delta\omega$, причем $\Delta\omega$ выбирается так, чтобы за время T образовался набег фазы, равный 2π, т. е.

$$\Delta\omega T = 2\pi.$$

Тогда на протяжении посылки мгновенный сдвиг фазы $\delta\psi$ будет принимать все возможные значения. Интегрировать следует абсолютное значение произведения сигнала на весовую функцию, так что перед интегратором нужно включить двухтактный выпрямитель. При этих условиях мы будем иметь на выходе преобразо-

вателя для полезного сигнала следующую величину:

$$b = a_0 \int_0^T |\sin(\omega t + \psi)\sin(\omega + \Delta\omega)t|\,dt =$$

$$= a_0 \frac{4}{\pi^2} \int_0^T \left[1 - \sum_{k=1}^{\infty} \frac{1}{4k^2-1} \cos 2k(\omega t + \psi)\right]\left[1 - \sum_{k=1}^{\infty} \frac{1}{4k^2-1} \times \right.$$

$$\left. \times \cos 2k(\omega + \Delta\omega)t\right] dt.$$

Если положить $\omega T = n\pi$, то $b = 4a_0/\pi^2$. Эту величину следует сравнить со значением $b_{max} = 1a_0/2$, получаемым при синхронном

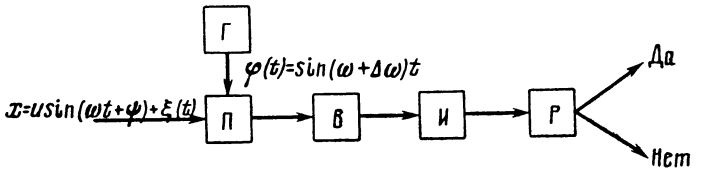


Рис. 46

детектировании (т. е. в случае, когда фаза известна). Таким образом, $b/b_{max} = 8/\pi^2 = 0{,}81$, а для отношения сигнал/помеха имеем $\rho/\rho_{max} = (b/b_{max})^2 = 0{,}66$, т. е. уменьшение на 34%. Этот проигрыш больше, чем при применении описанного выше многоканального метода. Зато схема приемника проще; она показана на рис. 46 и отличается от схемы рис. 15 только наличием выпрямителя В перед интегратором.

Рассмотрим теперь случай, когда неизвестны частота и фаза. Проще всего при таких условиях поместить решающее устройство после детектора. Однако при этом среднее значение помехи не равно нулю, и накопление выгоды не дает. Иначе говоря, нет смысла производить интегрирование после детектора, верность получится такая же, как при однократном отсчете (см. § 8).

Существует, однако, возможность обработки сигнала до детектора. Эта возможность заключается в том, что в качестве весовой функции берется принятый сигнал, сдвинутый на некоторое время τ. Такой метод приема называется автокорреляционным.

Мы исследуем свойства автокорреляционного приемника и получим оценку для отношения сигнал/помеха, введя ряд упрощений. Некоторые из них довольно грубы, но все они оправданы стремлением получить результат, правильный по порядку величины и по возможности прозрачный по форме. Пусть преобра-

зователь приемника выполняет следующую операцию:

$$y = \int_0^{T-\tau} [s(t) + \xi(t)]\,[s(t-\tau) + \xi(t-\tau)]\,dt. \qquad (14.3)$$

Раскрывая скобки, получим, полагая $\tau \ll T$,

$$y = \int_0^T s(t)\,s(t-\tau)\,dt + \int_0^T s(t)\,\xi(t-\tau)\,dt + \int_0^T \xi(t)\,s(t-\tau)\,dt + \int_0^T \xi(t)\,\xi(t-\tau)\,dt = E_{ss} + E_{s\xi} + E_{\xi s} + E_{\xi\xi}. \qquad (14.4)$$

Здесь E_{ss} — полезный сигнал, являющийся неслучайной функцией τ; остальные величины случайны. Особенностью данного случая является то, что нельзя разделить выходную величину y на два слагаемых, одно из которых зависит только от сигнала, а другое — только от помехи. Этому препятствуют перекрестные члены $E_{\xi s}$ и $E_{s\xi}$. Поэтому для условных вероятностей ошибок мы должны записать:

$$p_0(1) = p\{E_{\xi\xi} > y_0\},$$
$$p_1(0) = p\{E_{ss} + E_{s\xi} + E_{\xi s} + E_{\xi\xi} < y_0\},$$

где y_0 — пороговое значение. Однако мы сразу упростим дело, отнеся перекрестные члены к помехе в обоих случаях, т. е. положим, как и ранее,

$$y_0 = \frac{1}{2} E_{ss}$$

и

$$p_{\text{ош}} = p\left\{|E_{\xi\xi} + E_{s\xi} + E_{\xi s}| > \frac{1}{2} E_{ss}\right\} = p\left\{|\eta| > \frac{1}{2} b\right\}.$$

Отношение сигнал/помеха будет определяться как

$$\rho = b^2/M\eta^2. \qquad (14.5)$$

Найдем

$$M\eta^2 = M(E_{\xi\xi} + E_{s\xi} + E_{\xi s})^2 = M(E_{\xi\xi}^2 + E_{s\xi}^2 + E_{\xi s}^2 + 2E_{\xi\xi}E_{s\xi} + 2E_{\xi\xi}E_{\xi s} + 2E_{s\xi}E_{\xi s}).$$

Отбрасывая корреляционные члены, возьмем в качестве приближения

$$M\eta^2 \simeq M(E_{\xi\xi}^2 + E_{s\xi}^2 + E_{\xi s}^2). \qquad (14.6)$$

Для первого члена имеем

$$ME_{\xi\xi}^2 = (ME_{\xi\xi})^2 + DE_{\xi\xi},$$

и в качестве нижней оценки можно взять

$$ME_{\xi\xi}^2 > (ME_{\xi\xi})^2 \simeq B_\xi^2(\tau)\,T^2, \qquad (14.7)$$

где $B_\xi(\tau)$ — функция корреляции помехи [1]. Средние значения $E_{\xi\xi}$ и $E_{\xi s}$ равны нулю, а средние квадраты равны дисперсии, которую можно считать для обоих членов одинаковой. Тогда

$$M(E_{s\xi}^2 + E_{\xi s}^2) \simeq 2D\int_0^T s(t)\,\xi(t-\tau)\,dt \simeq 2P_\xi \tau_0 E_s, \qquad (14.8)$$

где τ_0 — интервал корреляции помехи; $P_\xi = B_\xi(0)$ — мощность помехи; E_s — энергия сигнала. Вывод этого соотношения дан в § 9.

Если сигнал имеет вид

$$s = a_0 \sin \omega t,$$

то

$$b = E_{ss} = a_0^2 \int_0^T \sin \omega t \sin \omega (t-\tau)\,dt =$$

$$= \frac{a_0^2 T}{2}\left[\cos \omega\tau\left(1 - \frac{\sin 2\omega T}{2\omega T}\right) - \sin \omega\tau \frac{1-\cos 2\omega T}{2\omega T}\right].$$

При $2\omega T \gg 1$

$$b = E_{ss} \simeq \frac{1}{2} a_0^2 T \cos \omega\tau = E_s \cos \omega\tau. \qquad (14.9)$$

Подставляя (14.6) и (14.9) в (14.5), получаем

$$\rho = \rho(\tau) \simeq \frac{E_s^2 \cos^2 \omega\tau}{B_\xi^2(\tau)\,T^2 + 2P_\xi \tau_0 E_s} = \frac{\cos^2 \omega\tau}{\frac{1}{\rho^2} k_\xi^2(\tau) + 2\frac{\tau_0}{T}\cdot\frac{1}{\rho_0}}, \qquad (14\ 10)$$

где $\rho_0 = P_s/P_\xi = E_s/E_\xi$, $k_\xi(\tau) = B_\xi(\tau)/P_\xi$ — соответственно отношение сигнал/помеха на входе приемника и нормированная функция корреляции помехи. Итак, ρ есть функция τ.

С увеличением τ функция корреляции $k_\xi(\tau)$ убывает, и при $\tau > \tau_0$ первым членом в знаменателе (14. 10) можно пренебречь. Тогда

$$\rho(\tau) \simeq \frac{1}{2}\cdot\frac{T}{\tau_0}\rho_0 \cos^2 \omega\tau. \qquad (14.11)$$

Теперь мы можем сделать одно из двух: либо найти максимальное значение, соответствующее $\cos \omega\tau = 1$, и тогда будем иметь

$$\rho(\tau_{\max}) = \frac{1}{2}\cdot\frac{T}{\tau_0}\rho_0, \qquad (14.12)$$

[1] Некоторые подробности по поводу $D(E_{\xi\xi})$ приведены в Добавлении IV.

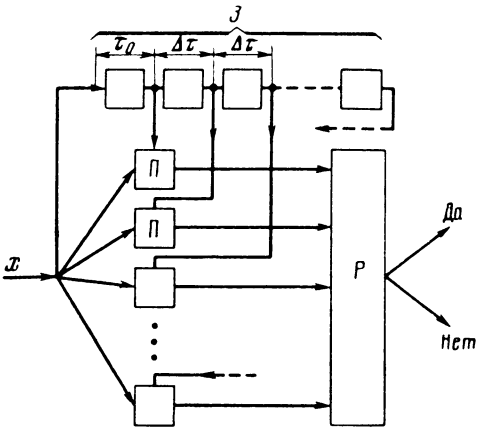


Рис. 47

либо усреднить (14. 11) по значениям cos за период и получить вдвое меньшую величину

$$\rho_{ср} = \frac{1T}{\tau_0}\rho_0. \qquad (14.13)$$

Заметим, что (14. 12) дает величину ρ, вдвое меньшую, чем при синхронном приеме, а (14. 13) — вчетверо меньшую (см. формулу (9. 7)). Такой (по меньшей мере) проигрыш в результате незнания частоты и фазы.

Схема автокорреляционного приемника показана на рис. 47. Здесь имеется линия задержки З. Ее звенья подобраны так, чтобы наименьшая (начальная) задержка была τ_0, а в последующих звеньях шаг задержки $\Delta\tau_0$ должен позволить с небольшой потерей найти наибольшее значение b_{max} (соответствующее $\cos\omega_0 t=1$), либо среднее значение $b_{ср}$. Здесь применимы соображения, высказанные выше по поводу схемы рис. 45. Конечно, должны быть известны пределы, в которых заключена возможная частота сигнала.

§ 15. Восстановление непрерывного сигнала

Задача восстановления непрерывного сигнала ставится следующим образом: требуется так обработать поступающую на вход приемника смесь сигнала и помехи, чтобы получить на выходе приемника функцию, наименее отличающуюся от сигнала с точки зрения выбранного критерия верности.

Специфическая трудность заключается в том, что о сигнале известно только, что он принадлежит некоторому ансамблю. Иначе говоря, сигнал рассматривается как реализация некоторого случайного процесса, и, следовательно, сигнал может описываться только вероятностными характеристиками этого случайного процесса. Такими характеристиками могут служить распре-

деления или моменты распределений, или же те или иные их преобразования. Наиболее употребительной характеристикой является спектр мощности или функция корреляции, связанные между собой парой преобразований Фурье:

$$G(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} B(\tau) e^{-j\omega t} d\tau, \tag{15.1}$$

$$B(\tau) = \frac{1}{2} \int_{-\infty}^{\infty} G(\omega) e^{j\omega\tau} d\omega. \tag{15.2}$$

Что касается критерия верности, то обычно пользуются критерием квадратичного уклонения. Мы поставим себе целью отыскание коэффициента передачи оптимального фильтра, т. е. пассивного четырехполюсника, отклик которого на входной сигнал, смешанный с шумом, дает наилучшее квадратичное приближение к переданному сигналу.

Пусть входной сигнал

$$x(t) = s(t) + \xi(t). \tag{15.3}$$

Обозначим через $y(t)$ отклик фильтра. В общем случае возможна задержка сигнала на некоторое время α, так что сравнивать между собой нужно $y(t)$ и $s(t-\alpha)$. Но мы положим сначала, что задержка не допускается, и определим уклонение (погрешность) как

$$\varepsilon(t) = y(t) - s(t). \tag{15.4}$$

Отсюда сразу следует, что искомый коэффициент передачи фильтра есть вещественная величина. Действительно, в общем случае

$$K \mid K \mid e^{j\varphi}, \quad \alpha = d\varphi/d\omega$$

и если $\alpha=0$, то φ равно постоянной. Но эта постоянная может быть только нулем, так как нельзя получить неравный нулю фазовый сдвиг, одинаковый для всех частот.

Средний квадрат ошибки может быть выражен через спектр мощности ошибки

$$M\varepsilon^2 = \int_0^{\infty} G_{\varepsilon}(\omega)\, d\omega \tag{15.5}$$

и дело сводится к нахождению этого спектра. Спектр ошибки есть спектр разности $y(t) - s(t)$. Составим сначала выражение для функции корреляции

$$B_{\varepsilon}(\tau) = M\{[y(t) - s(t)][y(t-\tau) - s(t-\tau)]\} = M[y(t)y(t-\tau)] + \\ + M[s(t)s(t-\tau)] - M[s(t)y(t-\tau)] - M[y(t)s(t-\tau)]$$

или

$$B_\varepsilon(\tau) = B_y(\tau) + B_s(\tau) - B_{sy}(\tau) - B_{ys}(\tau). \qquad (15.6)$$

Так как преобразование Фурье (15. 1) линейно, то для спектра ошибки имеем

$$G_\varepsilon(\omega) = G_y(\omega) + G_s(\omega) - G_{sy}(\omega) - G_{ys}(\omega). \qquad (15.7)$$

В этих формулах B_{sy} и B_{ys} — функции взаимной корреляции s и y, а G_{sy} и G_{ys} — соответствующие спектры, называемые взаимными спектральными плотностями.

Выражение для G_y легко найти. Мы имеем вообще

$$G_y = |K|^2 G_x. \qquad (15.8)$$

Но $x = s + \xi$ (см. (15. 3)). Так как s и ξ статистически независимы (точнее, некоррелированны, так что $B_{s\xi} = 0$), то

$$G_x = G_s + G_\xi \qquad (15.9)$$

и

$$G_y = |K|^2 (G_s + G_\xi). \qquad (15.10)$$

Остается найти взаимные спектральные плотности G_{sy} и G_{ys}. Фурье-преобразование для G_{sy} на основании (15. 1) можно записать в виде

$$G_{sy} = \frac{1}{\pi} \int_{-\infty}^{\infty} B_{sy}(\tau)\, e^{-j\omega\tau} d\tau$$

или

$$G_{sy} = \frac{1}{\pi} \int_{-\infty}^{\infty} M[s(t)\, y(t-\tau)]\, e^{-j\omega\tau}\, d\tau. \qquad (15.11)$$

Отклик фильтра выражается интегралом

$$y(t) = \int_{-\infty}^{t} x(\tau)\, g(t-\tau)\, d\tau = \int_{-\infty}^{\infty} x(t-\sigma)\, g(\sigma)\, d\sigma \qquad (15.12)$$

(после замены переменной $\sigma = t - \tau$ нижний предел получается равным нулю. Он заменен на $-\infty$, так как все равно $g(-t) \equiv 0$ для любой физически осуществимой системы). Функция $g(t)$ есть импульсная реакция фильтра, связанная с коэффициентом передачи $K(\omega)$ парой преобразований Фурье

$$K(\omega) = \int_{-\infty}^{\infty} g(t)\, e^{-j\omega t} dt, \qquad (15.13)$$

$$g(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{\infty}K(\omega)\,e^{j\omega t}d\omega. \qquad (15.14)$$

Подставляя (15. 12) в (15. 11), получаем

$$G_{sy}=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}e^{-j\omega\tau}d\tau\int\limits_{-\infty}^{\infty}g(\sigma)\,M[x(t-\tau-\sigma)\,s(\sigma)]\,d\sigma=$$

$$=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}e^{-j\omega\tau}d\tau\int\limits_{-\infty}^{\infty}g(\sigma)\,B_{sx}(\tau+\sigma)\,d\sigma=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}g(\sigma)\,d\sigma\int\limits_{-\infty}^{\infty}B_{sx}(\tau+\sigma)\,e^{-j\omega\tau}d\tau=$$

$$=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}g(\sigma)\,e^{j\omega\sigma}d\sigma\int\limits_{-\infty}^{\infty}B_{sx}(z)\,e^{-j\omega z}dz.$$

Используя (15. 1) и (15. 13), находим

$$G_{sy}=K^{*}G_{sx}. \qquad (15.15)$$

Аналогично можно найти

$$G_{ys}=KG_{sx}. \qquad (15.16)$$

Здесь K и K^{*} — комплексно-сопряженные величины. Но мы условились, что K — вещественная величина. Поэтому $K^{*}=K$ и

$$G_{sy}=G_{ys}=KG_{sx}. \qquad (15.17)$$

Далее

$$B_{sx}(\tau)=M[s(t)\,x(t-\tau)]=M\{s(t)[s(t-\tau)+\xi(t-\tau)]\}=B_{s}(\tau)$$

(так как $B_{s\xi}=0$). Следовательно,

$$G_{sx}=G_{s}$$

и окончательное выражение для спектра погрешности на основании (15. 10) и (15. 15) принимает вид

$$G_{\varepsilon}=K^{2}(G_{s}+G_{\xi})+G_{s}-2KG_{s}. \qquad (15.18)$$

Теперь нужно найти такое значение K, которое дает минимальную квадратичную погрешность. Для этого достаточно минимизировать G_{ε}. Перепишем (15. 18) в виде

$$G_{\varepsilon}=\left(K\sqrt{G_{s}+G_{\xi}}-\frac{G_{s}}{\sqrt{G_{s}+G_{\xi}}}\right)^{2}+\frac{G_{s}G_{\xi}}{G_{s}+G_{\xi}}. \qquad (15.19)$$

Первый член положителен и зависит от K, второй член — заданная величина. Ясно, что наименьшее значение G получится тогда, когда первый член равен нулю. При этом условии

$$K_{\text{опт}}=\frac{G_{s}}{G_{s}+G_{\xi}}. \qquad (15.20)$$

Это и есть решение нашей задачи; формула (15. 20) дает выражение искомого коэффициента передачи фильтра через заданные спектры сигнала и помехи.

Возвращаясь к более общему случаю, когда допускается задержка на время α, получаем

$$K_{\text{опт}} = \frac{G_s}{G_s + G_\xi} e^{-j\omega\alpha}. \qquad (15.21)$$

Модуль коэффициента передачи, т. е. амплитудно-частотная характеристика, остается без изменений.

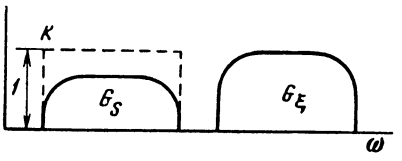


Рис. 48

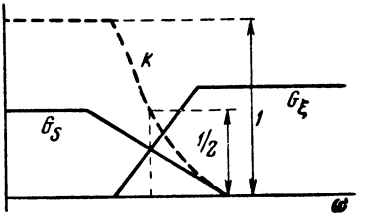


Рис. 49

Если примем оптимальный фильтр с коэффициентом передачи (15. 20), то, как следует из (15. 19),

$$G_{\varepsilon\min} = \frac{G_s G_\xi}{G_s + G_\xi}$$

и средний квадрат погрешности равен

$$M\varepsilon^2_{\min} = \int_0^\infty \frac{G_s G_\xi}{G_s + G_\xi}\, d\omega. \qquad (15.22)$$

Если помеха отсутствует, то $G_\xi = 0$, $\varepsilon = 0$ и $K = 1$, т. е. фильтра не требуется. Если спектры G_s и G_ξ не перекрываются, т. е. при любом значении частоты либо G_s, либо G_ξ равно нулю, то их произведение равно нулю для всех частот, и в этом случае помеха может быть полностью устранена, т. е. $\varepsilon = 0$, при условии, что $K = 1$ в полосе частот, занимаемой спектром сигнала (рис. 48).

В действительности, спектры сигнала и помехи всегда в той или иной мере перекрываются и погрешность не равна нулю. При этом фильтр с оптимальной характеристикой (15. 20) пропускает различные частоты с тем меньшим весом, чем больше отношение G_ξ/G_s при данной частоте. Так, например, если спектры сигнала и помехи перекрываются, как показано на рис. 49, то частотная характеристика фильтра должна иметь вид, показанный на том же рисунке штриховой линией.

Если спектры сигнала и помехи занимают одну и ту же полосу и вдобавок отношение сигнал / помеха мало, то удовлетворитель-

ное восстановление сигнала при помощи фильтрации становится невозможным.

Действительно, при $G_s \ll G_\xi$ имеем

$$M\varepsilon^2_{\min} \simeq \int_0^\infty G_s d\omega = P_s,$$

т. е. средний квадрат (мощность) погрешности сравнивается с мощностью самого сигнала.

Можно получить лучший результат, если прибегнуть к так называемому предыскажению. Передача с предыскажением

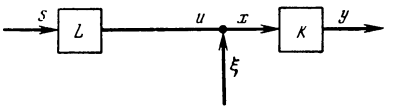


Рис. 50

состоит в том, что на передающей стороне сигнал s пропускается через фильтр с коэффициентом передачи L. На выходе этого фильтра получается видоизмененный сигнал, который поступает в канал. На выходе канала имеем сигнал $x=u+\xi$. Этот сигнал поступает на фильтр с коэффициентом передачи K, на выходе которого получается сигнал y. Схема системы показана на рис. 50.

Нас по-прежнему интересует разность $\varepsilon=y-s$. Выражения (15. 6)—(15. 8) остаются в силе, равно как и (15. 5) и (15. 16). Действуя совершенно так же, как при выходе этих последних двух формул, получим в рассматриваемом случае $G_{sx}=L^*G_s$, так что $G_{sy}=K^*G_{sx}=K^*L^*G_s$. Для G_{ys} получается сопряженная величина $G_{ys}=KLG_s$.

Мы будем полагать, что KL — действительная величина, так что $G_{sy}=G_{ys}=KLG_s$. Подставляя эту величину в (15. 7), находим

$$G_\varepsilon = (KL-1)^2 G_s + K^2 G_\xi. \qquad (15.23)$$

Найдем оптимальный коэффициент K, дающий минимум среднеквадратичной погрешности, т. е. функционалу

$$M\varepsilon^2 = \int^\infty G_\varepsilon d\omega.$$

Это — вариационная задача. Чтобы решить ее, составим уравнение Эйлера для подынтегральной функции[1]

$$F[K(\omega), \omega] = G_\varepsilon,$$

т. е. $\partial F/dK = 0$. Это дает

$$K = LG_s/L^2G_s + G_\xi. \qquad (15.24)$$

[1] Это вырожденное уравнение, т. е. F не зависит от K'. Условия на границах, т. е. при $\omega=0$ и $\omega=\infty$, удовлетворяются вследствие общих свойств функций K и L как коэффициентов передачи.

Теперь найдем оптимальное выражение для L. Будем решать вариационную задачу как изопериметрическую, поставив дополнительное условие,

$$P_u = \int_0^\infty G_u d\omega = \int_0^\infty L^2 G_s d\omega = \text{const}, \tag{15.25}$$

и составим уравнение Эйлера для

$$F_1[L^2(\omega), \omega] = G_\varepsilon + \lambda^2 L^2 G_s = \frac{G_s G_\xi}{L^2 G_s + G_\xi} + \lambda^2 L^2 G_s.$$

Приравнивая нулю производную $\partial F_1/\partial L^2$, находим

$$L^2 = \frac{1}{\lambda}\sqrt{\frac{G_\xi}{G_s}} - \frac{G_\xi}{G_s}. \tag{15.26}$$

Подставляя (15.24) и (15.26) в (15.23), получаем

$$G_\varepsilon = \lambda\sqrt{G_s G_\xi}. \tag{15.27}$$

Остается найти постоянную λ. Для этого воспользуемся условием (15.25), в которое подставим значение L^2 из (15.26). Это даст

$$\lambda = \frac{1}{P_u + P_\xi}\int_0^\infty \sqrt{G_s G_\xi}\, d\omega, \tag{15.28}$$

и, таким образом,

$$G_\varepsilon = \frac{\sqrt{G_s G_\xi}}{P_u + P_\xi}\int_0^\infty \sqrt{G_s G_\xi}\, d\omega. \tag{15.29}$$

Теперь мы можем вычислить наименьшую среднеквадратичную погрешность, получаемую при оптимальном выборе L и K (по формулам (15.24), (15.26) и (15.28)). Для этого нужно проинтегрировать (15.29) и мы получаем окончательный результат

$$M\varepsilon^2_{\min} = \int_0^\infty G_\varepsilon d\omega = \frac{1}{P_u + P_\xi}\left(\int_0^\infty \sqrt{G_s G_\xi}\, d\omega\right)^2. \tag{15.30}$$

Заметим, что оптимальные коэффициенты передачи L и K не являются взаимно обратными величинами. Они связаны соотношением

$$KL = 1 - \lambda\sqrt{G_\xi/G_s},$$

которое вырождается в $KL=1$ только при отсутствии помехи, т. е. когда фильтры вообще не нужны.

Заметим еще, что, согласно неравенству Буняковского,

$$M\varepsilon^2_{\min} \leqslant \frac{1}{P_u + P_\xi} \int_{\Delta\omega} G_\xi d\omega \int_{\Delta\omega} G_s d\omega = \frac{\hat{P}_s \hat{P}_\xi}{P_u + P_\xi}. \qquad (15.31)$$

Знак равенства относится к случаю, когда G_s и G_ξ постоянны в полосе перекрытия $\Delta\omega$ (т. е. при тех частотах, при которых ни G_s, ни G_ξ не равны нулю). В знаменателе стоят полные мощности, в числителе — мощности, приходящиеся на полосу перекрытия, т. е., например,

$$\hat{P}_s = \int_{\Delta\omega} G_s d\omega.$$

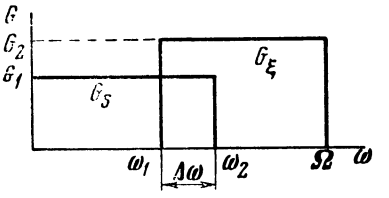


Рис. 51

С целью пояснить различие между фильтрацией с предыскажением и без него, рассмотрим простейший пример, а именно: пусть

$$G_s(\omega) = G_1 = \text{const} \quad [0 < \omega < \omega_2],$$
$$G_\xi(\omega) = G_2 = \text{const} \quad [\omega_1 < \omega < \Omega],$$

где $\omega_2 - \omega_1 = \Delta\omega$ — полоса перекрытия, а Ω — общая полоса, занимаемая сигналом и помехой (рис. 51).

Тогда по формуле (15. 22)

$$M\varepsilon^2_{\min} = \frac{\hat{P}_s \hat{P}_\xi}{\hat{P}_u + \hat{P}_\xi}.$$

Сравнивая это выражение с (15. 31), мы видим, что предыскажение дает выигрыш в

$$\frac{P_u + P_\xi}{\hat{P}_u + \hat{P}_\xi} \geqslant \frac{\Omega}{\Delta\omega} \cdot \frac{L^2 G_1 \frac{\omega_2}{\Omega} + G_2\left(1 - \frac{\omega_1}{\Omega}\right)}{G_1 + G_2}$$

раз, т. е. выигрыш тем больше, чем меньше относительная ширина полосы перекрытия.

§ 16. Мультипликативная помеха

Мы занимались до сих пор исключительно аддитивной помехой. Необходимо хотя бы вкратце разобрать некоторые вопросы, относящиеся к мультипликативной помехе.

Прежде всего заметим, что мультипликативную помеху всегда можно свести с эквивалентной аддитивной. Это обстоятельство во многом упрощает исследование действия мультипликативной помехи.

В самом деле, если записать

$$x = s\vartheta = \vartheta_0 s + \xi_э, \tag{16.1}$$

где s — переданный сигнал; x — принятый сигнал; ϑ — стационарный случайный процесс, выражающий мультипликативную помеху; ϑ_0 — его среднее значение; $\xi_э$ — эквивалентная аддитивная помеха, то сразу получаем

$$\xi_э = s(\vartheta - \vartheta_0) = s\zeta. \tag{16.2}$$

После этого можно пользоваться результатами, полученными ранее для аддитивной помехи, беря ее равной $\xi_э$.

Здесь надо, однако, указать на особенности помехи $\xi_э$. Самое существенное ее свойство состоит в том, что она представляется произведением случайного процесса $\zeta = \vartheta - \vartheta_0$ на детерминированную функцию времени s, представляющую сигнал. Это значит, что $\xi_э$ есть нестационарный процесс, и что все его распределения и их моменты зависят от времени. Практически это означает, что для получения осмысленных результатов нужно после усреднения по множеству прибегать к повторному усреднению по времени.

Затем нужно напомнить, что ϑ есть случайный процесс с не нулевым средним

$$M\vartheta = \vartheta_0 \neq 0,$$

и что по смыслу мультипликативной помехи $\vartheta > 0$. Таким об разом, ϑ имеет распределение, ограниченное снизу. Процесс $\zeta = \vartheta - \vartheta_0$ имеет уже нулевое среднее и соответственно смещенное распределение. Все остальные сведения об этом процессе можно почерпнуть только из опыта путем статистической обработки его результатов. Отметим еще, что ϑ и ζ — безразмерные величины, тогда как аддитивная помеха $\xi_э$ имеет размерность сигнала s.

Найдем эквивалентное отношение сигнал/помеха для случая мультипликативной помехи. Выразим это отношение как $\rho_0 = \vartheta_0^2 s^2 / \overline{D\xi}_э$. Мы имеем

$$D\xi_э = Ds\zeta = s^2 D\zeta; \quad \overline{D\xi}_э = \overline{s^2 D\zeta} \left(= \overline{s^2} D\vartheta\right)$$

и, следовательно,

$$\rho_0 = \frac{\vartheta_0^2}{D\zeta} \left(= \frac{\vartheta_0^2}{D\vartheta}\right), \tag{16.3}$$

т. е. отношение сигнал/помеха определяется только средним значением и дисперсией процесса, характеризующего мультипликативную помеху.

Рассмотрим действие метода накопления в случае мультипликативной помехи. Мы имеем

$$y = s\vartheta_1 + s\vartheta_2 + \ldots + s\vartheta_n = \sum_{i=1}^{n} s\vartheta_i = \sum s(\zeta_i + \vartheta_{0i}).$$

Предполагается, что один и тот же сигнал s передается по n каналам, в которых мультипликативные помехи независимы. Средняя интенсивность сигнала, поступающего по i-му каналу, определяется величиной ϑ_{0i}. Введя среднее по всем каналам значение

$$\vartheta_0 = \frac{1}{n}\sum_{i=1}^{n} \vartheta_{0i},$$

можем переписать выражение для принятого сигнала y в виде

$$y = \vartheta_0 ns + \sum s\zeta_i = b + \eta.$$

В этом выражении первый член представляет полезный сигнал, а второй — помеху. Отношение сигнал/помеха запишется в виде

$$\rho = \frac{b^2}{D\eta} = \frac{\vartheta_0^2 n^2 \overline{s^2}}{D\sum s\zeta_i}.$$

Так как ζ_i предполагаются статистически независимыми, то

$$D\sum s\zeta_i = s^2 D\sum \zeta_i = s^2 n D\zeta,$$

и отношение сигнал/помеха равно

$$\rho = \vartheta_0^2 n / D\zeta = n\rho_0.$$

Итак, выигрыш от применения метода накопления тот же самый, что и при аддитивной помехе.

Разберем вопрос об оптимальном приеме. Представляя действие приемника операцией интегрирования с весом, запишем

$$y = \int_0^T \varphi(t)\, x(t)\, dt = \int_0^T \varphi(t)\, \vartheta(t)\, s(t)\, dt =$$

$$= \int_0^T \varphi(t)\, [\zeta(t) + \vartheta_0]\, s(t)\, dt = \vartheta_0 \int_0^T \varphi(t)\, s(t)\, dt +$$

$$+ \int_0^T \varphi(t)\, \zeta(t)\, s(t)\, dt = b + \eta.$$

Здесь b — полезный сигнал (на входе решающего устройства), η — помеха. При полностью известном сигнале s наивыгоднейший вид весовой функции есть $\varphi = s$.

Тогда

$$b = \vartheta_0 \int_0^T s^2(t)\,dt = E_s, \quad \eta = \int_0^T \zeta(t)\, s^2(t)\,dt$$

и отношение сигнал/помеха

$$\rho = \frac{b^2}{D\eta} = \frac{\vartheta_0^2 E_s^2}{D\eta}.$$

Если мультипликативная помеха представляется медленным (по сравнению с $s(t)$) процессом, то можно полагать, что за время T величина $\zeta(t)$ заметным образом не меняется и может рассматриваться как постоянный случайный множитель, так что

$$\eta \simeq \zeta(t) \int_0^T s^2(t)\,dt = E_s \zeta.$$

Тогда

$$D\eta \approx E_s^2 D\zeta$$

и для отношения сигнал/помеха получаем выражение (16. 3)

$$\rho = \frac{\vartheta_0^2}{D\zeta}.$$

На практике мультипликативная помеха возникает во всех случаях, когда параметры системы передачи претерпевают случайные изменения во времени. В сущности так обстоит дело во всех реальных системах, но иногда флюктуации параметров хотя и существуют, но практически неощутимы. В других же случаях, когда случайные изменения коренным образом перестраивают весь механизм передачи, она может стать вовсе невозможной; так обстоит, например, дело с суточными и сезонными изменениями условий передачи коротких радиоволн.

Отдельного упоминания заслуживает явление замирания (фединг). Интерференционный механизм этого явления чрезвычайно чувствителен к незначительным изменениям условий распространения. В грубых чертах дело заключается в том, что волна, посланная передатчиком, достигает антенны приемника, следуя одновременно по нескольким разным путям (так называемое многопутевое или многолучевое распространение). Вследствие разности длин различных путей возникают соответственные разности фаз, и волны, прибывшие различными путями, интерферируют между собой. А так как самые пути представляют собой случайные образования, все время изменяющиеся с изменениями состояния атмосферы, то в результате интенсивность принимаемого сигнала претерпевает глубокие изменения вплоть до полного пропадания на некоторое время. Если представить себе простей-

шую модель с двумя лучами равной интенсивности, то пропадание сигнала вследствие интерференции будет происходить уже при разности ходов, равной половине длины волны, на коротких волнах это составляет порядок десятка метров. При большем числе лучей характер явления определяется распределением суммы синусоидальных колебаний со случайными фазами и (амплитудами), к тому же и число слагаемых (т. е. число путей или лучей) также случайно. Мы не будем вдаваться ни в подробности теории, ни в анализ обширного экспериментального материала, накопленного к настоящему времени. Укажем лишь на простые и эффективные методы борьбы с замиранием. Применяемые методы борьбы с замиранием следует отнести к разновидностям метода накопления. Суть дела состоит в том, что стремятся образовать несколько каналов по возможности с независимыми замираниями. Чаще всего ограничиваются двумя каналами; это уже дает заметный эффект, хотя теория показывает, и опыт это подтверждает, что дальнейшее увеличение числа каналов дает ощутительный дополнительный выигрыш.

Сигналы нескольких каналов можно использовать по-разному. Можно просто их сложить — это будет метод накопления в его классической форме. Но часто прибегают к другому приему — к автоматическому подключению того канала, сигнал которого в данный момент больше. Операцию обработки нескольких сигналов в общем виде можно представить как суммирование с весом. Легко видеть, что при простом накоплении весовые коэффициенты одинаковы, а при выборе максимального сигнала все весовые коэффициенты равны нулю, кроме одного. Можно выбирать весовые коэффициенты так, чтобы удовлетворить некоторому критерию оптимальности. Так, можно потребовать максимизации отношения сигнал/помеха для суммарного сигнала. Впрочем преимущество такой оптимальной системы перед системой с простым накоплением, как показывают расчеты, невелико.

Один из методов борьбы с замиранием состоит в применении приема на разнесенные антенны. Если рассмотреть напряженность поля в месте приема как функцию времени и двух пространственных координат (в горизонтальной плоскости), то окажется, естественно, что процессы в двух фиксированных точках протекают тем более сходно, чем эти точки ближе. Раздвигая точки наблюдения, можно найти такое наименьшее расстояние между ними, на котором изменения напряженности поля можно уже считать практически некоррелированными. Это расстояние называется интервалом (пространственной) корреляции. Разнос антенн и определяется интервалом корреляции. Практика показывает, что на коротких волнах (частоты порядка 10 *Мгц*) отношение разноса d к длине волны λ должно быть порядка 10.

Другой метод состоит в разнесении по частоте, т. е. в передаче одного и того же сигнала на двух различных несущих. Суть дела в следующем: замирание обусловлено интерференцией,

зависящей от фазовых соотношений, а на другой частоте при тех же разностях ходов фазовые соотношения будут совершенно иные. Можно искать интервал частотной корреляции, т. е. наименьший интервал по шкале частот между двумя несущими частотами, обеспечивающий практически некоррелированное замирание по обоим каналам. Оказывается, что этот интервал невелик — достаточен относительный разнос двух частот порядка 10^{-3} (т. е. на интервал порядка 1 *кгц*).

Кроме передачи на двух (или более) синусоидальных несущих колебаниях для борьбы с замиранием предлагалось использование в качестве переносчика шума с подходящим образом выбранной шириной полосы. Другими словами, берется переносчик не с линейчатым, а со сплошным спектром определенной ширины. Эта возможность пока на практике не использовалась.

Заканчивая этот параграф, нельзя не отметить, что самым естественным методом устранения мультипликативной помехи является применение автоматической регулировки усиления (АРУ). Можно рассматривать АРУ как корректор для мультипликативной помехи. Идеальное устройство АРУ, действие которого можно было бы представить как умножение сигнала на $1/\vartheta$, где ϑ — мультипликативная помеха, позволило бы полностью от нее избавиться. Но, к сожалению, кроме мультипликативной помехи имеется всегда и аддитивная. Обозначая через z сигнал на выходе, можем записать

$$x = \vartheta s + \xi, \quad z = \frac{1}{\vartheta} x = s + \frac{\xi}{\vartheta},$$

т. е. в результате действия идеального АРУ мы хотя и получаем сигнал постоянной интенсивности, но с флюктуирующей по интенсивности аддитивной помехой. Отношение сигнал/помеха АРУ изменить не может (в отличие от описанных выше методов с разделением каналов).

§ 17. Помеха, коррелированная с сигналом

При обсуждении методов борьбы с помехой предполагают обычно сигнал и помеху статистически независимыми, или, что является менее жестким условием, не коррелированными между собой.

Однако во многих случаях это не так, и тогда полученные ранее результаты требуют пересмотра.

Иногда помеха не только коррелирована с сигналом, но полностью им обусловлена. Во избежание недоразумений нам нужно условиться о терминологии, и, в частности, ввести отчетливое различие между понятиями помехи и искажения сигнала.

Искажением мы будем называть всякое детерминированное (неслучайное) преобразование сигнала, которое нам известно либо на основании теоретических данных, либо в результате пря-

мого эксперимента. Простейший вид искажений определяется нелинейной зависимостью между воздействием x на входе и откликом y на выходе некоторого безреактивного четырехполюсника

$$z = f(x). \tag{17.1}$$

Так как функция f предполагается известной, то искажение может быть устранено, если взять дополнительный корректирующий четырехполюсник с характеристикой

$$z = f^{-1}(y),$$

где f^{-1} — функция, обратная f. Однако необходимым условием выполнимости такого рода коррекции является однозначность функции f и наличие всюду конечной производной.

Смысл этого условия легко уяснить на примере квантования. Квантование можно представить как результат нелинейного преобразования (17. 1) со ступенчатой характеристикой, показанной на рис. 52. Ясно, что восстановить исходный сигнал по квантованному нельзя, хотя характер произведенного преобразования в точности известен.

В более общем случае для системы с памятью, например в системе с реактивными элементами, способными накапливать энергию, преобразование сигнала выражается уже не функцией (17. 1), а некоторым оператором. Так, линейные искажения сигнала могут быть выражены линейным оператором вида

$$y(t) = \int_{-\infty}^{t} x(\tau)\, g(t-\tau)\, d\tau, \tag{17.2}$$

где $g(t)$ — импульсная реакция системы, т. е. отклик системы на единичный импульс. Линейные искажения описываются обычно на частотном (спектральном) языке, на котором к тому же особенно просто формулируются правила построения корректирующих устройств. Коррекция становится невозможной, если некоторые участки спектра выпадают вовсе.

Чтобы покончить с вопросом об искажениях, отметим еще специальный вид искажения, известный под названием попутного потока. Речь идет о возмущении сложной структуры, образующемся в кабеле (или волноводе) в результате многократных отражений несущей сигнал волны от случайных неоднородностей кабеля. Эти неоднородности (обусловленные как производством, так и прокладкой) можно характеризовать некоторой случайной функцией, представляющей зависимость того или иного физического параметра от пространственной координаты, отсчитываемой вдоль оси кабеля. Но для данного кабеля мы имеем дело лишь с одной из реализаций этой функции, т. е. с детерминированной функцией. Поэтому построение корректора для попутного потока в принципе возможно (хотя практическое его осуществление наталкивается на значительные затруднения).

При наличии искажений мы всегда можем представить искаженный сигнал в виде

$$y = x + \delta x.$$

Но слагаемое δx отличается от аддитивной помехи тем, что если $x \equiv 0$, то и $\delta x \equiv 0$, каков бы ни был оператор системы.

Все предыдущие рассуждения и примеры нужны были только для того, чтобы уточнить наше понимание помехи. Мы будем называть помехой всякое возмущение, не связанное с сигналом посредством детерминированного функционального оператора. Таким образом, наличие в зависимости между возмущением и

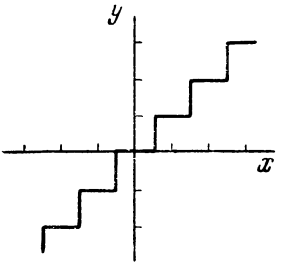


Рис. 52

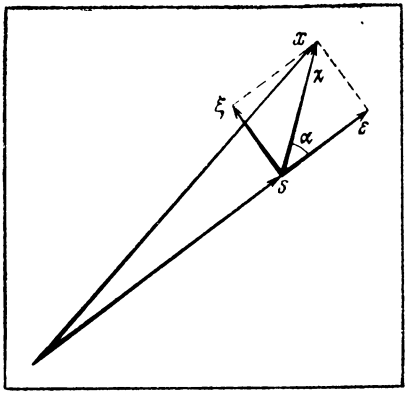


Рис. 53

сигналом хотя бы слабой степени случайности послужит для нас основанием квалифицировать это возмущение как помеху. Так, например, если неоднородности, порождающие попутный поток, не остаются постоянными, а случайным образом изменяются во времени, то в этих условиях попутный поток должен рассматриваться уже не как искажение, а как помеха.

Итак, искажение сигнала и наложение статистически независимой помехи — это два крайних случая. Мы рассмотрим вкратце один промежуточный случай — случай аддитивной помехи, коррелированной с сигналом.

Начнем с геометрической модели рис. 53, где $\boldsymbol{s}$ — вектор передаваемого сигнала; $\boldsymbol{\chi}$ — вектор помехи. Оба вектора образуют угол α, косинус которого равен нормированному коэффициенту корреляции

$$\cos\alpha = \frac{\boldsymbol{s}\boldsymbol{\chi}}{\|\boldsymbol{s}\| \cdot \|\boldsymbol{\chi}\|} = k_{s\chi} = k. \qquad (17.3)$$

Для наших целей удобно разложить помеху $\boldsymbol{\chi}$ на две составляющие, $\boldsymbol{\xi}$ и $\boldsymbol{\varepsilon}$ (рис. 53), из которых первая ортогональна сигналу $\boldsymbol{s}$, а вторая совпадает по направлению с $\boldsymbol{s}$, так что можно записать

$$\boldsymbol{\varepsilon} = \gamma \boldsymbol{s}, \qquad (17.4)$$

где γ — множитель, постоянный для данной реализации. Так как ξ и ε представляют собой проекции χ на два взаимно перпендикулярных направления, то имеем

$$\|\varepsilon\| = \|\chi\| \cos\alpha = \|\chi\| k, \tag{17.5}$$

$$\|\xi\| = \|\chi\| \sin\alpha = \|\chi\| \sqrt{1-k^2}. \tag{17.6}$$

Для принятого сигнала можем записать

$$x = s + \chi = s + \varepsilon + \xi = (1+\gamma)s + \xi,$$

и дело сводится, как видим, к тому, что в качестве аддитивной помехи выступает только некоррелированная составляющая ξ, а наличие корреляции увеличивает (если $k > 0$) полезный сигнал. Отношение сигнал/помеха запишется в виде

$$\rho = \frac{(1+\gamma)^2 \|s\|^2}{\|\xi\|^2} = \frac{(1+\gamma)^2 \|s\|^2}{(1-k^2)\|\chi\|^2},$$

или

$$\rho = \frac{(1+\gamma)^2}{1-k^2} \rho_0, \tag{17.7}$$

если обозначить

$$\rho_0 = \frac{\|s\|^2}{\|\chi\|^2}. \tag{17.8}$$

Коэффициент ϑ определяется с помощью (17. 4) и (17. 5); он равен

$$\gamma = \frac{\|\chi\|}{\|s\|} k = \frac{k}{\sqrt{\rho_0}}, \tag{17.9}$$

так что окончательно для отношения сигнал/помеха имеем

$$\rho = \frac{(1 + k/\sqrt{\rho_0})^2}{1-k^2} \rho_0. \tag{17.10}$$

Мы ввели разложение $\chi = \gamma s + \xi$, основываясь непосредственно на наглядной геометрии рис. 53. Но если не прибегать к геометрическим представлениям, то нужно доказать следующее предложение:

Функцию $\gamma(t)$ такую, что $\overline{\chi} = 0$, $\overline{\chi s} \neq 0$, где $s(t)$ — заданная функция ($\overline{s} = 0$), можно представить в виде $\chi(t) = \gamma s(t) + \xi(t)$, причем составляющая $\xi(t)$ будет не коррелирована (т. е. $\overline{\xi s} = 0$), если выбрать $\gamma = \overline{\chi s}/\overline{s^2}$, где черта означает усреднение по реализации.

Доказательство очень просто. Запишем

$$\xi = \chi - \gamma s,$$

умножим обе части на s и возьмем средние

$$\overline{\xi s} = \overline{\chi s} - \overline{\gamma s^2}.$$

Но левая часть должна равняться нулю. Поэтому

$$\gamma = \frac{\overline{\chi s}}{\overline{s^2}}$$

или

$$\gamma = \sqrt{\frac{\overline{\chi_2}}{\overline{s^2}}} \cdot \frac{\overline{\chi s}}{\sqrt{\overline{s^2}\,\overline{\chi^2}}},$$

что соответствует (17. 9). Если χ — эргодический процесс, то усреднение по реализации можно заменить усреднением по множеству.

Теперь мы повторим все приведенные выше выкладки в терминах пространства C_{L^2} непрерывных функций, заданных на интервале. Иначе говоря, рассмотрим действие приемника, выполняющего над принятым сигналом операцию интегрирования с весом. Соответствующий линейный функционал имеет вид

$$y = \int_0^T \varphi(t)\, x(t)\, dt = \int_0^T \varphi(t)\,[s(t) + \chi(t)]\, dt =$$

$$= \int_0^T \varphi(t)\,[s(t) + \gamma s(t) + \xi(t)]\, dt =$$

$$= (1+\gamma) \int_0^T \varphi(t)\, s(t)\, dt + \int_0^T \varphi(t)\, \xi(t)\, di = b + \eta.$$

Оптимальный прием получается при $\varphi(t) = s(t)$. При этом

$$b = (1+\gamma) \int_0^T s^2(t)\, dt = (1+\gamma) E_s,$$

$$\eta = \int_0^T s(t)\, \xi(t)\, dt.$$

Так как $\xi(t)$ и $s(t)$, по нашему предположению, не коррелированы, то $M\eta = 0$ и $D\eta = M\eta^2$. Для этой последней величины ранее было получено

$$D\eta \approx \frac{\tau_0}{T} E_s E_\xi,$$

так что отношение сигнал/помеха равно

$$\rho = \frac{b^2}{D\eta} = \frac{(1+\gamma)^2 T E_s}{\tau_0 E_\xi} = 2FT \frac{(1+\gamma)^2 E_s}{E_\xi}.$$

Связь между E_ξ и E_χ найдем из соотношения

$$E_\chi = \int_0^T \chi^2 dt = \int_0^T (\xi + \varepsilon)^2 dt = E_\xi + \gamma^2 E_s =$$

$$= E_\xi + \frac{k^2}{\rho_0} E_s = E_\xi + k^2 E_\chi,$$

откуда

$$E_\xi = (1 - k^2) E_\chi$$

и, наконец,

$$\rho = 2FT \frac{1 + k/\sqrt{\rho_0}}{1 - k^2} \rho_0.$$

Легко видеть, что при $k \to 0$, т. е. в случае некоррелированной помехи, мы получаем уже хорошо известный из предыдущего результат

$$\rho = 2FT\rho_0.$$

При $k \to 1$ отношение сигнал/помеха стремится к бесконечности, что означает, что помеха, как таковая, исчезает — она превращается в некоторую добавку к полезному сигналу.

Общее заключение, к которому мы приходим, состоит в том, что наличие корреляции между аддитивной помехой и сигналом увеличивает отношение сигнал/помеха за счет уменьшения собственно помехи (т. е. некоррелированной составляющей помехи).

§ 18. Обнаружение сигнала как статистическая задача

За последнее время в теории приема сигналов применяются методы математической статистики, проблемы приема трактуются с точки зрения теории статистических решений и теории игр. Такой подход позволяет ставить и решать вопросы построения приемных устройств с большой общностью, а главное, с полной ясностью в отношении критериев оптимальности. Применению статистических методов к проблемам приема сигналов посвящена обширная литература. В этом параграфе мы ограничимся кратким изложением основных понятий применительно к простейшей задаче обнаружения полностью известного сигнала. Как мы увидим, в предыдущем изложении уже были использованы некоторые понятия теории статистических решений, однако без применения терминологии этой теории.

Рассмотрим простейшую так называемую двухальтернативную ситуацию, когда происходит одно из двух событий, A или B. Мы производим наблюдения, совокупность которых в математической статистике называется выборкой. Характер этих наблюдений таков, что по ним нельзя с достоверностью установить, какое из двух событий, A или B, в действительности имело место.

Известны лишь вероятностные связи между событиями и наблюдениями.

Наша задача состоит в том, чтобы по результатам наблюдений принять одну из двух гипотез: гипотезу H_A о том, что произошло событие A, или гипотезу H_B о том, что произошло событие B. Обе гипотезы взаимно исключают друг друга, так как A и B образуют полную систему событий. Выбор одной из гипотез — это и есть то, что называется статистическим решением.

В дальнейшем мы будем говорить о критериях и о правилах решения. Под критерием понимаются некоторые общие условия, которым должен удовлетворять выбор гипотезы, т. е. решение. Эти условия часто имеют экстремальный характер, т. е. требуется, чтобы решение минимизировало или максимизировало те или иные величины. Под правилом понимается описание конкретной процедуры, которую нужно выполнить, чтобы получить удовлетворяющее данному критерию решение.

В простейшем случае критерий может состоять в том, что из двух гипотез выбирается более правдоподобная. Иначе говоря, выбирается та гипотеза, которая с большей вероятностью является правильной.

Для решения используются апостериорные вероятности событий, т. е. условные вероятности событий A и B при условии, что получены наблюдения $y_1, y_2, \ldots$, образующие выборку $y=(y_1, y_2, \ldots, y_n)$. Для апостериорных вероятностей на основании формулы Бейеса имеем

$$\begin{aligned} p(A/y) &= \frac{p(y/A)\,p(A)}{p(y/A)\,p(A)+p(y/B)\,p(B)}, \\ p(B/y) &= \frac{p(y/B)\,p(B)}{p(y/A)\,p(A)+p(y/B)\,p(B)}, \end{aligned} \tag{18.1}$$

где $p(A)$ и $p(B)$ — безусловные так называемые априорные вероятности событий A и B. Составим отношение апостериорных вероятностей

$$\frac{p(A/y)}{p(B/y)} = \frac{p(y/A)\,p(A)}{p(y/B)\,p(B)}. \tag{18.2}$$

Условные вероятности $p(y/A)$ и $p(y/B)$ представляют собой вероятности того, что вектор y попадает в область

$$dV = dy_1 dy_2 \ldots dy_n$$

при условии, что имеет место событие A или событие B. Таким образом, можно выразить вероятности через соответствующие плотности

$$p(y/A) = w(y/A)\,dV, \quad p(y/B) = w(y/B)\,dV.$$

Подставляя в (18. 2) и сокращая на dV, получаем

$$\frac{p(A/y)}{p(B/y)} = \frac{w(y/A)\,p(A)}{w(y/B)\,p(B)}. \tag{18.3}$$

Для понимания дальнейшего важно заметить, что в то время как априорные вероятности $p(A)$ и $p(B)$ — просто числа, условные вероятности $p(y/A)$ и $p(y/B)$ и соответствующие плотности $w(y/A)$ и $w(y/B)$ — случайные величины. Поэтому в качестве аргумента плотностей взята сама случайная величина y, а не текущая переменная шкалы, по которой она измеряется.

Теперь можно сформулировать правило решения: если $p(A/y) > p(B/y)$, т. е.

$$\frac{p(A/y)}{p(B/y)} > 1, \tag{18.4}$$

то принимается гипотеза H_A, т. е. считается, что произошло событие A; в противном случае принимается гипотеза H_B.

Входящее в (18. 3) отношение

$$\Lambda = \frac{w(y/A)}{w(y/B)} \tag{18.5}$$

называется отношением правдоподобия [1]. Введя это определение, можем, используя (18. 2) и (18. 3), сформулировать правило решения в виде

$$\Lambda > \frac{p(B)}{p(A)} = \Lambda_0, \tag{18.6}$$

и дело сводится, таким образом, к нахождению отношения правдоподобия и к сличению его значения с постоянным пороговым значением Λ_0, зависящим от априорных вероятностей. Если события A и B равновероятны, т. е. $p(A) = p(B)$, то правило решения принимает вид

$$\Lambda > 1. \tag{18.7}$$

Результат наблюдения y можно в самом общем случае представить вектором в пространстве наблюдений Y соответствующего числа измерений. Тогда задача выбора гипотезы получает наглядную геометрическую интерпретацию: пространство наблюдений Y делится на две области: Y_A и Y_B. Если y попадает в Y_A ($y \in Y_A$), то принимается гипотеза H_A; если же Y попадает в ($y \in Y_B$), то H_B. Область Y_A называется областью принятия гипотезы H_A;

---

[1] Иногда этот термин относят к отношению апостериорных вероятностей и называют

$$\frac{w(y/A)\, p(A)}{w(y/B)\, p(B)} = \tilde{\Lambda}$$

обобщенным отношением правдоподобия. Очевидно,

$$\tilde{\Lambda} = \frac{p(A)}{p(B)} \Lambda.$$

Для величины Λ встречаются также термины коэффициент правдоподобия, вероятностное отношение.

область Y_B — областью отвергания гипотезы, или критической областью. Правило решения с геометрической точки зрения сводится к установлению границы между областями Y_A и Y_B. Эта граница представляет собой поверхность в пространстве Y, называемую решающей поверхностью.

Применительно к задаче обнаружения сигнала события A и B соответствуют наличию и отсутствию сигнала. Результат наблюдения y есть принимаемый сигнал; он представляет собой либо сигнал плюс шум, либо только шум. Приемник выдает решение, принимая одну гипотезу и отвергая вторую. Собственные области, о которых говорилось ранее, это и есть области принятия соответствующих гипотез.

При различении двух сигналов события A и B соответствуют посылке сигнала s_1 или сигнала s_2. Случай различения многих сигналов представляет собой уже многоальтернативную ситуацию и здесь не будет рассматриваться.

Введем теперь вероятности ошибок. Пусть $A=C_1$ означает наличие сигнала, $B=C_0$ — отсутствие сигнала. Пусть, далее, H_1 и H_0 — гипотезы о наличии и отсутствии сигнала, а Y_1 и Y_0 — области принятия соответствующих гипотез. Принятый сигнал обозначим через y. При наличии двух возможных событий и соответственно двух гипотез возможны следующие четыре случая:

Действительное событие	Принятый сигнал	Выбранная гипотеза	Решение
C_1	$y \in Y_1$	H_1	Q_{11} — правильное
C_1	$y \in Y_0$	H_0	Q_{01} — ошибочное
C_0	$y \in Y_0$	H_0	Q_{00} — правильное
C_0	$y \in Y_1$	H_1	Q_{10} — ошибочное

Итак, возможны ошибки двоякого рода. Решение Q_{01} означает, что мы считаем сигнал отсутствующим, когда в действительности он есть; такая ошибка называется пропуском сигнала. Если же принять решение Q_{10}, т. е. что сигнал есть, когда на самом деле его нет, то в этом случае ошибка называется ложной тревогой (эта терминология заимствована из области радиолокации). Вероятности этих двух видов ошибок выражаются через условные распределения принятого сигнала и априорные вероятности $p(C_0)$ и $p(C_1)$ следующими формулами:

$$\begin{aligned} p(Q_{01}) &= p(C_1) \int\limits_{Y_0} w(y/C_1)\, dy, \\ p(Q_{10}) &= p(C_0) \int\limits_{Y_1} w(y/C_0)\, dy. \end{aligned} \tag{18.8}$$

В дальнейшем для сокращения письма мы воспользуемся обозначениями

$$p(C_1) = p_1, \quad p(C_0) = p_0,$$
$$w(y/C_1) = w_1(y), \quad w(y/C_0) = w_0(y).$$

Мы будем также применять сокращенные обозначения для условных вероятностей пропуска сигнала и ложной тревоги, а именно:

$$\int_{Y_0} w_1(y)\,dy = \beta,$$
$$\int_{Y_1} \omega_0(y)\,dy = \alpha, \tag{18.9}$$

так что (18.8) записывается в виде

$$p(Q_{01}) = p_1\beta,$$
$$p(Q_{10}) = p_0\alpha. \tag{18.10}$$

Теперь заметим, что ошибочные решения Q_{01} и Q_{10}, т. е. пропуск сигнала и ложная тревога, могут иметь в зависимости от обстановки совершенно различные последствия. Эти последствия можно выразить некоторыми весовыми коэффициентами, приписываемыми каждому из ошибочных решений и называемых потерями. Мы обозначим потери при ошибочных решениях Q_{01} и Q_{10} соответственно через L_{01} и L_{10}*.

Тогда можно с учетом вероятностей ввести средние ожидаемые потери, так называемый риск

$$r = L_{01}p(Q_{01}) + L_{10}p(Q_{10}) = L_{01}p_1\beta + L_{10}p_0\alpha. \tag{18.11}$$

Понятие риска образует совершенно естественный критерий для выбора гипотез, а именно: правило решения должно минимизировать риск.

Преобразуем (18.11) следующим образом:

$$r = \int_{Y_0} L_{01}p_1w_1(y)\,dy + \int_{Y} L_{10}p_0w_0(y)\,dy =$$
$$= \int_{Y} L_{01}p_1w_1(y)\,dy + \int_{Y_1} [L_{10}p_0w_0(y) - L_{01}p_1w_1(y)]\,dy,$$

так как $Y = Y_0 U Y_1$. ** Но первый интеграл берется по всему множеству Y, т. е. по области всех возможных значений y. Поэтому по условию нормировки вероятностей

$$r = L_{01}p_1 + \int_{Y} [L_{10}p_0w_0(y) - L_{01}p_1w_1(y)]\,dy.$$

---

* Нельзя, разумеется, указать общего правила для выбора значений потерь. Эти величины назначаются произвольно на основе широкой оценки данной ситуации. Иногда приписывают некоторые условные (может быть, отрицательные) потери и правильным решениям, но мы этого делать не будем.

** Обозначение объединения множеств.

Первое слагаемое положительно. Поэтому для минимизации риска нужно выбрать область Y_1 так, чтобы подынтегральная функция была отрицательна, т. е.

$$L_{10}p_0w_0(y) < L_{01}p_1w_1(y),$$

или

$$L_{01}p_1w_1(y)/L_{10}p_0w_0(y) > 1,$$

или, наконец,

$$\Lambda > L_{10}p_0/L_{01}p_1 = \Lambda_0, \qquad (18.12)$$

где $\Lambda = w_1(y)/w_0(y)$ — по-прежнему отношение правдоподобия.

Как видим, правило, (18.12) представляет собой дальнейшее обобщение правила (18. 6) и отличается от последнего множителем, равным отношению потерь. Если считать потери при обоих видах ошибки равными, т. е. положить $L_{01}=L_{10}$, то (18. 12) переходит в (18. 6). Правило минимального риска (18. 12) называют также правилом Бейеса.

В том случае, когда априорные вероятности p_1 и p_0 неизвестны, может оказаться целесообразным применение минимаксного критерия. Идея заключается в том, чтобы минимизировать максимальный возможный риск, откуда и происходит сокращенное название критерия.

Риск (18. 11), получающий наименьшее значение при условии (18. 12), зависит от априорных вероятностей, во-первых, потому, что эти вероятности входят в качестве множителей в выражение для $p(Q_{10})$ и $p(Q_{01})$, а во-вторых, потому, что от априорных вероятностей зависит граница областей Y_0 и Y_1.

При $p_1 \to 0$ (т. е. при $p_0 \to 1$), а равно и при $p_1 \to 1$ (т. е. при $p_0 \to 0$), риск r также стремится к нулю. Это получается формально из выражения

$$r = L_{01}p_1 \int_{Y_0} \omega_1(y)\,dy + L_{10}p_0 \int_{Y_1} \omega_0(y)\,dy,$$

если учесть, например, что при $p_1 \to 1$ размер области Y стремится к нулю, и наоборот. Но и по существу совершенно ясно, что если одна из априорных вероятностей равна единице, а другая нулю, то положение полностью определено, и никакого риска при выборе гипотезы нет.

Итак, наименьший риск имеет максимум при некотором значении $p_1^*=1-p_0^*$, не равном ни нулю, ни единице. Очевидно, что если мы выберем это критическое значение и определим пороговое значение отношения правдоподобия как

$$\Lambda^* = L_{10}p_0^*/L_{01}p_1^*,$$

то действительный риск при любом отличном от p_1^* значении p_1 не превзойдет риска, подсчитанного для $\Lambda_0=\Lambda_0^*$. Таким образом, мы рассчитываем на наихудший случай и выбираем наивыгоднейшие соотношения именно для этого случая. Конечно, если бы в дей-

ствительности p_1 отличалось от p_1^* и если бы это было нам заранее известно, то мы могли бы уменьшить риск, но ведь минимаксный критерий как раз и применяется тогда, когда априорные вероятности известны.

В теории приема (главным образом применительно к локации) употребляются критерии, оперирующие непосредственно с вероятностями ошибок. Эти критерии известны под названием критериев наблюдателей. Подразумевается критерий, которым руководствуется, принимая решение, некоторый гипотетический наблюдатель. Конечно, наблюдатель — не обязательно человек; скорее это автоматическое приемное устройство, действующее согласно определенной инструкции, реализующей критерий данного вида наблюдателя.

Простейший наблюдатель — это так называемый идеальный наблюдатель, или наблюдатель Зигерта—Котельникова. Критерий идеального наблюдателя минимизирует среднюю вероятность ошибок, приписывая пропуску сигнала и ложной тревоге одинаковый вес. Таким образом, минимизируется величина

$$p_{\text{ор}} = p(Q_{01}) + p(Q_{10}) = p_1\beta + p_0\alpha.$$

Подставляя сюда выражения вероятностей ошибок из (18. 9) и действуя, как выше при выводе формулы (18. 12), находим, что правило решения идеального наблюдателя выражается формулой (18. 6).

Наблюдатель Неймана—Пирсона различает оба рода ошибок. Критерий наблюдателя Неймана—Пирсона состоит в том, что при поддержании вероятности ложной тревоги ниже некоторого наперед заданного значения ε минимизируется вероятность пропуска сигнала (или, что то же, максимизируется вероятность правильного обнаружения). При этом априорные вероятности p_0 и p_1 не задаются, так что речь идет об условных вероятностях пропуска сигнала и ложной тревоги.

Итак, для нахождения правила решения согласно критерию Неймана—Пирсона нужно минимизировать

$$\beta = \int_{Y_0} w_1(y)\, dy$$

при дополнительном условии

$$\alpha = \int_{Y_1} w_0(y)\, dy = \varepsilon$$

(очевидно, следует взять наибольшее допустимое значение вероятности ложной тревоги, так как при этом можно получить наименьшую вероятность пропуска сигнала). Мы имеем дело с вариационной задачей с подвижной границей в многомерном, а может быть, и бесконечномерном пространстве Y. Задача будет упрощена, если

мы перейдем от многомерной переменной y к одномерной переменной Λ.

Замена переменных производится на основании равенств

$$\begin{aligned} w_1(y)\,dy &= w_1(\Lambda)\,d\Lambda, \\ w_0(y)\,dy &= w_0(\Lambda)\,d\Lambda. \end{aligned} \tag{18.13}$$

Правые и левые части этих равенств выражают одну и ту же условную вероятность принять сигнал y при условии C_1 (первая строка) или C_0 (вторая строка). Пространство Y при переходе от y к Λ преобразуется в числовую ось значений Λ, на которой значение Λ_0 представляет границу областей соответствующих гипотез. Таким образом, для условных вероятностей ошибок можно записать

$$\begin{aligned} \beta &= \int\limits_{Y_0} w_1(y)\,dy = \int\limits_{0}^{\Lambda_0} w_1(\Lambda)\,d\Lambda, \\ \alpha &= \int\limits_{Y_1} w_0(y)\,dy = \int\limits_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda. \end{aligned} \tag{18.14}$$

Вероятность ложной тревоги α равна заданной величине ε. Следовательно, заданным является и пороговое значение Λ_0, которое определяется из равенства

$$\alpha = \int\limits_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda = \varepsilon. \tag{18.15}$$

Для нахождения относительного минимума вероятности пропуска сигнала β воспользуемся методом множителей Лангранжа.

Вероятности ошибок зависят только от Λ_0. Составим выражение

$$\frac{d\beta}{d\Lambda_0} + k\frac{d\alpha}{d\Lambda_0} = 0.$$

Подставляя значение вероятностей ошибок из (18.14) и выполняя дифференцированные по пределу, получаем

$$w_1(\Lambda_0) - kw_0(\Lambda_0) = 0 \tag{18.16}$$

или

$$w_1(\Lambda_0)/w_0(\Lambda_0) = k.$$

Но на основании (18.13)

$$w_1(\Lambda)/w_0(\Lambda) = w_1(y)/w_0(y) = \Lambda$$

и, следовательно,

$$w_1(\Lambda_0)/w_0(\Lambda_0) = \Lambda_0,$$

так что β имеет относительный минимум при $k=\Lambda_0$, где Λ_0 определяется из (18. 15). Таким образом, требования критерия Неймана—Пирсона удовлетворены.

Заметим, что путем перехода от y к Λ могли бы быть получены и правила для ранее рассмотренных критериев. Для примера проделаем заново вывод правила Бейеса.

Критерий требует в этом случае минимизации риска (18. 11)

$$r = L_{01}p(Q_{01}) + L_{10}p(Q_{10}),$$

причем учитываются априорные вероятности.

Подставляя в (18.11) значения α и β из (18. 14), можем записать

$$r = L_{01}p_1 \int_0^{\Lambda_0} w_1(\Lambda)\,d\Lambda + L_{10}p_0 \int_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda$$

или

$$\frac{r}{L_{01}p_1} = \int_0^{\Lambda_0} w_1(\Lambda)\,d\Lambda + k\int_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda = F(\Lambda_0),$$

где $k = L_{10}p_0/L_{01}p_1$.

Для нахождения минимума запишем

$$\frac{dF(\Lambda_0)}{d\Lambda_0} = w_1(\Lambda_0) - kw_0(\Lambda_0) = 0,$$

что в точности совпадает с (18. 16). Таким образом, взвешенная сумма $F(\Lambda_0)$ имеет минимум при $k=\Lambda_0$, и мы получили правило Бейеса.

Как видим, правила решения для всех рассмотренных до сих пор критериев сводятся к сличению отношения правдоподобия Λ с некоторым пороговым значением Λ_0, и мы можем для удобства обозрения свести наши результаты в следующую таблицу:

Таблица 18.1

Наименование критерия	Условие критерия	Пороговое значение отношения правдоподобия
Идеальный наблюдатель (Зигерт—Котельников)	$\min\,[p_{\text{ср}} = p(Q_{10} + p(Q_{01})]$	$\Lambda_0 = p_0/p_1$
Минимальный риск (Бейес)	$\min\,[r = L_{10}p(Q_{10}) + L_{01}p(Q_{01})]$	$\Lambda_0 = L_{10}p_0/L_{01}p_1$
Минимакс	$\min r_{\max}$	$\Lambda_0 = L_{10}p_0^*/L_{01}p_1^*$, где p_1^* находится из $\partial r/\partial p_1 = 0$
Нейман—Пирсон	$\min\limits_{\alpha \leqslant \varepsilon} \beta$	Λ_0 находится из $\int_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda = \varepsilon$

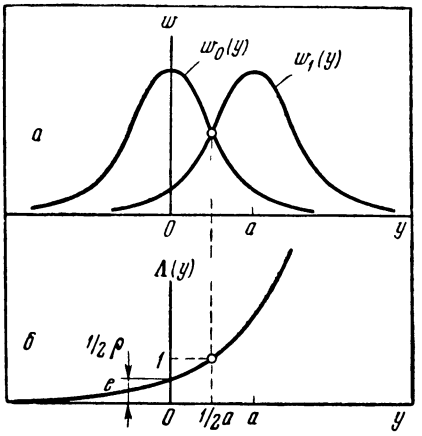


Рис. 54

Для иллюстрации покажем применение этих критериев и правил на простейшем примере.

Пусть задача состоит в обнаружении постоянного сигнала α при аддитивной помехе ξ с нормальным распределением, средним значением, равным нулю, и дисперсией σ^2. Метод приема — однократный отсчет, так что выборку y представляет одномерная величина y. Этот случай был рассмотрен в § 8. Для условных плотностей имеем

$$w_0(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-y^2/2\sigma^2} \tag{18.17}$$

(это — плотность распределения помехи ξ)

$$w_1(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(y-a)^2/2\sigma^2} \tag{18.18}$$

(так как добавление постоянной смещает распределение, не изменяя его характера).

Деля (18. 18) на (18. 17), находим отношение правдоподобия

$$\Lambda(y)=e^{\frac{a}{2\sigma^2}(2y-a)}=e^{\rho\left(\frac{y}{a}-\frac{1}{2}\right)}, \tag{18.19}$$

где $\rho=a^2/\sigma^2$ — отношение сигнал/помеха. Графики функций (18.17)—(18. 19) показаны на рис. 54, *а* и *б*.

Теперь применим различные критерии и соответствующие правила. Пусть априорные вероятности равны, и ошибкам приписываются равные веса. В этом случае пороговое значение (см. (18. 17)) $\Lambda_0=1$ и (18. 19) дает граничное значение $y_0=a/2$. Применяя критерий идеального наблюдателя, требующий минимизации средней ошибки, имеем (см. (18. 6))

$$\Lambda_0=p_0/p_1$$

и для нахождения y_0 нам нужно задаться не только отношением априорных вероятностей, но и отношением сигнал/помеха. Из (18.19) находим общее соотношение

$$y_0 = a\left(\frac{1}{\rho}\ln\Lambda_0 + \frac{1}{2}\right). \qquad (18.20)$$

Пусть $\Lambda_0 = p_0/p_1 = 10$, $\rho = 5$. Тогда

$$y_0 = a\left(\frac{1}{5}\ln 10 + \frac{1}{2}\right) = 0{,}96\,a.$$

Применим теперь критерий минимального риска. Пороговое значение в этом случае (см. (18.12))

$$\Lambda_0 = L_{10}P_0/L_{01}P_1.$$

Пусть $L_{10}/L_{01} = 0{,}01$. При тех же значениях p_0/p_1 и ρ, что и выше, находим

$$y_0 = a\left(\frac{1}{5}\ln\frac{1}{10} + \frac{1}{2}\right) = 0{,}04\,a.$$

Перейдем к критерию Неймана—Пирсона. В нашем одномерном случае удобно сразу записать вместо (18.15)

$$\alpha = \int_{y_0}^{\infty} w_0(y)\,dy = \varepsilon.$$

Подставляя выражение для плотности $\omega_0(y)$ из (18.17), получаем для нахождения y_0 уравнение

$$\frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{2}\rho}\,\frac{y_0}{a}\right)\right] = \varepsilon,$$

где

$$\Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}dz -$$

функция Лапласа. На рис. 55 изображены графики зависимости y_0/a от ε при нескольких значениях ρ. Значение Λ_0 в случае надобности может быть определено на основании формулы (18.19).

Сделаем еще расчет для минимаксного критерия. Чтобы найти критические значения, возьмем выражение для риска

$$r = L_{01}p_1\int_0^{\Lambda_0} w_1(\Lambda)\,d\Lambda + L_{10}p_0\int_{\Lambda_0}^{\infty} w_0(\Lambda)\,d\Lambda \qquad (18.21)$$

и продифференцируем его по p_1, учитывая, что $p_0 = 1 - p_1$ и $\Lambda_0 = L_{10}p_0/L_{01}p_1$. Принимая, кроме того, во внимание равенство

$$\Lambda_0 = \frac{w_1(\Lambda_0)}{w_0(\Lambda_0)},$$

найдем по выполнении вычислений наибольший (минимальный) риск

$$r^* = L_{01} \int_0^{\Lambda_0^*} w_1(\Lambda)\, d\Lambda = L_{10} \int_{\Lambda_0^*}^{\infty} w_0(\Lambda)\, d\Lambda. \qquad (18.22)$$

Это — общая формула: она же служит для нахождения Λ_0^*. В нашем одновременном случае удобнее вернуться к переменной y и записать вместо (18. 21)

$$r = L_{01} \int_{-\infty}^{y_0} w_1(y)\, dy + L_{10} \int_{y_0}^{\infty} w_0(y)\, dy \qquad (18.23)$$

и соответственно

$$r^* = L_{01} \int_{-\infty}^{y_0^*} w_1(y)\, dy = L_{10} \int_{y_0^*}^{\infty} w_0(y)\, dy. \qquad (18.24)$$

Подставляя выражения плотностей из (18. 17) и (18. 18), получаем

$$r^* = \frac{1}{2} L_{01}\left[1 + \Phi\left(\sqrt{\frac{1}{2}\rho}\left(\frac{y_0^*}{a} - 1\right)\right)\right] =$$

$$= \frac{1}{2} L_{10}\left[1 - \Phi\left(\sqrt{\frac{1}{2}\rho}\,\frac{y_0^*}{a}\right)\right].$$

Беря прежние значения $\rho=5$, $L_{10}=1$, $L_{01}=100$ и решая это уравнение (например, методом корней), получаем

$$\frac{y_0^*}{a} = -0{,}11, \quad \Lambda_0^* = e^{\rho\left(\frac{y_0^*}{a} - \frac{1}{2}\right)} = 0{,}047,$$

$$p_1^* = \frac{1}{1 + \frac{L_{01}}{L_{10}} \Lambda_0^*} = 0{,}174$$

и минимальный риск

$$r^* \simeq 0{,}61.$$

Для наглядности на рис. 56 изображена зависимость риска от априорной вероятности p_1, вычисленная по соответствующей формуле (18. 23):

$$r = \frac{1}{2}\left\{L_{01} p_1\left[1 + \Phi\left(\sqrt{\frac{1}{2}\rho}\left(\frac{y_0}{a} - 1\right)\right)\right] + \right.$$

$$\left. + L_{10} p_0\left[1 - \Phi\left(\sqrt{\frac{1}{2}\rho}\left(\frac{y_0}{a}\right)\right)\right]\right\}$$

(напомним еще раз, что y_0 зависит от p_1).

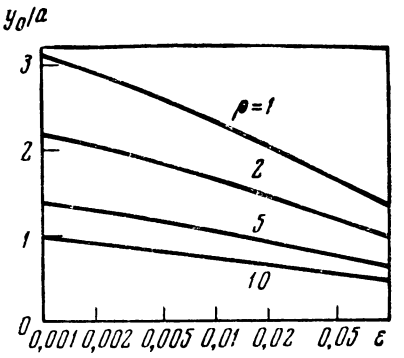


Рис. 55

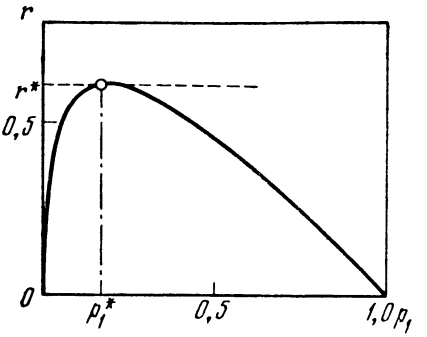


Рис. 56

Таким образом, мы продемонстрировали на одном простом примере все четыре правила, приведенные в табл. 18.1.

Свойства приемника часто описывают посредством так называемой рабочей характеристики. Эта характеристика представляет собой зависимость условной вероятности правильного обнаружения от условной вероятности ложной тревоги. Для этих вероятностей имеем

$$p(H_1/C_0) = p_{10} = \int_{\Lambda_0}^{\infty} w_0(\Lambda)\, d\Lambda = x \quad (=\alpha), \tag{18.25}$$

$$p(H_1/C_1) = p_{11} = \int_{\Lambda_0}^{\infty} w_1(\Lambda)\, d\Lambda = y, \tag{18.26}$$

где x и y — соответственно абсцисса и ордината рабочей характеристики, заданные параметрически (через параметр Λ_0). Условные вероятности p_{10} и p_{11} связаны с безусловными вероятностями решений Q_{10} и Q_{11} следующими соотношениями:

$$p(Q_{10}) = p(C_0)\, p(H_1/C_0) = p_0 p_{10},$$

$$p(Q_{11}) = p(C_1)\, p(H_1/C_1) = p_1 p_{11}.$$

Наклон касательной к рабочей характеристике равен значению Λ_0 в точке касания, в чем легко убедиться, составив производную

$$\frac{dy}{dx} = \frac{dp_{11}/d\Lambda_0}{dp_{10}/d\Lambda_0} = \frac{w_1(\Lambda_0)}{w_0(\Lambda_0)} = \Lambda_0. \tag{18.27}$$

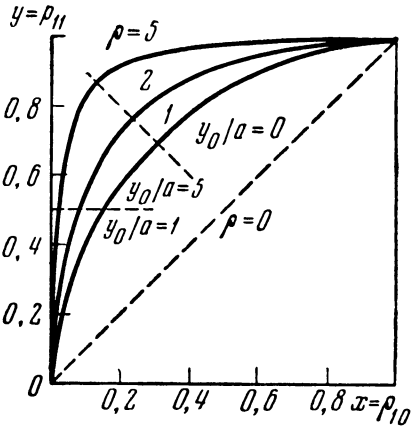


Рис. 57

В одномерном случае рассмотренного выше примера вместо (18. 25) и (18. 26) можно записать

$$p_{10} = \int_{y_0}^{\infty} w_0(y)\,dy = \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{2}\rho}\,\frac{y_0}{a}\right)\right],$$

$$p_{11} = \int_{y_0}^{\infty} w_1(y)\,dy = \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{2}\rho}\left(\frac{y_0}{a} - 1\right)\right)\right].$$

По этим формулам вычислены рабочие характеристики (рис. 57). Легко сообразить, что точки кривых, соответствующие значениям $y_0/a=0$; 0,5; 1 лежат на прямых $x=0{,}5$, $y=1-x$, $y=0{,}5$; эти прямые нанесены на график. Асимптотическим выражением рабочей характеристики при $\rho \to 0$ является прямая $y=x$, также отмеченная на графике.

Общий характер кривых иллюстрирует тот факт, что повышение вероятности правильного обнаружения влечет за собой увеличение вероятности ложной тревоги. Чем больше отношение сигнал/помеха, тем больше вероятность правильного обнаружения, достигаемая при заданной вероятности ложной тревоги. Выбрав на кривых точку, отвечающую поставленным требованиям, находим значение Λ_0 на основании формулы (18. 27).

Как видим, рабочие характеристики особенно удобны применительно к критерию Неймана—Пирсона. Для графического определения Λ_0 при $p_{10} \ll 1$ нужно, конечно, построить рабочие характеристики в подходящем масштабе.

§ 19. Последовательный анализ

В предыдущем изложении статистического подхода к проблеме обнаружения сигнала предполагалось, что в нашем распоряжении имеется совокупность наблюдений (выборка)

$$y = (y_1, y_2, \ldots, y_n).$$

Объем выборки предполагался заданным. Конечно, вероятность (т. е. вероятность принятия правильного решения) возрастает с объемом выборки. Но, с другой стороны, увеличение объема выборки достается нам не даром. Так, если y_i представляет собой последовательные отсчеты, то увеличение их числа сопровождается соответствующим увеличением времени наблюдения.

С этой точки зрения большой интерес представляет метод последовательного анализа, что заданная верность достигается при наименьшем среднем числе наблюдений.

Сущность метода заключается в том, что проверка гипотез производится на каждом этапе наблюдения, т. е. при получении каждого очередного отсчета y_i. При этом пространство наблюдений делится не на две области, как во всех ранее описанных статистических методах, а на три Y_0, Y_1 и Z. Если $y \in Y_0$, то принимается гипотеза H_0; если $y \in Y_1$, то H_1; если же $y \in Z$, то решение не принимается, и наблюдение продолжается. Таким образом, области Y_0 и Y_1 — это области принятия соответственно гипотез H_0 и H_1, а Z — область неопределенности (называемая иногда нулевой зоной). В общем случае границы областей Y_0, Y_1 и Z могут меняться на каждом этапе наблюдения.

Наблюдение продолжается до тех пор, пока вектор y не попадет в Y_0 или Y_1, после чего и принимается соответствующее решение. Мы не будем излагать подробности теории, укажем лишь, что с вероятностью единица процесс конечен; что необходимое число отсчетов n есть случайная величина, среднее значение которой меньше, чем для любого статистического метода, имеющего дело с постоянной выборкой.

При последовательном анализе применяется критерий отношения вероятностей. Правило применения этого критерия состоит в том, что на каждом этапе составляется отношение правдоподобия

$$\Lambda^{(m)} = \frac{w_1(y^{(m)})}{w_0(y^{(m)})},$$

где m — номер последнего отсчета, т. е.

$$y^{(m)} = (y_1, y_2, \ldots, y_m).$$

Величина $\Lambda^{(m)}$ сличается с двумя постоянными пороговыми значениями, A и B $(A > B)$. Если $\Lambda^{(m)} \leqslant B$, то принимается гипотеза H_0; если $\Lambda^{(m)} \geqslant A$, то принимается гипотеза H_1; если же $B < \Lambda^{(m)} < A$, то испытание продолжается, т. е. берется $(m+1)$-й отсчет, находится $\Lambda^{(m+1)}$, с которым поступают точно так же, и

так далее, пока после некоторого n-го отсчета положение не определится. Таким образом, числовая ось значений Λ делится на три отрезка, представляющие три вышеупомянутые области, как показано на рис. 58. На этом же рисунке показан примерный ход наблюдения, т. е. значения $\Lambda^{(m)}$, получаемые на первых и последних этапах наблюдения, заканчивающегося принятием гипотезы H_1.

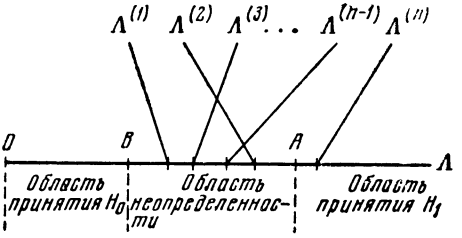


Рис. 58

Если значения y_i статистически независимы, то

$$w(y^{(m)}) = w(y_1)\,w(y_2)\ldots w(y_m) = \prod_{i=1}^{m} w(y_i)$$

и

$$\Lambda^{(m)} = \prod_{i=1}^{m} \frac{w_1(y_1)}{w_0(y_i)} = \prod_{i=1}^{m} \lambda_i.$$

Поэтому удобнее рассматривать не $\Lambda^{(m)}$, а его логарифмы

$$\ln \Lambda^{(m)} = \sum_{i=1}^{m} \ln \frac{w_1(y_i)}{w_0(y_i)} = \sum_{i=1}^{m} z_i$$

и сравнить эту величину соответственно с $\ln A$ и $\ln B$.

Теперь нужно выяснить, как выбираются пороговые значения A и B. Заданными являются условные вероятности α и β ложной тревоги и пропуска сигнала. Дело сводится, таким образом, к установлению зависимости между A и B, с одной стороны, и α и β— с другой.

Запишем условие принятия гипотезы H_1

$$\Lambda = \frac{w_1(y)}{w_0(y)} \geqslant A \tag{19.1}$$

или

$$w_1(y) \geqslant A w_0(y). \tag{19.2}$$

Это условие относится к любой выборке, попадающей в область Y_1. Мы можем поэтому проинтегрировать обе части неравенства (19.2) по этой области

$$\int_{Y_1} w_1(y)\,dy \geqslant A \int_{Y_1} w_0(y)\,dy. \tag{19.3}$$

Но интеграл в правой части есть условная вероятность α ложной тревоги. Интеграл в левой части, выражает условную вероятность правильного обнаружения сигнала, т. е. принятия гипотезы H_1 при наличии сигнала. Но если вероятность пропуска сигнала есть β, то вероятность его правильного обнаружения есть $1-\beta$. Следовательно, (19. 3) можно переписать в виде

$$1-\beta \geqslant A\alpha$$

или

$$A \leqslant \frac{1-\beta}{\alpha}, \tag{19.4}$$

и мы определили, таким образом, верхнюю границу значения A.

Аналогично, беря условие принятия гипотезы H_0,

$$\Lambda = \frac{w_1(y)}{w_0(y)} \leqslant B,$$

переписывая его в виде

$$W_1(y) \leqslant B w_0(y)$$

и интегрируя обе части неравенства по области Y_0, находим

$$B \geqslant \frac{\beta}{1-\alpha}, \tag{19.5}$$

и этим соотношением определена нижняя граница для B [1].

Вообще говоря, как A, так и B являются функциями аргументов α и β. Но оказывается, что в качестве хорошего приближения для A и B можно взять их граничные значения, соответствующие знакам равенства в (19. 4) и (19. 5).

Для доказательства положим, что вместо истинных значений A и B мы взяли граничные значения

$$A' = \frac{1-\beta}{\alpha}, \quad B' = \frac{\beta}{1-\alpha}.$$

Но $A \leqslant A'$, $B \geqslant B'$. Поэтому при таком выборе границ мы получим измененные значения вероятностей α и β. Отметив эти измененные значения штрихами, можем записать

$$\frac{1-\beta'}{\alpha'} \geqslant A' = \frac{1-\beta}{\alpha}, \quad \frac{\beta'}{1-\alpha'} \leqslant B' = \frac{\beta}{1-\alpha}.$$

Но α и β — обычно малые по сравнению с единицей величины, так что изменениями заданных вероятностей вследствие замены истинных значений A и B их граничными значениями A' и B'

---

[1] Формулы (19. 4) и (19. 5) относятся к значениям $\alpha < 0{,}5$, $\beta < 0{,}5$. Но описанным методом можно получить и более общие формулы: $A \leqslant \max((1-\beta)/\alpha, \beta/(1-\alpha))$, $\beta \geqslant \min(\beta/(1-\alpha), (1-\beta)/\alpha)$.

можно пренебречь во всех случаях, представляющих практический интерес.

Выведем теперь выражение для среднего числа отсчетов. Процедура последовательного анализа с применением критерия отношения вероятностей состоит, как уже говорилось, в том, что на каждом этапе, т. е. при получении каждого очередного m-го отсчета, составляется величина

$$\ln \Lambda^{(m)} = \sum_{i=1}^{m} z_i, \tag{19.6}$$

которая сравнивается с $\ln A$ и $\ln B$. Испытание прекращается на n-м отсчете, если $\ln \Lambda^{(n)}$ выходит за указанные пределы, т. е. если

$$\ln \Lambda^{(n)} \geqslant \ln A \quad \text{или} \quad \ln \Lambda^{(n)} \leqslant \ln B. \tag{19.7}$$

Пренебрежем влиянием перехода за граничные значения, т. е. будем полагать, что при n-м отсчете $\ln \Lambda^{(n)}$ принимает в точности одно из граничных значений $\ln A$ или $\ln B$. Тогда среднее значение n может быть получено из следующего рассуждения.

Представим сумму Σz_i случайного числа случайных величин в виде

$$\sum_{i=1}^{n} z_i = n \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Второй сомножитель есть среднее арифметическое.

Усредним обе части равенства, полагая, что среднее произведения равно произведению средних; получаем [1]

$$M \sum_{i=1}^{n} z_i = Mn \cdot Mz. \tag{19.8}$$

С другой стороны, сумма Σz_i, как мы условились, принимает одно из граничных значений, а именно, $\ln A$ или $\ln B$. При отсутствии сигнала верхняя граница достигается с вероятностью α, а нижняя — с вероятностью $1-\beta$. Таким образом, среднее значение суммы есть

$$M \sum_{i=1}^{n} z_i = \alpha \ln A + (1-\beta) \ln B. \tag{19.9}$$

Приравнивая (19.8) и (19.9), находим искомое среднее значение числа отсчетов

$$Mn = \frac{\alpha \ln A + (1-\beta) \ln B}{Mz}. \tag{19.10}$$

[1] Условие (достаточно широкое), при котором этот результат справедлив, а также доказательство более общей теоремы (для случая, когда $Mz_i = a_i$) приведены в статье А. Н. Колмогорова и Ю. В. Прохорова «О суммах случайного числа случайных слагаемых», УМН, 1949, № 4.

Это равенство приближенное, так как получено путем замены знака неравенства в (19. 7) знаком равенства, т. е. в результате пренебрежения влиянием перехода за граничные значения.

Нам остается рассмотреть пример. Пусть по-прежнему речь идет об обнаружении постоянного сигнала a при наличии гауссовой аддитивной некоррелированной помехи с дисперсией σ^2 и нулевым средним значением. Тогда имеем

$$w_0(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2},$$

$$w_1(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-a)^2/2\sigma^2},$$

$$\lambda_i = \frac{w_1(y_i)}{w_0(y_i)} = e^{-(a^2+2ay_i)/2\sigma^2}$$

или

$$\lambda_i = e^{\rho_0\left(\frac{y_i}{a} - \frac{1}{2}\right)},$$

где введено обозначение $\rho_0 = a^2/\sigma^2$. Для выборки из m отсчетов имеем

$$\Lambda^{(m)} = \prod_{i=1}^{m} \lambda_i = e^{\rho_0\left(\frac{1}{a}\sum y_i - \frac{m}{2}\right)}. \qquad (19.11)$$

Найдем среднее число отсчетов, требуемое для принятия решения. У нас

$$z_i = \ln\lambda_1 = \rho_0\left(\frac{y_i}{a} - \frac{1}{2}\right).$$

При отсутствии сигнала средний отсчет равен нулю. Поэтому

$$Mz = \frac{1}{2}\rho_0.$$

Подставляя в (19. 10) это значение, а также приближенные значения $A \approx 1/\alpha$, $B \approx \beta$, получаем

$$M(n\rho_0) = 2\left[(1-\beta)\ln\frac{1}{\beta} - \alpha\ln\frac{1}{\alpha}\right]. \qquad (19.12)$$

При $\alpha = \beta = p$ формула упрощается и принимает вид

$$M(n\rho_0) = 2(1-2p)\ln\frac{1}{p}. \qquad (19.12а)$$

Пусть, например, $\alpha = \beta = 0{,}01$. Тогда

$$M(n\rho_0) = 2 \cdot 2{,}3 \cdot 0{,}98 \cdot 2 \approx 9.$$

Сопоставим этот результат с числом повторений n_0 при обычном методе накопления. При применении этого метода мы сличаем

сумму n_0 отсчетов с пороговым значением $n_0 a/2$. При этом вероятность ошибки

$$\alpha = \beta = \frac{1}{2}\left[1 - \Phi\left(\sqrt{\frac{1}{8} n_0 \rho_0}\right)\right].$$

Выбирая $\alpha = \beta = 0{,}01$, т. е. $\Phi\left(\sqrt{n_0\rho_0/8}\right) = 0{,}98$, находим по таблицам $\sqrt{n_0\rho_0/8} = 1{,}64$, откуда $n_0\rho_0 = 21{,}6$, что в 2,4 раза больше, чем среднее значение для последовательного анализа.

Чтобы сделать сопоставление более наглядным, преобразуем условие принятия решения для последовательного анализа. В исходном виде это условие есть

$$\begin{aligned} &\Lambda^{(n)} > A \text{ или} \\ &\Lambda^{(n)} < B. \end{aligned} \tag{19.13}$$

Из (19.11) получаем

$$\ln \Lambda^{(n)} = \rho_0\left(\frac{1}{a}\sum y_i - \frac{n}{2}\right)$$

или

$$a\left(\frac{1}{n\rho_0}\ln \Lambda^{(n)} + \frac{1}{2}\right) = \frac{1}{n}\sum y_i = \eta_n.$$

Таким образом, условие (19.13) можно представить в виде

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n} y_i \begin{cases} > a\left(\frac{1}{n\rho_0}\ln A + \frac{1}{2}\right) = u, \\ < a\left(\frac{1}{n\rho_0}\ln B + \frac{1}{2}\right) = v. \end{cases} \tag{19.14}$$

Для метода накопления решение принимается при

$$\eta_{n_0} = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i \begin{cases} > \frac{1}{2}a \\ < \frac{1}{2}a. \end{cases} \text{ или} \tag{19.15}$$

Как видно из (19.14), пороговые значения для случайной величины $\eta_n = \frac{1}{n}\sum y_i$ (которая есть среднее арифметическое значение n отсчетов) оказываются переменными: они зависят от числа отсчетов, асимптотически приближаясь к $a/2$ при $n \to \infty$. Это показано на графике рис. 59, на котором построены кривые

$$\frac{u}{a} = \frac{1}{n\rho_0}\ln A + \frac{1}{2} \quad \text{и} \quad \frac{v}{a} = \frac{1}{n\rho_0}\ln B + \frac{1}{2}$$

при $\alpha = \beta = 0{,}01$. Заштрихованная область между u/a и v/a есть область неопределенности. На этом же рисунке пунктиром нане-

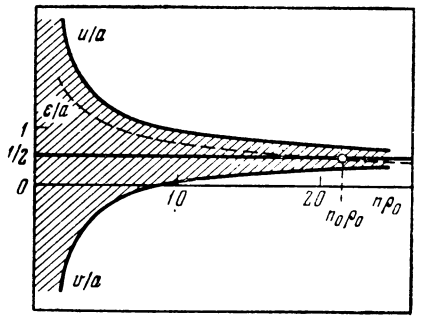


Рис. 59

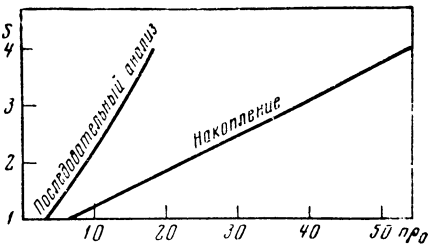


Рис. 60

сена кривая значений, которых с вероятностью 0,01 не превосходит случайная величина

$$\eta_n = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Из соотношения

$$p\{\eta_n < \varepsilon\} = \frac{1}{2}\left[1 + \Phi\left(\sqrt{\frac{n\rho_0}{2}} \cdot \frac{\varepsilon}{a}\right)\right] = 0{,}01$$

находим

$$\varepsilon/a = 2{,}33/\sqrt{n\rho_0}.$$

Пересечение этой кривой с горизонтальной прямой с ординатой $^1/_2$ дает требуемое число повторений n_0 для простого накопления.

На рис. 60 дано прямое сравнение числа отсчетов, требуемого при последовательном анализе (в среднем) и при простом накоплении для получения заданных значений вероятности правильного решения. Предполагается, что $\alpha = \beta$, так что вероятность ошибки подсчитана по формуле (19.12а). Интересно сравнить асимптотические выражения для последовательного анализа и накопления. При накоплении вероятность ошибки

$$p = \frac{1}{2}\left[1 - \Phi\sqrt{\frac{1}{8}n_0\rho_0}\right] \sim \frac{1}{\sqrt{\frac{\pi}{2}n_0\rho_0}}\, e^{-\frac{1}{8}|n_0\rho_0},$$

откуда верность

$$S = \lg\frac{1}{p} \sim 0{,}098 + \frac{1}{2}\lg(n_0\rho_0) + 0{,}054 n_0\rho_0.$$

При высокой верности можно пренебречь первыми двумя членами; это дает

$$n_0\rho_0 \sim 18{,}4S \quad (= 2{,}3 \cdot 8S). \tag{19.16}$$

Из формулы (19. 12а) получаем для высокой верности

$$M(n\rho_0) \sim 4{,}6S \quad (= 2{,}3 \cdot 2S). \tag{19.17}$$

Сопоставляя (19. 16) и (19. 17), видим, что в пределе верность, достигаемая посредством последовательного анализа, в четыре раза больше, чем при простом накоплении.

Идея последовательного анализа находит непосредственное практическое применение в системах передачи сигналов с переспросом, о которых говорится в следующем параграфе.

§ 20. Передача с переспросом

При осуществлении системы передачи информации иногда оказывается возможным организовать, кроме прямого канала (по которому информация передается от передатчика к приемнику), еще и обратный канал (по которому информация может передаваться от приемника к передатчику). Такой обратный канал может быть использован для повышения верности передачи.

Прежде всего заметим, что прямой и обратный каналы могут существенно различаться по своим свойствам. Приведем пример, когда прямой канал соединяет маломощный передатчик, установленный на борту какого-либо летательного аппарата, и наземный стационарный приемник. Бортовой передатчик передает важную информацию, но имеет малую мощность, антенны ограниченного размера и т. п. При этих условиях передача может оказаться недостаточно надежной. Обратный же канал может быть образован между бортовым приемником и наземным передатчиком, для которого энергетические ресурсы, а также веса и габариты аппаратуры практически неограниченны. Поэтому обратный канал при описанной ситуации может быть значительно более надежным, и при исследованиях системы с обратной связью обычно полагают, что обратный канал действует безошибочно [1].

Существует большое число способов использования обратного канала. Системы, в которых применяется передача по обратному каналу, носят общее название систем с обратной связью. Краткий обзор этих систем дан в следующем параграфе. Здесь же мы рассмотрим в простейшем варианте так называемую систему с переспросом.

Сущность этой системы состоит в том, что в сомнительных случаях приемник посылает по обратному каналу специальный сигнал переспроса. По этому сигналу передатчик повторяет передачу того элемента сигнала, к которому относится переспрос. Можно различать систему с ограниченным переспросом, когда сомнительный элемент сигнала повторяется не более чем r раз, и систему с неограниченным переспросом, когда сигнал повторяется столько раз, сколько потребуется, чтобы приемник мог вывести о сигнале окончательное решение.

[1] Это допущение обосновано еще и тем, что по обратному каналу передаются только сигналы переспроса, т. е. значительно меньшее количество информации, нежели по прямому каналу. Поэтому условия работы обратного канала более благоприятны.

В системе с переспросом приемник имеет решающее устройство с областью неопределенности, так что сомнительным признается сигнал, попавший не в одну из собственных областей, а в область неопределенности. Сущность процесса повторений в результате переспросов сводится к описанному в предыдущем параграфе последовательному анализу.

Выведем соотношения, показывающие повышение верности, достигаемое применением переспроса.

Обозначим собственные области сигналов через Y_k ($k=1, 2, \ldots, N$, где N — число различных сигналов), а область неопределенности — через Z. В общем случае все эти области могут меняться при каждом очередном повторении сигнала. В частности, при ограниченном переспросе на последнем этапе должно быть принято окончательное решение; поэтому при последнем повторении решающее устройство перестраивается так, что область неопределенности Z отсутствует и области Y_k заполняют все пространство сигналов.

Найдем вероятность правильного приема сигнала y_k после r-кратного повторения. Сигнал будет принят правильно, если:

1) при первой же передаче $y_k \in Y_k$;

2) при первой передаче $y_k \in Z$, а при повторной $y_k \in Y$;

3) при первой и повторной передаче $y_k \in Z$, а при передаче в третий раз $y_k \in Y$ и так далее.

Таким образом, вероятность правильного приема равна сумме вероятностей указанных несовместимых событий, т. е.

$$p_{\text{пр}} = p(y_k^{(0)} \in Y_k^0) + p(y_k^{(0)} \in Z^{(0)},\ y_k^{(1)} \in Y_k^{(1)}) + \ldots + p(y_k^{(0)} \in Z^{(0)},\ y_k^{(1)} \in Z^{(1)}, \ldots, y_k^{(r-1)} \in Z^{(r-1)},\ y_k^{(r)} \in Y_k^{(r)}). \qquad (20.1)$$

Здесь верхний индекс указывает номер повторения (или переспроса); $y_k^{(0)}$ означает сигнал, переданный в первый раз.

Все последующее относится к k-му сигналу. Поэтому нижний индекс при y и Y в дальнейшем опущен.

Если в системе отсутствует память [1] (т. е. если при приеме очередного сигнала ранее принятые сигналы не учитываются), то события, состоящие в попадании сигнала в ту или иную область при очередном повторении, независимы. При этом совместные вероятности в (20. 1) могут быть заменены произведениями простых вероятностей

$$p_{\text{пр}}^{(r)} = p(y^{(0)} \in Y^{(0)} + p(y^{(0)} \in Z^{(0)})\, p(y^{(1)} \in Y^{(1)}) + \ldots + p(y^{(0)} \in Z^{(0)})\, p(y^{(1)} \in Z^{(1)}) \ldots p(y^{(r-1)} \in Z^{(r-1)})\, p(y^{(r)} \in Y^{(r)}). \qquad (20.2)$$

[1] Это допущение упрощает классическую схему последовательного анализа (§ 19), в котором выборка включает все предшествующие наблюдения.

Эта формула годится и в том случае, когда при каждом повторении передаваемый (а следовательно, и принимаемый) сигнал определенным образом изменяется. Если же передаваемый сигнал при повторениях остается неизменным, то принимаемый сигнал представляет собой все время одну и ту же случайную величину.

Положим также, что области Y и Z остаются неизменными на протяжении всего процесса, за исключением последнего повторения, для которого область отсутствует. Обозначая

$$Y^{(0)} = Y^{(1)} = \ldots = Y^{(r-1)} = Y', \quad Y^{(r)} = Y,$$

$$Z^{(0)} = Z^{(1)} = \ldots = Z^{(r-1)} = Z,$$

получаем вместо (20. 2)

$$p_{\text{пр}}^{(r)} = p(y \in Y')\,[1 + p(y \in Z) + p^2(y \in Z) + \ldots + p^{r-1}(y \in Z)] + p(y \in Y)\,p^r(y \in Z)$$

или

$$p_{\text{пр}}^{(r)} = p(y \in Y')\,\frac{1 - p^r(y \in Z)}{1 - p(y \in Z)} + p(y \in Y)\,p^r(y \in Z). \qquad (20.3)$$

При неограниченном переспросе, т. е. при $r \to \infty$, это дает

$$p_{\text{пр}}^{(\infty)} = \frac{p(y \in Y')}{1 - p(y \in Z)}. \qquad (20.4)$$

На практике часто встречается случай однократного переспроса ($r=1$); для этого случая на основании (20. 3) имеем

$$p_{\text{пр}}^{(1)} = p(y \in Y') + p(y \in Z)\,p(y \in Y). \qquad (20.5)$$

Из формулы (20. 3)—(20. 5) уже видно, что верность, выражаемая вероятностью правильного приема, может быть повышена путем переспроса при надлежащем выборе областей. Так, например, формула (20. 4) показывает, что верность правильного приема при неограниченном переспросе может быть сделана сколь угодно близкой к единице путем увеличения области неопределенности Z (при соответствующем уменьшении областей Y'). Влияние вероятности переспроса может быть сделано особенно наглядным, если ввести

$$p_{\text{пр}}^{(0)} = p(y \in Y), \qquad (20.6)$$

т. е. вероятность правильного приема без переспроса (при однократной передаче), и составить следующее отношение, характеризующее увеличение вероятности правильного приема с возрастанием числа переспросов

$$\mu(r) = \frac{p_{\text{пр}}^{(r+1)} - p_{\text{пр}}^{(0)}}{p_{\text{пр}}^{(1)} - p_{\text{пр}}^{(0)}} = \frac{1 - p^{r+1}(y \in Z)}{1 - p(y \in Z)}. \qquad (20.7)$$

При $r \to \infty$

$$\mu \to \frac{1}{1 - p(y \in Z)}. \qquad (20.7\text{а})$$

Строение отношения в левой части подобрано так, что отношение это оказывается зависящим только от вероятности переспроса $p\,(y \in Z)$. Как видим, числитель (20. 7) стремится к единице при увеличении числа переспросов, а знаменатель стремится к нулю при увеличении вероятности переспроса, так что отношение μ может быть сделано сколь угодно большим. Разумеется, повышение верности дается нам недаром, так как при повторной передаче затрачиваются дополнительная энергия и время.

Пусть при первой передаче сигнал имеет энергию E_0. Если принятый сигнал $y \in Z$, то он повторяется с энергией $E^{(1)}$; если он снова попадает в область неопределенности, то повторяется с энергией $E^{(2)}$ и т. д. Математическое ожидание суммарной энергии при r повторениях можно записать в виде

$$ME(r) = E_0 + E^{(1)} p\,(y^{(0)} \in Z) + E^{(2)} p\,(y^{(0)} \in Z,$$
$$y^{(1)} \in Z) + \ldots + E^{(r)} p\,(y^{(0)} \in Z,\ y^{(1)} \in Z, \ldots, y^{(r-1)} \in Z).$$

Если сигнал остается при повторениях неизменным, т. е.

$$y^{(0)} = y^{(1)} = \ldots = y^{(r-1)} = y, \quad E^{(0)} = E^{(1)} = \ldots = E^{(r)} = E_0,$$

то

$$ME(r) = E_0 [1 + p\,(y \in Z) + p^2\,(y \in Z) + \ldots + p^r\,(y \in Z)]$$

или (см. (20. 7))

$$ME(r) = E_0 \frac{1 - p^{r+1}\,(y \in Z)}{1 - p\,(y \in Z)} = \mu E_0. \tag{20. 8}$$

Соотношение (20. 8) раскрывает смысл величины μ — это, как видим, не что иное, как среднее число передач, т. е. среднее число повторений плюс единица.

Итак, средняя энергия, затрачиваемая при передаче с ограниченным переспросом, возрастает прямо пропорционально отношению μ, характеризующему повышение верности. Соответственно возрастает и затрачиваемое на передачу время.

Рассмотрим пример. Пусть, как в предыдущем параграфе, передается постоянный сигнал a с пассивной паузой при аддитивной гауссовой помехе ξ мощностью σ^2. Тогда имеем $N=2$, т. е. всего два сигнала, $y_1 = \xi$, $y_2 = a + \xi$.

Распределения для этих сигналов равны соответственно

$$w_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}, \quad w_2(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-a)^2/2\sigma^2}.$$

Разметим области, как показано на рис. 61. Будем иметь

$$p\,(y \in Z) = p\left\{\frac{1}{2}(1-\alpha) < \frac{y}{a} < \frac{1}{2}(1+\alpha)\right\},$$

$$p\,(y_1 \in Y_1) = p\left\{-\infty < \frac{y_1}{a} < \frac{1}{2}\right\},$$

$$p(y_2 \in Y_2) = p\left\{\frac{1}{2} < \frac{y_2}{a} < \infty\right\},$$

$$p(y_1 \in Y_1') = p\left\{-\infty < \frac{y_1}{a} < \frac{1}{2} - \alpha\right\},$$

$$p(y_2 \in Y_2') = p\left\{\frac{1}{2}(1+\alpha) < \frac{y_2}{a} < \infty\right\}.$$

Вычисляя вероятности по соответствующим распределениям, получаем

$$p(y \in Z) = \frac{1}{2}\{\Phi[(\alpha+1)z_0] + \Phi[(\alpha-1)z_0]\},$$

$$p(y_1 \in Y_1) = p(y_2 \in Y_2) = \frac{1}{2}[1 + \Phi(z_0)], \tag{20.9}$$

$$p(y_1 \in Y_1') = p(y_2 \in Y_2') = \frac{1}{2}\{1 - \Phi[(\alpha-1)z_0]\},$$

где $z_0 = \sqrt{\rho_0/8}$.

Для системы с неограниченным переспросом находим, подставляя (20.9) в (20.4) и (20.7а)

$$p_{\text{пр}} = \frac{1}{1 + \frac{1-\Phi(x_2)}{1-\Phi(x_1)}}, \tag{20.10}$$

$$\mu = \frac{2}{[1-\Phi(x_1)] + [1-\Phi(x_2)]}, \tag{20.11}$$

где обозначено для краткости

$$x_1 = (\alpha-1)z_0, \quad x_2 = (\alpha+1)z_0.$$

Вместо вероятности правильного приема $p_{\text{пр}}$ можно рассматривать вероятность ошибки $p_{\text{ош}}$:

$$p_{\text{ош}} = 1 - p_{\text{пр}} = \frac{1}{1 + \frac{1-\Phi(x_1)}{1-\Phi(x_2)}}. \tag{20.12}$$

При высокой верности, т. е. при $p_{\text{ош}} \ll 1$,

$$p_{\text{ош}} \approx \frac{1-\Phi(x_2)}{1-\Phi(x_1)}. \tag{20.13}$$

Эффективность системы с переспросом можно оценить, сравнив ее с системой с простым накоплением, при котором на решающее устройство поступает сумма n экземпляров сигнала. В такой системе (см. § 9) вероятность ошибки выражается формулой

$$p_{\text{н}} = \frac{1}{2}[1 - \Phi(\sqrt{n}\, z_0)]. \tag{20.14}$$

На рис. 62, *а* построена зависимость верности

$$S = \lg \frac{1}{p}$$

от среднего числа передач μ для фиксированного значения $z_0 = 1{,}12$ ($\rho_0 = 10$). На этом же графике показана кривая верности для системы с накоплением, причем n отложено по оси абсцисс в той же шкале, что и μ. На рис. 62, *б* представлена зависимость между μ и α при $\rho_0 = 10$; $z_0 = 1{,}12$, а функциональная шкала для α построена под рис. 62, *а*.

График показывает, что рассматриваемая нами простейшая система с переспросом не всегда выгоднее системы с накоплением. Для различных значений z_0 получаются различные соотношения.

Это наводит на мысль построить комбинированную систему, в которой накопление сочетается с переспросом. Сигнал накапливается путем n_1 повторений, а затем производится переспрос. Таким образом, во все предыдущие соотношения войдет вместо z_0 аргумент

$$z = \sqrt{n_1}\, z_0 = \sqrt{n_1 \rho_0 / 8}.$$

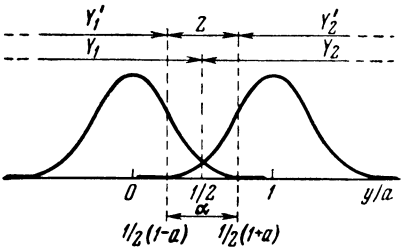


Рис. 61

Величина ρ_0 — отношение сигнал/помеха — предполагается заданной. Параметрами, характеризующими комбинированную систему, являются величины α и n_1 (или α и z). Эти параметры нужно подобрать так, чтобы комбинированная система была в определенном смысле оптимальной.

Естественным является требование, чтобы при заданной энергии сигнала достигалась наивысшая верность, или, наоборот, чтобы заданная верность достигалась при наименьшей затрате энергии. Комбинированную систему, удовлетворяющую этому условию, мы назовем оптимальной.

Если энергия однократно передаваемого сигнала обозначена через E_0, то в системе с накоплением полная энергия равна nE_0, в простой системе с неограниченным переспросом полная энергия (в среднем) равна μE_0, а в комбинированной системе полная энергия равна $\mu n_1 E_0$. Эту последнюю величину и нужно минимизировать при заданной верности для нахождения оптимальной системы. Подробности вычисления вынесены в Добавление VI, а результат представлен в графической форме на рис. 63. По оси ординат отложена верность. По оси абсцисс отложена величина относительной полной энергии, т. е.

$$\rho_0 \frac{E}{E_0} = \mu n_1 \rho_0 = 8z^2.$$

На том же графике построена для сравнения кривая верности для простого накопления *3*. Аргумент этой зависимости

$$\rho_0 \frac{E}{E_0} = n\rho_0$$

отложен в той же шкале по оси абсцисс. Штрихом обозначена кривая рис. 62, *а*, перенесенная на рис. 63 с соответствующим изменением масштаба абсцисс.

Все формулы и графики годятся и в том случае, когда приемник выполняет интегрирование на интервале T. В этом случае вместо ρ_0 нужно подставить $\rho = 2FT_{\rho_0}$.

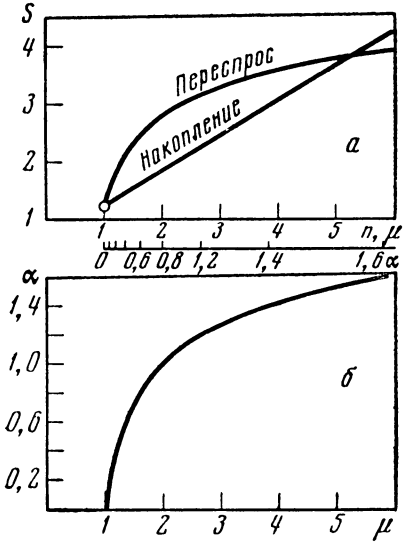


Рис. 62

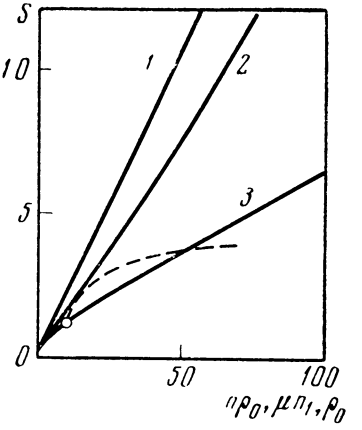


Рис. 63

Графики показывают значительное преимущество оптимальной системы (*2*) по сравнению с системой с простым накоплением. Так, например, при $\rho_0 E/E_0 \approx 70$ верность возрастает с 4,8 до 10,8, т. е. вероятность ошибки уменьшается в миллион раз.

Нужно, однако, заметить, что рассмотренная оптимальная система все же уступает системе, в которой осуществляется классическая схема последовательного анализа (*2*). Преимущество последней состоит в том, что она использует все принятые сигналы, т. е. является системой с памятью, тогда как рассматриваемая здесь оптимальная система памяти не имеет. Отсутствие памяти, конечно, упрощает систему, но и потеря верности при заданной энергии оказывается значительной. Для сравнения на том же рис. 63 построен график верности для классической схемы последовательного анализа по формуле (см. § 19)

$$n\rho_0 = 2(1 - 2p)\ln\frac{1}{P},$$

которая при высокой верности переходит асимптотически в $np_0 \sim$ $\sim 4{,}6S$. График рис. 63 повторяет в более широком интервале сопоставление, данное на рис. 60.

§ 21. Системы с обратной связью (обзор)

Как уже говорилось, обратная связь в системах передачи информации может быть использована по-разному. Для систематического обзора разновидностей систем с обратной связью удобно начать с рассмотрения общей схемы рис. 64. На этой схеме информация от источника, имеющая произвольный характер, по-

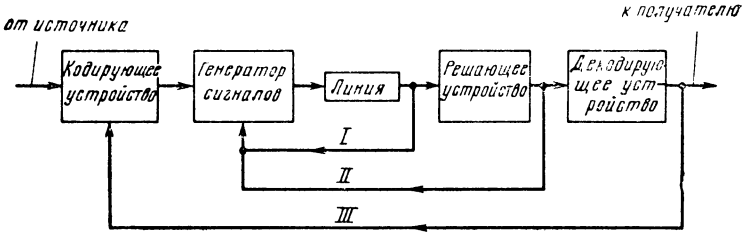


Рис. 64

ступает в кодирующее устройство, где она преобразуется в дискретную форму, т. е. в последовательность дискретных символов. Далее следует генератор сигналов, вырабатывающий сигналы с теми или иными физическими различиями. Набор различных сигналов на выходе генератора сигналов находится в определенном соответствии с набором символов (алфавитом) кода. В частности, генератор сигналов может представлять собой совокупность генератора переносчика и модулятора. Сигналы передаются по линии, где на них воздействует помеха. На приемной стороне после того или иного преобразователя (на схеме не показан; он может осуществлять фильтрацию, накопление и т. п. операции) сигналы поступают на решающее устройство, функция которого состоит в том, чтобы определить, какой из возможных сигналов в действительности был передан. После этого сигналы поступают в декодирующее устройство, восстанавливающее информацию в ее первоначальной форме.

Обратный канал может соединять различные точки схемы, т. е. обратной связью может быть охвачена та или иная часть системы рис. 64. Три главнейших случая занумерованы на рисунке римскими цифрами.

I. Обратная связь охватывает только линию. При этом по обратному каналу передаются сведения о сигнале, поступившем на приемный конец до того, как об этом сигнале принято какое-либо решение (predecision feedback) [1].

---

[1] Здесь ниже в скобках приводится терминология, принятая в американской литературе.

II. Обратная связь включена после решающего устройства и передает по обратному каналу сведения о принятых решениях.

III. Обратная связь охватывает всю систему, включая кодирующее и декодирующее устройства.

Таким образом, случаи II и III объединяются по тому признаку, что на приемной стороне обратная связь подключена после решающего устройства (postdecision feedback).

В случае I обратная связь контролирует линию, или, точнее говоря, часть системы, включающую линию и элементы оконечной аппаратуры, в которых совершаются те или иные непрерывные преобразования. Роль обратной связи при этом состоит в том, чтобы при тех или иных случайных изменениях параметров этой части системы производить соответствующие изменения в характере передаваемых сигналов. Так, например, при изменениях уровня аддитивной помехи может потребоваться соответствующее изменение мощности сигнала; при изменениях частотных характеристик должен соответственно меняться спектральный состав сигналов; можно изменять темп передачи или даже вовсе прекращать на время передачу, если на линии создаются особо неблагоприятные условия. При этом на передающей стороне должны предусматриваться соответствующие органы воздействия на источник сигналов (регуляторы, частотные корректоры и т. п.), управляемые сигналами, поступающими по обратному каналу. Примером специального применения обратной связи такого рода может служить система метеорной связи, в которой сеанс передачи начинается и кончается по сигналам с приемной стороны, где ведется постоянное наблюдение за изменяющимися условиями связи.

Таким образом, случай I можно назвать обратной связью, контролирующей условия передачи, или, короче, контролирующей линию. Тогда случаи II и III можно назвать обратной связью, контролирующей сигналы.

В системах с такого рода обратной связью на основании поступающих по обратному каналу сведений о принятых сигналах производится либо повторная передача сомнительных сигналов, либо передача данных о необходимых исправлениях. При этом в зависимости от способа действия системы с обратной связью, контролирующей сигналы, делятся на системы со сравнением и с переспросом.

В системах со сравнением (information feedback system) приемник передает по обратному каналу сведения о том, какие сигналы приняты. Передатчик сличает эти сведения с тем, что фактически было передано. При наличии расхождений передатчик передает неправильно принятые сигналы вновь или сообщает необходимые поправки, уведомляя об этом приемник специальным сигналом. В системах с переспросом (decision feedback system) приемник сам выделяет сомнительные или заведомо неверно принятые сигналы и посылает по обратному каналу сигнал переспроса. По этому сигналу передатчик повторяет сигналы, к которым отно-

сится переспрос. Пример простейшей системы с переспросом был рассмотрен в предыдущем параграфе.

Таким образом, различие между системами со сравнением и системами с переспросом сводится к тому, что в первых активная роль принадлежит передатчику, а во вторых — приемнику. В системах со сравнением решение о необходимости повторения или исправления тех или иных сигналов применяется на передающей стороне на основании информации о принятых сигналах. В системах с переспросом это решение принимается на приемной стороне, а передатчик лишь пассивно выполняет требования со стороны приемника.

Нагрузка обратного канала в системах с переспросом обычно значительно меньше, нежели в системах со сравнением, так как в первых передаются сигналы переспроса только в сомнительных случаях, тогда как во вторых передаются сведения о всех принятых сигналах. Даже в том случае, когда в системе с переспросом дается обратный сигнал, подтверждающий или отвергающий прием каждого переданного сигнала, объем информации, передаваемой по обратному каналу, невелик, так как дело сводится к передаче двоичного обратного сигнала («да» или «нет»).

В случае, когда в обратном канале возможны ошибки, возникает специфический для систем с переспросом недостаток, состоящий в пропадании элементов сигнала или возникновении лишних. Если послан сигнал переспроса, требующий повторения и если в результате ошибки этот сигнал воспринят на передающей стороне, как подтверждение правильности приема, то передатчик продолжает передачу, тогда как приемник ожидает повторения. Если же, наоборот, по обратному каналу послан сигнал подтверждения, который ошибочно воспринят как сигнал переспроса, то передатчик повторяет предыдущий элемент сигнала, тогда как приемник воспринимает его как следующий элемент последовательности. Если между элементами сигнала существуют достаточно тесные статистические взаимосвязи, то такого рода ошибки могут быть обнаружены. В противном же случае они остаются незамеченными и могут совершенно исказить всю принимаемую последовательность сигналов. Одно из средств борьбы с этим явлением состоит во включении в передаваемую последовательность контрольных сигналов, разделяющих последовательность на достаточно короткие отрезки. Можно также передавать контрольные сигналы, сопровождающие всякое повторение, с целью отличить его от основной последовательности.

Строение сигналов и устройство приемника в системе с переспросом должны быть таковы, чтобы ошибка при передаче по прямому каналу могла быть обнаружена. Для этого применяются, в частности, специальные коды, обнаруживающие ошибку. Подробнее о корректирующих кодах говорится в следующих параграфах. Здесь же упоминаем лишь для примера две простые системы, уже давно применяемые на практике. Первая из них исполь-

зует так называемый «код 3 из 7». Это — двоичный семизначный код, все кодовые комбинации которого имеют по три единицы. Всего таких комбинаций $C_7^3=35$. Приемник делает соответствующую проверку, признает принятый сигнал неверным и требует его повторения всякий раз, когда в кодовой комбинации число единиц оказывается отличным от трех. Во второй системе применяется код с произвольным основанием, и после каждой кодовой комбинации передается в качестве контрольного числа сумма цифр в этой кодовой комбинации. Приемник заново вычисляет сумму цифр и сверяет ее с контрольным числом.

Различные варианты систем с переспросом можно классифицировать по следующим независимым признакам:

1. Ограниченный или неограниченный переспрос. В первом случае допускается не более чем r повторений, во втором сигнал повторяется до тех пор, пока не будет с уверенностью принят.

2. Наличие или отсутствие памяти. В первом случае так или иначе учитываются ранее принятые сигналы, во втором принимается во внимание только очередной.

3. Постоянная или изменяющаяся настройка передатчика и (или) приемника. В первом случае характер сигнала и пороговые значения (границы областей Y и Z) остаются неизменными на протяжении всего процесса приема (за исключением последнего этапа при ограниченном переспросе). Во втором при каждом повторении меняется характер сигнала и (или) настройка приемника.

4. Переспрос по отдельным элементам сигнала или по группам элементов. Во втором случае переспрос охватывает группу из n элементов, и сигнал переспроса либо указывает место сомнительного элемента в этой группе, либо требует повторения всей группы.

Что касается систем со сравнением, то их разновидности могут различаться по следующим признакам.

1. Полное или неполное сравнение. В первом случае по обратному каналу передается принятый сигнал. Во втором случае все множество передаваемых сигналов разделяется на несколько подмножеств, и по обратному каналу передается лишь номер подмножества, к которому приемник относит полученный сигнал. В системе с неполным сравнением нагрузка обратного канала меньше, но ошибка в отождествлении принятого сигнала внутри данного подмножества остается незамеченной.

2. Сбрасывание или накопление. В первом случае передатчик не учитывает в дальнейшем те сигналы, которые он счел сомнительными, и просто повторяет их. Во втором случае ранее принятые по каналу обратной связи сигналы запоминаются и информация о них так или иначе используется при последующей передаче, что позволяет сократить время, затрачиваемое на повторение.

3. Ограниченная или неограниченная коррекция. В первом случае число повторения или корректирующих сигналов для исправления данного элемента или группы элементов ограничено. Во втором случае корректирование производится до тех пор, пока по

каналу обратной связи не поступит сигнал, извещающий об исправлении всех ошибок. При ограниченной коррекции вероятность ошибки больше, но потери времени меньше; кроме того, исключается возможность перегрузки запоминающих устройств.

В заключение этого краткого обзора отметим, что возможны комбинированные системы, в которых обратная связь используется как для переспроса, так и для сравнения.

Большой интерес представляют системы, в которых действие обратной связи сочетается с применением тех или иных корректирующих кодов. Такое сочетание дает богатые возможности построения систем передачи информации с особо высокой верностью.

§ 22. Корректирующие коды; общие соображения

Одним из мощных современных средств борьбы с помехами является применение корректирующих кодов. Корректирующими называются коды, позволяющие обнаруживать и исправлять ошибки, происходящие при передаче из-за влияния помех. Корректирующие коды начали усиленно разрабатываться с 1950 г. С тех пор сделано очень много; теоретический уровень исследований по кодам сильно возрос, а число публикаций увеличивается с каждым днем. Появились уже и обобщающие монографии. Интересно отметить, что на протяжении сравнительно небольшого времени общие представления о существе дела менялись довольно заметно. В этом параграфе излагаются простейшие предварительные рассуждения о возможностях корректирующих кодов.

Ошибки при передаче кодированного сигнала сводятся к тому, что некоторые из переданных символов заменяются другими — неверными. Говорят о различной кратности ошибок, имея при этом в виду число q искаженных символов в пределах одной кодовой комбинации.

Идея возможности обнаружения ошибок (т. е. констатации факта их наличия в принятой кодовой комбинации) крайне проста. Она состоит в том, что (в равномерном блочном коде) для передачи используются не все $N_0=m^n$ возможных кодовых комбинаций, а лишь часть их

$$N < N_0. \tag{22.1}$$

Здесь m — основание кода, т. е. число различных символов; n — значность кода, т. е. число символов в кодовой комбинации. В дальнейшем мы будем говорить только о двоичных кодах ($m=2$), символы которых обозначаются 0 и 1.

Используемые в данном коде комбинации часто называют разрешенным, а остальные N_0-N неиспользуемых комбинаций — запрещенными. Если в результате ошибок переданная (разрешенная) комбинация превращается в одну из запрещенных, то тем самым и обнаруживается наличие ошибок. Вместе с тем ясно, что

если совокупность ошибок в данной кодовой комбинации превращает ее в какую-либо другую разрешенную, то в этом случае ошибки не могут быть обнаружены.

Таким образом, всякий код, удовлетворяющий единственному условию (22.1), способен обнаружить ошибки в N (N_0—N) возможных случаях из общего числа NN_0*.

Доля обнаруживаемых ошибочных комбинаций составляет

$$\frac{N(N_0-N)}{NN_0}=1-\frac{N}{N_0}. \qquad (22.2)$$

Почти так же просто обстоит дело и с исправлением ошибок. Для использования данного кода в качестве исправляющего нужно произвести разбиение множества $\{B_j\}$ (j=1, 2, . . ., N_0—N) запрещенных комбинаций на N непересекающихся подмножеств M_k. Каждое из подмножеств M_k приписывается одной из кодовых комбинаций A_k. Способ приема состоит в том, что если принята комбинация $B_j \in M_k$, то считается, что передана комбинация A_k. При таком способе приема ошибки исправляются, однако, не все, и наша очередная задача состоит в том, чтобы выяснить, какую долю возможных ошибок способен исправить данный код.

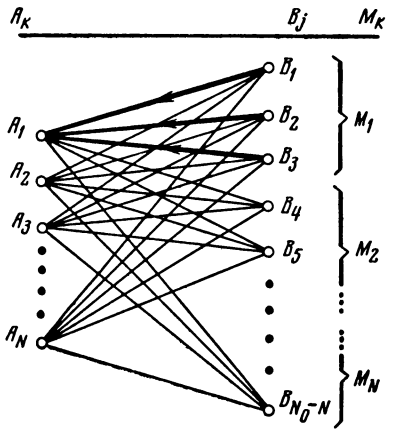


Рис. 65

Для наглядности положение представлено в виде графа на рис. 65. Слева изображена совокупность передаваемых (разрешенных) кодовых комбинаций $\{A_k\}$, образующих данный код. Справа — все возможные запрещенные комбинации $\{B_j\}$, в которые в результате различного сочетания ошибок могут перейти кодовые комбинации $\{A_k\}$. Здесь же показано разбиение множества $\{B_j\}$ на подмножества M_k. Способ приема состоит, как уже говорилось, в том, что если принимается комбинация B_j, попадающая в подмножество M_k, то считается, что передана комбинация A_k (на рис. 65 это показано стрелками для M_1). Если комбинация B_j действительно образовалась из A_k, то ошибка исправлена. Следовательно, при произвольном разбиении на подмножества M_k ошибка всегда исправляется в N_0—N случаях. Но общее число возможных переходов, как видно из рис. 65, составляет N (N_0—N). Это есть число

* Говоря об общем числе возможных случаев, мы имеем в виду, что может передаваться любая из N_0 разрешенных комбинаций и что каждая из них может превратиться в любую из N_0 возможных кодовых комбинаций.

обнаруживаемых ошибочных комбинаций. Отношение числа исправляемых ошибочных комбинаций к числу обнаруживаемых комбинаций составляет

$$\frac{N_0 - N}{N(N_0 - N)} = \frac{1}{N}. \qquad (22.3)$$

Этот простой результат имеет совершенно общий характер; он относится к любому коду, удовлетворяющему единственному условию (22.1).

Итак, всякий код при условии (22.1) может применяться в качестве исправляющего; то или иное использование его потенциальной исправляющей способности зависит от способа приема, т. е. выбора разбиения $\{M_k\}$. Выбор разбиения в свою очередь определяется условиями применения кода.

Имеющиеся здесь возможности мы поясним на примере. Прежде всего введем очень удобный вектор ошибки e_i, записываемый в виде двоичного числа той же значности, что и кодовые комбинации A_k. Нули в e_i означают, что ошибка в соответствующих позициях A_k не происходит; единицы стоят на позициях, которые искажаются при передаче, так что число единиц в e_i равно кратности ошибок q. Принятая комбинация B_{ik} получается путем сложения A_k и e_i по модулю два [1], т. е.

$$B_{ik} = A_k + e_i^*.$$

Векторы e_i представляют все возможные сочетания ошибок, так что общее число e_i составляет $N_0 - 1$.

В качестве примера возьмем код с $N=4$, $n=4$. При этом $N_0 = 2^4 = 16$: число исправляемых ошибочных комбинаций $N_0 - N = 12$. Примем в качестве кода первый попавшийся набор кодовых комбинаций

$$A_1 = 0001, \quad A_2 = 0101, \quad A_3 = 1110, \quad A_4 = 1111. \qquad (22.4)$$

Составим полную кодовую таблицу, включающую как разрешенные комбинации кода A_k, так и все комбинации B_{ik}, образующиеся в результате ошибок.

В таблице имеются пробелы на тех местах, где образуются искаженные комбинации, совпадающие с одной из разрешенных. Эти ошибки неисправимы. Например,

$$B_{31} = l_3 + A_1 = 0100 + 0001 = 0101 = A_2.$$

Таким образом, в каждом столбце сохранены только $N_0 - N$ запрещенных комбинаций. Всего, следовательно, в табл. 22.1

---

[1] Напомним, что сложение по модулю два выполняется по следующему правилу: $0+0=0$, $0+1=1$, $1+0=1$, $1+1$.

* Множество $\{B_{ik}\}$ есть множество всех ошибочных комбинаций; индекс i пробегает все значения от 1 до $N_0 - 1$, индекс k — от 1 до N, так что число элементов множества $\{B_{ik}\}$ равно $N(N_0 - 1)$. Таково число ячеек в нижеследующей табл. 22.1. В этой таблице множество $\{B_{ik}\}$ запрещенных комбинаций занимает один столбец.

имеется $N\,(N_0-N)=48$ комбинаций с обнаруживаемыми ошибками, из которых $1/N={}^1/_4$ (т. е. $N_0-N=12$) могут быть исправлены.

Таблица 22.1

e_i	A_k				
	0001	0101	1110	1111	q
0001 0010 0100 1000	0000 0011 — 1001	0100 0111 — 1101	— 1100 1010 0110	— 1101 1011 0111	1
0011 0101 1001 0110 1010 1100	0010 0100 1000 0111 1011 1101	0110 (0000) 1100 0011 — 1001	1101 1011 0111 1000 0100 0010	1100 1010 0110 1001 — 0011	2
0111 1011 1101 1110	0110 1010 1100 —	0010 — 1000 1011	1001 — 0011 (0000)	1000 0100 0010 —	3
1111	—	1010	—	(0000)	4

Выбор исправляемых комбинаций в нашей власти. Так, например, естественнее потребовать, чтобы средняя вероятность ошибки была наименьшей. Разбиение кода, удовлетворяющее этому требованию, зависит от статистики ошибок. Если ошибки в каждом символе независимы, то вероятность ошибок убывает с повышением их кратности q, и для уменьшения средней вероятности ошибки следует в первую очередь исправлять ошибки низшей кратности. Этим и определяется в данном случае выбор разбиения на подмножества M_k. Практически разбиение можно построить, пользуясь табл. 22.1, следующим образом: оставить исправляемую запрещенную комбинацию в данном столбце в желательном месте, т. е. в строках, относящихся к малым кратностям, и вычеркнуть ту же комбинацию во всех остальных столбцах. Для примера в табл. 22.1 оставлена в первой строке первого столбца комбинация 0000; в остальных столбцах эта комбинация возникает при ошибках более высокой кратности; эта комбинация вычеркивается (взята в скобки в табл. 22.1). Действуя подобным образом, получаем разбиение, приведенное в табл. 22.2.

Таблица 22.2

M_1	M_2	M_3	M_4	q
0000	0100	1100	1101	1
0011	0111	1010	1011	
1001	—	0110	—	2
0010	—	1000	—	

Таблица 22.3

M_1	M_2	M_3	M_4	q
1101	—	0111	—	2
0110	1000	1001	0100	3
1100	1011	0011	0010	
—	1010	—	0000	4

Таким образом исправляется 10 комбинаций с одиночной ошибкой и две — с двойной. Тройные и четверные ошибки не исправляются.

Если статистика ошибок такова, что, наоборот, большую вероятность имеют ошибки высокой кратности, то разбиение нужно соответствующим образом изменить. Применяя тот же прием, но с оставлением исправляемых комбинаций в нижних строках табл. 22.1, получим табл. 22.3.

При таком разбиении код исправляет две комбинации с двойными ошибками, восемь — с тройными и две — с четверными. Предположим, наконец, что мы желаем исправлять преимущественно комбинации с двойными ошибками. В этом случае можно построить разбиение, представленное в табл. 22.4.

Таблица 22.4

M_1	M_2	M_3	M_4	q
0010	0110	1101	1100	2
0100	0000	1011	1010	
10001	0011	0111	1001	

Итак, исправляется 12 комбинаций с двойными ошибками. Ни одиночные, ни тройные, ни четверные ошибки не исправляются. На рис. 66, *а*, *б* и *в* представлено распределение числа исправляемых комбинаций по кратности ошибок для табл. 22.2, 22.3 и 22.4 соответственно.

С геометрической точки зрения выбор подмножеств M_k есть не что иное, как выбор собственных областей в пространстве сигналов. Для кодированных сигналов это пространство дискретно, т. е. представляет собой не континуум, а конечное множество дискретных точек. В частности, для двоичного кода множество сигнальных точек совпадает с вершинами n-мерного куба; каждая кодовая комбинация представляет собой запись координат соответствующей вершины. На рис. 67 изображено условное представление совокупности кодовых точек на плоскости чертежа. Выбраны две разрешенные комбинации A_1 и A_2.

На рис. 67, *а* в собственные области входят ближайшие к данной разрешенной комбинации кодовые точки, т. е. точки, образованные ошибками низкой кратности. На рис. 67, *б*, наоборот, собственные области включают наиболее удаленные точки, определяющие комбинации с ошибками высокой кратности.

В предыдущем примере, взяв первый попавшийся код, мы стремились по-разному использовать его исправляющую способность. Это нам удалось в смысле общей тенденции, но не полностью. Так, например, желая исправить ошибки высшей кратности, мы получили разбиение, исправляющее две комбинации с двойными ошибками, в то время как четыре комбинации с тройными ошибками остались неисправленными. Поэтому естественно поставить вопрос о построении кода, удовлетворяющего наперед поставленным требованиям.

Пусть для примера ищется четырехзначный код из четырех комбинаций ($n=4$, $N=4$), исправляющий все двойные ошибки, встречающиеся попарно. Это соответствует векторам ошибки:

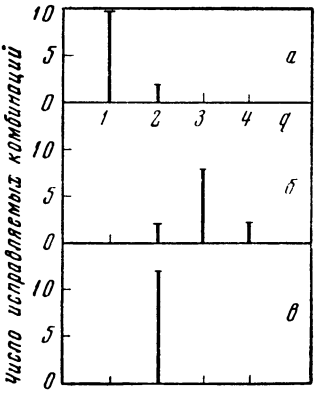


Рис. 66

$$e_1 = 0011, \quad e_2 = 0110, \quad e_3 = 1100.$$

Комбинаций с парными двойными ошибками может быть всего 12, так что в принципе требуемый код может существовать. Его можно отыскать простым подбором при помощи таблицы $B_{ik} = e_i + A_k$, содержащей столько строк, сколько задано векторов e_i и $N_0 = 2^n$ столбцов (табл. 22.5).

Подбор производится так: первая кодовая комбинация выбирается произвольно; пусть это будет 0001 (второй столбец). Эта комбинация вычеркивается во всех трех строках таблицы; вычеркиваемые комбинации отмечены $[\ ]_1$. Кроме того, вычеркиваются все B_{ik}, занятые во втором столбце, вычеркиваемые комбинации отмечены $(\ )_1$. В результате из игры выходят столбцы 3, 5, 8, 12, 14 и 15.

Выберем следующую кодовую комбинацию 0101 в свободном шестом столбце. Эта комбинация, отмеченная $[\ ]_2$, вычеркивается во всех трех строках таблицы; вычеркиваются также комбинации B_{ik}, занятые в шестом столбце — эти комбинации отмечены $(\ )_2$. Из дальнейшего подбора исключаются столбцы 1, 4, 7, 10, 11 и 16. Следующую кодовую комбинацию 1000 берем в свободном девятом столбце, и т. д. Искомый код и его разбиение представлены в табл. 22.6.

Мы получили код с требуемыми свойствами простым подбором. В дальнейшем будут рассмотрены некоторые конструктивные методы построения кодов. Но к методу подбора отнюдь не следует относиться пренебрежительно.

В настоящее время метод подбора во многих случаях признается не только целесообразным, но и прогрессивным. К тому же этот метод в некоторых случаях оказывается единственно возможным. Говоря о подборе, мы имеем в виду, конечно, работу не «вручную»,

а на электронных вычислительных машинах, действующих по некоторой разумным образом составленной программе.

В предыдущих примерах мы брали код определенной значности, и различным образом использовали его исправляющую способность. В практике чаще может встретиться другая постановка задачи: задается число передаваемых сообщений, т. е. число кодовых комбинаций N, задаются подлежащие исправлению ошибки.

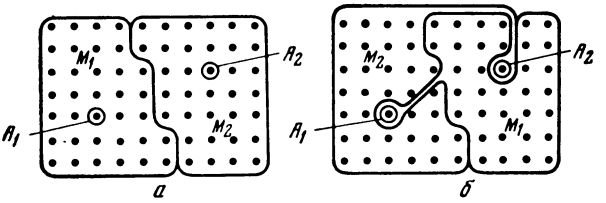


Рис. 67

Искомой является значность кода. Основываясь на предыдущем, можно получить простые формулы для нахождения значности кода, удовлетворяющего поставленным требованиям.

Пусть задано E векторов e_i, каждый из которых представляет определенное сочетание ошибок. Число исправляемых кодовых комбинаций будет EN, так как сочетание ошибок e_i исправляется во всех N комбинациях кода. С другой стороны, код может исправить не более N_0-N ошибочных комбинаций. Следовательно, должно быть

$$EN \leqslant N_0 - N, \tag{22.5}$$

откуда $N_0/(1+E)$.

Таблица 22.5

	1	2	3	4	5	6	7	8
e_i	A_k							
	0000	$\boxed{1}$ 0001	0010	0011	0100	$\boxed{2}$ 0101	0110	0111
0011	$(0011)_2$	0010	$[0001]_1$	0000	$(0111)_1$	0110	$[0101]_2$	$(0100)_3$
0110	$(0110)_2$	0111	$(0100)_3$	$[0101]_2$	$(0010)_1$	0011	0000	$[0001]_1$
1100	1100	1101	$(1110)_3$	1111	1000_3	1001	1010	$(1011)_3$
	9	10	11	12	13	14	15	16
e_i	A_k							
	$\boxed{3}$ 1000	1001	1010	1011	$\boxed{4}$ 1100	1101	1110	1111
0011	1011	1010	$(1001)_2$	$[1000]_3$	1111	$(1110)_3$	$(1101)_1$	1100
0110	1110	1111	1100	$(1101)_1$	1010	$(1011)_3$	$[1000]_3$	$(1001)_2$
1100	0100	$[0101]_2$	$(0101)_2$	$(0111)_1$	0000	$[0001]_1$	$(0010)_1$	$(0011)_2$

В качестве примера рассмотрим случай, когда требуется исправить все ошибки кратности q. Их общее число пусть будет λ_q, а число подлежащих исправлению ошибочных кодовых комбинаций равно $N\lambda_q$. Таким образом, согласно (22.5),

$$\frac{N_0}{1+\lambda_q} \geqslant N. \qquad (22.6)$$

Левая часть неравенства зависит от n, так как $N_0=2^n$, $\lambda_q=C_n^q$.

Выбираем наименьшее целое n, при котором неравенство (22.6) удовлетворяется. Если требуется, чтобы код исправлял ошибки нескольких кратностей q, то формуле (22.6) можно придать более общий вид [1]

$$\frac{N_0}{1+\sum\limits_q \lambda_q} \geqslant N, \qquad (22.7)$$

где сумма берется по всем заданным значениям q.

Таблица 22.6

e_i	A_k			
	0001	0101	10000	1100
0011	0010	0110	1011	1111
0110	0111	0011	1110	1010
1100	1101	1001	0100	0000

Рассмотрим простейший пример. Пусть задано $N_1=4$ и пусть искомый код должен исправлять все одиночные ошибки, т. е. ошибки с кратностью $q=1$. Число одиночных ошибок $\lambda_1=C_n^1=n$ (так как ошибка может встретиться в любой позиции n-значной кодовой комбинации). По формуле (22.6)

$$\frac{2^n}{1+n} \geqslant N=4.$$

Составляя вывод

n	2	3	4	5	6	7
$2^n(1+n)$	1,33	2	3,2	5,33	9,2	16

мы видим, что значность кода должна быть $n=5$. При этом неравенство (22.7) удовлетворяется с запасом (левая часть равна 5,33 вместо требуемых 4). Вследствие этого исправляются не только все одиночные ошибки, но и несколько комбинаций с ошибками более высокой кратности. Действительно, при $N=4$ и $n=5$ имеется всего $Nn=20$ комбинаций с одиночными ошибками. Вместе с тем число исправляемых комбинаций равно $N_0-N=32-4=28$, так что кроме 20 комбинаций с одиночными ошибками исправляется еще восемь комбинаций, безразлично каких именно, так как они исправляются сверх задания. Пример кода $N=4$, $n=5$ дан в табл. 22.7.

Как видим, кроме заданных одиночных ошибок исправляются еще две двойные.

[1] Интересно отметить, что подобного рода соотношения были получены Хэммингом из совершенно других соображений.

Таблица 22.7

e_i	A_k				
	00000	00111	11001	11110	q
00001	00001	00110	11000	11111	
00010	00010	00101	11011	11100	
00100	00100	00011	11101	11010	1
01000	01000	01111	10001	10110	
10000	10000	10111	01001	01110	
01010	01010	01101	10011	10100	
10010	10010	10101	01011	0110	2

Знак равенства в (22.7) получается только при

$$1+n=2^i \quad (i=2,\ 3,\ 4,\ \ldots).$$

Так, при $i=3$, $n=7$ получаем код, для которого $N_0=2^7=128$, а число используемых комбинаций, определяемое из соотношения (22.7), равно

$$N=\frac{N_0}{1+n}=\frac{2^7}{2^3}=2^4=16.$$

Этот же код может, в случае надобности, исправлять (при соответствующем разбиении) вместо всех одиночных ошибок — все шестикратные (так как их столько же, сколько одиночных).

Мы переходим теперь к обсуждению некоторых общих вопросов. К их числу относится вопрос о требованиях к корректирующему коду с точки зрения условий его применения.

Корректирующие коды применяются, вообще говоря, для повышения верности передачи сигналов при наличии помех, вызывающих случайные искажения кодированных сигналов. Но нужно уточнить, что мы понимаем под верностью. Критерии качества кода могут быть различными в зависимости от статистики, а также от содержания и назначения передаваемых сообщений.

Простейшая ситуация имеется в том случае, когда все сообщения, а следовательно, и соответствующие им кодовые комбинации равновероятны и равноправны. Естественно потребовать в этом случае, чтобы применение корректирующего кода уменьшало среднюю вероятность ошибки

$$p_{\text{ср}}=\sum_q p_q\lambda_q \qquad (q=1,\ 2,\ \ldots,\ n). \tag{22.8}$$

С точки зрения этого критерия тот код будет наилучшим, который обеспечивает заданную среднюю вероятность ошибки при наименьшей значности или дает наименьшую вероятность ошибки при заданной значности. Это — наиболее распространенная точка

зрения; такому критерию соответствует большинство результатов теории кодов.

Но возможны и другие критерии. Если известны априорные вероятности p_k появлений комбинаций A_k, то нужно ввести вероятность p_{qk} появления ошибки кратности q в каждой из комбинаций A_k. Тогда средняя вероятность ошибки будет

$$p_{\text{ср}} = \sum_k p_k \sum_q p_{qk}\lambda_q \quad (k = 1,\ 2,\ \ldots,\ n), \tag{22.9}$$

и можно потребовать, чтобы код минимизировал эту величину.

На практике сообщения, входящие в ансамбль, неравноценны. Например, если ансамбль представляет собой набор команд, то последствия от неправильного исполнения различных команд могут быть совершенно различными. Проще говоря, одни команды могут быть значительно важнее других. В таком случае естественно ввести весовые коэффициенты L_k, выражающие потери при ошибке в приеме кодовой комбинации A_k. Усредняя вероятности ошибок с весами L_k, получим риск

$$r = \sum_k L_k p_k \sum p_{qk}\lambda_q, \tag{22.10}$$

и можно требовать, чтобы код минимизировал эту величину.

Если распределение априорных вероятностей $\{p_k\}$ неизвестно, то можно проварьировать наименьший риск $r_{\min}$ по всем распределениям и выбрать наихудший случай, т. е. такое распределение $\{p_k^*\}$, которое дает максимальное значение минимальному риску. Это значение будет

$$r_{\min} = \sum_k L_k p_k^* \sum_q p_{qk}\lambda_q. \tag{22.11}$$

Возможна практическая ситуация, когда требования к коду состоят в том, чтобы вероятность ошибок в одних кодовых комбинациях была минимизирована, в то время как вероятность ошибки в других не превосходила заданной величины. Например, кодовые комбинации могут быть разбиты на две группы

$$k = 1,\ 2,\ \ldots,\ s \quad \text{и} \quad k = s+1,\quad s+2,\ \ldots,\ n,$$

и требование состоит в минимизации величины

$$p' = \sum_{k=1}^{s} p_k \sum_{q=1}^{n} p_{qk}\lambda_q \tag{22.12}$$

при условии

$$p'' = \sum_{k=s+1}^{n} p_k \sum_{q=1}^{n} p_{qk}\lambda_q \leqslant \varepsilon. \tag{22.13}$$

Легко заметить, что все перечисленные варианты требований к коду аналогичны различным критериям статистической теории обнаружения, а именно: условие минимизации средней вероятности (22.9) аналогично критерию идеального наблюдения Ко-

тельникова—Зигерта; формула (22. 8) есть вырожденный случай (22. 9) при равных априорных вероятностях; условие минимизации (22.10) есть критерий бейесова риска; (22. 11) выражает минимаксный критерий, а (22. 12) и (22. 13) представляют нечто вроде критерия Неймана—Пирсона.

В направлении применения перечисленных выше критериев теория кодов еще мало развита, хотя принципиальных затруднений здесь не предвидится. В качестве примера покажем построение разбиения ранее использованного кода (22. 4), потребовав, чтобы кодовая комбинация A_1 передавалась с высокой верностью по сравнению с остальными. Именно потребуем, чтобы при передаче A_1 исправлялись все одиночные и все двоичные ошибки (разумеется, за вычетом неисправимых). Разбиение производится описанным выше способом с той разницей, что теперь мы сохраняем комбинации B_{ik} не только в заданных строках, но и в заданном столбце. Комбинация B_{11} неисправима. Мы получаем разбиение, представленное в табл. 22. 8.

Таблица 22.8

M_1	M_2	M_3	M_4	
0000	—	1100	—	1
0011	—	1010	—	
1001	—	0110	—	
0010	—	—	—	2
0100	—	—	—	
1000	—	—	—	
0111	—	—	—	
1011	—	—	—	
1101	—	—	—	

Итак, для A_1 исправляются три одиночные ошибки из четырех возможных и все шесть двойных. Для A_3 исправляются три одиночные ошибки. Ошибки в комбинациях A_2 и A_4 вовсе не исправляются. Напомним, что в этом примере мы берем не специально подобранный, а какой попало код. Подбирается только разбиение.

Другой важный вопрос, которого мы до сих пор совершенно не касались — это вопрос о декодировании. Операция декодирования реализует возможности корректирующего кода. Декодирование состоит в установлении номера переданной кодовой комбинации путем некоторой обработки принятой кодовой комбинации. Таким образом, задача состоит в том, чтобы отнести принятую кодовую комбинацию к одному из подмножеств M_k.

Важно представить себе, из каких элементарных действий слагается операция декодирования, так как от этого зависит как время, затрачиваемое на декодирование, так и сложность декодирующего устройства. Простота процедуры декодирования является настолько существенным достоинством кода, что на практике предпочтение будет отдано менее эффективному коду, если он легко декодируется. Поэтому для современных исследований по теории кодов характерна тенденция к отыскиванию корректирующих кодов с наиболее простым декодированием.

Совершенно универсальный метод декодирования, пригодный для какого угодно кода, состоит в сличении принятой кодовой комбинации со всеми N_0 возможными комбинациями. При этом принятая комбинация совпадает либо с одной из N рабочих комбинаций кода, либо с одной из N_0—N запрещенных комбинаций, отнесенных согласно разбиению к одному из подмножеств M_k. Схема, осуществляющая декодирование по этому принципу, может быть составлена из переключателей на m направлений, соединения которых воспроизводят строение кодового дерева. На рис. 68 показана такая схема для двоичного кода ($m=2$) при $n=4, N=4$. Кружочками обозначены переключатели (реле) на два направления; при появлении нуля включается направление влево, при появлении единицы — вправо. Сигнал появляется на одном из N_0 зажимов, расположенных на штриховой линии. Для примера на рисунке показано прохождение комбинации 0110. Ниже штриха показаны соединения, соответствующие выбранному разбиению. На рис. 68 изображены соединения для кода с разбиением согласно табл. 22.6. В результате сигнал появляется на одном из N выходных зажимов в нижнем ряду.

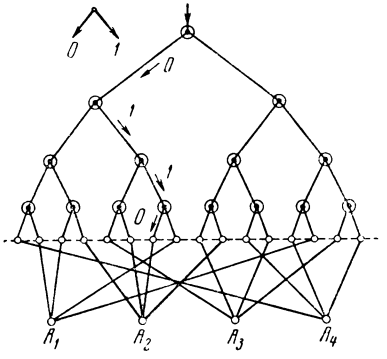


Рис. 68

Легко видеть, что число реле равно 2^n-1. Таким образом, схема описанного типа хотя и является универсальной, но практически пригодна только для кодов небольшой значности. Между тем в настоящее время находят применение коды, значность которых достигает порядка 10^3. Отсюда и вытекает необходимость изыскания более простых принципов декодирования. Процедура декодирования может быть упрощена, если задаться этой целью при конструировании кода.

За последнее время появился целый ряд специальных кодов, декодируемых по относительно простым правилам. Краткое описание некоторых из этих кодов дано в последующих параграфах.

§ 23. Исправляющая способность и кодовое расстояние

Мы не ввели до сих пор одной употребительной характеристики корректирующего кода, а именно, расстояний между кодовыми комбинациями.

Расстояние между парой кодовых комбинаций выражает различие между ними. Наименьшее расстояние для данного кода мы будем называть кодовым расстоянием.

Принято пользоваться метрикой

$$d_{ij} = \sum_{k=1}^{n} | x_{ik} - x_{jk} |.$$

Для двоичного кода расстояние просто равно числу знаков, в которых одна комбинация отличается от другой. С геометрической точки зрения это означает число ребер единичного куба, отделяющих одну вершину куба от другой. Определенное таким образом расстояние часто называют хэмминговым расстоянием.

Можно также заметить, что для двоичного кода расстояние между двумя комбинациями равно числу единиц в сумме этих комбинаций по модулю два. Например:

$$\begin{array}{r} 1\,0\,1\,1\,0\,1\,0\,0\,1 \\ +\;1\,1\,0\,1\,1\,1\,0\,1\,0 \\ \hline 0\,1\,1\,0\,1\,0\,0\,1\,1 \end{array}$$

и расстояние равно пяти. Замена в одной из кодовых комбинаций единицы нулем (или наоборот) изменяет расстояние на единицу (т. е. вызывает смещение кодовой комбинации, равное единице); перестановка (перемена местами) единицы и нуля изменяет расстояние на две единицы.

Корректирующие коды были первоначально созданы для обнаружения и исправления независимых ошибок. Величина кодового расстояния играет в этом случае существенную роль, как мы сейчас увидим.

В случае независимых ошибок вероятность ошибки убывает с возрастанием кратности. Пусть вероятность замены в каждом отдельном знаке единицы нулем или, наоборот, обозначена через p_0. Тогда вероятность одиночной ошибки, т. е. ошибки в одном каком-либо знаке n-значной кодовой комбинации, равна

$$p_1 = np_0 (1 - p_0)^{n-1}.$$

Вероятность двойной ошибки

$$p = \frac{1}{2} n (n - 1) p_0^2 (1 - p_0)^{n-2}$$

и вероятность ошибки кратности q *

$$p_q = C_n^q p_0^q (1 - p_0)^{n-q}. \qquad (23.1)$$

Таким образом, в случае независимых ошибок следует в первую

* Эта формула получается следующим образом. Предполагается, что ошибка имеется в q знаках (вероятность p_0^q), остальные же $n-q$ знаков должны быть правильны (вероятность этого есть $(1-p_0)^{n-q}$). Так как ошибки независимы, то вероятности этих событий перемножаются. Множитель C_n^q есть число перестановок, т. е. число возможных положений q ошибочных знаков в пределах n-значной комбинации.

очередь обнаруживать и исправлять ошибки низшей кратности, как более вероятные.

Если кодовое расстояние d равно единице, то это значит, что одиночная ошибка может превратить разрешенную комбинацию в другую разрешенную. При этом одиночная ошибка остается необнаруженной.

Не нужно, однако, думать, что при $d=1$ код вообще не обладает никакой корректирующей способностью Кодовое расстояние d определено как наименьшее расстояние. Более полное представление о свойствах кода можно почерпнуть из матрицы расстояний d_{ij} (i, $j=1,2,3,\ldots, N$).

Возьмем, к примеру, код

$$A_1=000, \quad A_2=001, \quad A_3=010, \quad A_4=111. \tag{23.2}$$

Матрица расстояний имеет следующий вид:

j	1	2	3	4
1	0	1	1	3
2		0	2	2
3			0	2
4				0

Таким образом, данный код не обнаруживает только те одиночные ошибки, которые вызывают переходы $A_1 \rightleftarrows A_2$ и $A_1 \rightleftarrows A_3$. Все же остальные одиночные ошибки обнаруживаются. Очевидно, что для обнаружения всех одиночных ошибок необходимо выполнить условие $d \geqslant 2$, а для обнаружения всех ошибок кратности q_d требуется

$$d \geqslant q_d + 1. \tag{23.3}$$

Ошибка может быть не только обнаружена, но и исправлена, если содержащая ошибку комбинация все же остается ближе к той, из которой она произошла, чем к любой другой разрешенной комбинации. Заметим, что кратность ошибки q есть смещение кодовой комбинации, выраженное в тех же единицах, в которых измеряется расстояние. Таким образом, должно быть

$$q_c < \frac{1}{2} d$$

или, принимая во внимание, что q и d — целые числа,

$$d \geqslant 2q_c + 1. \tag{23.4}$$

Таким образом, приведенный выше код (23. 2) не способен исправить одиночную ошибку. Наличие в таблице $d_{14}=3$ означает лишь, что при передаче комбинации $A_4=111$ одиночная ошибка не позволяет спутать ее с $A_1=000$.

Рассуждая аналогичным образом, можно прийти к следующему. более общему результату: для того чтобы код мог исправлять все ошибки кратности $\leqslant q_c$ и одновременно обнаруживать все ошибки

кратности $\leqslant q_d$ (при $q_d \geqslant q_c$), достаточно,чтобы кодовое расстояние удовлетворяло условию

$$d \geqslant q_c + q_d + 1. \tag{23.4а}$$

Приведенные соображения позволяют строить простейшие корректирующие коды, обнаруживающие и исправляющие одиночные ошибки. Рассмотрим код

$$A_1 = 00, \quad A_2 = 01, \quad A_3 = 10, \quad A_4 = 11. \tag{23.5}$$

Этот двухзначный код использует все возможные комбинации и не может обнаруживать ошибки, так как $d=1$. Припишем к комбинациям (23.5) один знак, так что образуются новые трехзначные комбинации

$$A_1 = 000, \quad A_2 = 011, \quad A_3 = 101, \quad A_4 = 110. \tag{23.6}$$

Для этого кода все расстояния равны 2*, и код способен обнаружить все одиночные ошибки. Чтобы получить код, исправляющий все одиночные ошибки, придется добавить еще не менее двух знаков (см. § 22), например, повторить первые два знака. Получится

$$A_1 = 00000, \quad A_2 = 01101, \quad A_3 = 10110, \quad A_4 = 11011. \tag{23.7}$$

Матрица расстояний имеет следующий вид:

	1	2	3	4
1	0	3	3	4
2		0	4	3
3			0	3
4				0

Следовательно, код (23. 7) способен исправить все одиночные ошибки (так как $d=3$).

В связи с нахождением оптимальных кодов возникает следующая задача: найти наибольшее число N (d) кодовых комбинаций n-значного двоичного кода, расстояние между которыми не менее d. Общее решение этой задачи неизвестно; некоторые частные результаты даны ниже:

d	N
1	2^n
2	2^{n-1}
3	$\leqslant \dfrac{2^n}{1+n}$
4	$\leqslant \dfrac{2^n - 1}{n}$
. . . . .	. . . . .
$2k+1$	$\leqslant \dfrac{2^n}{1 + C_n^1 + C_n^2 + \ldots + C_n^k}$ **

* Геометрическая модель этого кода — тетраэдр. Это — единственный двоичный симплексный код.

** Этот предел указал Хэмминг (BSTJ, 1950, v. 29, p. 47). Позднее его улучшили Вакс (Ire Trans., IT-5, 1959, p. 168) и Грэй (IRE Trans., IT-7, 1961, p. 270).

Последняя формула может быть получена из соотношения (22. 7), если принять во внимание связь между d и q, выражаемую формулой (23. 4).

Отметим теперь, что код (23. 6) образован из кода (23. 5) по простому правилу, а именно: к комбинациям кода (23. 5) приписывается в качестве третьего знака 0 или 1 с таким расчетом, чтобы число единиц в новой кодовой комбинации было четным. Отсюда следует, что обнаружение ошибки может производиться проверкой на четность: все разрешенные комбинации кода (23.6) имеют четное число единиц, а все запрещенные (т. е. содержащие одиночную ошибку) — нечетное число единиц. Технически проверку на четность удобно осуществить суммированием по модулю два цифр кодовой комбинации. Сумма равна нулю, когда число единиц четно, и единице — в противном случае. Принцип проверок на четность можно распространить и на случай исправления ошибок, однако делать это в рамках элементарного изложения данного параграфа не имеет смысла.

Мы перейдем теперь к случаю приемника со стиранием (часто говорят «канал со стиранием» — erasure channel). Особенность приемника со стиранием состоит в том, что решающее устройство имеет область неопределенности, в которую попадают все сомнительные сигналы. Решающее устройство выдает при этом специальный символ, заменяющий сомнительный элемент сигнала. Последний оказывается, таким образом, «стертым». Это означает, что если передача ведется кодом с основанием m, то на выходе решающего устройства имеется алфавит из $m+1$ символов, так как к m элементам кода добавляется еще символ стирания. Так, при передаче двоичным кодом на выходе решающего устройства появляется один из трех символов: 0,1 и символ стирания θ.

Оказывается, что восстановить стертые знаки в известном смысле легче, чем исправить ошибочные. Это обусловлено, вообще говоря, тем, что местонахождение стертых знаков известно, так как оно обозначено символом стирания θ, тогда как местоположение ошибок неизвестно, и каждый из знаков 0 или 1 может быть как верным, так и неверным. Поэтому иногда оказывается выгодным ввести искусственно стирание относительно большого числа знаков, с тем, чтобы оставшиеся были верны с высокой вероятностью.

Посмотрим, каковы возможности корректирующего кода в применении к восстановлению стертых знаков. Обозначим кратность стирания, т. е. число стертых знаков в пределах одной кодовой комбинации через t. Будем полагать, что стирание отдельных знаков — события независимые. Кроме того, предположим, что все нестертые знаки правильны. Тогда необходимое для восстановления стертых знаков кодовое расстояние можно найти из следующих рассуждений.

Приняв кодовую комбинацию с t стертыми знаками, образуем укороченный код, в котором во всех комбинациях исходного

кода вычеркнуты позиции, стертые в принятой комбинации. Укороченный код будет иметь значность $n-t$*. Для того чтобы укороченные комбинации можно было отличить друг от друга, расстояние между ними должно быть по меньшей мере единица (т. е. они должны различаться хотя бы в одном знаке). Но полные (неукороченные) комбинации могут различаться между собой еще в $t' \leqslant t$ вычеркнутых знаках. Таким образом, для сохранения различимости кодовых комбинаций при стирании не более t знаков кодовое расстояние должно удовлетворять условию

$$d \geqslant t + 1. \tag{23.8}$$

Сравнение (23. 8) и (23. 4) показывает, что код с данным кодовым расстоянием восстанавливает стертые символы с большей кратностью t, нежели кратность q_c исправляемых ошибок. В этом и состоит точный смысл высказанного выше положения о том, что восстанавливать стертые символы легче, чем исправлять ошибки [1].

Так, например, при $d=2$ код способен только обнаружить (но не исправить) одиночную ошибку. Но тот же код может восстановить одиночное стирание. При $d=3$ код исправляет одиночную ошибку, но восстанавливает два стертых символа, и т. д.

Для того чтобы код мог одновременно исправлять q_c ошибок и восстанавливать t стертых символов, кодовое расстояние должно быть

$$d \geqslant 2q_c + t + 1. \tag{23.9}$$

Можно отметить еще один специальный случай, возникающий при передаче с пассивной паузой. При этом 1 означает наличие сигнала (посылку), 0 — отсутствие сигнала (паузу). Предположим, что характер помехи таков, что 1 может превратиться в 0, но не наоборот. Так будет обстоять дело, например, при мультипликативной помехе, когда уровень сигнала в посылке упадет ниже порогового значения, установленного в решающем устройстве. Конечно, можно рассматривать возможную замену единицы нулем как обычную ошибку, и тогда справедливо условие (23. 4). Но можно указать более выгодный способ приема. Способ этот состоит в образовании укороченного кода путем вычеркивания тех позиций, на которых в принятой комбинации оказались нули. После этого принятая комбинация отождествляется с той кодовой комбинацией, которая в укороченном виде состоит из одних единиц.

При описанном способе приема достаточно, чтобы кодовое расстояние удовлетворяло условию

$$d \geqslant 2r - s + 2, \tag{23.10}$$

---

* С геометрической точки зрения это означает проектирование n-мерного пространства сигналов на подпространство с числом измерений $n-t$.

[1] Сравнение (23. 8) и (23. 3) показывает, что при заданном расстоянии наибольшая кратность восстанавливаемых стираний равна кратности обнаруживаемых ошибок.

где r — наибольшее число единиц, заменившихся нулями; s —наименьшая разность чисел единиц в других кодовых комбинациях.

Вывод этого соотношения вынесен в Добавление V.

§ 24. Систематические коды

Современная теория кодов представляет собой высокоразвитую алгебраическую теорию. В ней широко используются матрицы, векторные пространства, группы, кольца и поля. В рамках нашего краткого изложения нет возможности последовательно применять эту специальную теорию. Мы воспользуемся ее результатами и отчасти терминологией. Однако многие положения придется привести в виде готовых правил, а не в качестве требующих доказательства теорем, каковыми они в действительности являются.

В этом параграфе будет рассмотрен специальный класс кодов, построение которых производится по определенным правилам. Закономерности, заложенные в структуре этих кодов, позволяют производить декодирование более простым способом, нежели прямое сличение принятой кодовой комбинации с полной кодовой таблицей.

Условимся, что речь будет идти только о двоичных кодах. Вместо термина «кодовая комбинация» мы будем в дальнейшем пользоваться термином кодовый вектор. Вектор, состоящий из одних нулей, будет называться нулевым вектором. Число единиц в кодовом векторе будет называться его весом. Таким образом, расстояние любого вектора от нулевого измеряется его весом.

Мы будем называть систематическим n-значный код, состоящий из $N=2^k$ кодовых векторов. Из n символов, образующих кодовый вектор, k символов являются информационными, а остальные $n-k$ — избыточными. Наличие избыточных символов определяет корректирующую способность кода и позволяет проверять наличие ошибок, по обнаружении исправлять их. Поэтому эти $n-k$ символов называют контрольными, или проверочными. Во всех кодовых векторах систематического кода проверочные символы занимают те же позиции. Для систематического кода применяется обозначение: (n, k) код.

Построение систематического кода производится следующим образом. В качестве первого берется нулевой вектор. Затем составляется производящая матрица G, имеющая k строк и n столбцов. В качестве строк берутся любые ненулевые линейно-независимые n-значные векторы, отстоящие друг от друга не менее чем на заданное кодовое расстояние. Линейно-независимыми называются вектора, для которых выполняется условие

$$c_1v_1 + c_2v_2 + \ldots + c_kv_k \neq 0. \qquad (24.1)$$

Здесь v_1, v_2, . . . — кодовые векторы, c_1, c_2, . . . могут принимать значения 0 или 1. Сложение здесь и на протяжении всего пара-

графа имеется в виду по модулю два. Заметим, что сложение и вычитание по модулю два равносильны.

Остальные $2^k - k - 1$ кодовых вектора получаются как линейные комбинации (взвешенные суммы) векторов, входящих в производящую матрицу. Весовыми коэффициентами могут быть 0 и 1, так что дело сводится к суммированию строк производящей матрицы во всевозможных сочетаниях, сначала попарно, затем по три и т. д. до суммы всех k строк.

После этого составляется множество векторов, ортогональных векторам кода. Если кодовый вектор v образован символами $a_1, a_2, \ldots, a_n$, то ортогональный вектор u и будет состоять из символов $b_1, b_2, \ldots, b_n$, удовлетворяющих условию

$$vu = \sum a_i b_i = 0, \tag{24.2}$$

причем это условие должно быть выполнено для всех v и для всех u.

Теперь можно составить матрицу H, порождающую множество векторов u. Эта матрица имеет $n-k$ строк и n столбцов. Ее строками являются любые линейно-независимые комбинации u. Матрица H называется проверочной.

Покажем описанные выше операции на примере. Пусть строится (5.3) код с минимальным расстоянием 2. Этот код может обнаружить все одиночные ошибки и исправить некоторые из них. Составим производящую матрицу в виде

$$G = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Таким образом, мы имеем уже четыре кодовые вектора: нулевой и три вектора, соответствующие строкам матрицы G. Суммируем первую и вторую строки матрицы. Это дает вектор 01110. Затем суммируем вторую и третью, первую и третью, и наконец, все три строки матрицы. В результате получаем следующий код:

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
00000	00011	01101	11010	01110	10111	11001	10100

(24.3)

Заметим, что производящая матрица имеет в качестве строк линейно-независимые векторы v_2, v_3 и v_4. Можно было бы построить производящую матрицу из любой другой тройки ненулевых векторов, кроме следующих сочетаний: $v_2v_3v_5$, $v_2v_4v_7$, $v_2v_6v_8$, $v_3v_4v_6$, $v_3v_7v_8$, $v_4v_5v_8$, $v_5v_6v_7$. Эти сочетания линейно-зависимы. Так, например

$$v_2 + v_4 + v_7 = 00011 + 11010 + 11001 = 0.$$

Построим множество ортогональных векторов u. Составим суммы (24. 2) для векторов v_2, v_8 и v_3. Для вектора v_2 имеем

$$0b_1 + 0b_2 + 0b_3 + 1b_4 + 1b_5 = 0$$

или, короче,

$$b_4 + b_5 = 0.$$

Для v_8 и v_3 получаем аналогично

$$b_1 + b_3 = 0, \quad b_2 + b_3 + b_5 = 0.$$

(Мы выбрали v_2, v_8 и v_3 потому, что два из них имеют наименьший вес, равный двум, а третий, v_3, дает такое соотношение между b_i, которое вместе с двумя первыми полностью определяет положение.)

Итак, для любого u должно быть

$$b_5 = b_4, \quad b_3 = b_1, \quad b_2 = b_3 + b_5.$$

Перебирая все удовлетворяющие этим условиям сочетания нулей и единиц, получаем

$$\begin{array}{cccc} u_1 & u_2 & u_3 & u_4 \\ 00000 & 01011 & 10111 & 11100 \end{array} \qquad (24.4)^*$$

Производящую матрицу для этого (т. е. исправляющую матрицу) можно записать в виде

$$H = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Перейдем теперь к вопросам декодирования (т. е. приема с исправлением ошибок). Начнем с составления полной кодовой таблицы.

Полная кодовая таблица в стандартном расположении составляется следующим образом. В первой строке выписываются все кодовые векторы, начиная с нулевого. В первом столбце второй строки записывается какой-либо вектор e_1, с наименьшим весом из числа не входящих в первую строку. Далее вторая строка заполняется путем суммирования вектора первого столбца e_1 с каждым из кодовых векторов. Аналогично составляется третья строка и т. д., всего $m = 2^{n-k}$ строк. В результате в таблице оказываются перечисленными все 2^n возможных n-значных векторов (табл. 24. 1):

Таблица 24.1

v_1	v_2	v_3	. . .	v_n
e_1	$e_1 + v_2$	$e_1 + v_3$	. . .	$e_1 + v_n$
. . . . .	. . . . .	. . . . .	. . .	. . . . .
e_{m-1}	$e_{m-1} + v_2$	$e_{m-1} + v_3$	. . .	$e_{m-1} + v$

* Заметим, что совокупность векторов u можно рассматривать как систематический, а именно $(n, n-k)$ код.

Таблица 24.2

00000	00011	01101	11010	01110	10111	11001	10100
00001	00010	01100	11011	01111	00110	11000	10101
00100	00111	01001	11110	01010	10011	11101	10000
01000	01011	00101	10010	00100	11111	10001	11100

Например, для кода (24. 3) получаем табл. 24. 2.

Общий метод декодирования состоит в том, что, приняв некоторый вектор v_x, мы отыскиваем его в таблице и отождествляем с тем кодовым вектором, который возглавляет данный столбец. Так, если $v_x = e_i + v_k$, то переданным вектором считается v_k. Операция исправления ошибки состоит в том, что к вектору v_x добавляется вектор e_i согласно равенству

$$v_x + e_i = v_k.$$

Таким образом, векторы e_i — это не что иное, как векторы ошибок, упоминавшиеся в § 22.

Только что описанный метод декодирования также упоминался ранее. Он является вполне универсальным, но чрезвычайно громоздким. Систематические коды позволяют осуществить декодирование значительно более простым методом, к описанию которого мы и переходим.

Воспользуемся множеством ортогональных векторов u. Если раньше мы определяли связи между составляющими этих векторов b_i с целью составления матрицы H, то теперь, наоборот, установим совершенно аналогичным путем связи между составляющими a_i кодовых векторов. Для этого служит условие ортогональности (24. 2), из которого по известным b_i можно найти интересующие нас связи между a_i.

Так, для рассмотренного примера получаем на основании (24. 2) (выписывая ненулевые элементы строк матрицы H)

$$a_2 + a_4 + a_5 = 0, \quad a_1 + a_2 + a_3 = 0.$$

Эти соотношения выражают не что иное, как проверки на четность. Если все проверки удовлетворяются, т. е. дают нуль, значит, ошибки нет. Если какая-либо проверка не удовлетворяется, т. е. дает единицу, это указывает на наличие ошибки. Пусть, например, принят вектор 00110. Проверочные равенства дают: первое — 1, второе — 1. Следовательно, ошибочным является второй символ, и исправленный вектор есть 01110. Однако в этом примере мы нарочно взяли ошибку во втором символе, которая может быть исправлена. Но данный код, имеющий кодовое расстояние, равное двум, не может исправить все одиночные ошибки. Его возможности легче всего уяснить из таблицы проверок (табл. 24. 3), которая показывает, какие символы охватываются данной проверкой.

Из табл. 24.3 видно, что рассматриваемый код может исправить ошибку во втором символе; что касается остальных ошибок, то проверки могут лишь установить наличие ошибок в одной из двух пар символов: a_1 и a_3 или a_4 и a_5, так как обе эти пары входят в таблицу совершенно одинаково. Для исправления всех ошибок необходимо, чтобы каждый символ появлялся в неповторяющейся комбинации строк таблицы проверок. Так оно и будет для кода с кодовым расстоянием $d=3$.

Итак, один из способов декодирования основан на применении проверок на четность. Существует и другой метод декодирования систематических кодов. Он основан на вычислении исправляющего вектора C (parity check vector, corrector, syndrome). Составляющие c_j этого вектора представляют собой скалярные произведения принятого вектора v_x на строки исправляющей матрицы H, т. е. на векторы u_j ($j=1, 2, \ldots, n-k$), т. е.

$$c_j = u_j v_x \qquad (24.5)$$

Таблица 24.3

№ проверки	№ символа				
	1	2	3	4	5
1		X		X	X
2	X	X	X		

и вектор C представляется $(n-k)$-значным двоичным числом [1]

$$C = (c_1, c_2, \ldots, c_{n-k}).$$

Если ошибок нет, то $C=0$ (см. 24. 2). Если же имеется ошибка, то некоторые из c_j имеют ненулевое значение. При этом существенно, что c_j имеет одно и то же значение для всех векторов данной строки полной кодовой таблицы. Действительно, для i-й строки кодовой табл. 23.1 имеем

$$e_i \quad e_i + v_2 \quad e_1 + v_3 \ldots e_i + v_n,$$

умножив эту строку скалярно на какой-либо из векторов u_j, получим

$$u_j e_i \quad u_j e_i + u_j v_2 \quad u_j e_i + u_j v_3 \ldots u_j e_i + u_j v_n.$$

Но вторые члены равны нулю в силу ортогональности векторов u и v. Следовательно, для всей i-й строки имеем

$$c_{ji} = u_j (e_i + v_k) = u_j e_i.$$

На этом основан метод декодирования, состоящий из следующих операций:

1. Для принятого вектора v_x вычисляется исправляющий вектор C по формуле (24. 5).

[1] Конечно, можно и вычисление составляющих исправляющего вектора трактовать как проверки на четность. Вообще, любая сумма по модулю дает либо нуль, либо единицу в зависимости от того, четно или нечетно число суммируемым единицам.

2. По найденному C определяется вектор ошибок e.

3. Ошибка, если она есть, исправляется путем добавления к принятому вектору v_x вектора ошибок e.

Поясним эту процедуру на примере кода (24. 3). Вычисление дает значения C, приведенные в табл. 24. 4 вместе с соответствующими значениями e. Пусть принят вектор $v_x=01001$. По формуле (24. 5) и с помощью матрицы H (стр. 400) получаем

$$c_1=(01011\cdot 01001)=0, \quad c_2=(11100\cdot 01001)=1$$

и, таким образом, $C=01$. Этому соответствует $e=00100$. Исправление производим, суммируя v_x и e:

$$v_x+e=01001+00100=01101=v_3.$$

Таким образом, сущность метода сводится к тому, что заранее устанавливается однозначная связь между вектором ошибки e, знания которого достаточно для исправления ошибки, и исправляющим вектором C, легко вычисляемым по принятому вектору v_x.

Таблица 24.4

e	C
00001	11
00100	01
01000	10

В заключение этого параграфа отметим возможность построения систематического (n, k) кода с определенным и удобным расположением символов, а именно: первые k символов — информационные, последние $n-k$ символов — контрольные [1]. Производящая матрица в этом случае строится так: составляется тождественная матрица $k\times k$, т. е. матрица с k строками и k столбцами, имеющая единицы по главной диагонали, а все остальные символы — нули. К этой матрице справа приписываются $(n-k)\times k$ матриц двоичных чисел.

Приведем пример:

$$G=\begin{pmatrix}1&0&0&0&1\\0&1&0&1&0\\0&0&1&1&1\end{pmatrix}.$$

Эта матрица порождает код

00000 00111 01010 01101 10001 10110 11011 11100

Как видим, первые три символа представляют собой все трехзначные двоичные числа. Два последних символа являются контрольными; легко получить для нахождения этих символов соотношения

$$a_4=a_2+a_3, \quad a_5=a_1+a_3.$$

Преимущество описанного кода в том, что его очень удобно составить. В исправляющей способности мы, конечно, ничего

[1] Заметим, что термин систематический применяется иногда именно к этой разновидности кодов.

не выигрываем. Так, таблица проверок на четность на основании приведенных соотношений между a_i составляется в виде табл. 24.5: т. е. ошибка в третьем символе исправляется, а в других случаях указывается лишь, находится ошибка в паре a_1a_5 или в паре a_2a_4.

Надо заметить также, что и в других систематических кодах можно локализовать информационные и контрольные символы; например, в коде (24.3) можно рассматривать $a_2a_3a_4$ как информационные символы, а a_1a_5 — как контрольные.

Таблица 24.5

№ проверки	№ символа				
	1	2	3	4	5
1	X		X		X
2		X	X	X	

§ 25. Циклические коды

В поисках более простой техники кодирования и декодирования в самое недавнее время были созданы так называемые циклические коды. Эти коды продолжают усиленно разрабатываться, появляются новые их разновидности, обладающие ценными свойствами, так что класс циклических кодов представляется многообобщающим.

С точки зрения современной теории циклический код определяется как идеал в линейной коммутативной алгебре A_n полиномов по модулю x^n—1 над полем коэффициентов.

Эта формулировка приведена только для того, чтобы оправдать полный отказ от попытки изложения теории циклических кодов в рамках нашего беглого очерка.

На основании того, что будет сказано в этом параграфе о циклических кодах, нельзя установить теоретически те или иные свойства этих кодов; все приведенные ниже немногочисленные результаты придется взять на веру. Нельзя также научиться составлять схемы кодирующих и декодирующих устройств. Однако, рассмотрев приведенные ниже примеры, можно все же оценить специфику техники циклических кодов. Это может быть полезным для читателя, который, получив здесь превоначальное представление о циклических кодах, займется ими в дальнейшем на должном теоретическом уровне, начав с изучения соответствующих разделов общей алгебры.

Циклические коды относятся к числу систематических. Основное свойство циклических кодов, определяющее их название, состоит в том, что если кодовый вектор $v=(a_0, a_1, a_2, \ldots, a_{n-1})$ принадлежит коду V, то и вектор v', получаемый из v циклической перестановкой составляющих, т. е. $v'=(a_{n-1}, a_0, a_1, \ldots, a_{n-1})$, также принадлежит коду V.

В теории циклических кодов принято представлять векторы в форме полиномов, а именно

$$v(x)=a_0+a_1x+a_2x^2+\ldots+a_{n-1}x^{n-1}.$$

Такое представление удобно хотя бы потому, что упомянутая выше циклическая перестановка есть результат простого умножения данного полинома на x. Действительно,

$$xv(x) = x(a_0 + a_1x) + \ldots + a_{n-1}x^{n-1}) =$$
$$= a_0x + a_1x^2 + \ldots + a_{n-2}x^{n-1} + a_{n-1}x^n.$$

Но в последнем члене нужно заменить x^n на единицу — это и есть то, что называется «полиномы по модулю x^n-1».

Таким образом,

$$xv(x) = a_{n-1} + a_0x + a_1x^2 + \ldots + a_{n-2}x^{n-1} = v^1(x).$$

Полином $g(x)$ степени $n-k$, на который делится без остатка двучлен $1+x^n$, называется производящим полиномом. Производящая матрица G имеет в качестве строк векторы, соответствующие

$$g(x), \quad xg(x), \ldots, x^{k-1}g(x).$$

Исправляющая матрица H, т. е. производящая матрица векторов, ортогональных кодовым, строится на основе вектора

$$h(x) = \frac{1+x^n}{g(x)}.$$

Образуя последовательность векторов $h(x)$, $xh(x)$, . . ., $x^{n-k-1}h(x)$ и записывая их составляющие в обратном порядке, получаем строки матрицы H.

Поясним все это на примере кода (7; 4), взяв в качестве производящего полинома

$$g(x) = 1 + x + x^3$$

(т. е. $g_0=g_1=g_3=1$, $g_2=0$). Обычная запись этого вектора будет 01101. Чтобы получить семизначный ($n=7$) кодовый вектор, припишем к вектору 1101 еще три нуля. Производя затем циклический сдвиг, получим

$$\begin{matrix} g(x) \\ xg(x) \\ x^2g(x) \\ x^3g(x) \end{matrix} \begin{pmatrix} 1101000 \\ 0110100 \\ 0011010 \\ 0001101 \end{pmatrix} = G.$$

Остальные кодовые векторы получаются, как обычно, суммированием строк производящей матрицы во всех возможных комбинациях.

Применительно к циклическим кодам принято (хотя это вовсе не обязательно) считать информационными символами последние k символов (соответствующие высшим степеням x), а контрольными первые $n-k$ символов. При обработке кодовых векторов начинают с информационных символов, т. е. считывание кодового вектора

идет справа налево. Это условие нужно иметь в виду, так как оно соблюдается во всем последующем изложении.

Если теперь перечислить все кодовые векторы в такой последовательности, чтобы информационные символы расположились в порядке возрастающих двоичных чисел (при чтении справа налево), то кодовая таблица принимает вид (табл. 25.1):

Таблица 25.1

№	Кодовый вектор		№	Кодовый вектор	
	контрольный знак	информационный знак		контрольный знак	информационный знак
1	110	1000	9	011	1001
2	011	0100	10	110	0101
3	101	1100	11	000	1101
4	111	0010	12	010	0011
5	001	1010	13	100	1011
6	100	0110	14	001	0111
7	010	1110	15	111	1111
8	101	0001			

Для вектора

$$h(x)=\frac{1+x^n}{g(x)}$$

находим [1] $h(x)=1+x+x^2+x^4$

[1] Необходимо пояснить, что действия над векторами производятся по правилам арифметики по модулю два, в котором вычитание равносильно сложению. Так, из равенства $x^n-1=0$ получаем $x^n=1$. Прибавив к правой и левой частям по единице, получаем $x^n+1=1+1=0$. Таким образом, вместо двучлена x^n-1 можно ввести двучлен x^n+1 или $1+x^n$, как мы и сделали. Приведем далее порядок деления и умножения. Он основан на том, что $x^k+x^k=x^k(1+1)=0$. Деление $1+x^n$ на $g(x)$ выполняется так:

$$\begin{array}{r} 1+x^7 \\ +\,1+x+x^3 \\ \hline x+x^3+x^7 \\ +\,x+x^2+x^4 \\ \hline x^2+x^3+x^4+x^7 \\ x^2+x^3+x^5 \\ \hline x^4+x^5+x^7 \\ +\,x^4+x^5+x^7 \end{array} \qquad \begin{array}{l} 1+x+x^3 \\ \hline 1+x+x^2+x^4 \end{array}$$

Умножение $g(x)\,h(x)$:

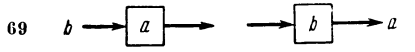
$$(1+x+x^3)(1+x+x^2+x^4)=1+x+x^3+x+x^2+$$
$$+x^4+x^2+x^3+x^5+x^4+x^5+x^7=1+x^7.$$

и соответствующая матрица имеет вид

$$H = \begin{pmatrix} 0010111 \\ 0101110 \\ 1011100 \end{pmatrix}.$$

Обратимся теперь к вопросам техники кодирования и декодирования.

Цикличность перестановок, определяющая строение рассматриваемых кодов, лежит в основе техники кодирования и декодирования. Эта техника применяет сдвигающие регистры в форме триггерных цепочек с теми или иными обратными связями. Ос-

Рис. 69 

новными элементами схем являются, во-первых, триггерная ячейка, и, во-вторых, суммирующая (по модулю два) ячейка. На всех нижеследующих схемах триггера ячейка обозначена квадратиком, суммирующая ячейка кружком со знаком + внутри.

Действие триггерной ячейки состоит в том, что при дискретном воздействии на нее она меняет свое состояние. Каждое такое изменение мы будем называть шагом. Под состоянием понимается символ, содержащийся внутри ячейки. Пусть в ячейке содержится символ a (ячейка находится в состоянии a). При появлении на входе символа b состояние ячейки меняется: символ b входит в ячейку и остается в ней до следующего шага (ячейка переходит в состояние b), а символ a вытесняется из ячейки и появляется на выходе. Изменение состояния ячейки, соответствующее одному шагу, показано схематически на рис. 69.

Назначение кодирующего устройства состоит в образовании векторов данного корректирующего кода. На вход кодирующего устройства подается последовательность k информационных символов. Кодирующее устройство добавляет по определенному правилу $n-k$ контрольных символов и выдает на выход полный кодовый вектор, составляющийся из n символов.

Кодирующее устройство для циклических кодов может осуществляться по-разному. Ниже описаны два варианта.

В первом варианте применяется k-ступенный регистр по схеме рис. 70. На этой схеме, кроме упомянутых выше элементов, имеются еще белые кружочки, означающие простое умножение на приписанное рядом постоянное число, а также ключи, замыкающие или размыкающие цепи, в которые они введены.

Действие схемы происходит в следующем порядке. Сначала ключ K находится в положении 1. На вход подаются информационные символы. После k последовательных шагов информационные символы занимают все k ячеек регистра. Затем ключ K перекладывается в положение 2, замыкая цепь обратной связи,

и регистр совершает еще n шагов, при каждом из которых на выходе появляется очередной символ кодового вектора. Уже при первом из этих шагов в первой (считая справа) ячейке формируется первый контрольный символ. После этих шагов в первой (считая справа) ячейке формируется первый контрольный символ. После $n-k$ шагов весь кодовый вектор сформировался; $n-k$ символов выданы на выход, остальные K символов еще находятся в регистре.

Теперь ключ K возвращается в исходное положение 1 и в регистр вводится K информационных символов следующего кодо-

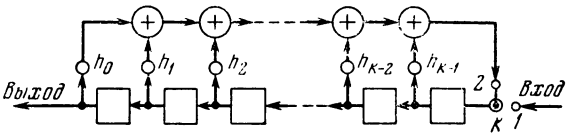


Рис. 70

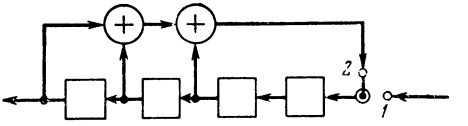


Рис. 71

вого вектора. При этом оставшиеся в регистре k символов предыдущего вектора выводятся наружу.

Поясним эту процедуру на примере ранее рассмотренного (7. 4) кода. Имеем $h(x)=1+x+x^2+x^4$, и схема кодирующего устройства принимает вид рис. 71. Рассмотрим работу схемы с момента, когда информационные символы уже введены в регистр. Пусть последовательность информационных символов будет 1010. В табл. 25.2 показано шаг за шагом формирование кодового вектора

Таблица 25.2

							1	0	1	0	
						1	0	1	0	0	
					1	0	1	0	0	1	$n-k$
				1	0	1	0	0	1	1	
			1	0	1	0	0	1	1	—	
		1	0	1	0	1	0	0	—	—	
	1	0	1	0	1	0	0	1	—	—	k
1	0	1	0	0	1	1	—	—	—	—	

Полученный вектор соответствует № 10 в табл. 25.1. Черточками отмечены постепенно освобождающиеся ячейки, занимаемые информационными символами следующего вектора.

Во втором варианте кодирующего устройства применяется $(n—k)$-ступенный регистр по схеме рис. 72. Действие схемы таково. Вначале ключ K_1 находится в положении *1*, а ключ K_2 замкнут. Информационные символы, подаваемые на вход, во-первых, непосредственно передаются на линию, а во-вторых, поступают в кодирующее устройство, где за k шагов образуются $n—k$ контрольных символов. После этого ключ перекладывается в положение *2*, соединяя регистр с выходом, а ключ K_2 в цепи обратной связи размыкается. Затем регистр делает еще $n—k$ шагов, выдавая контрольные символы на выход.

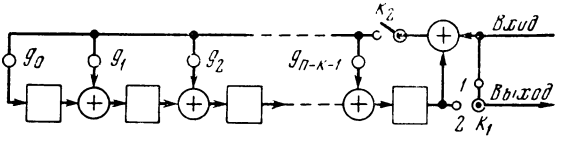


Рис. 72

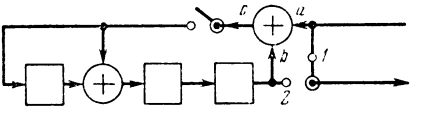


Рис. 73

Заметим, что схема рис. 72 есть (несколько видоизмененная) схема деления заданного полинома на полином $g(x)$, на чем и строится теоретическое обоснование этой схемы.

Покажем работу схемы на примере того же кода, что и ранее. Так как

$$g(x) = 1 + x + x^3,$$

то схема принимает вид, показанный на рис. 73. Пусть на вход подается последовательность информационных символов 1001. В нижеследующей табличке показано шаг за шагом формирование контрольных символов в регистре. Исходное состояние ячеек регистра определяется тремя нулями. Левый столбец — последовательность информационных символов a. Правый столбец — символы b, вытесняемые при каждом шаге из регистра. Таким образом, в цепь обратной связи поступает сумма (по модулю два) $c=a+b$

a	0	0	0	b
1	1	1	0	0
0	0	1	1	0
0	1	1	1	1
1	0	1	1	1

Следующие три шага выводят контрольные символы из регистра, образуя кодовый вектор 1001110 (№ 9 из табл. 25.1). Сравнивая оба рассмотренных варианта кодирующих устройств, можно заметить, что по сложности оборудования (определяемой

в данном случае числом триггерных ячеек) предпочтение может быть отдано тому или иному варианту в зависимости от соотношения между n и k. В нашем примере $k=4$; $n-k=3$, и второй вариант требует только трех ячеек вместо четырех по первому варианту.

Обратимся теперь к декодирующему устройству. Его главное назначение состоит в том, чтобы обнаружить ошибки и исправить их (в пределах исправляющей способности кода, разумеется).

Для исправления достаточно указать местоположение ошибочного знака. Мы и ограничимся рассмотрением этой операции, так как все остальное (т. е. собственно исправление, устранение избыточных контрольных знаков и выдача информационных знаков в надлежащей последовательности) выполняется элементарно.

В качестве основы декодирующего устройства может быть использована та же схема рис. 72, которая была выше рассмотрена в качестве кодирующей. Действие этой схемы в составе декодирующего устройства сводится к тому, что принятый вектор вводится в нее n последовательными шагами. Если ошибок нет, то содержимое регистра состоит из одних нулей. Если же в тех или иных ячейках оказываются единицы, то это указывает на наличие ошибки. Предполагая, что имеется одиночная ошибка, можно определить ее местоположение. Для этого, отключив вход, заставляют регистр делать последовательные шаги. Номер шага, на котором в регистре появляется комбинация из единицы в первой ячейке и нулей во всех остальных, и есть номер ошибочного знака.

Покажем эту процедуру на примере все того же (7; 4) кода (этот код имеет наименьшее расстояние $d=3$, а потому способен исправить одиночную ошибку). Схема показана на рис. 74. Будем сначала вводить в нее кодовый вектор без ошибки, например, № 13 из табл. 25.1.

Получим следующее:

a				b
1	1	0	0	0
1	1	1	0	0
0	0	1	1	0
1	0	1	1	1
0	1	1	1	1
0	1	0	1	1
1	0	0	0	1

Пусть теперь принята последовательность 0111001. В этом случае имеем

a				b
0	0	0	0	0
1	1	0	0	0
1	1	1	0	0
1	1	1	1	0
0	1	0	1	1
0	1	0	0	1
1	1	1	0	0

Итак, имеется ошибка. Будем теперь сдвигать образовавшееся сочетание 110

1	1	0	
0	1	1	0
1	1	1	1
1	0	1	1
1	0	0	1

Требуемое сочетание 100 образовалось в регистре на четвертом шаге. Это значит, что ошибка содержится в четвертом символе принятой последовательности. Исправленная последовательность есть 0110001 (№ 6 из табл. 25.1).

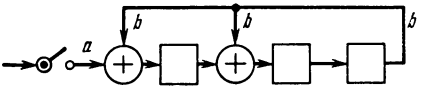


Рис. 74

Наши примеры относились к коду, исправляющему одиночную ошибку. Но циклические коды широко применяются в настоящее время для исправления так называемых серийных ошибок, или пакетов ошибок. В этом применении они оказываются особенно эффективными.

Серийные ошибки возникают в результате воздействия на канал передачи помех импульсного характера, длительность которых больше длительности одного символа. При этих условиях ошибки уже независимы; они возникают пачками, общая длительность которых соответствует длительности помехи.

Серийная ошибка (burst) длиной b определяется вектором ошибки e, в котором все единицы заключены в последовательности b символов при условии, что крайние символы этой последовательности — единицы.

Так, серийные ошибки длиной 4 могут выглядеть следующим образом ($n=10$): 0001011000, 0100100000, 0000011110 и т. п.

Конечно, всякий корректирующий код может быть использован для обнаружения и исправления серийных ошибок. Однако нужно позаботиться о том, во-первых, чтобы эффективность кода (измеряемая отношением числа информационных символов k к общему числу символов в кодовом векторе n) была как можно выше, а во-вторых, чтобы техника кодирования и декодирования была по возможности проста. Имея в виду требования для обнаружения и исправления серийных ошибок, предпочитают пользоваться специально сконструированными кодами. К числу таких кодов относятся циклические коды Файра (Fire).

Коды Файра имеют производящий полином вида

$$g(x) = p(x)(x^c + 1),$$

где $p(x)$ — непроводимый полином степени m (непроводимым называется полином, не делящийся ни на какой полином степени меньше m).

Приведем без доказательства следующие свойства кодов Файра: значность кода n есть общее наименьшее кратное показателя c и порядка корней полинома $p(x)$, равного $l=2^m-1$; число проверочных символов $n-k=c+m$ — число информационных символов $k=n-c-m$ *.

Далее код исправляет одиночную серию длиной $\leqslant b_c$ и одновременно обнаруживает серию длиной $\leqslant b_d$ при условиях

$$c \geqslant b_c + b_a - 1, \quad m \geqslant b_c.$$

Если код применяется только для обнаружения серийных ошибок, то он способен обнаружить серию длиной $\leqslant b$ при условии $c+m \geqslant b$. Приведем в качестве примера код с производящим полиномом

$$g(x) = (x^3 + x + 1)(x^{10} + 1).$$

Имеем $m=3$, $c=10$, $l=2^3-1=7$, $n=ec=70$ (так как 7 — простое число), $n-k=c+m=13$, $k=n-c-m=57$. Наибольшая длина исправляемой серии $b_c=m=3$. При этом наибольшая длина обнаруживаемой серии $b_d=c+1-b_c=8$. Если код применяется только для обнаружения, то наибольшая длина обнаруживаемой серии равна $c+m=13$.

Первая строка производящей матрицы для рассматриваемого кода имеет вид

$$1101000000110110\overbrace{0\ldots\ldots}^{57\text{ нулей}}0.$$

Последующие строки получаются, как было описано выше, путем циклической перестановки символов.

Следует заметить, что в современной практике находят применение коды такого типа с числом символов n, превосходящим 10^3.

Коды Файра сравнительно просто реализуются. Как и для всех циклических кодов, здесь применяется техника сдвигающих регистров, причем число ступеней как в кодирующем, так и в декодирующем устройствах равно числу проверочных символов, т. е. $n-k=c+m$.

§ 26. Непрерывные коды

Все рассмотренные до сих пор коды относятся к числу так называемых блочных кодов. Их строение определяется тем, что последовательность передаваемых символов разделена на блоки по n символов в каждом. Образование блока, т. е. кодирование, и обработка его на приемной стороне с целью обнаружения и исправления ошибок, т. е. декодирование, — операции, производимые с каждым блоком в отдельности.

---

* Представляет интерес общая оценка числа проверочных символов (n, k) кода, необходимого для исправления серии длиной $\leqslant b$. Вывод такой оценки дан в Добавлении IX.

За последнее время появились особого вида коды, в которых кодирование и декодирование представляют собой операции, непрерывно совершаемые над последовательностями символов. Деление на блоки при этом отсутствует. Такого рода коды мы назовем непрерывными.

Рассмотрим в качестве примера простейший из рекуррентных кодов Хагельбаргера (Habelbarger). В этих кодах проверочные символы размещены между информационными так, что на каждые n символов непрерывно передаваемой последовательности приходится k информационных символов. Введем для такого кода

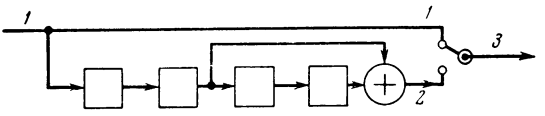


Рис. 75

обозначение (k/n). Это отношение непосредственно выражает эффективность кода.

Простейшим является код (1/2), в котором за каждым информационным символом следует проверочный символ. Этот код способен исправить серийную ошибку длиной $\leqslant b$ при условии, что две соседние серии разделены между собой защитным промежутком $\geqslant 3b+1$.

Это значит, что между последним символом данной серии и первым символом следующей серии должно быть не менее $3b+1$ безошибочных символов. Наибольшая длина серии b кратна n; таким образом, код (1/2) может исправлять серии с наибольшей длиной 2, 4, 6 . . .

Возьмем в качестве примера код (1/2), исправляющий серийные ошибки длиной $\leqslant 4$. Схема кодирующего устройства показана на рис. 75. На вход подается последовательность информационных символов. Синхронный коммутатор выдает на выход поочередно информационные символы и контрольные символы, вырабатываемые сдвигающим регистром со «связью вперед». Число ступеней в этом регистре равно $b=4$. Рассмотрим действие кодирующего устройства подробно. Пусть на вход подается последовательность информационных символов [1]

$$10110111001\ldots \qquad (1)$$

В регистре с учетом связи и суммирования образуется последовательность проверочных символов

$$00100110101\ldots \qquad (2)$$

[1] Различные последовательности символов, упоминаемые далее, обозначены курсивными цифрами. Эти же обозначения сохранены и на соответствующих рисунках.

Образование этой последовательности удобно проследить при помощи таблички:

0	0	0	0	
1	0	0	0	0
0	1	0	0	0
1	0	1	0	1
1	1	0	1	0
0	1	1	0	0
1	0	1	1	1
1	1	0	1	1
1	1	1	0	0
0	1	1	1	1
0	0	1	1	0
1	0	0	1	1

В левом столбце — последовательность информационных символов (1). В правом столбце последовательность проверочных сим-

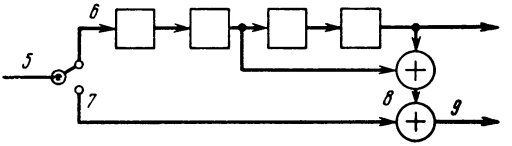


Рис. 76

волов (2), элементы которой получаются суммированием второй и четвертой позиций предыдущей строки.

Коммутатор выдает на выход последовательность, состоящую из первого информационного символа, затем первого проверочного, затем второго информационного и т. д. Выходная последовательность имеет вид

$$10001110001111100100 11. \qquad (3)$$

Это и есть последовательность информационных символов, закодированная рекуррентным кодом (1/2).

Обратимся к декодирующему устройству. Оно состоит из двух частей. Первая часть вырабатывает исправляющую последовательность, во второй — производится самое исправление.

Схема для получения исправляющей последовательности показана на рис. 76. При помощи такого же синхронного коммутатора, как на рис. 75, принятая последовательность (5) разделяется на последовательность информационных символов (6) и последовательность проверочных символов (7). Первая поступает на регистр, в точности повторяющий схему для получения последовательности проверочных символов в кодирующем устройстве рис. 75. Поэтому если ошибок нет, то выработанная регистром последовательность (8) в точности совпадает с последовательностью (7) проверочных символов, и их суммирование дает последовательность (9), состоящую из одних нулей. Если же имеются ошибки, то последовательность (9) и есть исправляющая последовательность. Рассмотрим ее строение.

Возьмем наихудший случай — серию длиной b. Заметим прежде всего, что такая ошибка поражает только $b/2$ информационных символов и столько же контрольных. Начнем с момента, когда первый ошибочный символ серии появляется на входе схемы рис. 76. Регистр содержит пока безошибочные информационные символы. Поэтому первые $b/2$ шагов дают в последовательности (9) положение ошибок в проверочных символах. На этом серия заканчивается, и последовательность (7) содержит в дальнейшем лишь безошибочные проверочные символы. В то же время следующие $b/2$ шагов выводят на выход ошибочные информационные символы из первого полурегистра, так что в последовательности (9) отображается положение соответствующих ошибок. Следующие $b/2$ шагов повторно выводят ошибочные информационные символы, на этот раз уже из второго полурегистра.

Итак, последовательность (9) содержит:

1. Единицы на местах ошибок в проверочных символах.
2. Со сдвигом на $b/2$ символов — единицы на местах ошибок в информационных символах.
3. То же, что и 2 — со сдвигом еще на $b/2$ символов.

Покажем все это на примере. Пусть произошла серийная ошибка

$$00000101100000000000\ldots \tag{4}$$

Складывая (3) и (4), получим принятую последовательность

$$10001011101111100100 11\ldots \tag{5}$$

Коммутатор разделяет (5) на информационные символы

$$1011\underline{1}111001\ldots \tag{6}$$

и проверочные символы

$$0001\underline{0}110101. \tag{7}$$

Последовательности (6) и (7) содержат ошибочные символы (ср. (1) и (2)); эти символы подчеркнуты. Регистр схемы рис. 76 выдает последовательность

$$00100100001\ldots, \tag{8}$$

которая в сумме с (7) дает исправляющую последовательность

$$00110010100\ldots \tag{9}$$

На основании сказанного выше эту последовательность можно представить в виде следующей суммы:

$$\begin{array}{l} 00110000000\ldots \\ \ldots000010000\ldots \\ \ldots\ldots0000100\ldots \\ \hline 00110010100\ldots \end{array}$$

Таким образом, последовательность информационных символов (6) содержит ошибку в пятом символе. Этого уже достаточно для исправления ошибки. Остается автоматизировать процесс исправления. Это достигается при помощи схемы рис. 77, представляющей собой непосредственное продолжение схемы рис. 76. Действие исправляющей схемы мы опишем без пояснений.

Ячейка «НЕ» заменяет в (9) единицы нулями и наоборот. Последовательности (11) и (12) представляют собой результат сдвига (9) на $b/2$ и на b символов соответственно. Ячейка «И» выдает на выход единицу только тогда, когда на все три ее входа поступают единицы. Мы имеем

$$11001101011\ldots \qquad (10)$$

$$..001100101\ldots \qquad (11)$$

$$....0011001\ldots \qquad (12)$$

$$....000000100000\ldots \qquad (13)$$

Остается прибавить (13) к должным образом сдвинутой последовательности информационных символов (14) и получить исправленную последовательность (15)

$$\ldots0000001000000\ldots \qquad (13)$$

$$\ldots10111111001\ldots \qquad (14)$$

$$10110111001\ldots \qquad (15)$$

Как видим, на пути информационных символов имеется всего $3b/2$ регистровых ячеек. Это соответствует $3b$ символам в передаваемой последовательности. Чтобы вывести все ошибочные символы из схемы, требуется защитный промежуток из $3b+1$ символов.

Сложность оборудования для рассмотренного кода легко оценивается числом регистровых ячеек. Именно: кодирующее устройство содержит b ячеек, а декодирующее устройство $2{,}5b$ ячеек, из которых к исправляющей схеме относятся $1{,}5b$ ячеек.

Мы рассмотрели (1/2) рекуррентный код. Он чрезвычайно прост как по идее, так и по технической реализации. Его недостаток — малая эффективность, равная 1/2. Но возможно на основании той же общей идеи строить рекуррентные коды $(n-1)/n$, т. е. коды со сколь угодно близкой к единице эффективностью. Схемы кодирующих и декодирующих устройств для этих кодов усложняются. В частности, входная последовательность информационных символов при помощи коммутатора расщепляется на $n-1$ последовательностей. Не разбирая устройства и действия этих схем, приведем лишь данные, позволяющие оценить их сложность. Пусть $L(n)$ есть $1+\log n$ или ближайшее большее целое. Тогда

число регистровых ячеек в кодирующем устройстве (речь идет, конечно, о некоторых конкретных схемах) составляет

$$\frac{b}{n}[L(n)(n-1)+1],$$

а в декодирующем устройстве

$$\frac{b}{n}[(n^2-n+1)L(n)-n].$$

Защитный промежуток при применении $((n-1)/n)$ кода должен быть не менее $nbL(n-1)$ символов. Так, например, для кода (4/5) при $b=5$ число ячеек кодирующего устройства равно 17, декодирующего 79 ($L(n)=4$). Защитный промежуток $\geqslant 99$ символов. Как видим, повышение эффективности до значения 0,8 покупается ценой довольно значительного усложнения схем. Для сравнения заметим, что код (1/2) с эффективностью 0,5 требует при $b=6$ всего шесть ячеек в кодирующем устройстве и 15— в декодирующем при защитном промежутке всего 19 символов. Таким образом, только специальные требования в отношении эффективности могут заставить отдать предпочтение коду $((n-1)/n)$ перед простейшим кодом (1/2).

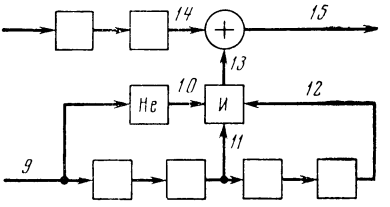


Рис. 77

Описанные выше рекуррентные коды могут с успехом применяться не для исправления, а лишь для обнаружения серийных ошибок. Такое использование кода целесообразно в сочетании с применением того или иного вида обратной связи (см § 21). Обнаружение основано на выработке исправляющей последовательности. Как уже говорилось, при отсутствии ошибок эта последовательность состоит из одних нулей. Она не будет содержать единиц также и в случае необнаруженных ошибок.

Рассмотрим вкратце возможность обнаружения серийных ошибок. Схема передачи, состоящая из кодирующего и декодирующего устройств, показана в общем виде на рис. 78. Декодирующее устройство имеет два выхода: на один поступают информационные символы, на второй — исправляющая последовательность, по наличию единиц в которой можно заключить о наличии ошибок. В схемах кодирующего и декодирующего устройств имеются идентичные преобразователи, вырабатывающие последовательность проверочных символов. Передаточная функция этих преобразователей обозначена через $F(x)$. Положим для простоты (можно показать, что это нисколько не влияет на общность результата), что исходная последовательность информационных символов состоит из одних нулей. Тогда принятая последовательность будет содержать единицы только за счет ошибок.

Так как серийная ошибка охватывает конечное число символов, то ее можно представить полиномом. Обозначим вектор ошибки, заданный в форме полинома, через $E(x)$, причем пусть $E_и$ означает сочетание ошибок в информационных символах, $E_п$ — в проверочных. Если

$$E_п = FE_и, \tag{26.1}$$

то исправляющая последовательность состоит только из нулей (см. схему декодирующего устройства), следовательно, ошибка остается необнаруженной.

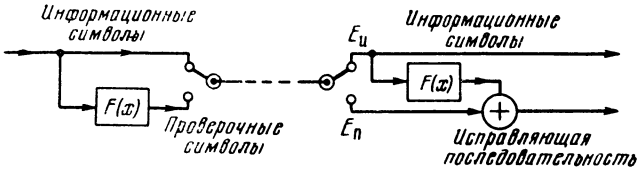


Рис. 78

Пусть $F(x)$ выражено отношением двух взаимно простых полиномов, т. е.

$$F(x) = \frac{f(x)}{g(x)}, \tag{26.2}$$

причем степень полинома $f(x)$ равна s, а полинома $g(x)$ равна r.

Тогда из (26.1) и (26.2) получаем

$$E_п g = E_и f \tag{26.3}$$

или

$$\frac{E_п}{f} = \frac{E_и}{g}, \tag{26.4}$$

т. е. $E_п$ должно делиться на f, а $E_и$ — на g. Это возможно лишь при условии, что степени $E_п$ и $E_и$ не меньше s и r соответственно. Но степени $E_п$ и $E_и$ — это не что иное, как уменьшенные на единицу числа ошибочных символов в последовательностях проверочных и информационных символов соответственно.

Отсюда следует заключение: описанная система обнаруживает все серии, поражающие не более r информационных символов и не более s проверочных символов.

Но нужно заметить, что даже при невыполнении этого условия ошибка может быть обнаружена. Для этого, чтобы ошибка осталась незамеченной, необходимо выполнение равенства (26.1) (или равносильных ему (26.3) или (26.4)). Таким образом, мы определили лишь нижнюю грань обнаруживающей способности кода.

Приведем пример. Пусть

$$F(x) = \frac{1}{x^5 + x^4 + 1},$$

т. е.

$$f(x)=1, \quad g(x)=x^5+x^4+1.$$

Получаемый код способен обнаружить все серийные ошибки, поражающие не более пяти соседних информационных символов. Схема кодирующего и декодирующего устройств для данного частного случая показана на рис. 79, *а* и *б*. Защитный промежуток определяется емкостью регистра. Так как на каждый информационный символ приходится один проверочный, то длина защитного промежутка составляет 10 символов.

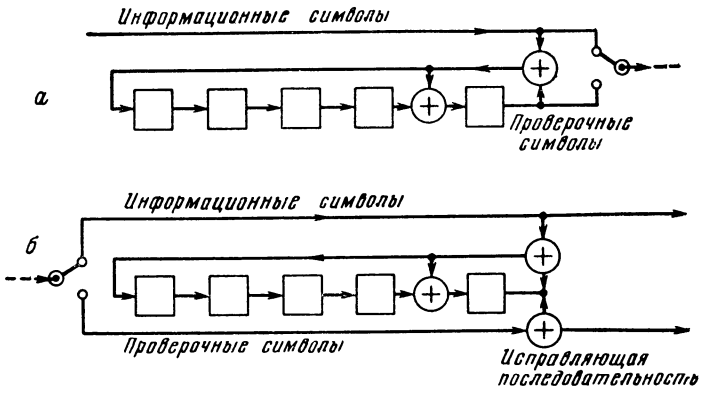


Рис. 79

Нужно еще заметить, что специальным выбором полиномов $f(x)$ и $g(x)$ можно установить связь рекуррентного кода с некоторым циклическим кодом и сделать на этом основании дополнительные заключения об обнаруживающей способности.

§ 27. Корректирующие коды (обзор)

Усилия многочисленных исследователей, занимающихся проблемой корректирующих кодов, привели за последнее время к созданию большого числа различных кодов. Исследования продолжаются; количество публикаций непрерывно возрастает; растет список известных кодов. Надо полагать, что этот процесс вскоре замедлится, так как для каждой области применения будут найдены относительно наилучшие коды, наилучшие как с точки зрения наибольшей эффективности, так и по простоте технической реализации.

Однако пока такой подбор еще не произошел, и для того, чтобы ориентироваться в многообразии известных в настоящее время кодов, необходимо стремиться к установлению некоторого порядка, облегчающего обозрение. С этой целью мы рассмотрим классификацию, схема которой изображена на рис. 80. Ниже следуют пояснения по признакам деления и по некоторым разновидностям кодов. Рассматриваются только двоичные коды, хотя многие коды предложены в общем виде (т. е. для любого основания).

Для удобства здесь повторены вкратце сведения о кодах, уже упоминавшихся в предыдущих параграфах.

Все корректирующие коды (речь идет, понятно, о кодах, известных в настоящее время; эта оговорка относится ко всему последующему) можно разделить на два класса — блочные и непрерывные. Последние появились совсем недавно и быстро развиваются.

Блочные коды — коды, в которых каждому сообщению (или элементу сообщения) сопоставляется блок из n символов (кодовая комбинация, кодовый вектор). Возможны неравномерные блочные коды, имеющие блоки различной длины; такие коды возникают при статистическом кодировании, но за последнее время ими мало занимаются — основное внимание уделено равномерным кодам. Возможность обнаружения и исправления ошибок основана на том, что число кодовых векторов меньше, чем $N_0=m^n$, где m — основание; n — значность кода, т. е. длина блока. Ошибка обнаруживается, если принятый вектор не принадлежит коду.

Непрерывные коды (называемые также рекуррентными, конволюционными или ценными) представляют собой непрерывную последовательность символов, не подразделяемую на блоки. Процессы кодирования и декодирования также имеют непрерывный характер. Передаваемая последовательность образуется путем размещения в определенном порядке проверочных символов между информационными символами исходной последовательности.

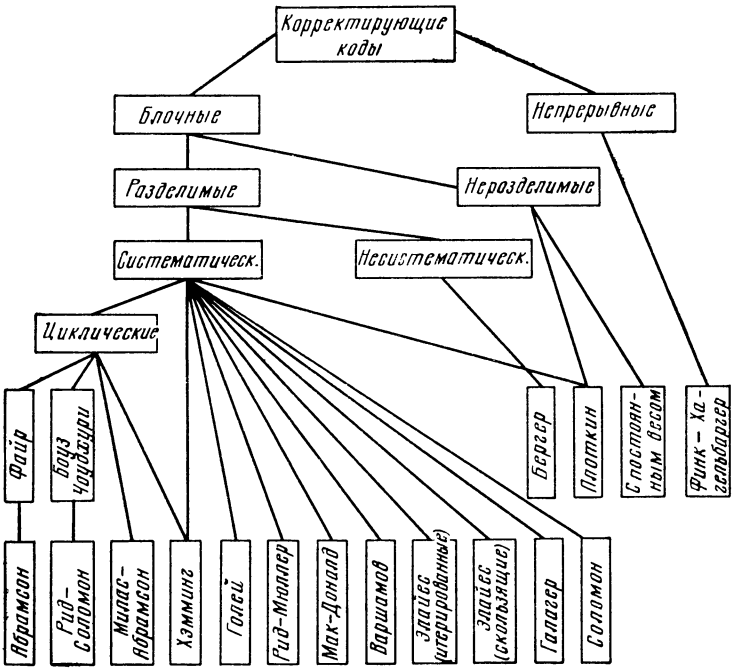


Рис. 80

Как блочные, так и непрерывные коды делятся на разделимые и неразделимые.

Разделимыми (separable) называются такие коды, в которых роль символов, входящих в состав блока (или непрерывной последовательности), может быть отчетливо разграничена. Одни символы, основные, являются информационными, другие же, добавляемые по тем или иным правилам, являются проверочными и служат для обнаружения и исправления ошибок.

Информационные и проверочные символы занимают во всех кодовых векторах одни и те же позиции. Обычное обозначение разделимых кодов (n, k) — коды, где n — значность кода: k — число информационных символов. Число кодовых векторов равно 2^k.

Неразделимые коды образуют в настоящее время немногочисленную группу. К ней относятся коды с постоянным весом и коды Плоткина.

Коды с постоянным весом — простые блочные коды, широко применяемые на практике. Суть дела определяется названием: все кодовые векторы этих кодов имеют одинаковый вес, т. е. одинаковое число единиц. Изменение этого числа указывает на наличие ошибки. Эти коды обычно используются лишь для обнаружения ошибок. Примером является известный код «3 из 7» — семизначный код с постоянным весом 3.

Коды Плоткина (Plotkin, 1960) отличаются тем, что по эффективности достигают теоретического предела Плоткина. Эти коды получаются из матриц Адамара. Матрица Адамара — квадратная матрица из элементов 1 и —1 с ортогональными строками. С помощью этих матриц могут быть построены коды с большим кодовым расстоянием, а именно (в зависимости от четности):

$$d \geqslant \begin{matrix} \frac{1}{2} n, \\ \frac{1}{2}(n-1). \end{matrix}$$

Если $d=2p$, $2d=n=4p$, то возможное число кодовых векторов равно $2n=8p$. Такой код образуется строками матрицы Адамара порядка n при замене — 1 на 0, а затем строками, получаемыми заменой всех единиц нулями, а нулей — единицами. Если $n=4p=2^r$, то код является групповым и совпадает с кодом Рида-Мюллера первого порядка (см. ниже). Если $n=4p$ не является степенью двух, то код не является групповым. Порядок m матриц Адамара должен быть кратен четырем, т. е. $m=4k$. Значения m, для которых доказано существование таких матриц (до $m=1000$), имеются в литературе.

Коды Плоткина имеют высокую исправляющую способность (большое d). Однако нет практических методов кодирования и декодирования, так что прикладное значение этих кодов пока невелико.

Разделимые коды делятся на систематические и несистематические. К числу несистематических разделимых кодов относятся коды Бергера (Berger, 1960) или коды с суммированием (Sum Code). Метод построения этих кодов состоит в том, что проверочные символы представляют запись суммы подблоков длиной l, на которые разделена последовательность информационных символов. При таком построении код способен обнаружить серийные ошибки с длиной серии, не превосходящей l. В другом варианте проверочные знаки представляют собой двоичную запись веса последовательности информационных символов. Этот код предназначен для обнаружения независимых ошибок. Коды Бергера с успехом могут применяться в асимметричных каналах.

Наиболее обширный класс среди разделимых кодов образуют систематические или линейные коды, являющиеся групповыми. (В двоичном случае всякий групповой код является систематическим. Название «групповой код» обусловлено тем, что код является групповым по отношению к операции сложения по модулю 2 (т. е. сумма mod 2 любых двух кодовых векторов также является кодовым вектором).

Основным отличием систематических кодов является то, что проверочные символы представляют собой различные линейные комбинации информационных символов. Этим определяется и способ декодирования, основанный на проверках линейных соотношений между символами, определяемых строением кода. В двоичном случае эти проверки сводятся к проверкам на четность (так как линейная комбинация символов дает либо 0, когда число единиц четно, либо 1 в противном случае). Этим обусловлено другое название систематических кодов: коды с проверкой на четность (parity check Codes). Совокупность результатов проверок позволяет обнаружить ошибки, определить ошибочные позиции и исправить ошибки.

Метод построения систематических кодов основан на применении производящей матрицы, строки которой — линейно-независимые базисные векторы. Все кодовые векторы получаются как линейные комбинации строк производящей матрицы (по модулю 2).

Ниже перечислены некоторые разновидности систематических кодов.

Кодами Хэмминга (Hamming, 1950) называются обычно

1) коды с расстоянием $d=3$, исправляющие все одиночные ошибки, и

2) коды с расстоянием $d=4$, исправляющие все одиночные ошибки и обнаруживающие двойные. Коды первого вида имеют проверочную матрицу, $n=2^r-1$ столбцов которой представляют собой все возможные ненулевые r-значные векторы. Коды второго вида получаются из первого добавлением одного проверочного символа, равного сумме (по модулю 2) всех остальных символов.

Метод декодирования — проверки на четность. Число проверок равно числу проверочных символов, т. е. $r=n-k$.

Код Голея (Golay, 1949)—(23; 12) — код, исправляющий все одиночные, двойные и тройные ошибки ($d=7$).

Код Рида—Мюллера (Reed, Muller, 1954) имеет следующие данные:

$$n=2^m, \quad k=1+\sum_{i=1}^{r} C_m^i, \quad n-k=1+\sum_{i=1}^{m-r-1} C_m^i, \quad d=2^{m-r},$$

где m и $r < m$ — произвольные числа; r называется порядком кода.

Код строится на основе матрицы с m строками, столбцы которой представляют собой все m-значные двоичные числа. Остальные кодовые векторы получаются перемножением строк во всех комбинациях по две, по три, . . ., по r. При $r=m-2$ получается код Хэмминга $d=4$. Декодирование производится на основании линейных соотношений между исходными информационными символами и составляющими кодового вектора. Всего для каждого информационного символа имеется 2^{m-r} независимых соотношений. Каждая ошибка нарушает лишь одно из этих соотношений, так что решение принимается «большинством голосов». Таким образом, могут быть исправлены все ошибки кратности $\leqslant 2^{m-r-1}-1$.

Код Мак-Доналда (Mac Donald, 1960) интересен тем, что он имеет наибольшее расстояние, соответствующее пределу Плоткина,

$$d \leqslant n \frac{2^{k-1}}{2^k-1}.$$

Используя свойства матрицы

$$C=M^T M$$

(где M — матрица k (2^k-1), столбцы которой содержат все возможные k-значные векторы, кроме нулевого, M^T — обращенная матрица, получаемая из M переменой мест строк и столбцов), можно показать, что наибольшим возможным расстоянием обладают коды, приведенные ниже:

n	2^k-1	2^k-2	2^k-3	2^k-2^u	2^k-2^u-1
d	2^{k-1}	2^{k-1}	$2^{k-1}-2$	$2^{k-1}-2^{u-1}$	$2^{k-1}-2^{u-1}-1$.

Здесь $u=2, 3, 4, \ldots, k-1$.

Код Варшамова (1957) строится по заданному расстоянию. Его производящая матрица имеет вид

$$G=(\lambda_k G'),$$

где λ_k — единичная матрица информационных символов; G' — подматрица, каждая строка которой содержит не менее $d-1$ еди-

ниц, а сумма любых i строк содержит не менее $d-i$ единиц. Число $r=n-k$ проверочных символов определяется из условия

$$1+\sum_{i=1}^{d-2} C_{n-1}^{i}<2^{r}.$$

Так, для $n=5$, $d=3$ имеем $2^r>5$, откуда $r=3$, $k=n-r=2$, и производящая матрица может иметь вид

$$G=\begin{pmatrix}10\ 011\\ 01\ 111\end{pmatrix}.$$

Итерированные (iterated) коды Элайеса (Elias, 1954) отличаются применением нескольких систем проверок. В простейшем случае все комбинации информационных символов записываются в прямоугольную таблицу. Одна проверка производится по строкам, вторая — по столбцам, например,

Информационные символы		
1101	1	
1010	0	Проверка строк
0110	0	Проверка строк
1011	1	
Проверка столбцов: 1010	0	Проверка проверок

Проверке подвергаются также и проверочные знаки. Так как определяются столбец и строка, в которой находится ошибочный символ, то такой код способен исправить все одиночные ошибки. Эта идея может быть обобщена на многомерные таблицы проверок.

Элайес показал, что итерированные коды обладают замечательным свойством (в отличие от всех других известных кодов), а именно: с увеличением значности вероятность ошибки стремится к нулю, тогда как скорость передачи стремится к конечному пределу (это доказано для симметричного двоичного канала).

Скользящие (Sliding parity check) коды Элайеса (Elias, 1955) имеют проверочную матрицу

$$H=(H'\lambda_r),$$

где λ_r — единичная матрица проверочных символов. Когда коэффициенты подматрицы H' берутся из последовательности $n-1$ двоичных цифр, $a_1, a_2, \ldots, a_{n-1}$. Первые k цифр образуют первую строку H', вторая строка составляется из k последовательных цифр, начиная со второй, т. е. $a_2, a_3, \ldots, a_{k+1}$ и т. д. Например, для (7.4) кода

$$H=\begin{pmatrix}a_1 & a_2 & a_3 & a_4 & 100\\ a_2 & a_3 & a_4 & a_5 & 010\\ a_3 & a_4 & a_5 & a_6 & 001\end{pmatrix}.$$

Таким образом, (n, k) код задается $n-1$ двоичными цифрами (вместо k^2). Фэно (Fano, 1961) называет эти коды конволюционными (convolution codes).

Коды Галагера (Gallager, 1962), названные автором кодами с малой плотностью проверок на четность (Lowdensity parity-check codes), декодируются при помощи проверок на четность, каждая из которых охватывает фиксированное небольшое число v принимаемых символов, а каждый символ участвует также в фиксированном и небольшом числе проверок u ($3 \leqslant u < v$). Проверочная матрица кода Галагера (обозначаемого (n, u, v)) содержит u единиц в каждом столбце и v единиц в каждой строке. Матрица разделена на u подматриц, каждая из которых имеет n/v строк. Первая подматрица имеет следующее строение:

$$\overbrace{111\ldots1}^{v}0000\ldots$$

$$\overbrace{000\ldots0}^{v}\,\overbrace{111\ldots1}^{v}0000\ldots$$

$$\overbrace{000\ldots0}^{2v}\,\overbrace{111\ldots1}^{v}000\ldots$$

$$\cdots\cdots\cdots\cdots$$

Остальные подматрицы получаются из первой случайной перестановкой столбцов (поэтому коды Галагера можно отнести к случайным, а не к конструктивным). Скорость передачи не менее $1-u/v$. Кодовое расстояние (после некоторого достаточно большого n) растет с увеличением n линейно; иначе говоря, отношение d/n при $n \to \infty$ стремится к постоянному пределу (тогда как у всех других известных систематических кодов это отношение стремится к нулю).

Процедура декодирования своеобразна. Она выполняется последовательными шагами. После первой проверки исправляется (т. е. заменяется на обратный) символ, входящий в наибольшее число невыполненных проверочных соотношений, после чего проверка производится повторно, пока все проверочные соотношения не будут удовлетворены.

Большую и важную группу систематических кодов составляют циклические коды. Их основное свойство состоит в том, что каждый вектор, получаемый из кодового вектора путем циклической перестановки составляющих, также является кодовым вектором. Принято описывать циклические коды при помощи порождающих полиномов $g(x)$ степени $r=n-k$. Умножение на x соответствует циклической перестановке на один шаг. При этом x^r заменяется на единицу. Это записывается так:

$$g(x) = g_0 + g_1 x + g_2 x^2 + \ldots + g_{r-1} x^{r-1},$$

$$xg(x) = g_{r-1} + g_0 x + g_1 x^2 + \ldots + g_{r-2} x^{r-1}.$$

Запись $g(x)=g_0+g_1x+\ldots+g_{r-1}x^{r-1}$ соответствует вектору $g=(g_0, g_1, \ldots, g_{r-1})$. В двоичном случае g_k могут принимать значения нуль или единица, так что, например, $g(x)=1+x^2+x^4+x^5$ соответствует вектору 101011. В качестве первой строки порождающей матрицы записываются составляющие вектора $g(x)$ с добавлением k нулей; остальные строки получаются из первой вышеописанной циклической перестановкой.

Операции кодирования и декодирования сводятся к умножению и делению полиномов по правилам двоичной арифметики. Эти операции легко реализуются технически при помощи сдвигающих регистров, составленных из триггерных ячеек. Получаются относительно простые схемы, в чем состоит одно из практических достоинств циклических кодов.

Циклические коды с успехом применяются как для обнаружения и исправления независимых ошибок, так и в особенности для обнаружения и исправления серийных ошибок.

К числу циклических относятся коды Файра (Fire, 1959). Эти коды порождаются полиномом

$$g(x)=p(x)(x^c+1),$$

где $p(x)$ — непроводимый полином степени m. Значность кода n равна общему наименьшему кратному показателя c и $e=2^m-1$, число проверочных символов $n-k=c+m$, число информационных символов $k=n-c-m$. Код предназначен для обнаружения и исправления серийных ошибок. Он исправляет одиночную серию длиной $\leqslant b_c$ и одновременно обнаруживает серию длиной $\leqslant b_d$ $(b_d > b_c)$ при условиях

$$c \geqslant b_c + b_d - 1, \quad m \geqslant b_c.$$

Если код применяется только для обнаружения серийных ошибок, то он способен обнаружить серию длиной $\leqslant b$ при условии

$$c+m \geqslant b.$$

Частным случаем кодов Файра являются коды Абрамсона (Abramson, 1959). Порождающий полином этих кодов имеет вид

$$g(x)=p(x)(x+1)$$

(т. е. коды Абрамсона совпадают с кодами Файра, если положить $c=1$). Значность кодов $n=2^m-1$, число проверочных символов $n-k=m+1$. Эти коды исправляют все одиночные и смежные двойные ошибки (т. е. серии длиной 2). Коды Абрамсона заслуживают отдельного упоминания потому, что они являются первыми циклическими кодами, исправляющими серийные ошибки.

Наилучшими являются в настоящее время циклические коды Боуза—Чоудхури (Bose, Chaudhuri, 1960). При описании построения этих кодов не удается избежать ссылок на общую алгебру.

Производящий полином $g(x)$ двоичного кода Боуза—Чоудхури имеет в качестве корней элементы поля Галуа $GF(2^m)$, представляемые последовательными степенями некоторого первообразного элемента α. Четные степени элемента α могут быть отброшены, так что последовательность корней полинома $g(x)$ имеет вид

$$\alpha,\ \alpha^3,\ \alpha^5,\ \ldots,\ \alpha^{2t-1}.$$

Полином, удовлетворяющий этому условию, порождает код, исправляющий все ошибки кратности $\leqslant t$.

Для построения полинома $g(x)$ можно образовать минимальные полиномы (или минимальные функции) $m_i(x)$ ($i=1, 3, 5, \ldots, 2t-1$). Минимальные полиномы определяются системой своих корней, которая имеет следующий вид

$$\alpha^i,\ \alpha^{2i},\ \alpha^{4i},\ \alpha^{8i},\ \ldots,\ \alpha^{2^S i}\ \ldots$$

Производящий полином есть общее наименьшее кратное минимальных полиномов до порядка $2t-1$ включительно. Степень $m_i(x)$ не выше m, так что степень $g(x)$ не выше mt. Таким образом, число проверочных символов не превосходит mt, а число информационных символов

$$k \geqslant 2^{m-1} - 1 - mt.$$

Так, например, при $m=4$, $n=2^4-1=15$.

$$m_1(x) = 1 + x + x^4,$$

$$m_3(x) = 1 + x + x^2 + x^3 + x^4,$$

$$m_5(x) = 1 + x + x^2,$$

$$g(x) = 1 + x + x^2 + x^4 + x^5 + x^8 + x^{10},$$

$$t = 3, \quad n - k = 10 < mt = 12.$$

Коды Боуза—Чоудхури можно строить также, беря за основу непервообразный элемент, например $\alpha_1 = \alpha^\nu$, где α — первообразный элемент. При этом произведение νn_1 кратно $2^m - 1 = n$, так что $n_1 < n$.

Для декодирования кодов Боуза—Чоудхури может применяться специально разработанная процедура определения местоположения ошибок.

Мы рассматривали двоичные коды Боуза—Чоудхури. В общем случае (для кодов с основанием q) коды образуются из элементов поля Галуа $GF(q^m)$. Если положить $m=1$, т. е. строить код из элементов $GF(q)$, то производящий полином будет иметь корни $\alpha,\ \alpha^2,\ \ldots,\ \alpha^{d-1}$, где d — кодовое расстояние. Минимальные полиномы вырождаются в разности $x-\alpha^i$, так что производящий полином принимает вид

$$g(x) = (x-\alpha)(x-\alpha^2)\ldots(x-\alpha^{d-1}).$$

Такие коды носят название Рида—Соломона (Reed, Solomon, 1960).

Двоичный код Рида—Соломона получится, если взять $q=2^s$. Это значит, что каждый элемент заменяется s-значной двоичной последовательностью. Если исходный код с основанием q исправляет ошибки кратности $\leqslant t$, то полученный из него двоичный код имеет $2ts$ проверочных символов (по $2t$ на каждый блок из s символов) из общего числа $n=s(2^s-1)$. Код может исправлять серийные ошибки длиной $\leqslant b=s(t-1)+1$. Так, например, при $s=5$, $t=3$ имеем $n=5(2^5-1)=155$, $n-k=2\cdot5\cdot3=30$, $b=5\cdot2+1=11$. Коды Рида—Соломона наряду с кодами Файра представляются в настоящее время наиболее подходящими для исправления серийных ошибок.

К циклическим относятся также коды Миласа—Абрамсона. Милас (Melas, 1960) показал, что исправляющие серийные ошибки порождаются полиномом, представляющим собой произведение двух неприводимых полиномов. Частный случай таких кодов, для которых

$$g(x)=(1+x+x^2)\,p(x),$$

где $p(x)$ — неприводимый полином четной степени >2, рассмотрен Абрамсоном (1960). Эти коды исправляют серии длиной $\leqslant 3$.

Упоминавшиеся выше коды Хэмминга также могут рассматриваться как циклические коды, порождаемые неприводимым полиномом $p(x)$.

В дополнение обзора циклических кодов следует упомянуть об укороченных циклических кодах. Они образуются путем приравнивания нулю некоторого количества информационных знаков (начиная с первого) и обладают всеми свойствами исходных циклических кодов в отношении исправления и обнаружения ошибок. Математически укороченные циклические коды описываются порождающими полиномами $g(x)$ по модулю некоторого полинома $f(x)$ степени n (n — длина укороченного кода), отличного от $1+x^n$ (по этому признаку их называют псевдоциклическими кодами).

В заключение нам остается вернуться к непрерывным кодам. Как уже говорилось, эта группа кодов возникла в самое последнее время и пока еще немногочисленна. Мы ограничимся упоминанием о кодах Финка—Хагельбаргера (Финк, 1955; Hagelbarger, 1959). В этих кодах на каждые n символов приходится $k=n-1$ информационных символов. Напомним, что речь идет о непрерывном коде, не подразделяемом на блоки, так что n — произвольное число. Поэтому код целесообразно обозначать отношением $(n-1)/n$, которое, кстати говоря, выражает скорость передачи. Простейший код (1/2) * имеет один проверочный символ на каждый информационный. Код предназначается специально для испра-

* Такой код был предложен Л. М. Финком в 1955 г.

вления серийных ошибок; он способен исправить любую серию длиной $\leqslant b$. Кодирование и декодирование производятся простыми схемами со сдвигающими регистрами. Задание длины серии b определяет емкость регистра. Так, для кода (1/2) число ячеек регистра в кодирующем устройстве равно b, а в декодирующем устройстве (включая схему исправления ошибок) — $2{,}5b$. Следующие друг за другом серии исправляются при условии, что между ними имеется защитный промежуток, т. е. определенное число безошибочных символов. Для кода (1/2) длина защитного промежутка $\geqslant 3b+1$. Можно осуществить коды $(n-1)/n$ со скоростью передачи, сколько угодно близкой к единице. Однако повышение скорости влечет за собой быстрое увеличение числа ячеек в регистрах и, кроме того, возрастание защитного промежутка.

Коды описанного типа могут применяться также только для обнаружения ошибок. В этом случае соответственно упрощается схема декодирующего устройства — она повторяет схему кодирующего устройства. Для кода (1/2), применяемого для обнаружения ошибок, как кодирующая, так и декодирующая схемы имеют регистры из b ячеек.

С помощью непрерывных кодов могут исправляться и независимые ошибки. Коды такого назначения были описаны Элайесом (1955) под названием конволюционных (Convolutional codes). Разработка таких кодов (под названием рекуррентных) продолжена Килмером (Kilmer, 1960—1961).

В работе 1961 г. Килмер приводит пример непрерывного кода, обеспечивающего наименьшую вероятность ошибки, какую только можно получить с помощью систематического кода с заданным интервалом взаимных связей между символами (в случае блочного кода этим интервалом является просто длина блока).

Все вышеописанные коды представляют собой определенные алгебраические структуры; методы декодирования этих кодов являются по существу алгебраическими.Но за последнее время возникло и развивается новое направление в области кодирования и декодирования. Это направление опирается на вероятностные идеи. Сюда относятся последовательное декодирование Возенкрафта и Рейфена (Wozencraft, Reiffen, 1961), пороговое декодирование Мэсси (Massey, 1963). Обстоятельное изложение новых идей и результатов вероятностного декодирования не укладывается в рамки нашего очерка.

ДОБАВЛЕНИЯ

I. Дисперсия интеграла от белого шума с ограниченной полосой. Требуется найти дисперсию интеграла

$$\eta = \int_0^T \xi(t)\,dt.$$

Процесс $\xi(t)$ имеет постоянную спектральную плотность в полосе частот $0 < \omega < \omega_c$, т. е.

$$G(\omega) = \begin{matrix} G_0 & [0 < \omega < \omega_c], \\ 0 & [\omega_c < \omega < \infty]. \end{matrix}$$

Функция корреляции

$$B(\tau) = \int_0^\infty G(\omega)\cos\omega\tau d\omega$$

равна в данном случае

$$B(\tau) = G_0 \int_0^{\omega_c} \cos\omega\tau d\omega = G_0 \frac{1}{\tau}\sin\omega_c\tau$$

или

$$B(\tau) = G_0\omega_c \frac{\sin\omega_c\tau}{\omega_c\tau} = P\frac{\sin\omega_c\tau}{\omega_c\tau},$$

где $P = G_0\omega_c$ — мощность процесса $\xi(t)$. Выражение для дисперсии имеет вид

$$D\eta = \int_0^T dt \int_0^T B(t - t_1)\,dt_1.$$

Подставляя сюда найденное выражение для $B(t)$, имеем

$$D\eta = P\int_0^T dt \int_0^T \frac{\sin\omega_c(t - t_1)}{\omega_c(t - t_1)}\,dt_1.$$

Заменим переменную, положив

$$x = \omega_c(t - t_1).$$

Тогда

$$D\eta = \frac{P}{\omega_c}\int_0^T dt \int_{\omega_c(t-T)}^{\omega_c t} \frac{\sin x}{x}\,dx = \frac{P}{\omega_c}\left\{\int_0^T Si(\omega_c t)\,dt - \right.$$

$$\left. - \int_0^T Si[\omega_c(t - T)]\,dt\right\} = \frac{P}{\omega_c}\left[\int_0^T Si(\omega_c t)\,dt + \int_0^{-T} Si(\omega_c t)\,dt\right] =$$

$$= \frac{2PT}{\omega_c}\left[Si(\omega_c T) + \frac{\cos\omega_c T - 1}{\omega_c T}\right] = \frac{1}{2}A_0 T f(\omega_c T),$$

где

$$A_0 = \frac{P}{F} = 2\pi G_0, \quad (\omega_c = 2\pi F),$$

$$F(\omega_c T) = \frac{2}{\pi}\left[Si(\omega_c T) + \frac{\cos \omega_c T - 1}{\omega_c T}\right].$$

График этой функции показан на рис. Д. I, из которого видно, что уже при $\omega_c T \simeq 15$ (т. е. при $FT \simeq 2{,}4$) функция $f(\omega_c T)$ достигает значения, равного единице, и в дальнейшем, колеблясь, асимптотически приближается к этому значению. Так как обычно $FT \gg 1$, то можно положить $f(\omega_c T) \simeq 1$ и

$$D\eta \simeq \frac{1}{2} A_0 T = \frac{1}{2} \cdot \frac{PT}{F} = \frac{E}{2F},$$

где $E = PT$ — энергия процесса $\xi(t)$ за время T.

В ряде случаев нужно найти дисперсию интеграла

$$\zeta = \int_0^T \xi^2(t)\, dt.$$

Приближенные выражения для распределения и дисперсии были найдены Райсом [1]. При большом FT асимптотическое выражение для дисперсии имеет вид

$$D\zeta \sim \frac{2P^2T}{F} = 2A_0E.$$

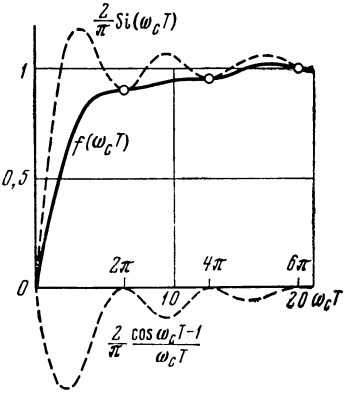


Рис. Д.1

Точные выражения для распределения и дисперсии при нескольких видах спектров помехи получены Слепяном.

II. Условие монотонного накопления. Всегда ли выгодно увеличение интервала интегрирования при накоплении? Точнее говоря, является ли отношение сигнал/помеха монотонно возрастающей функцией интервала T? Чтобы это было так, необходимо, оказывается, выполнение некоторого условия, накладываемого на характеристики помехи; это условие и названо для краткости условием монотонного накопления.

Преобразуем прежде всего выражение для дисперсии помехи

$$D\eta = D\int_0^T \xi(t)\, dt = M\int_0^T\int_0^T \xi(t)\,\xi(t_1)\, dt\, dt_1 = \int_0^T dt \int_0^T B(t - t_1)\, dt_1,$$

---

[1] *S. O. Ricc.* Filtered thermal noise-fluctuation of energy as a function of interval length. JASA, 14, 1943, N 4; *S. O. Rice.* Mathematical analysis of random noise. BSTJ, 1944, v. 23, N 3; 1945, v. 24, N 1. Русский перевод в сб.: Теория передачи электрических сигналов при наличии помех. ИЛ, 1953.

где $B(\tau)$ — функция корреляции помехи. Далее, интегрируя по частям, получаем

$$D\eta = 2\int_0^T (T-\tau)\,B(\tau)\,d\tau.$$

Отношение сигнал/помеха $\rho = a^2T^2/D\,\eta$. Рассмотрим величину, обратно пропорциональную ρ,

$$I = \frac{a^2}{2\rho} = \frac{1}{T^2}\int_0^T (T-\tau)\,B(\tau)\,d\tau.$$

Для того чтобы ρ было монотонно возрастающей функцией T, должно быть $\partial I/\partial T < 0$. [$0 < t < T$].

Выполнив дифференцирование, получаем искомое условие монотонного накопления

$$\int_0^T \left(1 - 2\frac{\tau}{T}\right) B(\tau)\,d\tau > 0.$$

Это — необходимое и достаточное условие. Можно найти ряд более простых (и соответственно более жестких) достаточных условий, например: при $\tau = T/2$ функция $B(\tau)$ должна менять знак; на интервале 0, T функция $B(\tau)$ должна монотонно убывать и т. п.

Условие монотонного накопления можно, разумеется, формулировать и на частотном (спектральном) языке.

III. Обнаружение при немонотонном распределении помехи. Наименьшая вероятность ошибки при обнаружении постоянного сигнала методом однократного отсчета получается при пороговом значении, определяемом из уравнения

$$w_0(x_0) = w_a(x_0), \qquad \text{(Д. III. 1)}$$

где w_0 и w_a — распределения помехи и суммы сигнала и помехи. Уравнение (д. III. 1) определяет значение x_0, при котором вероятность ошибки получает экстремальное значение; для того чтобы экстремум был минимумом, необходимо соблюдение дополнительного условия

$$w_a'(x_0) > w_0'(x_0). \qquad \text{(Д. III. 2)}$$

Это условие выполняется, если $w_0(x)$ есть монотонно убывающая функция $|x|$.

Мы рассмотрим здесь особенности случая, когда распределение помехи $w_0(x)$ немонотонно, в частности, когда $w_0(x)$ имеет максимум при $x \neq 0$. Пусть, к примеру, $w_0(x)$ есть четная функция, график которой изображен на рис. Д.II. Кривая $w_a(x) = w_0(x-a)$

пересекается с кривой $w_0(x)$ в трех точках, что соответствует трем решениям уравнения (Д. III. 1). Однако условию (Д. III. 2) удовлетворяют два из этих решений — точки *1* и *2*, отмеченные на рис. Д.II кружочками. Третье решение (точка *3* с абсциссой $x=a/2$; отмечена крестиком) дает не минимум, а максимум вероятности ошибки. Можно выбрать одно из двух решений, минимизирующих ошибку. Различие между ними состоит в соотношении условных вероятностей.

Если взять в качестве порогового значения абсциссу точки *1*, то получим малое значение $p_a(0)$ и большое значение $p_0(a)$.

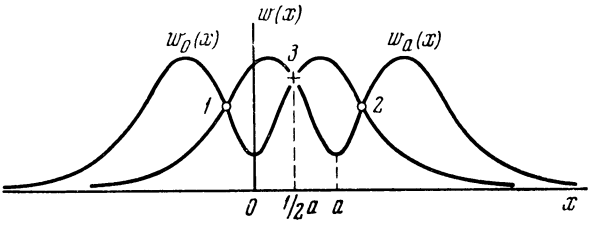


Рис. Д.II

Если же выбрать порог, соответствующий точке *2*, то получим, наоборот, $p_0(a) < p_a(0)$. Полная вероятность ошибки остается в обоих случаях одинаковой и равной

$$p_{\text{ош}} = \frac{1}{2}[p_0(a) + p_a(0)] \qquad \text{(Д. III. 3)}$$

(мы полагаем $p(0)=p(a)=1/2$).

Представляет интерес вырожденный случай дискретного распределения

$$\begin{aligned} w_0(x) &= \frac{1}{2}[\delta(x-b)+\delta(x+b)], \\ w_a(x) &= \frac{1}{2}[\delta(x-a-b)+\delta(x-a+b)]. \end{aligned} \qquad \text{(Д. III. 4)}$$

Физически это означает, что помеха представляет собой случайную функцию, принимающую с равной вероятностью одно из двух возможных значений $\pm b$. Такого рода помеха может получиться при идеальном ограничении сверху и снизу (клиппировании) произвольного случайного процесса. Ее можно также представить себе как посторонний двоичный сигнал с переменой знака. Заметим попутно, что распределение, вроде показанного на рис. Д. II, может быть результатом симметричного, но неидеального ограничения.

Распределения (Д. III. 4) изображены на рис. Д.III для двух возможных случаев

$$\text{I. } a > 2b,$$

$$\text{II. } a < 2b.$$

В первом случае следует выбрать пороговое значение в области A, определяемой неравенством

$$b < x_0 < a - b.$$

При этом обе условные вероятности, а следовательно, и полная вероятность ошибки равны нулю.

Во втором же случае выбор x_0 из области A' дает наибольшее значение $p_{\text{ош}}=1/2$.

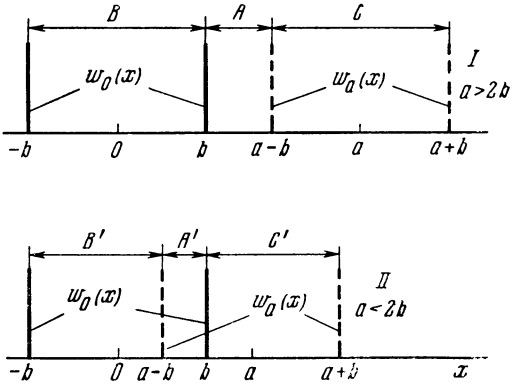


Рис. Д.III

Если же x_0 лежит в области B' или C', то вероятность ошибки равна $^1/_4$, а из двух условных вероятностей одна равна 0, а вторая — $^1/_2$. Сводка результатов дана в нижеследующей таблице:

Случай	Область	$p_0(a)$	$p_a(0)$	$p_{\text{ош}}$
I ($a > 2b$; $\rho > 4$)	A ($b < x_0 < a - b$) B ($-b < x_0 < b$) C ($a - b < x_0 < a + b$)	0 1/2 0	0 0 1/2	0 1/4 1/4
II ($a < 2b$; $\rho < 4$)	A' ($a - b < x_0 < b$) B' ($-b < x_0 < a - b$) C' ($b < x_0 < a + b$)	1/2 1/2 0	1/2 0 1/2	1/2 1/4 1/4

Совершенно ясно, что применение одного порогового значения не позволяет получить в рассматриваемом случае малой вероятности ошибки. Однако вероятность ошибки будет равна нулю, если выбрать в качестве собственной области сигнала интервалы, включающие в себя только точки $a \pm b$ и не включающие точек $\pm b$. Технически это осуществляется при помощи двусторонних ограничителей.

IV. О дисперсии величины $E_{\xi\xi}$. Мы имеем

$$E_{\xi\xi} = \int_0^T \xi(t)\,\xi(t+\tau)\,dt, \quad DE_{\xi\xi} = ME_{\xi\xi}^2 - (ME_{\xi\xi})^2.$$

Второй член равен

$$(ME_{\xi\xi})^2 = \left(\int_0^T M[\xi(t)\,\xi(t+\tau)]\,dt\right)^2 = B_\xi^2(\tau)\,T^2,$$

так что дело сводится к определению первого члена, для которого можно записать

$$ME_{\xi\xi}^2 = \int_0^T \int_0^T M[\xi(t)\,\xi(t+\tau)\,\xi(t_1)\,\xi(t_1+\tau)]\,dt\,dt_1.$$

Таким образом, для вычисления этой величины нужно располагать четырехмерным распределением помехи. Переходя к новой переменной $\vartheta = t - t_1$, можем записать

$$ME_{\xi\xi}^2 = \int_0^T dt \int_0^T B_u(\vartheta)\,d\vartheta,$$

где $B_u(\vartheta)$ — функция корреляции для случайного процесса

$$u(t) = \xi(t)\,\xi(t+\tau).$$

Меняя порядок интегрирования, можно получить следующее выражение (основанное на четности функции корреляции):

$$ME_{\xi\xi}^2 = 2\int_0^T (T-\vartheta)\,B_u(\vartheta)\,d\vartheta.$$

Если интервал корреляции $\vartheta_0 \ll T$, то

$$ME_{\xi\xi}^2 \simeq 2T\int_0^T B_u(\vartheta)\,d\vartheta.$$

Этот результат можно просто выразить через преобразование Фурье для $B_u(\vartheta)$, т. е. через спектр процесса $u(t)$:

$$G_u(\omega) = \frac{2}{\pi}\int_0^T B_u(\vartheta)\cos\omega\vartheta\,d\vartheta.$$

а именно:

$$ME_\xi^2 \simeq \pi G_u(0)\,T.$$

откуда следует, что $ME_{\xi\xi}^2$ (как и другие составляющие помехи) растет с первой степенью T.

Для вычисления $ME_{\xi\xi}^{2-}$ нужно знать либо $B_u(\vartheta)$, либо $G_u(\omega)$. Эти функции являются смешанными моментами второго порядка по отношению к процессу $u(t)$ и моментами четвертого порядка по отношению к процессу $\xi(t)$.

V. Кодовое расстояние для асимметричного двоичного канала. Пусть ошибка состоит в замене единицы нулем. Обратный переход исключен. Требуется найти кодовое расстояние, необходимое для исправления числа независимых ошибок, не превосходящего r. Обозначим переданную кодовую комбинацию через A, любую другую — через B, и принятую комбинацию (при передаче A) через C.

Пусть A содержит a_1 единиц, B a_2 единиц, C $a'=a_1-r'$ единиц, где $r'\leqslant r$. Прежде всего заметим, что для сравнения с C имеет смысл выбирать только те комбинации, число единиц в которых не превосходит $a'+r$. Поэтому под B понимаются в дальнейшем лишь такие комбинации, для которых

$$a' \leqslant a_2 \leqslant a' + r. \qquad (\text{Д. V. 1})$$

Очевидно также, что случай $a_2 < a'$ не представляет интереса, так как в этом случае B и C заведомо не могут совпадать.

Введем величину

$$\Delta a = a_2 - a' = (a_1 - a') + (a_2 - a_1) = r' \pm s,$$

где

$$s = |a_2 - a_1| \geqslant 0.$$

Так как $a_2 \geqslant a'$, то по крайней мере Δa единиц в B приходится против нулей в C, что составляет расстояние $\Delta a = r' \pm s$. Но для правильного приема необходимо и достаточно, чтобы расстояние между укороченными комбинациями B' и C' было

$$d(B'C') \geqslant 1,$$

т. е. против хотя бы одной из a' единиц в C должен стоять нуль в B. Но это значит, что необходимым и достаточным условием правильного приема является

$$d(BC) = r' \pm s + 1,$$

а так как $d(AC)=r'$, то для расстояния между A и B получаем условие

$$d(AB) \geqslant 2r' \pm s + 1. \qquad (\text{Д. V. 2})$$

Если $a_2 < a_1$ (т. е. $\Delta a = r' - s$), то r' ограничено лишь условием $r' \leqslant r$ и, следовательно, неравенство (Д. V. 2) выполняется для всех $r' \leqslant r$, только если

$$d(AB) \geqslant 2r - s + 1.$$

Если же $a_2 > a_1$ (т. е. $\Delta a = r' + s$), то r ограничено условием (Д. V. I), которое означает, что $\Delta a \leqslant r$ и, следовательно, $r' + s \leqslant r$ или $r' \leqslant r - s$. Но в этом случае условие (Д. V. 2)

$$d(AB) \geqslant 2r' + s + 1$$

будет выполняться для всех $r' \leqslant r - s$, только если

$$d(AB) \geqslant 2(r - s) + s + 1 = 2r - s + 1,$$

что совпадает со случаем $a_2 < a_1$.

Так как $d(AB)$ имеет ту же четность, что и s, равенства в последнем выражении не может быть, и его следует записать в виде

$$d(AB) > 2r - s,$$

либо в виде

$$d(AB) \geqslant 2r - s + 2. \qquad (Д. V. 3)$$

Кодовое расстояние, т. е. наименьшее расстояние между любой парой комбинаций, получим по этой же формуле, понимая под s наименьшую разность числа единиц.

Мы рассматривали предельный случай асимметричного канала, в котором возможен переход $1 \to 0$, вероятность же перехода $0 \to 1$ равна нулю. Имеется решение задачи для более общего случая, когда вероятности обоих переходов не равны нулю и не равны друг другу[1]. В этом случае различают ошибки, соответствующие обоим переходам, и вводят две кратности r_1 и r_0. Можно потребовать, чтобы код обнаруживал или исправлял обоего рода ошибки с кратностями, не превосходящими заданных. Для кодового расстояния получается (при $r_1 > r_0$)

$$d \geqslant 2(r_1 + r_0 + l + 1) - s \qquad (Д. V. 4)$$

при $2r_0 + l + 1 \leqslant s < r_1 + r_0 + l + 1$,

$$d \geqslant 2r_1 + l + 1 \qquad (Д. V. 5)$$

при $s < 2r_0 + l + 1$, где r_1 — кратность исправляемых ошибок $(1 \to 0)$; r_0 — кратность исправляемых ошибок $(0 \to 1)$, $r_1 + l$ — кратность обнаруживаемых ошибок $(1 \to 0)$; s — наименьшая разность чисел единиц в двух кодовых комбинациях. При этом обнаруживается также не более $r_0 + l$ ошибок $(0 \to 1)$. Положив в (Д. V. 4) $r_0 + l = 0$, получим (Д. V. 3).

VI. Нахождение оптимальной комбинированной системы с переспросом. Введем сокращенные обозначения $y_1 = 1 - \Phi(x_1)$, $y_2 = 1 - \Phi(x_2)$, где $x_1 = (\alpha - 1)\, z$; $x_2 = (\alpha + 1)\, z$; $z = \sqrt{n_1} z_0$; $z_0 = \sqrt{\rho_0/8}$. Для среднего числа передач имеем (см. (20. 11)) $\mu = 2/(y_1 + y_2)$. Вероятность ошибки равна (см. (20. 12)) $p = y_2/(y_1 + y_2)$. Полная энергия в комбинированной системе, осуществляющей как накопление (с кратностью n_1), так и переспрос (с кратностью μ),

[1] *W. H. Kim, C. V. Freiman.* IRF Trans. CT-6, 1969, N 71.

равна $E=n_1\mu E_0$. Задача состоит в том, чтобы минимизировать энергию E при заданной вероятности ошибки p. Величины E_0 и ρ_0 считаются заданными постоянными.

Переменными параметрами являются α (размер области неопределенности в решающем устройстве) и n_1 (кратность предварительного накопления).

Поставленная задача есть задача на нахождение условного экстремума; она решается методом неопределенных множителей Лагранжа.

Составим функцию

$$F = n_1\mu + \lambda p = \frac{2n_1 + \lambda y_2}{y_1 + y_2} = F(n_1, \alpha)$$

и приравняем нулю ее частные производные $\partial F/\partial n_1$ и $\partial F/\partial\alpha$. Это дает следующие два уравнения:

$$-2n_1(e^{-x_1^2} + e^{-x_2^2}) + \lambda(y_1 e^{-x_2^2} - y_2 e^{-x_1^2}) = 0,$$

$$2(y_1 + y_2 + \frac{z_0}{\sqrt{\pi n}}\{2n_1[(\alpha-1)e^{-x_1^2} + (\alpha+1)e^{-x_2^2}] - \quad \text{(Д. VI. 1)}$$

$$-\lambda[y_1(\alpha+1)e^{-x_2^2} - y_2(\alpha-1)e^{-x_1^2}]\} = 0.$$

Отсюда находим

$$\lambda = 2n_1\frac{e^{-x_1^2} + e^{-x_2^2}}{y_1 e^{-x_2^2} - y_2 e^{-x_1^2}}$$

и, подставляя в первое уравнение (Д. VI. 1), получаем

$$y_1 e^{x_1^2} - y_2 e^{-x_2^2} - \frac{1}{\sqrt{\pi}}(x_2 - x_1) = 0. \qquad \text{(Д. VI. 2)}$$

(Разумеется, не обязательно брать в качестве аргументов α и n; можно было бы взять за аргументы x_1 и x_2, что привело бы к тому же результату). Уравнение (Д. VI. 2) приходится решать численным или графическим методом. Получаются пары значений x_1 и x_2, по которым находятся

$$\alpha = \frac{x_2 + x_1}{x_2 - x_1}, \quad z = \sqrt{n_1}z_0 = \frac{1}{2}(x_2 - x_1)$$

и все остальные интересующие нас величины.

VII. Систематические коды с точки зрения теории групп. Систематические коды являются групповыми кодами. Ниже излагаются основные понятия теории групп, имеющие отношение к нашей теме.

Группой называется множество A, для которого определена некоторая операция над элементами множества — назовем ее сложением, — удовлетворяющая следующим условиям:

1. $(a+b)+c=a+(b+c)$ (условие ассоциативности).

2. $a+0=0+a=a$ (условие существования нейтрального элемента 0).

3. $a+(-a)=(-a)+a=0$ (условие существования элемента, противоположного данному).

Основное свойство группы состоит в том, что если даны два элемента $a \in A$ и $b \in A$, то

$$a+b=c \in A.$$

Условия 1, 2 и 3 называют групповыми аксиомами. Если кроме этих условий выполняется еще условие коммутативности

$$a+b=b+a,$$

то группа называется коммутативной или абелевой. Группа, состоящая из конечного числа элементов, называется конечной.

Групповая операция названа в определении сложением. Ее называют также умножением; тогда групповые аксиомы записываются в другой форме, а именно:

1. $(ab)\,c=a\,(bc)$.
2. $a1=1a=a$.
3. $a \cdot a^{-1}=a^{-1} \cdot a=1$.

Неважно, как называется групповая операция. Важно, что эта операция должна быть определена, т. е. дожно быть указано, чему равняется элемент c, получаемый в результате применения групповой операции к двум элементам a и b. Мы можем говорить о группах как на «аддитивном», так и на «мультипликативном» языке, хотя групповая операция, вообще говоря, не является ни обычным арифметическим сложением, ни арифметическим умножением.

Подгруппой называется подмножество B множества A, являющееся группой. Иначе говоря, сумма $c=a+b$ элементов $a \in B$ и $b \in B$ должна принадлежать подмножеству B, т. е. подмножество должно обладать основным свойством группы.

Группа может быть разложена по данной подгруппе следующим образом: пусть $b_1=0$, b_2, b_3, . . . — элементы подгруппы B. Составим таблицу, в первой строке которой выписана подгруппа B. В первом столбце второй строки записан какой-либо элемент $e_1 \in A$, не входящий в B, а в следующих столбцах — суммы этого элемента с находящимися в тех же столбцах элементами подгруппы B. В первом столбце третьей строки записывается элемент $e_2 \in A$, не содержащийся в первых двух строках, и т. д.

Таблица имеет вид

$$\begin{array}{cccc} 0 & b_2 & b_3 & b_4 \ldots \\ e_1 & e_1+b_2 & e_1+b_3 & e_1+b_4 \ldots \\ e_2 & e_2+b_2 & e_2+b_3 & e_2+b_4 \ldots \\ \cdots & \cdots & \cdots & \cdots \end{array}$$

Множества, представляемые строками этой таблицы, кроме первой, называются смежными классами подгруппы B, а элементы e_1, e_2 — главными элементами смежных классов.

При описанном построении таблицы в ней оказываются все элементы группы A, причем каждый из них встречается только по одному разу, т. е. входит только в один смежный класс.

Теперь мы можем говорить о систематических двоичных кодах в терминах теории групп.

Групповая операция, которой мы пользуемся, есть сложение по модулю два. Ее определение дается таблицей

	0	1
0	0	1
1	1	0

Группа — это множество всех кодовых векторов, так что кодовые векторы — это элементы группы. Нейтральным элементом группы является нулевой вектор. Группа конечна, так как число кодовых векторов $N=2^n$. Она является абелевой группой, так как сложение по модулю два — коммутативная операция.

Систематический код является подгруппой, так как любой кодовый вектор может быть получен суммированием двух других.

Полная кодовая таблица в стандартном расположении есть разложение множества всех векторов по данному коду на смежные классы. Главными элементами смежных классов являются векторы ошибок.

Установив, таким образом, групповые свойства систематических кодов, можно непосредственно использовать в теории кодов ряд известных результатов теории групп, что и делается во всех современных работах.

VIII. Одна кодирующая схема для циклического кода. Дан вектор h, представленный полиномом

$$h(x)=h_0+h_1x+h_2x^2+\ldots+h_kx^k.$$

По определению, вектор h должен быть ортогонален любому кодовому вектору. Если взять один из кодовых векторов

$$v(x)=a_0+a_1x+a_2x^2+\ldots+a_kx^k,$$

то любой другой, получаемый сдвигом на i шагов, будет

$$v'(x)=a_i+a_{i+1}x+a_{i+2}+\ldots+a_{i+k}x^k.$$

По условию ортогональности

$$hv'=\sum_{j=0}^{k}h_ja_{i+j}=0$$

или

$$\sum_{j=0}^{k-1}h_ja_{i+j}+h_ka_{i+k}=0.$$

Но $h_k=1$, так что

$$a_{i+k}=-\sum_{j=0}^{k-1} h_j a_{i+j}. \qquad (Д.VIII.1)$$

При помощи этого рекуррентного соотношения можно найти очередной коэффициент a_{i+k} по известным a_i, $a_{i+1}, \ldots a_{i+k-1}$ или, иначе говоря, найти $n-k$ контрольных символов по k информационным символам. Именно эту операцию в соответствии с формулой (Д.VIII.1) выполняет схема рис. 70, которая повторена с указанием содержимого регистра (рис. Д.IV).

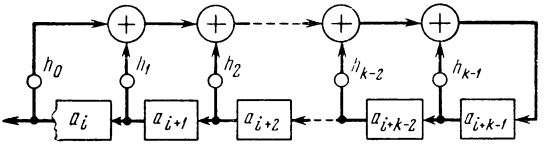


Рис. Д.IV

IX. Нижняя оценка числа проверочных символов в (n, k) коде, исправляющем серийные ошибки. Вывод основывается на формуле (22.5)

$$\frac{N_0}{1+E}\geqslant N.$$

Так как $N_0=2^n$, $N=2^k$, то эту формулу можно переписать в виде

$$2^{n-k}\geqslant 1+E$$

или

$$n-k\geqslant \log(1+E).$$

Здесь $n-k$ — число контрольных знаков, E — число векторов ошибок, т. е. всевозможных сочетаний, образующих подлежащие исправлению серийные ошибки. Задача и состоит в нахождении этого числа.

В пределах кодовой комбинации из n символов пачка из b символов может занимать $n-b+1$ различных положений; сама же пачка может представлять собой одно из 2^{b-2} сочетаний (так как два крайних символа заданы; по определению серийной ошибки — это единицы). Пачка из $b-1$ символов может занимать $n-b+2$ положений и представлять собой одно из 2^{b-3} возможных сочетаний и т. д. Наконец, пачка из двух символов состоит из двух единиц и может занимать одно из $n-1$ возможных положений. Следовательно,

$$E=\sum_{k=0}^{b-2}(n-1-k)\,2^k=(n-b+2)\,2^{b-1}-(n+1).$$

Если добавить еще и одиночную ошибку (n возможных положений), то

$$E = (n-b+2)\,2^{b-1} - 1$$

и мы получаем искомую оценку $n-k \geqslant b-1+\log(n-b+2)$. Так, для примера кода Файра, приведенного в тексте, мы имели $n-k = 13$, тогда как по данной оценке

$$n-k \geqslant 3-1+\log(70-3+2) = 8{,}1085.$$

Ближайшее целое число 9. Таким образом, фактическое число проверочных символов недалеко от устанавливаемого нашей нижней оценкой.

X. Более общий случай накопления: флуктуирующий сигнал. При подсчете выигрыша, даваемого накоплением, полагают обычно, что суммируемые экземпляры полезного сигнала совершенно одинаковы. В этом предположении выведены формулы (9.1) и (9.3). Но в действительных условиях сигнал может флюктуировать, и поэтому при вычислении отношения сигнал/помеха следует выражать мощность сигнала его средним квадратом.

Пусть на выходе накопителя имеем

$$x = \sum_k x_k = \sum_k (s_k + \xi_k) = \sum_k s_k + \sum_k \xi_k = b + \eta$$

и отношение сигнал/помеха $\rho = Mb^2/D\eta$ (напомним, что мы пишем везде $D\eta$ вместо $M\eta^2$, так как полагаем $M\xi$, а следовательно, и $M\eta$ равными нулю). Для знаменателя на основании (9. 2) и (9. 4) запишем

$$D\eta = D\left(\sum \xi_k\right) = n\sigma^2\left[1 + \frac{2}{n}\sum_{l=1}^{n-1}(n-l)\,k_\xi(l)\right] = n\sigma^2(1+\lambda_\xi).$$

Для числителя путем совершенно аналогичных выкладок получим

$$Mb^2 = M\left(\sum s_k\right)^2 = n\overline{s^2}\left[1 + \frac{2}{n}\sum_{l=1}^{n-1}(n-l)\,k_s(l)\right] = n\overline{s^2}(1+\lambda_s),$$

где

$$\overline{s^2} = \frac{1}{n}\sum s_k^2.$$

В этих выражениях k_ξ и k_s — нормированные коэффициенты корреляции для помехи и сигнала соответственно; λ_ξ и λ_s — сокращенные обозначения для сумм.

Итак, общее выражение для отношения сигнал/помеха на выходе накопителя принимает вид

$$\rho = \frac{\overline{s^2}}{\sigma^2}\cdot\frac{1+\lambda_s}{1+\lambda_\xi} = \rho_0\frac{1+\lambda_s}{1+\lambda_\xi}.$$

Если как помеха, так и сигнал полностью не коррелированы, т. е. $k_\xi = k_s = 0$, $\lambda_\xi = \lambda_s = 0$, то, как и следовало ожидать, $\rho = \rho_0$, т. е.

накопление ничего не дает. Если сигнал полностью коррелирован, то

$$k_в = 1, \quad \lambda_в = \frac{2}{n}\sum_{l=1}^{n-1}(n-l) = n-1, \quad 1+\lambda_в = n,$$

и мы получаем (9. 3)

$$\rho = \frac{n}{1+\lambda_\xi}\rho_0.$$

Литература

Бакут П. А., Большаков И. А., Герасимов Б. М. и др. Вопросы статистической теории радиолокации. «Сов. радио», 1964.

Башаринов А. Е., Флейшман Б. С. Методы статистического последовательного анализа и их приложения. «Сов. радио», 1962.

Блох Э. Л. Помехоустойчивость систем связи с переспросом. Изд-во АН СССР, 1963.

Вальд А. Последовательный анализ. Физматгиз, 1960.

Вайнштейн Л. А., Зубаков В. Д. Выделение сигналов на фоне случайных помех. «Сов. радио», 1960.

Возенкрафт Дж., Рейффен Б. Последовательное декодирование. ИЛ, 1963.

Вудворд Ф. М. Теория вероятностей и теория информации с применением в радиолокации. «Сов. радио», 1955.

Вулих Б. З. Введение в функциональный анализ. Физматгиз, 1958.

Гуткин Л. С. Теория оптимальных методов радиоприема при флуктуационных помехах. Госэнергоиздат, 1961.

Давенпорт В. Б., Рут В. Л. Введение в теорию случайных сигналов и шумов. ИЛ, 1960.

Зюко А. Г. Помехоустойчивость и эффективность систем связи. Связьиздат, 1963.

Котельников В. А. Теория потенциальной помехоустойчивости. Госэнергоиздат, 1956.

Левин Б. Р. Теория случайных процессов и ее применение в радиотехнике. «Сов. радио», 1960.

Левин Ю. С. Оптимальные фильтры и накопители импульсных сигналов. «Сов. радио», 1963.

Миддлтон Д. Введение в статистическую теорию связи «Сов. радио», 1962.

Питерсон У. Коды, исправляющие ошибки. «Мир», 1964.

Фалькович С. Е. Прием радиолокационных сигналов на фоне флюктуационных помех. «Сов. радио», 1961.

Фано Р. Передача информации. «Мир». 1965.

Финк Л. М. Теория передачи дискретных сообщений. «Сов. радио», 1963.

Хелстром К. Статистическая теория обнаружения сигналов. ИЛ, 1963.

Чернов Г., Мозес Л. Элементарная теория статистических решений. «Сов. радио», 1962.

Прием импульсных сигналов в присутствии шумов. Сб. пер. под ред. *А. Е. Башаринова* и *М. С. Александрова*. Госэнергоиздат, 1960.

Теория кодирования. Сб. пер. под ред. *Э. Л. Блоха*. «Мир», 1964.

Прием сигналов при наличии шума. Сб. пер. под ред. *Л. С. Гуткина*. ИЛ, 1960.

Коды с обнаружением и исправлением ошибок. Сб. пер. под ред. *А. М. Петровского*, ИЛ, 1956.

Передача цифровой информации. Сб. пер. ред. *С. И. Самойленко*. ИЛ, 1963.

Определение параметров случайных процессов. Сб. пер. под ред. *В. И. Чайковского*. Гос. изд. техн. лит. УССР, 1962.

СРАВНЕНИЕ НЕКОТОРЫХ ВОЗМОЖНОСТЕЙ ПЕРЕДАЧИ ПРОСТЫХ РИСУНКОВ

1. В служебных применениях фототелеграфии часто встречается случай, когда передаваемое изображение представляет собой простой рисунок, например, схему, синоптическую карту, и т. п.[1]

В этом случае требуется передать только конфигурацию линий, и неэффективность обычной системы фототелеграфа, способной передавать полутоновые изображения, занимающие все поле, становится совершенно очевидной.

Вопрос этот не нов; он ставился, в частности, в докладе Леба на лондонском симпозиуме 1952 г. [1]. Некоторые возможности были указаны в дискуссии; отметим в особенности выступление Бенджамина.

Ниже рассмотрены в общих чертах две возможности, основанные на использовании одномерной статистики черно-белого изображения; имеется в виду тот факт, что суммарная площадь линий много меньше площади всего поля изображения, откуда и вытекает неэффективность обычной системы передачи. Для рассматриваемого случая мы можем упростить дело, рассматривая линии как одномерные объекты, т. е. отвлекаясь от их толщины.

2. Одна из возможностей основана на применении обычной строчной развертки. Однако передается не значение каждого элемента изображения (черное или белое — так обстоит дело в обычной системе), а координаты точек пересечения линий рисунка со строкой.

Это в сущности классический пример передачи последовательности из двух символов с резко различными вероятностями, рассмотренный еще Шенноном ([2], стр. 34). Однако тут есть одно различие; в цитированном примере предлагается передавать расстояния между соседними точками пересечения, тогда как мы будем иметь в виду передачу координат (т. е. расстояний от начала строки) точек пересечения. Передача расстояний теоретически выгоднее, но нужно учесть, что координата каждой следующей

[1] В последующем имеются в виду следующие определения: рисунком называется изображение, информационное содержание которого определяется только конфигурацией линий; простым называется рисунок, общая длина линий которого мала по сравнению с наибольшей возможной (при данной разрешающей способности).

точки пересечения определяется суммированием всех предшествующих расстояний, а это значит, что в такой системе возможно накопление ошибки вдоль строки. Это соображение и заставляет отдать предпочтение системе с непосредственной передачей координат.

Простейшее использование статистики рисунка состоит в том, что назначается наибольшее значение n_0 числа точек пересечения. Это значит, что вероятность появления $n > n_0$ точек достаточно мала с точки зрения требований в отношении надежности данного вида связи.

Пусть точность передачи характеризуется числом N элементов разложения вдоль строки. Тогда объем сигнала, необходимый для передачи равномерным кодом координат n_0 точек равен (в двоичных единицах)

$$V = n_0 \log N^*.$$

При обычном же способе передачи объем сигнала на строку равен

$$V_0 = \log 2^N = N.$$

Таким образом, выигрыш, получаемый при применении описанной системы, можно оценить отношением

$$k_a = \frac{V_0}{V} = \frac{N}{n_0 \log N}. \qquad (1)$$

3. Другая возможность основывается на существенно ином принципе передачи. Развертка идет не по строкам, а вдоль линии рисунка. Рассмотрим точку, лежащую на линии рисунка, представив ее в виде элемента разложения (заштрихованный квадратик на рис. 1). Данный элемент окружен восьмью соседними; для того чтобы проследить дальнейший ход линии, достаточно указать, в какой из квадратиков идет линия. Для этого требуется, очевидно, сигнал объема

$$v = \log 8 = 3.$$

Для того чтобы определить полный объем сигнала для передачи всего рисунка, нужно задаться общей длиной образующих

---

* При любом способе передачи объем сигнала не менее информации, которая (на одну строку) равна

$$J = n \log \frac{N}{n_0} - (N - n_0) \log \left(1 - \frac{n_0}{N}\right)$$

или при $n_0 \ll N$ приблизительно

$$J \approx n_0 \left(\log N - \log \frac{n_0}{e}\right).$$

рисунок линий. Пусть эта длина, выраженная числом строк, составляет N_0. Тогда объем сигнала равен

$$V = vN_0N = 3N_0N,$$

а объем сигнала при обычном способе передачи (считая поле квадратным, т. е. принимая число строк равным N)

$$V_0 = \log 2^{N^2} = N^2.$$

Таким образом, выигрыш от применения данной системы, выражаемый аналогично предыдущему, составляет

$$k_в = V_0/V = N/3N_0. \qquad (2)$$

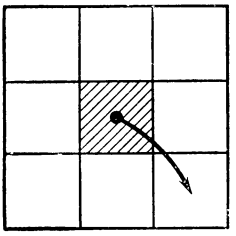


Рис. 1

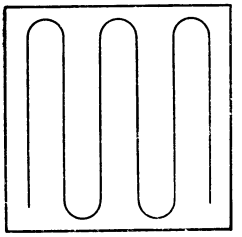


Рис. 2

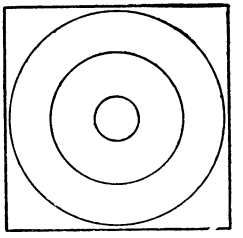


Рис. 3

4. Для того чтобы сделать возможным непосредственное сравнение обеих систем, нужно выяснить связь между n_0 и N. Интуитивно ясно, что обе величины связаны простой линейной зависимостью

$$n_0 = \alpha N_0.$$

Простейшие примеры показывают, что α имеет порядок единицы. Действительно, если представить себе рисунок в виде рис. 2, то этот результат получается непосредственно. Если же расположить линии, как на рис. 3, то элементарный подсчет дает $\alpha = 4/\pi$.

5. Теперь можно сравнить эффективность обеих систем, взяв отношение $k_в$ и k_a (формы (1) и (2))

$$\frac{k_в}{k_a} = \frac{N}{3N_0} \cdot \frac{n_0 \log N}{N} = \frac{\alpha}{3} \log N$$

или

$$\frac{k_в}{k_a} \approx \frac{1}{3} \log N. \qquad (3)$$

Таким образом, вторая система тем выгоднее, чем больше заданная точность, выражаемая числом элементов N (или, что то же, числом строк). Так, например, при $N = 512$, $\log N = 9$, вторая система втрое выгоднее первой.

6. Что касается абсолютного выигрыша, то, как показывают формулы (1) и (2), он оказывается для обеих систем тем большим, чем больше заданная точность и чем проще рисунок (т. е. чем меньше $n_0 \approx N_0$). Пусть, например, передается простой рисунок, для которого $n_0 \approx N_0 = 10$, $N = 512$. Тогда

$$k_a = 512/10 \log 512 \approx 5{,}7, \qquad k_в = 512/3 \cdot 10 \approx 17.$$

7. Сравнивая две описанные системы, нужно заметить, что первая использует статистику очень грубо, задавая наибольшее число пересечений n_0. Система не делает различия между более простыми и более сложными рисунками, будучи рассчитана на наихудший случай (т. е. на наиболее сложный рисунок). В этом, конечно, ее недостаток по сравнению со второй системой, в которой время передачи рисунка зависит от его сложности: это время прямо пропорционально общей длине линий N_0 (считая, естественно, что скорость передачи двоичных знаков задана и неизменна).

С другой стороны, при осуществлении второй системы неизбежно возникают специфические затруднения. Действие системы осуществляется просто лишь в том случае, когда рисунок образован одной непрерывной линией, начинающейся всякий раз в одной определенной точке и нигде не пересекающейся сама с собой. Это означает, что либо рисунок надо специально готовить, подчиняя его этому весьма стеснительному требованию, либо надо существенно усложнить систему: она должна самостоятельно отыскивать еще не воспроизведенные контуры рисунка. Эта задача, конечно, разрешима, но ясно наперед, что в систему должны быть добавлены два относительно сложных узла, а именно: 1) поисковое устройство и 2) блок памяти, удерживающей уже пройденные элементы рисунка. Вопрос о целесообразности построения такого устройства может быть решен только с учетом конкретных условий его возможного применения.

Еще одно замечание по поводу первой системы: хорошим способом повышения надежности является, по-видимому, повторная передача повернутого рисунка, т. е. передача с изменением направления развертки.

8. В предыдущем предполагалось весьма элементарное использование одномерной статистики. Первая система могла бы быть значительно улучшена, если бы передача велась на основе не наибольшего, а фактического числа точек пересечения. Однако это нарушило бы естественный ритм работы устройства, задаваемый ритмом развертки. Возможное решение состоит в том, что все координаты отсчитываются (по мере появления точек пересечения), запоминаются, а затем передаются для всего изображения в целом в некотором новом ритме. Эта схема эквивалентна в конце концов схеме с переменной скоростью развертки; разница состоит в том (что благоприятно для рассматриваемых систем), что не требуется восстановления истинного масштаба времени.

Что касается возможности использования многомерной статистики, то, не предпринимая серьезного исследования этого вопроса и ограничиваясь общими поясняющими замечаниями, можно сказать следующее. Рассмотрим вторую систему. Пользуясь многомерной статистикой, т. е. опираясь на знание предшествующего хода линии рисунка, мы могли бы, например, экстраполировать ее. Иначе говоря, мы могли бы с той или иной точностью предсказывать ее дальнейший ход. Так, если известно, что линия не имеет резких изломов, то ее направление может быть предсказано с определенной вероятностью. Если, например, дальнейшее направление линии можно предсказать с точностью до 180°, т. е. указать, в каком полукруге она расположится, то требуемый объем сигнала уменьшается на единицу; если можно указать квадрант, то объем сигнала сокращается на две единицы и т. д. В схеме рис. 1 это означает, что мы указываем не один из восьми квадратиков, а один из четырех или один из двух. Если же с достаточно высокой вероятностью может быть предсказан квадратик, через который линия должна пройти, то никакого сигнала для очередного шага вообще не нужно передавать. Таким образом, выигрыш от использования многомерной статистики в рассматриваемой системе может оказаться довольно значительным. Нужно, однако, помнить, что возможный выигрыш покупается ценой заметного усложнения устройства.

Литература

1. *J. Loeb.* Communication theory of transmission of simple drawings. Communication theory (W. Jackson ed.). London, Butterworth, 1953.
2. Теория передачи электрических сигналов при наличии шумов. Сб. пер. под ред. *Н. А. Железнова.* ИЛ, 1953.

ФОТОТЕЛЕГРАФ С ТОЧКИ ЗРЕНИЯ ТЕЛЕГРАФА

Эта заметка, имеющая методический характер, имеет целью рассмотреть вопрос о сравнительных возможностях фототелеграфа и телеграфа с точки зрения передачи текста.

Будет показано, почему телеграф имеет принципиально большую производительность, чем фототелеграф. Попутно разъясняются важные понятия, относящиеся к исправляющим кодам.

Телеграф имеет дело с условными кодовыми обозначениями букв, тогда как фототелеграф передает фактические очертания букв. На первый взгляд ситуация настолько различна, что прямое сопоставление обеих систем кажется невозможным. Но мы покажем сейчас, что общий подход к обеим столь несходным системам может быть установлен вполне непринужденно и естественно и что применение этого подхода позволяет легко прийти к некоторым общим заключениям.

Прежде всего ограничим себя рассмотрением черно-белого фототелеграфа. Это ограничение вполне естественно, раз мы собираемся рассматривать фототелеграф в применении к передаче текста.

Следующее важное свойство состоит в том, что фототелеграф имеет ограниченную и вполне определенную разрешающую способность. Это значит, что аппаратура не различает деталей, размер которых меньше определенной величины (обычно 0,2 *мм*). Таким образом, изображение можно представить себе разбитым на квадратики размером $0{,}2 \times 0{,}2$ *мм*. Каждый из этих элементарных квадратиков может быть либо белым, либо черным. Этим полностью определяются возможности черно-белого фототелеграфа.

Пусть передача сигнала состоит в том, что каждому черному элементу изображения соответствует посылка тока, а белому — пауза. Тогда фототелеграфный сигнал будет состоять из посылок и пауз, и в этом смысле он ничем не отличается от телеграфного сигнала.

Будем называть посылки и паузы знаками и посмотрим, сколько знаков требуется для передачи буквы по телеграфу и по фототелеграфу. В телеграфе применяется пятизначный код Бодо; для передачи буквы требуется пять знаков. Для подсчета числа знаков в случае фототелеграфа нужно задаться размером буквы. Пусть речь идет о мелкой машинописи, в которой каждая буква

вписывается в прямоугольник размером $1{,}5\times 2$ *мм*. Такой прямоугольник разбивается на $1{,}5\cdot 2/0{,}2\cdot 0{,}2=75$ элементарных квадратиков. Это и есть число знаков на одну букву в фототелеграфном сигнале при принятых условиях.

Однако заключение о том, что в условиях рассмотренного примера производительность телеграфа в $75/5=15$ раз больше, чем фототелеграфа, было бы неосновательным. Фототелеграф обладает большей помехоустойчивостью. Это всем известное обстоятельство обусловлено тем, что передается не сразу вся буква, а, так сказать, несколько ее сечений; несмотря на искажения одного сече-

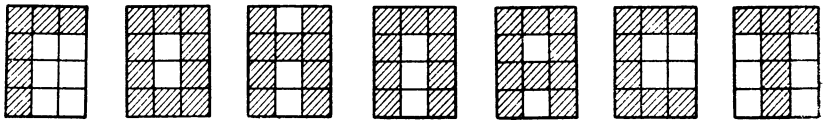


Рис. 1

ния, общее очертание буквы может быть воспринято правильно. При горизонтальной развертке с шагом строки 0,2 *мм* мы передаем $2/0{,}2=10$ горизонтальных сечений букв. Отсюда следует, что, желая сравнивать телеграф и фототелеграф, мы должны привести обе системы к одинаковым условиям в отношении помехоустойчивости. Для этого рассмотрим положение подробнее.

Рассмотрим построение букв в решетке с меньшим количеством клеток, т. е. с меньшей разрешающей способностью, например, $3\times 4=12$ клеток. При таком количестве клеток некоторые буквы невозможно изобразить [1]. Однако для наших целей это не имеет значения. Некоторые буквы (с наиболее простыми очертаниями) изображены по 12-клеточной системе на рис. 1.

Ошибки при передаче состоят в том, что из-за влияния помех (или вследствие несовершенства аппаратуры) посылка заменяется паузой (или наоборот) и соответственно черный элементарный квадратик заменяется белым (или наоборот). Такую ошибку будем называть одиночной. Если в двух клетках появились неверные значения, то имеем двойную ошибку и т. п. Легко сообразить, что перемещение черного квадратика в какую-либо другую клетку соответствует двойной ошибке.

Чем определяется помехоустойчивость передачи? Помехоустойчивость тем больше, чем меньше вероятность замены одной буквы другой — неверной. Эта вероятность в свою очередь тем меньше,

---

[1] Заметим, что вопрос о наименьшем числе клеток для удовлетворительного изображения всех букв алфавита приходится решать при конструировании светящихся надписей (для рекламы, на стадионах и т. п.), состоящих из отдельных ламп; можно считать, что в этом случае достаточно $5\times 8=40$ ламп на площадку, занимаемую одной буквой.

чем больше буквы различаются друг от друга. Количественной мерой этого различия может служить число различающихся знаков (т. е. число клеток черных для одной буквы и белых для другой, или наоборот). Так, буквы О и С (рис. 1) различаются двумя знаками, буквы О и Н — тремя.

В дальнейшем будем называть число различающихся знаков расстоянием [1]. Легко видеть, что расстояния между различными буквами неодинаковы. Для семи букв рис. 1 расстояния даны в табл. 1.

Таблица 1

Расстояния между буквами (рис. 1)

	Г	О	Н	П	Р	С	Т
Г	0	4	5	3	3	2	6
О		0	3	1	3	2	8
Н			0	2	5	5	9
П				0	2	3	9
Р					0	5	7
С						0	6
Т							0

Табл. 1 симметрична относительно диагонали; иными словами расстояние, например, от О до Н равно расстоянию от Н до О*. Рассматривая ее, мы видим, что различия между буквами, измеряемые расстояниями, неодинаковы. Обращает на себя внимание значительное отличие буквы Т от всех остальных букв. Это обусловлено тем, что Т имеет одну ножку посредине, тогда как остальные буквы имеют вертикальные части в крайних столбцах решетки.

Совершенно ясно, что легче всего смешать буквы, наименее отличающиеся друг от друга. Например, достаточно ошибки в одном знаке, чтобы буква О превратилась в П (расстояние 1); для этого нужно, чтобы в О нижний средний квадратик был заменен белым. Для перехода буквы Н в П достаточно двойной ошибки, переводящей горизонтальную перекладину в верхнее положение, и т. д. Буквы же, сильно отличающиеся от других, как, например, Т, оказываются более стойкими. На рис. 2 показан постепенный переход Т в П (расстояние 9) в результате специально подобранных ошибок. Цифры под рисунками означают число ошибок при передаче буквы Т и одновременно расстояние искаженного ри-

---

[1] Такая терминология принята в геометрической теории сигналов, в которой термин «расстояние» употребляется в своем обычном смысле (хотя и в применении к пространству с неевклидовой, вообще говоря, метрикой).

* Это кажется тривиальным, но ведь это ниоткуда не следует. Указанное свойство входит в аксиоматическое определение расстояния в пространстве с любой метрикой.

сунка от Т. Здесь отчетливо выступает один из основных принципов: получив искаженное очертание, мы будем считать его за Т или за П в зависимости от того, какая буква ближе. Так, искаженное изображение, отстоящее от Т на 4 единицы (т. е. различающееся в 4 знаках), все же похоже на Т, а отстоящее от Т на 6 единиц, — уже более похоже на П, от которого оно отстоит на 3 единицы [1].

В вопросе о возможном смешении букв играет роль не только расстояние, но и «направление». Возьмем, к примеру, букву С. Эта буква находится на одинаковом расстоянии, равном двум,

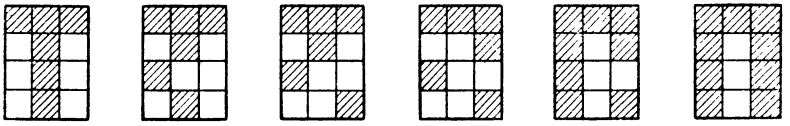


Рис. 2

от букв Г и О. На рис. 3 показано, что в результате одиночной ошибки изображение буквы С (рис. 3, *а*) делает один шаг «в направлении» буквы О (рис. 3, *б*) или «в направлении» буквы Г (рис. 3, *в*).

Имея в виду сравнение с телеграфом, введем цифровую запись букв. Для этого условимся отмечать черный элемент (посылку) единицей, а белый элемент (паузу) — нулем. Введя развертку в порядке, изображенном на рис. 4 (порядок развертки совершенно произволен и не влияет на ход рассуждения), получим цифровую запись букв рис. 1. Эта цифровая форма вполне пригодна для суждения о расстояниях. Так, например, мы видим, что все буквы начинаются группой из четырех единиц, кроме Т, у которой эта группа в середине.

Итак, представление букв изображениями в двенадцатиклеточной решетке равносильно применению двенадцатизначного телеграфного кода (табл. 2).

Предыдущие рассуждения вплотную подводят нас к важному заключению общего характера, а именно: мы получили бы наиболее помехоустойчивую фототелеграфную связь при условии, что буквы больше всего отличаются друг от друга, т. е. иначе говоря, что расстояния между буквами являются наибольшими возможными. Конечно, расстояние (выражаемое числом различающихся знаков) растет с увеличением числа клеток решетки, т. е. с увеличением разрешающей способности. При этом, однако, падает производительность, так как возрастает значность кода. Спраши-

[1] Эти рассуждения разъясняют понятие об исправляющем коде: если код таков, что, несмотря на ошибки, можно установить, что именно передавалось, то ошибки тем самым и исправляются.

Таблица 2

Представление букв (рис. 1) в цифровой записи

Г	Н	О	П
111110001000	111101001111	111110011111	111110001111

Р	С	Т
111110101110	111110011001	100011111000

вается, что можно сделать, не меняя значности кода? Возникает естественная мысль: отказаться от исторически сложившихся очертаний букв и создать искусственно новый набор очертаний, подобрав его так, чтобы удовлетворить требованию наибольших расстояний. Но в этом и состоит принцип телеграфной передачи, как сейчас станет ясно.

Рассмотрим обычный пятизначный код Бодо, часть кодовой таблицы которого приведена ниже (табл. 3).

Таблица 3

Код Бодо

А	Б	В	Г	Д	Е	Ж
10000	00110	01101	01010	11110	01000	00011

Наименьшее расстояние между двумя кодовыми комбинациями равно единице: это значит, что ошибка в одном знаке превращает одну букву в другую, причем ошибка остается незамеченной.

Можно построить простейший код, обнаруживающий одиночную ошибку, добавив к пятизначным кодовым комбинациям еще один знак — нуль или единицу — с таким расчетом, чтобы число единиц в кодовой комбинации было четным (нарушение этого условия в принятой комбинации и указывает на ошибку) (табл. 4).

Таблица 4

Шестизначный код, обнаруживающий одиночную ошибку

А	Б	В	Г	Д	Е	Ж
100001	001100	011011	010100	111100	010001	000110

Наименьшее расстояние стало теперь равным двум, так что этот телеграфный код при шести знаках не уступает двенадцатизначному фототелеграфному коду, рассмотренному выше.

Мы можем построить «изображения» букв, соответствующих этим шестизначным кодовым комбинациям, взяв шестиклеточную решетку и предполагая развертку по столбцам, как на рис. 4.

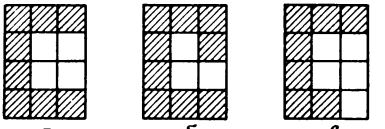


Рис. 3

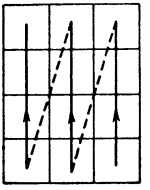


Рис. 4

На рис. 5 показано, как «выглядят» искусственные новые буквы (табл. 5).

Для данного кода возможны только четные расстояния. Наибольшее расстояние (не встречающееся в табл. 5), равно, очевидно, шести: таково, например, расстояние между кодовыми комбинациями 011011 и 100100, 101011 и 010100 и т. д.

Сравнивая табл. 1 и 5, мы видим, как равномерны и относительно велики расстояния в табл. 5* . Такое свойство таблицы расстояний является признаком хорошего кода.

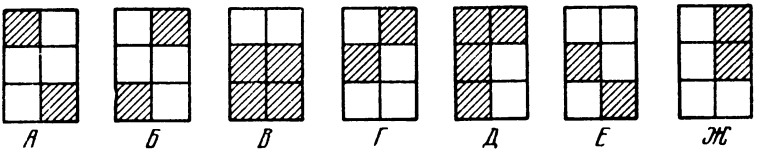


Рис. 5

Мы показали, что можно представить изображение буквы кодовой комбинацией, как в телеграфии, и обратно: представить кодовую комбинацию телеграфного кода соответствующим «изображением». Это позволило нам непосредственно сравнивать обе системы, и мы можем теперь сформулировать общее заключение. Это заключение таково: при выборе телеграфного кода мы ничем не связаны и можем строить наилучший код, обеспечивающий наибольшую производительность (т. е. имеющий наименьшую знач-

---

* Следует помнить, что табл. 1 относится к двенадцатизначному коду, а табл. 5 — всего лишь к шестизначному. Поэтому для непосредственного сравнения различных кодов удобнее пользоваться не абсолютными, а относительными расстояниями. Относительное расстояние определяется как отношение числа различающихся знаков к полному числу знаков; наибольшее значение этой величины есть, очевидно, единица.

Таблица 5

Расстояния между буквами (рис. 5)

	А	Б	В	Г	Д	Е	Ж
А	0	4	4	4	4	2	4
Б		0	4	2	2	4	2
В			0	4	4	2	4
Г				0	2	2	2
Д					0	4	4
Е						0	4
Ж							0

ность) при заданной помехоустойчивости. В случае же фототелеграфа существующие и не подлежащие изменению очертания букв приводят при прочих условиях к плохому коду, имеющему при заданной помехоустойчивости очень большую значность и дающему, следовательно, малую производительность.

Преимущество телеграфа перед фототелеграфом с точки зрения передачи текста является, таким образом, принципиальным.

ОПОЗНАНИЕ ОБРАЗОВ

1. Основные понятия. «Как мы узнаем идентичность черт человека, видим ли мы его в профиль, в три четверти или анфас? Как мы опознаем круг как круг, велик ли он или мал, близок или далек; лежит ли он в плоскости, перпендикулярной к линии, соединяющей глаз с центром, и виден как круг, или имеет какую-либо другую ориентацию, и виден как эллипс? Как мы видим лица, животных и географические карты в облаках или в пятнах испытания Роршаха? Все эти примеры относятся к глазу, но аналогичные вопросы распространяются и на другие чувства. . .» [1].

На эти вопросы дается следующий ответ: чувственное восприятие реального предмета подвергается некоторой обработке, состоящей в очистке, рафинировке и преобразовании, после чего обработанное таким образом восприятие сличается с запасенными в мозгу эталонами.

Эти эталоны мы и будем называть образами. Образ есть категория идеальная. Так, говоря «квадратный стол» или «круглая тарелка», мы представляем себе, что чувственное восприятие стола (тарелки) было путем надлежащей обработки доведено до абстрагированного состояния, допускающего прямое сличение с образом квадрата (круга).

Обработка и сличение составляют процесс опознания образа. Детали этого процесса неизвестны. Мы будем рассматривать проблему с точки зрения возможности опознания образов средствами техники.

2. Отношение к психологии. Конечно, опознание образов человеком есть процесс психологический. Проблема опознания составляет предмет многочисленных исследований и теоретических построений. Существует специальная теория, так называемая гештальт-теория, согласно которой форма (Gestalt) первична в процессе восприятия. Мы постараемся, однако, держаться подальше от психологических теорий и не будем пользоваться термином «гештальт». Психологические теории в их теперешнем виде имеют спекулятивный характер. Они представляли бы интерес с точки зрения техники лишь в том случае, если бы они в какой-то мере раскрывали механизм опознания образа; тогда можно было

искать целесообразные технические решения по аналогии с механизмом соответствующих функций мозга [1].

Но, удаляясь от психологических теорий, мы должны крепко держаться за психологические факты, руководствуясь в своих поисках ценнейшим материалом, добытым экспериментальной психологией.

3. Определения. Точные науки оперируют величинами, т. е. категориями, имеющими количественную меру. Невозможно обсуждать проблему опознания образов в техническом плане, не установив определений, содержащих количественную меру.

Мы попытаемся определить образ прежде всего как некоторое подмножество A_k множества A всех образов данного типа [2]. Последнее определяется свойствами восприятия, о чем речь будет ниже (см. п. 6). Что же касается подмножеств A_k, то ясно, что они должны быть непересекающимися. Однако условие

$$\bigcup_k A_k = A$$

не обязательно: могут остаться подмножества $A_i \subset A$, не являющиеся образами.

Итак, абстрагированное восприятие ω некоторого объекта отождествляется с образом A_k, если

$$\omega \in A_k. \tag{1}$$

Введем величины x, y, z, . . ., которые назовем признаками образа. Признаки являются количественно измеримыми величинами. Вся трудность и состоит в установлении величин, выбираемых в качестве признаков образа. Это — центральный вопрос обсуждаемой проблемы.

Условие (1) может быть выражено через значения признаков неравенством

$$f_k(x, y, z, \ldots) \geqslant 0. \tag{2}$$

Если признаки независимы, то (2) заменяется системой неравенств

$$\begin{aligned} x_{1_k} &\leqslant x \leqslant x_{2_k}, \\ y_{1_k} &\leqslant y \leqslant y_{2_k}, \\ z_{1_k} &\leqslant z \leqslant z_{2_k}. \\ &\ldots\ldots \end{aligned} \tag{3}$$

---

[1] За последнее время часто высказывалось мнение, возникшее в результате более близкого знакомства с устройством и действием центральной нервной системы с ее поражающим совершенством. Это мнение состоит в том, что, проектируя устройства, на которые возложены те или иные функции мышления, следует брать за образец принципы и схемы живого организма. Едва ли целесообразно всегда руководствоваться таким правилом. Живая природа имеет свою специфику, свои возможности и ограничения, а техника — свои. Достаточно напомнить принцип колеса, без которого немогло бы быть осуществлено множество наших машин и который по понятным причинам не находит себе применения в живой природе.

[2] Т. е. образов, обладающих одинаковыми признаками (см. ниже).

Неравенства (2) или (3) выражают критерий, на основании которого данный объект отождествляется с образом A_k, или, иначе, критерий опознания образа.

Критерий опознания должен быть необходимым и достаточным. Никакие умозрительные построения не позволяют (пока) судить о том, выполнено ли это требование. На этот вопрос ответ дает опыт.

Естественно рассматривать признаки x, y, z, ... как координаты некоторого n-мерного метрического пространства [1]. При такой геометрической трактовке образ представляется как область A_k в пространстве A. Аналитически область A_k определяется неравенством (2). В простейшем случае независимых признаков область вырождается в n-мерный параллелепипед, определяемый неравенствами (3).

Области A_k, а следовательно, и шкалы признаков x, y, z могут быть как дискретными, так и непрерывными (что соответствует счетным и несчетным множествам).

Легко видеть, что если принять приведенные выше определения, то проблема опознания образа не отличается по существу от проблемы классификации или расположения (filing) [2].

4. Примеры дискретного и непрерывного признаков. В пояснение общих определений приведем два примера.

Пример 1. Пусть множество A есть множество многоугольников. Это множество счетно: число вершин N многоугольника может быть только целым. Число вершин N является единственным признаком. Таким образом, всякий многоугольник при $N=3$ есть треугольник, при $N=4$ — четырехугольник и т. д.

Пример 2. Пусть множество A есть множество монохроматических цветов. Это множество несчетно. Общим и единственным признаком множества является частота (или длина волны) световых колебаний. Что мы назовем зеленым цветом (это является в данном примере образом A_k)? При изменении длины волны цвет изменяется непрерывно; поэтому мы можем совершенно произвольно определить образ зеленого цвета неравенством

$$500 < \lambda < 570 \ (\textit{ммкм}).$$

Нижняя граница соответствует голубовато-зеленому, верхняя желто-зеленому. Произвол относится к назначению границ; можно выбрать их так, что границы, скажем, желтого и зеленого примыкают друг к другу вплотную (перекрываться они не должны никоим образом — это противоречило бы основному условию

[1] Пространство это является метрическим, потому что в нем можно ввести удовлетворяющее обычным аксиомам расстояние. Однако метрика пространства образов определяется свойствами восприятия и может быть найдена только экспериментальным путем. Так (см. п. 7), фонема «И» ближе к «Ы», чем к «Э», но ниоткуда не следует, что, например, метрика формантной плоскости евклидова.

[2] Сходную точку зрения высказывают Таннер [2], Кларк и Фэрли [3] и другие исследователи.

непересечения подмножеств), но можно выбрать и более узкий интервал. Это будет означать либо, что мы отказываемся от опознания промежуточного цвета, либо, что мы рассматриваем этот промежуточный (желто-зеленый) цвет как самостоятельный цветовой образ.

Для полноты картины нужно отметить то очень важное обстоятельство, что несчетное одномерное множество монохроматических цветов может быть сведено к трехмерному конечному множеству (говоря о конечности мы имеем в виду только дискретную шкалу частот, но не интенсивностей). Речь идет о трехцветной системе, применяемой в цветной фотографии и полиграфии.

5. О преобразованиях при опознании формы. Одним из элементов психологической обработки восприятия является преобразование. Мы коснемся здесь некоторых частностей, относящихся к опознанию формы.

Если наблюдается квадрат, произвольно ориентированный, то предполагается, что его проекция на сетчатку может быть приведена к квадрату простейшими преобразованиями: переносом, изменением масштаба, сдвигом и поворотом. Такое мнение неоднократно высказывалось (см., например [1]), но вызывает серьезные возражения.

Прежде всего заметим, что, наблюдая квадрат в произвольной ориентации, мы можем признать за его проекцией лишь свойства параллелограмма: четыре попарно-параллельные стороны. Мы не можем установить, является ли наблюдаемый параллелограмм в действительности квадратом, если отсутствуют какие-либо дополнительные данные. Практически эти данные в большинстве случаев налицо. Так, например, если известна ориентация плоскости, на которой лежит плоский предмет (например, вырезанная из картона фигура на горизонтальном столе), то форма уверенно распознается. К этому нужно еще добавить способность зрения к оценке расстояний (аккомодация и конвергенция). Если же полагать, что указанные дополнительные сведения не даны, то мы сможем лишь отличить параллелограммы от неправильных четырехугольников, многоугольники от эллипсов и т. п.

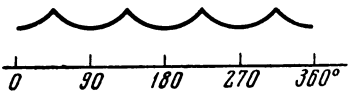
Второе замечание покажет, что простыми преобразованиями, перечисленными ранее, никак нельзя обойтись. Пусть наблюдатель находится внутри квадратной площадки. Обозреть ее контур он может, только повернувшись на 360°. При этом видимое отображение контура площадки будет, как на рисунке.

Из этого примера ясно, что для приведения восприятия формы к ее образу необходимо привлечь преобразования проективной геометрии. Нужно, впрочем, заметить, что во многих случаях отмеченные трудности не возникают. К таким случаям относятся: построение чертежей, проекция на экран, аэрофотосъемка по вертикали и ряд других.

6. **Множества образов.** Понятие образа распространяется на все виды восприятия, множество образов определяется свойствами

восприятия. Множество образов — это по существу совокупность существенно различных восприятий. Это формулировка не может рассматриваться как определение, она имеет лишь поясняющее значение. Положение станет более ясным, если мы перейдем к конкретным чувствам [1]. О форме мы здесь говорить не будем.

а) *Восприятие цвета.* Как уже говорилось, при восприятии цвета мы можем свести все к восприятию трех основных образов: красного, синего и зеленого. Не обязательно брать за основные названные три — можно взять и тройку дополнительных цветов: сине-зеленый, желтый и пурпуровый. Условия выбора таковы:



сумма трех цветов-образов должна давать белый; сумма любых двух цветов должна дать дополнительный к третьему.

б) *Восприятие звука.* Число образов здесь неопределенно велико. Например, любая совокупность музыкальных созвучий может рассматриваться как множество образов. С технической точки зрения, однако, особый интерес представляет совокупность фонем (фонетических элементов речи). Проблема опознания фонем имеет большое значение, и мы обсудим ее ниже в некоторых подробностях.

в) *Восприятие вкуса.* Установлено, что основных вкусовых образов всего четыре: сладкое, горькое, кислое и соленое.

г) *Восприятие запаха.* Обоняние — наименее изученное и поддающееся исследованию чувство. Однако имеется попытка выделить основные образы, образующие так называемую «призму запахов» (построение, аналогичное «цветовому треугольнику»).

д) *Осязательное восприятие.* Опыт показывает, что осязание различает четыре вида раздражения: холод, тепло, давление (собственно — осязание) и боль.

Физическая природа раздражителей в большей части известна; это не относится к вкусу и обонянию. Можно полагать, что вкусовой образ зависит от концентрации ионов определенного типа. По поводу обоняния ничего положительного сказать нельзя.

С точки зрения техники нас могут интересовать в настоящее время только некоторые зрительные образы (в частности, формы) и некоторые звуковые образы (в частности, фонемы).

7. Об опознании фонем. Фонемами называются фонетические элементы речи.

Этот термин применяется для обозначения образа; например, фонема, соответствующая гласному звуку ю, опознается как [*ju*], независимо от того, в каком языке она встречается, произносится

[1] Подробности можно найти в руководствах по экспериментальной психологии. См., например [4]:

она или поется; мужским или женским голосом или шепотом и т. д. Число фонем невелико; полагают, что оно составляет около четырех десятков. Конечно, в различных языках встречаются фонемы, специфичные только для этих языков или для группы родственных языков.

Простейшими для опознания оказались, как это уже давно было установлено, гласные фонемы. Их характерным признаком является наличие одного или двух максимумов в спектре — так называемых формантных областей. Частоты этих максимумов или, коротко, частоты формант и являются количественным признаком при опознании гласных фонем. По данным Л. А. Варшавского в русском языке имеются три одноформантные фонемы, критерий опознания которых может быть выражен неравенствами

$$\text{У}\ 200 < f < 400\ \text{гц}$$

$$\text{О}\ 400 < f < 700\ \text{»}$$

$$\text{А}\ 700 < f\ \ 1100\ \text{гц}.$$

Здесь границы областей примыкают вплотную; при этом форманты распознаются в среднем в 90% случаев.

Кроме того, имеются три двухформантные фонемы, для которых критерий имеет следующий вид:

$$\text{И}\begin{cases} 150 < f_1 < 300\ \text{гц} \\ 2000 < f_2 < 3500\ \text{»} \end{cases}$$

$$\text{Ы}\begin{cases} 200 < f_1 \quad 400\ \text{»} \\ 2000 < f_2 \quad 3000\ \text{»} \end{cases}$$

$$\text{Э}\begin{cases} 400 < f_1 \quad 600\ \text{»} \\ 1500 < f_2 \quad 2500\ \text{гц} \end{cases}$$

Иностранные исследователи полагают, что гласные фонемы английской речи имеют две форманты. При этом очень интересно отметить, что в некоторых работах применяется графическое представление [5], в точности соответствующее выбранным нами определениям. Именно, частоты двух формант f_1 и f_2 принимаются в качестве координат плоскости (двумерное пространство образов). Образы фонем изображаются областями на такой «формантной плоскости».

Что касается остальных фонем, то надежного критерия опознания для них еще не найдено и по этому поводу следует сделать несколько общих замечаний.

На протяжении многих десятков лет многочисленные исследователи с поразительным упорством искали признаки различных фонем в строении спектра. Иначе говоря, априори предполагалось, что признаки для всех фонем едины по своей природе. Конечно, это было бы очень удобно и приятно, но ведь ниоткуда не следует,

что это должно быть так, а неуспех всех усилий заставляет думать, что это и в самом деле не так. Не следует навязывать объективным вещам субъективных закономерностей нашего мышления. Множество образов гласных фонем образует одно пространство, а множество образов, скажем, взрывных фонем — другое пространство. Критерий ищется там, где его нет.

Конечно, вопрос о признаках и, следовательно, о критерии опознания согласных фонем остается до последнего времени открытым. Однако нет сомнения, что эти признаки будут найдены, и тем скорее, чем раньше мы откажемся от представления об универсальности спектрального подхода к проблеме.

Можно указать на одну возможность, представляющую несомненный интерес с точки зрения организации поисков. Речь идет о так называемой ограниченной речи. В этой специальной форме речевого сигнала, сохраняющей высокую степень разборчивости, а следовательно, существенные объективные признаки фонем, единственной объективной характеристикой является расположение нулей. Стало быть, признаки фонем могут содержаться только в тех или иных свойствах этого расположения, которое, разумеется, представляет собой значительно более простой объект для изучения, чем, скажем, осциллограммы обычной речи. Дальнейшее — в значительной степени дело удачи, так как признаки фонем нельзя найти умозрительно.

8. Транспозиция образов. Следует остановиться на одном свойстве образов, имеющем практическое значение: на возможности транспозиции образов. Здесь под транспозицией понимается перенос образа из одного вида чувственного восприятия в другое. Приведем примеры транспозиции.

а) Выпуклый шрифт для слепых (зрение → осязание).

б) Любая машина для чтения для слепых, преобразующая обычный печатный текст в условные сигналы, воспринимаемые осязанием или слухом. Было предложено очень много таких машин. Часто применяется развертка текста одновременно по нескольким полоскам в пределах одной печатной строки [1]. Так, например, Винер [1] описывает схему, в которой чтение осуществляется разверткой по трем полоскам, каждой из которых соответствует музыкальный тон определенной высоты. В результате получается сигнал в виде трезвучия, каждый из трех составляющих которого представляет собой импульсы, соответствующим образом модулированные по длительности.

в) Видимая речь (слух → зрение). Здесь на движущийся экран проектируется мгновенный спектр звука в координатах частота — время. Спектральная плотность мгновенного спектра отображается яркостью в данной точке. Подробные исследования показали [6],

---

[1] Этот принцип предложил, по-видимому, Фурнье д'Альб в 1914 г. Однако применяется и обычная развертка по вертикали в пределах одной строки текста [7].

что человек очень быстро научается «читать» видимую речь, различать индивидуальные голоса и т. п.

По-видимому, мы имеем дело с транспозицией образов чаще, чем может показаться на первый взгляд. Так, например, обычную нотную запись можно трактовать как транспозицию. Холод и тепло мы привыкли во многих случаях распознавать не осязанием, а по показаниям термометра. То же относится и ко многим другим измерениям величин, непосредственно воспринимаемых нашими чувствами. Развивая эту мысль, можно было бы сказать, что измерительные приборы значительно расширяют ассортимент образов, формирующихся в нашем сознании: эти образы могут относиться и к разного рода физическим воздействиям, воспринимаемым нами не непосредственно (т. е. не являющимися раздражителями для наших органов чувств), а с помощью тех или иных приборов.

9. Общие соображения по поводу опознания формы. Рассмотрим теперь несколько подробнее операции, из которых складывается опознание одного вида образов, а именно формы. Исходным материалом будем считать некоторое плоское изображение, например фотографическое, на котором запечатлены проекции на плоскость различных предметов, в том числе и интересующих нас. В числе первых операций в процессе опознания следует назвать выделение контуров (оконтурирование) и фильтрацию (очистку).

Фильтрация состоит в устранении интересующих нас объектов, второстепенных деталей, а также дефектов изображения. Эта операция может выполняться в несколько этапов. На первом этапе, когда качественные характеристики объектов еще не выявлены, грубая фильтрация может производиться по признаку размеров. Так, путем простого усреднения могут быть сняты все мелкие детали, а также дефекты. Эта операция может быть осуществлена также путем фильтрации высоких частот при обычной развертке изображения. Фильтрацией низких частот может быть снято крупное строение фона.

Затем следует операция оконтуривания, которая основана на дифференцировании. Точнее говоря, находятся линии высоких градиентов яркости, которые и принимаются за контуры объектов. Затем производится собственно опознание, которое состоит в обследовании контура на основе установленного критерия. После этого можно произвести повторную фильтрацию, которая состоит в устранении из числа опознанных контуров тех, которые нас не интересуют. И, наконец, выполняется операция, определяемая поставленной целью. Это может быть: подсчет числа избранных контуров, определение их индивидуальных, суммарных или средних размеров, каких-либо характеристик взаимного расположения и т. п.

При попытке хотя бы эскизного проектирования технических устройств, выполняющих указанные операции, возникает прежде-

всего вопрос об установлении количественного признака формы. Если речь идет об опознании N-угольника, то дело обстоит очень просто: нужно лишь выделить углы и пересчитать их. Если же требуется опознать более сложные образы, то изыскание признака сильно затрудняется.

Исследователи из Lincoln Laboratory (MIT) поставили себе для начала довольно скромную задачу: опознание печатных букв [8—10]. Они полагают, что следует стремиться свести количественный признак образа буквы к числу «сгустков» (blobs), т. е. мест сгущения черных элементов в изображении буквы. Нужно заметить, что в цитированных работах успешно осуществлено усреднение, оконтурирование, утончение и утолщение букв, выделение углов и т. п. Однако пока не совсем ясно, каким образом удастся свести описание конфигурации буквы (и притом любого шрифта, что является обязательным требованием) просто к числу сгустков.

Можно высказать следующее (конечно, не бесспорное) положение: на современном этапе едва ли целесообразно искать универсальный признак для опознания любых контуров [1]. Успех при решении конкретных задач может быть, по-видимому, скорее достигнут применением специализированных приемов, использующих индивидуальные особенности контуров данного типа. При этом надо отметить, что выбор признака тесно связан с техническими приемами его обнаружения, и более того: наличие определенных технических возможностей может обусловить выбор того или иного признака.

10. Некоторые технические возможности. В ряде иностранных работ процесс опознания формы, включая и подготовительные операции, возлагается на универсальную цифровую вычислительную машину [10, 11]. С этой целью черно-белое изображение при помощи обычной строчной развертки превращается в совокупность двоичных цифр и в таком виде вводится в машину. Машина, разумеется, может сделать с этой информацией все, что угодно, следуя соответствующей программе. В цитированных работах приведены очень интересные результаты обработки изображения машиной.

Нет сомнения, что применение вычислительной машины в предварительных исследованиях может дать и уже дало очень ценный материал. Но едва ли следует искать на этом пути окончательные технические решения. Не говоря уже о громоздкости и дороговизне универсальных вычислительных машин, следует отметить относительную сложность программ. Так, например [10] программа простого усреднения состоит из 300 команд, а выполнение операции (для изображения 90×90 элементов) требует 20 *сек*; оконтурирование занимает 700 командных регистров и продолжается 2 *мин*.

---

[1] Диннин [10] высказывает противоположное мнение.

Вообще говоря, иметь дело с изображением, представленным в виде решетки с черно-белыми клетками, принципиально невыгодно; выгоднее иметь дело непосредственно с контурами [1]. Поэтому можно ожидать успеха при применении не универсальных, а специализированных методов.

Один из таких методов основан на применении спиральной развертки с расположением центра развертки в центре симметрии фигуры (если он имеется); для этого должно предусматриваться поисковое устройство, способное распознавать симметрию и смещать надлежащим образом центр развертки. Пересечение линии развертки с контуром дает последовательность импульсов, специфичную для данной формы. Преимущество спиральной развертки состоит еще и в том, что она устраняет влияние ориентации контура.

Другой интересный метод состоит в применении следящих систем, заставляющих развертку следовать вдоль контура (см., например, дискуссию по докладу Леба [13]). Технически такая система может быть выполнена с применением круговой развертки малого радиуса. Пересечение линии развертки с контуром дает импульс, фаза которого указывает направление контура; центр развертки прогрессивно перемещается вдоль контура. Легко понять, что запись сигнала управления, т. е. значений фазы, содержит непосредственную характеристику контура: если фаза не меняется, — контур представляет собой прямую; если фаза меняется скачком на угол α, то контур содержит угол α между двумя отрезками; если производная фаза постоянна, то контур имеет постоянную кривизну, т. е. является дугой окружности и т. д. Очевидно, однако, что при применении этого метода неизбежно возникнут топологические затруднения. В простейшей форме метод применим только к универсальным кривым. Во всех же остальных случаях потребуются дополнительные устройства для поиска необследованных участков контура (например, при наличии ветвлений) и для запоминания уже пройденных участков.

11. Ближайшие практические применения. Проблема опознания образов имеет, как легко понять, большой общенаучный интерес. Но нужно также иметь в виду, что на различных этапах решения этой проблемы исследования должны дать ряд практических результатов, некоторые из которых сами по себе имеют немалое значение. Ниже коротко перечисляются некоторые из таких результатов.

А. Распознание букв

Возможность объективного автоматического распознания букв (и цифр) печатного текста непосредственно используется для следующих целей:

1. Построение читающей машины для слепых.

[1] В этом смысле концепция Мэк-Лахлана [12] представляется бесперспективной, она соответствует «мозаике» так называемой эмпирической теории в психологии.

2. Автоматический ввод данных в вычислительные машины (без перевода на перфоленту).

3. Автоматическая передача и переприем телеграмм.

Эта очень важная в технико-экономическом отношении задача до сих пор не имеет удовлетворительного решения.

4. Введение программ в управляющие машины с программным управлением и введение программ в вычислительные машины.

5. Построение приставок-датчиков для машин-переводчиков.

Б. Выделение контуров

Успешное выполнение этой операции, являющейся по отношению к распознанию образов вспомогательной, сразу позволяет решить следующие технические задачи:

1. Построение фототелеграфных систем, предназначенных для передачи простых рисунков, т. е. изображений, информационное содержание которых определяется конфигурацией линий (толщина которых не играет роли). В этом случае возможна передача значительно более экономными методами, чем обычная строчная развертка (например, пантографические системы).

2. Построение системы контурного телевидения. Эта система позволяет не только очень значительно сократить объем телевизионного сигнала, но в применении к так называемому служебному телевидению может иметь серьезные преимущества перед обычной системой, дающей полное полутоновое изображение.

В. Распознание фонем

Решение проблемы объективного распознания фонем позволяет сразу осуществить целый ряд устройств, как-то:

1. Автоматический стенограф, т. е. устройство, записывающее речь в читаемой форме (если не буквами, то хотя бы фонетическими знаками).

2. Автоматы, исполняющие команды и подаваемые голосом.

3. Система передачи речи, основанная на передаче номера фонемы. Такая система (называемая также синтетической, фонемным вокодером и т. п.) наиболее близка к теоретически оптимальной системе передачи речи. Она позволила бы сократить объем телефонного сигнала в десятки раз.

В системе такого рода стирается грань между телефоном и телеграфом. Можно, например, представить себе гибридную систему с микрофоном на передающей стороне и буквопечатающим (или знакопечатающим) аппаратом на приемной стороне. Во всяком случае документальная запись телефонного разговора является одним из непосредственных применений этой возможности.

12. Заключение. Проблема опознания образов относительно нова. Проблема эта имеет чисто кибернетический характер. Ее решение, хотя бы частичное, сулит возникновение новых возможностей, представляющих несомненный интерес с технической точки зрения. Поэтому этой проблеме стоит уделить серьезное внимание.

Очень важно с самого начала по возможности ясно ставить вопрос и формулировать задачу в терминах точных наук. С этой

точки зрения установление определений, содержащих количественную меру, имеет решающее значение. В настоящей статье сделана попытка ввести такие определения для образа, признаков образа и критерия опознания.

Литература

1. *N. Wiener.* Cybernetics. N. Y. — J. Wiley, 1948.
2. *W. P. Tanner.* Theory of recognition. — J. Acoust. Soc. America, 1956, v. 28, N 5.
3. *W. A. Clark, B. G. Farley.* Generalization of pattern recognition in a self-organizing system. — Proc. Western Joint Computer Conf., 1955.
4. *P. C. Вудвортс.* Экспериментальная психология. ИЛ, 1950.
5. *G. E. Peterson, H. L. Barney.* Control methods used in a study of the vowels. — J. Acoust. Soc. America, 1952, v. 29, N 2.
6. *R. K. Potter, G. A. Kopp, H. C. Green.* Visible speech. N. Y., Van Nostrand, 1947.
7. *V. K. Zworykin, L. E. Flory, W. S. Pike.* Letter reading machine. — Electronics, 1949, v. 21, № 6; см. также Electronics, 1946, v. 19, N 8.
8. *O. G. Selfridge.* Pattern recognition and learning information theory.(III London Sympos. C. Cherry (Ed.). London, Butterworth, 1956.
9. *O. G. Selfridge.* Pattern recognition and modern computers. — Proc. Western Joint Computer Conf., 1955.
10. *G. P. Dinneen.* Programming pattern recognition. — Proc. Western Joint Computer Conf., 1955.
11. *R. A. Kirsch, L. Cahn, L. C. Ray, G. H. Urban.* Experiments in processing pictorial information with a digital computer. — Proc. Eastern Joint Computer Conf. 1957; Pictorial information processed on a digital computer. Computer and Automation, 1958, v. 7, N 5.
12. *D. Mc Lachlan.* Description mechanics. — Information and Control, 1958, v. 1, N 3.
13. *J. Loeb.* Communication theory of transmission of simple drawings. Communication theory. (London Sympos., 1952) W. Jackson (Ed.). London, Butterworth, 1953.

О ПРИНЦИПАХ ПОСТРОЕНИЯ ЧИТАЮЩИХ МАШИН

1. Введение. За последнее время появилось много публикаций, посвященных читающим машинам, т. е. машинам, автоматически распознающим печатные и даже рукописные буквы и цифры.

Нет сомнений в том, что эта проблема представляет исключительный интерес не только с точки зрения весьма эффектных технических применений, но и потому, что в ней проявляется современная кибернетическая тенденция в технике, состоящая в возложении на машину все большего числа функций, до сих пор выполнявшихся только человеком.

Посвященная этой проблеме литература являет собой довольно пеструю картину. Имеется столько решений (осуществленных или предлагаемых), сколько авторов. Описываются схемы устройств и программы для вычислительных машин, причем и то и другое поражает своей сложностью, относительно богата патентная литература [1].

Такое положение характерно для любой новой отрасли в начальной стадии развития; можно было бы привести сколько угодно примеров из истории науки и техники. Однако имеется уже достаточно данных, чтобы поставить вопрос о том, каков фонд основных идей, определяющих научное содержание проблемы. Пришло время разобраться в накопленном материале, чтобы выяснить, действительным или кажущимся является различие между многими описанными системами. При этом следует стремиться отделить принципиальные основы от возможных вариантов технических решений, без чего очень затруднительно разобраться в существе дела. Настоящая статья представляет собой попытку сформулировать некоторые основные положения и сознательно ограничена принципиальной стороной дела; вопросы техники затрагиваются лишь по мере необходимости.

2. Определение читающей машины. Читающей машиной будет называться такая машина, которая автоматически распознает буквы, цифры или другие знаки печатного или рукописного текста, предъявленного машине. При этом существенно, что машина способна распознавать знаки из некоторого определенного ограниченного набора.

Машина может ошибаться, можно устроить машину так, что в сомнительных случаях она будет отказываться от опознания или

отмечать недостоверность результата. Относительное число ошибок и отказов может служить мерой ненадежности действия машин.

Технически машина оформляется так, что при наборе из m опознаваемых знаков выход машины образован m проводами, при этом для i-го знака на i-м проводе появляется электрический сигнал. Таким образом, машина в принципе способна прочесть данный знак и снова его отпечатать.

Легко видеть, что ряд устройств, предложенных в качестве читающих машин для слепых, не подпадают под это определение; таковы «оптофон» Фурнье д'Альба (1914 г.) и позднее описанные устройства [2, 3], представляющие собой лишь преобразователи зрительных образов в звуковые или осязательные. В этих устройствах функция опознания остается за человеком.

Так как машине совершенно безразлично употребление букв и цифр, то наша формулировка определяет в сущности машину, узнающую очертания. Мы сохраним для краткости термин «читающая машина».

3. Способ действия читающей машины. Действие любой читающей машины распадается на следующие основные операции: а) предъявление и осмотр; б) составление описания; в) сличение описания с эталонными, т. е. собственно опознание. Осмотр состоит в том, что предъявленное изображение воздействует на некоторый (обычно фотоэлектрический) датчик, вырабатывая соответствующий изображению электрический сигнал. Следующий и самый важный этап — составление описания. Под описанием мы понимаем некоторый сигнал, однозначно описывающий изображение наиболее подходящим для целей опознания образом. В принципе возможно (и желательно) совместить выработку описания с осмотром, но возможно и разделение этих операций. Что касается собственно опознания, то эта операция не представляет никакой принципиальной трудности. Сличение описания предъявленного изображения с эталонными описаниями может производиться либо по наибольшему уклонению, либо по среднему абсолютному, либо по среднему квадратичному. Этим собственно и исчерпывается ассортимент употребительных критериев. Особенно просто обстоит дело, если описания представлены двоичными числами; тогда операция сличения выполняется при помощи обычной электронной арифметической техники. Ясно, что читающая машина должна обладать той или иной формой памяти, в которой хранятся эталонные описания.

4. Развертка. В обширной группе читающих машин осмотр предъявленного изображения осуществляется при помощи так называемой развертки (сканирования). Суть дела состоит в том, что читающий луч движется по некоторой траектории, лежащей в плоскости изображения. Эта траектория может быть непрерывной, но может состоять и из отдельных сегментов. Так, например, обычная строчная развертка, применяемая в телевизионной и фототелеграфной технике, представляет собой совокупность па-

раллельных равностоящих отрезков прямых, покрывающих все поле изображения.

Результат развертки состоит в преобразовании функции двух переменных в функцию одной переменной. Это преобразование не является взаимно непрерывным (топологическим); множество точек плоскости изображения и точек линии развертки негомеоморфно. Но это усложнение снимается, если учесть, что любой осмотр возможен всегда с определенной разрешающей способностью или, что то же, с определенной точностью. Поэтому мы всегда можем себе представить, что на поле изображения наложена решетка (сеть) с клетками (ячейками) определенного конечного размера; в таком случае развертка определяет лишь порядок осмотра клеток, и между клетками изображения и элементами сигнала развертки устанавливается простое взаимно однозначное соответствие.

Весьма важное, и далеко не всеми учитываемое обстоятельство состоит в том, что если развертка непосредственно используется в качестве описания изображения, то выбор развертки может быть сделан некоторым оптимальным образом [1]. Эта мысль высказана Гловацким [4] и развита позднее Блохом [5], указавшим теоретико-информационный подход к решению проблемы.

5. Описание абсолютное и относительное. Для дальнейшего важно установить два вида описания в зависимости от его назначения.

Первый вид описания таков, что по нему можно восстановить описанный объект, разумеется, с той степенью точности, которая заложена в самом описании. Такое описание мы назовем абсолютным. Однако для целей опознания абсолютного описания не требуется. Достаточно такое описание, которое содержало бы лишь отличительные черты данного объекта от всех остальных, входящих в набор. Такое описание мы назовем относительным. Интуитивно ясно, что относительное описание может быть более экономным; на формальном доказательстве этого положения мы не останавливаемся. Таким образом, естественно применять для читающей машины относительное описание опознаваемых объектов.

Заметим, что абсолютное описание определяется свойствами отдельного объекта, а относительное — свойствами всей совокупности объектов, образующих набор.

6. Минимизация описания. Оптимальное описание при прочих равных условиях (каких именно, будет сказано ниже) должно быть как можно более коротким. Чем короче описание, тем, естественно, проще машина.

Нижнюю грань длины описания легко определить исходя из простейших теоретико-информационных соображений. Именно,

[1] Метод «зондов», предложенный Шприком [1], представляет собой интуитивную реализацию этой мысли; дело сводится к построению системы штрихов, подобранных так, чтобы их пересечения с контуром цифры позволили уверенно опознать ее.

если описание приведено к форме двоичного числа, то это число не может содержать меньше цифр, чем

$$J = \log_2 m,$$

где m — число равновероятных объектов в наборе.

Так, например, минимальное описание каждой из 32 букв представляется пятизначным двоичным числом. Это число в сущности выражает номер буквы; такое представление соответствует обычному пятизначному телеграфному коду

Рассмотрим некоторые возможности минимизации абсолютного описания. Пусть для определенности речь идет о черно-белом изображении, вписанном в решетку из M клеток. Считая, что каждая клетка может быть только черной или белой (1 или 0), будем иметь для числа возможных изображений

$$N = 2^M,$$

откуда

$$J = \log N = M$$

(считая все изображения равновероятными). В действительности, существуют ограничения [6], которые по существу сводят к нулю вероятность некоторых изображений. Так, если известно, что каждое изображение состоит только из p черных клеток из общего числа M, то

$$N = \frac{M!}{p!\,(M-p)!}.$$

Еще более серьезное ограничение мы введем, если предположим, что изображение образует контур, т. е. что черные клетки вплотную примыкают друг к другу, образуя непрерывную цепочку. В этом случае для указания места каждой очередной черной клетки достаточно трех двоичных единиц (так как каждая клетка окружена восемью соседними и $\log_2 8 = 3$), и если контур состоит из p черных клеток, то информация в описании этого контура будет равна [7]

$$J = 3p.$$

Дальнейшее сокращение описания получится, если считать тождественными контуры, отличающиеся друг от друга только положением относительно решетки. И, наконец, определенные и может быть существенные ограничения накладываются спецификой заданных очертаний, хотя пока и неясно, как эту специфику учитывать.

Переходя к относительному описанию, можно отметить, что самый первый шаг дает уже значительную экономию по сравнению с абсолютным описанием. В уже упоминавшейся работе Гловацкого [4] применяется простая строчная развертка, полученные с ее помощью абсолютные описания объединяются в дерево, из которого изымаются все этажи, не содержащие ветвлений (т. е. различий).

Аналогичная задача ставится в работе Вада и др. [8], но решается несколько иначе: составляются все попарные разности абсолютных описаний (полученных также путем строчной развертки); выделяется клетка, для которой в этих разностях наибольшее число единиц; отбрасываются все описания, содержащие эту клетку; операция повторяется с оставшимися описаниями, пока все описания ее выйдут из игры. Отмеченные таким образом клетки образуют определяющую решетку. Так, для алфавита 73 буквы при 120-клеточной решетке определяющая решетка содержит всего 44 клетки.

Общая задача о минимизации описаний как абсолютных, так и относительных, не только не решена, но пока еще и не поставлена с достаточной ясностью и полнотой. Между тем, несомненно, что это одна из основных проблем теории узнающих машин. Можно полагать, что ее решение возможно на теоретико-информационной основе.

7. Развертка и координатная система. Как уже говорилось, развертка состоит в последовательном осмотре изображения движущимся лучом. Траекторию движения луча мы называем линией развертки. Линия развертки аналитически задается в некоторой системе координат, связанной с полем изображения. Простейшие виды развертки непосредственно связаны с определенными координатными системами, так что линия развертки воспроизводит координатную сетку. Так, обычная телевизионная развертка связана с прямоугольной системой координат, применяемая иногда спиральная развертка — с полярной системой, развертка барабанного фототелеграфного аппарата — с цилиндрической системой.

Другие специальные развертки вроде применяемой в упомянутом выше методе зондов имеют особого вида траектории, которые, однако, также должны быть заданы в определенной системе координат, связанной с полем изображений.

Существует один вид развертки, существенно отличный от других. Речь идет о так называемой следящей развертке, сущность которой состоит в том, что луч, управляемый специальной следящей системой, следует вдоль контура изображения. Этот вид развертки непосредственно связан, как заметил Э. Л. Блох, с так называемыми естественными координатами.

Уравнение любой кривой может быть задано в виде

$$\omega = f(s), \qquad \text{(а)}$$

где $\omega = 1/\rho$ — кривизна кривой в точке s, а s — длина дуги кривой. Таким образом, в качестве координатной оси используется сама кривая. Путем интегрирования из основного соотношения (а) можно получить

$$v = F(s) + v_0,$$

где v — наклон кривой в точке s. Такой способ задания кривой обладает свойствами, весьма интересными с точки опознания.

Нужно лишь отметить, что при определении координат точки в независимой (т. е. не связанной с изображением) системе координат происходит накопление ошибки, так как эти координаты получаются путем интегрирования с переменным верхним пределом[1].

8. **Топологическое описание.** Из предыдущего видно, что один из способов описания изображения состоит в применении той или иной развертки. Сигнал развертки и является описанием. Такой способ описания можно назвать геометрическим. Существует, однако, другой, принципиально отличный способ описания изображения, состоящий в перечислении топологических признаков описываемого очертания.

Топологическими элементами могут быть замкнутые контуры различной связности и узлы различной кратности. Кроме того, в топологическое описание входят характеристики взаимных связей между элементами.

Связность контуров определяется числом взаимопересечений (например, 0 и 8); кратность узлов определяется числом лучей, выходящих из узла. Так, узел первой кратности есть конец, второй — угол, пример узла третьей кратности — буква Y, четвертой — буква X.

Топологическое описание является весьма общим. Нужно, однако, указать, что оно безусловно недостаточно для опознания букв и цифр, так как многие знаки топологически тождественны. Достаточно указать на цифры 6 и 9, отличающиеся друг от друга только поворотом на 180°. Из этого следует, что при пользовании топологическим описанием обязательно привлечение некоторых элементов геометрического описания, отнесенных к определенной координатной системе.

9. **Инвариантность описания.** Задача опознания очертаний в применении к чтению букв и цифр ставится с таким условием, чтобы опознание было возможно при любом шрифте, любом размере и по возможности любой ориентации знаков. Это требование, разумеется, сильно усложняет дело. (Если бы речь шла об опознании стандартного шрифта, то задача могла бы решаться прямым сличением знаков. Такие устройства и предлагались ранее, но они не представляют интереса, и мы не будем ими заниматься).

При таком условии первостепенный интерес приобретает вопрос об инвариантности описания по отношению к тем или иным преобразованиям. Рассматривая свойства различных описаний, мы легко приходим к следующим заключениям:

I. Геометрическое описание (посредством развертки):

а) простая развертка — описание изменяется при всех линейных (афинных) преобразованиях, как-то: перенос, растяжение, сдвиг, поворот;

---

[1] Ситуация аналогична той, которая имеется в инерциальных навигационных системах, основанных на нахождении координат путем двукратного интегрирования ускорений.

б) следящая развертка — описание через наклон зависит от растяжения, сдвига и поворота, но не зависит от переноса; описание через кривизну не зависит от переноса и поворота.

II. Топологическое описание оказывается инвариантным не только по отношению ко всем афинным преобразованиям, но и при любых непрерывно однозначных (топологических) преобразованиях, поскольку они не изменяют топологических признаков очертания.

С практической точки зрения из этого следует, что следящая развертка представляет значительный интерес, так как она дает описание, инвариантное по отношению к преобразованиям, чаще всего возникающим в действительных условиях. Что касается различий в шрифте, то они могут быть в принципе учтены определенными допусками на некоторое среднее описание. Имея в виду опять-таки практическую сторону дела, следует учитывать, что хотя топологическое описание обладает высокой степенью неизменяемости, но техника получения этого описания пока сложна, о чем можно судить по имеющимся публикациям [9, 10]. К тому же топологическое описание само по себе недостаточно, о чем говорилось выше.

10. Опознание одновременное и ступенчатое. Если подлежащие опознанию объекты характеризуются признаками, каждый из которых может либо быть налицо (1), либо отсутствовать (0), то описание каждого объекта может быть выражено n-значным двоичным числом.

В более общем случае признак может принимать q различных дискретных значений, и тогда описание представляется n-значным числом, записанным по системе с основанием q. В еще более общем случае, когда шкалы признаков непрерывны, описание состоит из наблюденных значений признаков, а эталонное описание состоит из набора интервалов [11]. Однако для наших целей достаточно рассмотреть простейший двоичный случай.

Для опознания нужно: 1) сличить описание предъявленного объекта с эталонными описаниями, запасенными в памяти машины, и 2) выдать на i-й выходной провод сигнал опознания. Если набор состоит из m объектов, то требуется столько же операций сличения (если только процесс сличения не прерывается, когда найдено соответствующее эталонное описание; при такой процедуре число операций сличения сокращается в среднем вдвое). Если число признаков (а следовательно, и двоичных цифр в описании) равно n, то при наивыгоднейшем описании имеем $m=2^n$.

Но можно производить операцию сличения шаг за шагом, сличая поочередно по одной позиции двоичных чисел. При таком порядке можно воспользоваться декодирующей схемой в виде релейной пирамиды, в точности соответствующей кодовому дереву, объединяющему все эталонные описания. Число операций будет при этом всего лишь n, т. е. число операций при втором, ступенчатом, порядке опознания будет сокращено в $m/n=m/\log m$ раз. К тому же

при первом (одновременном) способе операции сличения производятся с n-значными числами, а при втором (ступенчатом) способе — с однозначными. Отсюда следует, что ступенчатый способ приводит к заметному сокращению времени, затрачиваемого на опознание, и к упрощению машины.

Преимуществом ступенчатого способа опознания является также и то, что он без малейших затруднений позволяет пользоваться описаниями разной длины. Но у ступенчатого способа имеется и принципиальный недостаток. Задача опознания совершенно аналогична задаче приема кодированных сигналов. Известно, что прием кодовой комбинации «в целом» (что соответствует одновременному способу опознания) обеспечивает бо́льшую помехоустойчивость, нежели «посимвольный» метод приема (соответствующий ступенчатому способу опознания). Подробности по этому поводу можно найти в работах по теории передачи сигналов (см., например,[12]). Эти соображения надо учитывать при оценке надежности опознания, хотя на данном этапе возможность упрощения аппаратуры играет, по-видимому, главную роль.

11. Надежность опознания. Для обсуждения вопроса о надежности в общем виде удобно воспользоваться геометрической моделью; геометрические представления широко применяются в теории кодов, имеющей прямое отношение к обсуждаемым здесь вопросам. Сущность геометрического представления состоит в том, что последовательность чисел представляется вектором в n-мерном пространстве; каждое число последовательности рассматривается как проекция вектора на соответствующую координатную ось. Опознание будет тем более надежно, чем более различаются между собой описания подлежащих опознанию объектов. Нужно подчеркнуть, что речь идет именно о различии описаний, а не самих объектов, так как объекты нам заданы; нужно выбрать способ описания так, чтобы в описаниях были выявлены и подчеркнуты имеющиеся различия объектов.

С геометрической точки зрения это означает, что концы векторов, представляющих описания, должны быть как можно более удалены друг от друга. Таким образом, понятие расстояния является основным при обсуждении вопроса о надежности. (Такого рода соображения, впрочем, довольно тривиальные, имеются у Фрэнкла [13]; аналогичные соображения высказывались и ранее [14]; с большой полнотой геометрическая концепция развита в работах Мешковского [12, 15].

Теперь становится ясным, каким требованиям должно удовлетворять описание: кроме требования минимальности, обсуждавшегося в п. 6, должно быть удовлетворено также требование наибольшего минимального расстояния между любой парой описаний. Пока неясно, каким общим методом можно достичь желаемого результата; более того, неясно, можно ли сформулировать задачу максимизации расстояний как математическую задачу. Очевидно, что эти важные вопросы нужно решить как можно скорее.

Если описание дано в форме двоичного числа, то геометрия получается очень простой: описания располагаются по вершинам n-мерного куба. Расстояния измеряются в метрике Хэмминга [16] как число ребер куба, разделяющих две данные вершины. С арифметической точки зрения это соответствует числу разрядов, в которых различаются между собой два двоичных числа. Увеличение расстояния возможно в этом случае лишь за счет увеличения числа измерений, т. е. числа признаков, по которым различаются опознаваемые объекты. Положение в точности такое же, как при применении исправляющих кодов, теория которых здесь полностью приложима.

12. Препарирование объектов. Все предыдущие рассуждения предполагали некоторые идеализированные объекты. Так, имея в виду печатные знаки, мы предполагали, не оговаривая этого, что знаки образованы безупречными линиями разумно малой толщины и представляют собой черный рисунок на белом фоне, причем коэффициент отражения света от черного и белого составляет соответственно 0 и 1. Но реальные печатные знаки не таковы; их очертания имеют рваные края; имеются разрывы в тонких штрихах; имеются дефекты, обусловленные как неровной поверхностью бумаги, так и самой печатью, имеются помехи, обусловленные грязью и пылью, контрастность печати может быть невелика. Особо следует отметить случай, когда вследствие особенностей технологии изображение имеет зернистое строение, причем размер зерна может достигать размеров мелких, но характерных деталей изображения. Ясно, что при таких обстоятельствах даже вполне разумная система описания и опознания может оказаться настолько ненадежной, что практически не сможет применяться. Поэтому реальные объекты приходится подвергать предварительной обработке или препарированию, так что описание относится не к реальному объекту, а к его препарату.

Препарирование может включать утончение линий, сглаживание контуров, устранение мелких дефектов печати, устранение мелких пятен постороннего происхождения, повышение контраста и т. п. Препарирование может составлять отдельную операцию, следующую за осмотром при помощи обычной развертки, но можно полностью или частично включить эту операцию в осмотр.

Дальнейшие подробности носят уже чисто технический характер и здесь не рассматриваются. Заметим лишь, что в ряде работ операция препарирования возлагается на вычислительную машину [17, 18].

13. Замечание о самообучении читающих машин. В некоторых работах говорится о том, что читающая машина может быть самообучающейся. Но при ближайшем рассмотрении выясняется, что речь идет о самообучении лишь в сравнительно простом и ограниченном смысле. Именно, различается лишь способ запасания в памяти машины эталонных описаний и в связывании каждого

из этих описаний с наименованием объекта. В несамообучающихся машинах эталонные описания определенных объектов заранее заготовляются человеком и вкладываются в машину в готовом виде. Самообучающаяся же машина сама составляет описание и производит опознание под контролем человека-оператора; от последнего машина получает сигналы одобрения или неодобрения [19]. Под действием этих сигналов описание корректируется, пока не будет достигнута достаточная надежность. На этом процесс обучения заканчивается, и дальнейшего участия оператора уже не требуется. Можно полагать, что в такой постановке вопрос об обучении или самообучении имеет главным образом методологический интерес.

14. Заключительные замечания. В предыдущем изложении содержатся многие элементы классификации читающих машин по принципиальным признакам. Однако с составлением классификационных схем и таблиц, вообще говоря, очень полезных, следует, по-видимому, повременить, пока многие нерешенные вопросы не придут в ясность.

Хотя имеется уже несколько действующих моделей читающих машин, но необходимость разработки принципиальной стороны дела остается очевидной. Нахождение оптимальных решений позволит упростить технику, которая пока что несообразно громоздка. Так, английская машина ERA (Electronic Reading Automaton) состоит, судя по фотографии [20], из пяти (!) порядковых шкафов. Можно ожидать значительного упрощения аппаратуры в будущих разработках. В частности, можно полагать, что такому упрощению будет способствовать полный или частичный отказ от применения арифметической техники электронных счетных машин.

Литература

1. *K. Steinbuch.* Automatische Zeichenerkennung. — Nachrichtentechn. Z., 1958, Bd. 11, N 4.
2. *Н. Винер.* Кибернетика или управление и связь в животном и машине. — «Сов. радио», 1958.
3. *V. K. Zworykin, L. E. Flory.* Rading aid for the blind. — Electronics, 1946, v. 19, N 8.
4. *A. Glovazky.* Determination of redundancies in a set of patterns. IRE Trans. on Information Theory. 1956, IT-2, N 4.
5. *Э. Л. Блох.* К вопросу о минимальном описании. Радиотехника, 1960, т. 15, № 2.
6. *D. Mc Lachland.* Description mechanics. — Information and Control, 1958, v. 1, N 3.
7. *А. А. Харкевич.* Сравнение некоторых возможностей передачи простых рисунков. — Электросвязь, 1958, № 5.
8. *H. Wada, S. Takahashi, T. Iijima.* An electronic reading machine. — UNESCO NS ICIPIH 1959, v. 13.
9. *R. L. Grimsdale, F. H. Sumner, C. J. Tunis.* A system for the automatic recognition of patterns. — Proc. Inst. Electr. Engrs, 1959, ser. B, v. 106, N 26.

10. *H. Sherman.* A quasi topological method for recognition of line patterns. — UNESCO NS ICIPIH, 1959, v. 5.
11. *А. А. Харкевич.* Опознание образов. — Радиотехника, 1959, т. 14, № 5.
12. *К. А. Мешковский.* Анализ одной схемы приема двоичных сигналов. — Электросвязь, 1958, № 12.
13. *S. Frankel.* Information-theoretic aspects of character reading. — UHESCO NS ICIPIH, 1959, v. 2.
14. *А. А. Харкевич.* Фототелеграф с точки зрения телеграфа. — Электросвязь, 1959, № 5.
15. *К. А. Мешковский.* Помехоустойчивые коды. — Электросвязь, 1957, № 8.
16. *R. W. Hamming.* Error detecting and error correcting codes. — Bell System Techn. J., 1950, v. 29, N 2.
17. *G. P. Dinneen.* Programming pattern recognition. — Proc. Western Joint Computer Conf., 1955.
18. *O. G. Selfridge.* Pattern recognition and modern computers. — Proc. Western Joint Computer Conf., 1955.
19. *W. K. Taylor.* Pattern recognition by means of automatic analogue apparatus. — Proc. Inst. Electr. Engrs, 1959, v. 106, ser. B, N 26.
20. Electronic reading automat. — Ser. Engineer, 1956, v. 203; см. также. «Developments in electronic reading machines». — Brit. Communs Electronics, 1957, v. 4, N 5.

О ВЫБОРЕ ПРИЗНАКОВ ПРИ МАШИННОМ ОПОЗНАНИИ

1. Проблема опознания. Опознание состоит в отнесении предъявленного объекта к одному из конечного числа заранее установленных классов. Опознание производится на основании результатов наблюдения.

Представление объекта со всей полнотой и точностью, определяемыми способом наблюдения и разрешающей способностью средств наблюдения, назовем абсолютным описанием (так, при фотографировании животного получается описание его внешнего вида в одной проекции с точностью, определяемой разрешающей способностью объектива и эмульсии; фотографирование в рентгеновских лучах даст описание скелета и отчасти внутренних органов).

Для опознания абсолютного описания не требуется. Более того, для получения практического решения, как правило, необходимо перейти от абсолютного описания к некоторому более экономному представлению объекта, достаточному для опознания, т. е. для отнесения объекта к определенному классу. Такое представление состоит в перечислении некоторых характерных свойств, которые назовем признаками. Совокупность признаков назовем относительным описанием.

Проблема опознания естественно распадается на две части. Первая состоит в установлении признаков, характерных для каждого из классов. Вторая часть состоит в отнесении данного объекта к тому или иному классу на основании его признаков, определенных из результатов наблюдения. Эта вторая часть проблемы — собственно опознание — решается в настоящее время на основе хорошо разработанной теории статистических решений и не вызывает затруднений принципиального характера. Что же касается первой части проблемы, т. е. выбора признаков, то здесь пока ничего определенного нет, и можно полагать, что основное научное содержание проблемы опознания как раз и состоит в настоящее время в установлении оснований для выбора признаков.

2. Признаки. В некоторых случаях подходящие признаки более или менее очевидны. Так, если множество многоугольников разделено на класс треугольников, класс четырехугольников и т. д., то признаком класса является число вершин. Если из множества (нормальных) людей выделено два класса —взрослые и дошкольники, то единственным и достаточным признаком может служить

рост. Представителей монгольской и индоевропейской рас можно довольно уверенно различить по конструкции верхнего века. Для опознания письменных знаков представляется естественным положить в основу топологический критерий; если записать, например, знаки 0, 8, X, Y, P, то легко видеть, что их признаками являются наличие и кратность узлов и замкнутых контуров.

Однако в большом числе случаев признаки совершенно неочевидны. В качестве примера можно привести звуки речи. Они доступны объективному наблюдению и могут быть зафиксированы в виде осциллограмм звукового давления, представляющих абсолютное описание звуков речи. Эти осциллограммы имеют настолько сложную структуру, что их рассмотрение не дает никаких прямых указаний на существенные признаки тех или иных звуков. Между тем, такие признаки существуют; они просты и необыкновенно стойки. Это подтверждается тем, что человек уверенно распознает звуки речи в самых неблагоприятных условиях и при самом различном их произнесении.

Проблема опознания занимает сейчас многих исследователей. Литература по опознанию уже довольно обширна, и число публикаций растет с каждым днем. Очень поучительно проследить, как выбираются признаки. Оказывается, что выбор производится совершенно произвольно, по наитию. Конечно, метод проб в науке не исключен. Но если поиск не направляется какими-либо общими соображениями, то метод проб может оказаться слишком расточительным, а в сложных случаях вероятность случайного успеха настолько мала, что задача становится практически неразрешимой.

Нужно искать возможность определения признаков посредством некоторой регулярной процедуры.

3. Постановка задачи. *а. Абсолютное описание.* Примем, что абсолютное описание выражено значениями некоторой функции, определенной на многомерном интервале. Так, описание звука речи может быть представлено значениями давления как функции времени на интервале, определяемом длительностью звука; описание плоского монохроматического изображения представляется значениями яркости как функции двух пространственных координат, определенной в пределах поля изображения. Множество значений функции, образующих описание, может быть конечным; это зависит как от свойств объекта (например, функции с ограниченным спектром), так и от конечной разрешающей способности средств наблюдения. Число значений функции, входящих в описание, обычно очень велико (например, для звуков порядка 10^3, для изображений порядка 10^6). Для дальнейшего числа это не существенно; оно может быть и бесконечно большим.

Удобно пользоваться геометрическим представлением, состоящим в том, что каждый объект на основании его абсолютного описания изображается вектором (или точкой) в n-мерном пространстве наблюдений H (пространстве объектов). Для дальнейшего

достаточно, чтобы в этом пространстве было определено скалярное произведение (т. е. чтобы пространство H было гильбертовым).

б. Собственные области классов; эталоны. Класс определяется множеством объектов, обладающих некоторой общностью свойств. Так как в действительности имеются лишь сведения о конечном числе образцов — представителей данного класса или, иначе говоря, конечная выборка, то понятие класса как категории, к которой должны быть отнесены и другие, неизвестные образцы, является понятием асимптотическим. Класс относится к выборке примерно так же, как вероятность события к экспериментально наблюдаемой частоте его.

Если воспользоваться геометрическими представлениями, то выборка изобразится пучком векторов, а класс — некоторой областью A в пространстве H, в которой располагаются концы всех векторов выборки и в которую попадут векторы любых других объектов того же класса. Эту область назовем собственной областью класса (рис. 1). Граница собственной области уточняется асимптотически по мере увеличения выборки. Какой-либо другой класс может быть представлен своей выборкой, изображаемой другим пучком векторов с соответствующей собственной областью B.

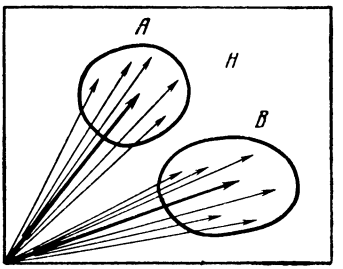


Рис. 1

Основное свойство класса по определению состоит в некотором сходстве его элементов. С геометрической точки зрения, это сходство должно выражаться в том, что точки (концы векторов) каждого класса располагаются более или менее кучно и что собственные области классов разнесены в пространстве H. Это предположение по существу есть то, что М. А. Айзерман и его сотрудники называют гипотезой компактности. Речь идет о том, что элементы одного класса в некотором смысле более сходны между собой, чем с элементами другого класса.

Введем ограничения, довольно жесткие, но зато сильно облегчающие дальнейшие рассуждения. Именно, будем полагать, что поверхности, ограничивающие собственные области, являются, во-первых, непересекающимися, во-вторых, выпуклыми и, в-третьих, гладкими. Первое означает, что любые две поверхности могут иметь не более одной общей точки — точки касания; второе — что каждая поверхность лежит всеми своими точками по одну сторону плоскости, проведенной через точку касания; третье — что для любой точки поверхности существует только одна касательная плоскость.

Класс удобно иногда характеризовать посредством типичного представителя, который назовем эталоном класса. Эталон может быть определен путем некоторого усреднения по классу. Можно

представить дело так, что любой образец данного класса записывается функцией $f(x, \theta)$, где x — детерминированный (может быть, многомерный) аргумент, а θ — случайный параметр. Эталон находится усреднением по этому параметру. (На рис. 1 эталоны изображены жирными линиями).

в. Признаки. Примем, что признаки должны выражаться некоторыми числами, зависящими от функции, представляющей объект. Короче говоря, признаки — это некоторые функционалы.

Разнообразие возможных функционалов совершенно необозримо. Поэтому введем с самого начала одно очень существенное ограничение. Именно, будем полагать, что признак выражается линейным функционалом. Это — очень жесткое условие; его принятие оправдано только стремлением к построению регулярной процедуры нахождения признаков.

Наиболее общий вид линейного функционала есть, как известно, скалярное произведение. Итак, введем определение

$$c_i = x\varphi_i, \tag{1}$$

где c — признак, i — номер признака ($1 \leqslant i \leqslant m$), x — вектор, представляющий предъявленный объект, φ_i — неизвестный пока вспомогательный вектор.

С геометрической точки зрения скалярное умножение соответствует проектированию вектора x на направление φ. Будем полагать, что вектор φ нормирован, так что $\|\varphi\| = 1$; на этом основании будем отождествлять скалярное произведение (1) с проекцией x на φ.

4. Пространство признаков. Относительное описание объекта состоит, по предположению, из m признаков. Взяв численные значения признаков за координаты, можно теперь изобразить объект вектором (или точкой) в m-мерном пространстве, которое назовем пространством признаков и обозначим через Π. С точки зрения техники опознания существенно, что $m \ll n$, т. е. размерность пространства признаков Π много ниже размерности пространства наблюдений H. Проекции векторов одного класса на направление φ_i лежат в некотором одномерном интервале Δ_{ij}, где j — номер класса ($1 \leqslant j \leqslant N$). Иначе говоря, Δ_{ij} есть проекция на φ_i собственной области j-класса в пространстве. Совокупность m интервалов Δ_{ij} образует в пространстве Π параллелепипед, являющийся собственной областью m-го класса в пространстве Π. Задача состоит в том, чтобы найти если не наименьшее, то разумно малое число признаков, обеспечивающее опознание.

Для опознания необходимо, чтобы любая пара классов различалась хотя бы по одному признаку, т. е. чтобы для классов $j=a$ и $j=b$ интервалы Δ_{ia} и Δ_{ib} не перекрывались хотя бы при одном значении i. С точки зрения геометрии пространства Π это значит, что m-мерные параллелепипеды, представляющие собственные области классов, не должны пересекаться.

В пояснение сказанного на рис. 2 в пространстве H изображены собственные области четырех классов. Области касаются одна другой, это, очевидно, — наихудший случай, который имеется в виду и в дальнейшем. Направление φ_1 выбрано так, чтобы разделились классы 1 и 4; направление φ_2 — чтобы разделились классы 1 и 2. Если ограничиться определенными таким образом двумя признаками c_1 и c_2, то в пространстве признаков получится картина, показанная на рис. 3. Классы 2 и 4 при этом остаются неразделенными.

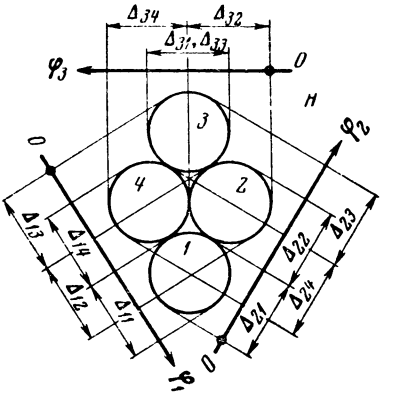


Рис. 2

Добавим теперь в пространстве H третье направление проектирования φ_3 (рис. 2), т. е. введем третий признак. Пространство H станет трехмерным; собственные области всех четырех классов уже не пересекаются (рис. 4), и безошибочное опознание возможно. Рис. 3 можно рассматривать как проекцию трехмерной картины рис. 4 на плоскость c_1c_2.

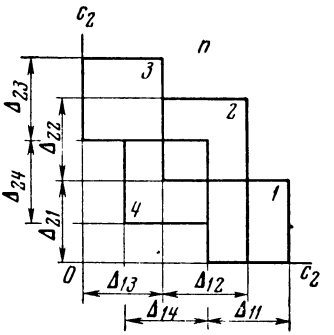


Рис. 3

Рассмотрим вопрос о необходимом числе признаков. Число признаков m зависит как от числа классов, так и от взаимного расположения собственных областей в пространстве H. Но можно легко найти нижнюю и верхнюю грани числа признаков. Нижняя грань есть $m=1$. Это наименьшее значение достигается в том случае, когда все собственные области располагаются в одну цепочку таким образом, что все касательные плоскости параллельны. Верхняя грань достигается при таком расположении, когда каждая область касается всех

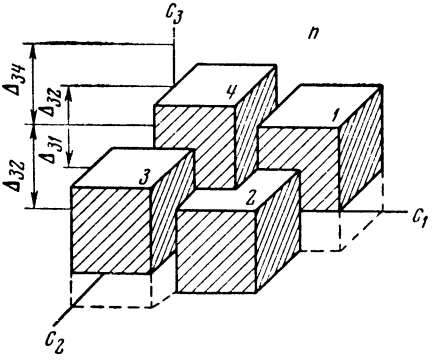


Рис. 4

остальных. Тогда для попарного разделения всех классов нужно провести $^1/_2\, N\,(N-1)$ касательных плоскостей и взять столько же нормальных к этим плоскостям направлений проектирования. Итак,

$$1 \leqslant m \leqslant \frac{1}{2} N(N-1). \qquad (2)$$

Как видно, число признаков m, или, что то же, размерность пространства признаков, не зависит от размерности n пространства наблюдений. Можно, однако, заметить, что вышеупомянутое наихудшее расположение (когда собственные области в пространстве H располагаются по вершинам n-мерного симплекса) осуществимо при условии

$$N \leqslant n+1,$$

которое, впрочем, выполняется во всех практических случаях. Те или иные особенности расположения собственных областей могут заметно сократить необходимое число признаков. Так, для $N=4$ верхняя граница $m=6$, тогда как для расположения рис. 2 достаточно $m=3$.

5. Связь с нумерацией; двоичные признаки. Как известно, самый экономичный способ обозначить элементы некоторого конечного множества состоит в том, чтобы занумеровать их. Если число элементов есть N, то номер будет выражен m_0-разрядным числом, где $m_0=\log_q N$ (q — основание системы счисления) или ближайшее большее целое.

В рассматриваемом случае относительное описание выражается m-значным числом, каждая цифра которого может иметь разное основание $q \leqslant N$. Так, для расположения рис. 2 первая и вторая цифры записаны по четверичной системе, а третья — по троичной (по числу различных интервалов, отмеченных на соответствующих направлениях проектирования). Итак, относительное описание состоит из двух четверичных цифр и одной троичной, или, переводя в употребительную двоичную систему, всего из 5585 двоичных цифр. Между тем при прямой нумерации четырех классов нам хватило бы двух двоичных цифр. Спрашивается, при каких условиях достигается теоретический минимум $\log_2 N$? Легко сообразить, что при рассматриваемом способе образования признаков этот минимум достигается в том случае, когда области классов образуют кубическую решетку. Для примера на рис. 5 изображены восемь классов, области которых расположены указанным образом. Достаточно двух признаков; иначе говоря, относительное описание представлено двухразрядным числом, первая цифра которого четверичная, а вторая — двоичная. Всего получается три двоичные цифры, что и соответствует минимуму, достигаемому при нумерации.

В связи с упоминанием о двоичной системе интересно рассмотреть способ образования признаков, приводящий к относительному

описанию в форме двоичного числа. Способ этот состоит в проведении касательных плоскостей и нормальных к ним направлений проектирования; однако теперь берутся только два интервала — справа и слева от точки пересечения с касательной плоскостью. Первому интервалу приписывается цифра 1, второму — 0. Это показано на рис. 6 для тех же четырех классов, что и на рис. 2. Дело сводится к последовательности дихотомий; оказывается, что в данном случае их требуется пять, так что описания получаются в виде пятиразрядных двоичных чисел (интересно отметить, что

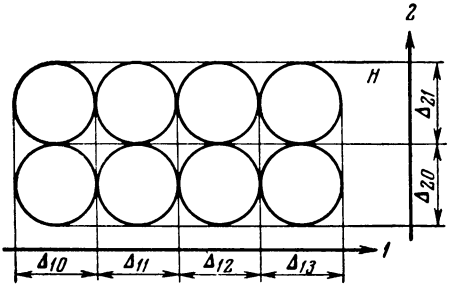


Рис. 5

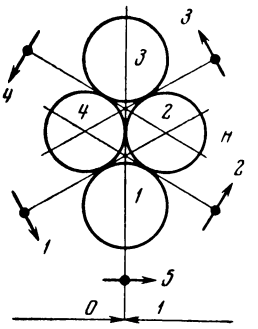


Рис. 6

ранее описанным способом мы получили 5585 двоичных цифр вместо 5). Однако описание данного класса выражается не одним двоичным числом, а одним из нескольких, как показано в таблице и на том же рис. 6 (таблица).

6. Блок-схемы узнающих машин. Процедура, описанная в п. 4, реализуется при помощи устройства, представленного схематически на рис. 7. Для определенности имеется в виду случай $N=4$, $m=3$, соответствующий расположению рис. 2.

1	2	3	4
10010	01001	01100	00010
10011	11001	01101	01010
	01011		00110
	11011		01110
1 и 2	2, 3 и 5	3 и 4	4, 1 и 5
10	100	10	100

Абсолютное описание предъявленного объекта x со входа поступает одновременно на m проекторов Пр. Каждый из них выполняет операцию скалярного умножения (1) и вырабатывает соответствующий признак. Признаки c_i попадают далее на квантующие устройства КУ, которые определяют, в каком из интервалов Δ_{ij} оказывается каждый из признаков. В рассматриваемом случае все квантующие устройства одинаковы; шкала квантования приведена на рис. 7 справа. Затем имеется N схем совпадения И с m входами каждая. Схема И дает сигнал на выходе только при наличии сигналов на всех ее входах. Появление сигнала на выходе одной из схем И завершает процесс опознания.

При применении метода дихотомий, приводящего к двоичной кодовой таблице, можно воспользоваться схемой, воспроизводящей строение кодового дерева. Такого рода схема изображена на рис. 8 применительно к кодовой таблице рис. 6. Предполагается, что проектор Пр производит последовательно требуемые дихотомии, а квантующее устройство КУ имеет один пороговый уровень, так что на выходе КУ получаются двоичные кодовые числа. Эти числа поступают на схему дерева, в узлах которого имеются переключатели на два направления. Каждый ярус приводится в действие очередным разрядом двоичного числа.

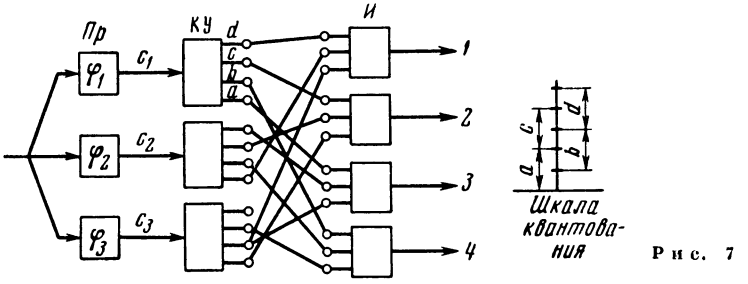


Рис. 7

Можно построить более компактные схемы с регистрами, учитывая то обстоятельство, что кодовые числа каждого класса имеют общие разряды, указанные в таблице и на рис. 6.

7. Примерная процедура. Теперь нужно представить себе процедуру нахождения признаков в практической обстановке, определяемой наличием некоторого количества экспериментальных данных. Искать положение касательных плоскостей путем вычислений (хотя бы при помощи машины) затруднительно, учитывая высокую размерность пространства наблюдений. Это становится и вовсе невозможным, если пространство H бесконечномерно. Поэтому рассмотрим процедуру именно в этом случае: будем считать, что пространство H есть пространство непрерывных функций $f_k(t)$, заданных на интервале $(0, T)$. Скалярное произведение и расстояние определяются формулами

$$f_1 f_2 = \int_0^T f_1(t) f_2(t)\, dt,$$

$$d_{12} = \|f_1 - f_2\| = \sqrt{\int_0^T [f_1(t) - f_2(t)]^2\, dt}.$$

Выбор признака, разделяющего два класса, сводится к выбору направления проектирования, нормального к общей касательной плоскости. Возможна следующая процедура.

1. Находятся эталоны классов путем усреднения по имеющимся образцам.

2. Берется произвольная пара классов $j=1$ и $j=2$. В качестве нулевого приближения весовой функции берется разность эталонов, т. е.

$$\varphi_1^0(t) = f_1(t) - f_2(t).$$

Это выражение является окончательным, когда вектор $f_1 - f_2$ нормален к касательной плоскости (например, когда собственные области сферичны). В общем случае это не так; возможные формы

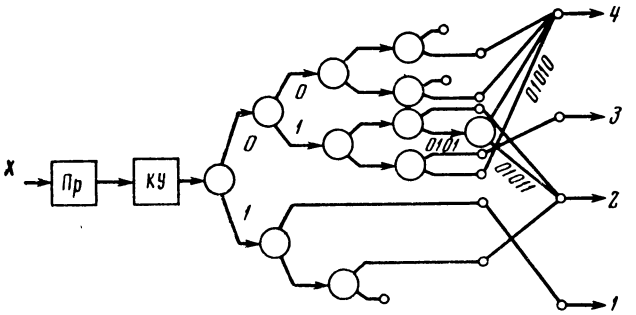


Рис. 8

и расположение собственных областей, а также требуемое направление проектирования показаны на рис. 9.

3. Составляются все проекции на φ_1^0 и отмечаются наибольшие и наименьшие значения для каждого класса. Этим определяются интервалы Δ_{11} и Δ_{12}.

4. Если при проектировании на φ_1^0 интервалы Δ_{11} и Δ_{12} перекрываются, то производится уточнение направления проектирования. Можно поступить так: сохранив f_1 в качестве опорной точки, перебрать все элементы $f_2 l$ класса 2, находящиеся от f_1 примерно на одном и том же расстоянии d_{12}, т. е. лежащие вблизи сферической поверхности, намеченной на рис. 9 штрихом. Проектируя оба класса на каждое из направлений, выбираем в качестве окончательного то из них, для которого Δ_{11} и Δ_{12} не перекрываются. Таким образом, классы 1 и 2 разделены. На это же направление проектируются и все остальные классы.

5. Берется любая следующая пара классов $j=3$ и $j=4$, для которых интервалы Δ_{13} и Δ_{14} оказываются перекрывающимися. Для этой пары классов процедура повторяется, начиная с п. 1, в том же порядке, как и для первой пары.

Процедура продолжается, пока все классы не будут разделены; т. е. пока любая пара классов не будет различаться без перекрытия хотя бы по одному признаку. Из сказанного ранее следует, что эта процедура всегда конечна.

Заключение. Описанная методика не является, конечно, ни универсальной, ни оптимальной в каком бы то ни было смысле.

Но она содержит по крайней мере элементы регулярного подхода к проблеме выбора признаков, к чему и сводится смысл всего изложенного. Возникает целый ряд вопросов, например: как быть, если собственные области невыпуклы? Что нужно делать, если собственные области пересекаются? Как подойти к выбору нелинейных признаков? При каких предпосылках можно рассчитывать на успех эвристического выбора признаков? Каковы общие пути минимизации относительного описания? Не следует ли подвергать пространство наблюдений какому-либо предварительному преобразованию; при каких условиях такое преобразование может быть полезным; как его выбирать?

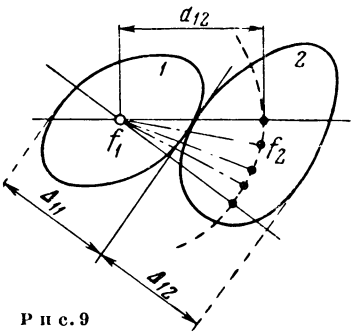


Рис. 9

Эти и многие другие вопросы должны составить предмет дальнейших исследований.

Добавление. При обсуждении изложенных соображений М. С. Пинскер заметил, что иногда (в зависимости от конфигурации) может оказаться выгодным рассечь собственную область на части и обращаться с каждой из этих частей как с самостоятельной областью. При этом число классов возрастает, а число признаков

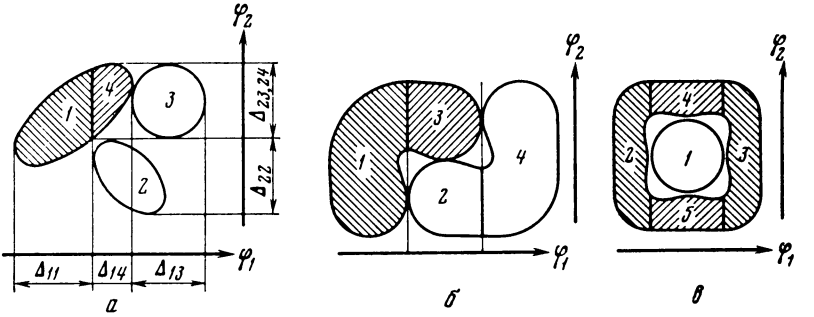


Рис. 10

уменьшается. Эта возможность поясняется рис. 10. На рис. 10,*а* заштрихованная область разделена на части 1 и 4. В результате этого для опознания достаточно двух признаков. Тот же прием можно применить и к невыпуклым областям, как показано на рис. 10,*б*, и даже в случае, когда одна область полностью охватывается другой, как на рис. 10,*в*.

О ЦЕННОСТИ ИНФОРМАЦИИ

1. Введение. Теория информации в ее теперешнем виде игнорирует смысл информации и тем более ценность информации для получателя. Это обстоятельство, вытекающее из принятых определений, настойчиво подчеркивается во всех руководствах по теории информации.

Количество информации определяется как мера неопределенности данной ситуации. Если число возможных равновероятных исходов составляло вначале P_0, а после получения информации сократилось до P_1, то количество полученной информации определяется как

$$J = \log_2 P_0 - \log_2 P_1 = \log_2 \frac{P_0}{P_1}. \qquad (1)$$

2. Задачи с определенной целью. Представляется, однако, возможным обсуждать вопрос о ценности информации в рамках существующей теории, для чего нужно лишь немного видоизменить принятые определения. Все последующее рассуждение основано на представлении о том, что информация собирается для достижения некоторой определенной цели. В пояснение рассмотрим несколько примеров.

а) *Расследование.* Цель, стоящая перед следователем, состоит в том, чтобы обнаружить и изобличить преступника. После ознакомления с основными фактами составляется обычно несколько версий; число подозреваемых может быть велико. Информация, поступающая в виде свидетельских показаний, прямых и косвенных улик, уменьшает число версий и первоначальную неопределенность.

б) *Игра в карты.* Цель игры — выигрыш. Большинство карточных игр основано на том, что вначале неизвестно, какие карты имеются на руках у партнера. По ходу партии информация об этом постепенно возрастает, так что последние ходы опытный игрок основывает уже на полном или почти полном знании чужих карт. Этот пример ясно показывает, что обсуждаемые вопросы могли бы с самого начала трактоваться в терминах теории игр.

в) *Охота по следу.* Цель охоты — настичь зверя. Информация о его пути и состоянии черпается из оставляемых зверем следов (речь идет не только об отпечатках лап на почве); эти следы должны

быть распознаны и отличны от следов других зверей; след может быть утерян и должен быть найден вновь.

г) *Разработка вакцины.* Цель разработки — получение действенной и безопасной вакцины (или вообще лечебного средства). Информация получается в результате длительной серии экспериментов, сперва над подопытными животными, а затем в клинических условиях.

д) *Стрельба по неподвижной цели с корректировкой.* В этом случае информация получается в виде данных о местах фактических попаданий; на основе этой информации исправляется первоначальная наводка.

е) *Стрельба по движущейся цели.* Имеется в виду стрельба с упреждением, т. е. с предвычислением точки встречи. Информация состоит из данных о положении цели в предшествующие моменты времени. Первые разности дают скорость, вторые — ускорение, и по всем этим данным можно с той или иной достоверностью предсказать траекторию цели.

Подобного рода примеров можно привести сколько угодно; все они сходны в том отношении, что информация собирается для достижения некоторой определенной цели. Информация ценна, поскольку она способствует достижению поставленной цели. Одна и та же информация может иметь различную ценность, если рассматривать ее с точки зрения использования для различных целей. Так, сообщение о погоде имеет значительную ценность для охотника, но не представляет обычно никакого интереса для игрока в карты.

Нужно, впрочем, заметить, что не все случаи получения информации укладываются в эту простую схему. Так, например, те виды информации, которые вызывают эмоции, в частности эстетические, остаются пока вне рассмотрения.

3. Определение ценности информации. Ограничимся рассмотрением той категории случаев, в которых цель, ради достижения которой собирается информация, может быть ясно определена. Тогда ценность информации может быть, естественно, выражена через приращение вероятности достижения цели. Если до получения информации эта вероятность равнялась p_0, а после получения информации она стала равна p_1, то ценность полученной информации в указанном выше смысле можно определить [1] как

$$\tilde{J} = \log_2 p_1 - \log_2 p_0 = \log_2 \frac{p_1}{p_0}. \qquad (2)$$

Выбор логарифмической меры определяется обычным условием аддитивности. Сравнивая (2) и (1), мы видим, что оба определения совпадают, если считать $p_0 = 1/P_0$, $p_1 = 1/P_1$.

Таким образом, ценность информации измеряется в единицах информации, и, более того, различие между обеими категориями

[1] Такое же определение вводит Вудворд для величины, называемой им «прирост информации» (information gain) (см. [1, гл. 3, § 3]).

вообще как бы стирается. Однако определение (2) является в некотором смысле более общим. Соотношение (1) предполагает наличие только одного благоприятного исхода из общего числа P_0 или P_1; вообще же говоря, число благоприятных исходов может быть и больше одного. Вероятность же определяется как отношение числа благоприятных исходов к общему их числу. Таким образом, (2) можно рассматривать как результат нормировки числа исходов. В пояснение приведем три схемы рис. 1, *а*, *б* и *в*, на которых приняты одинаковые значения $P_0=2$, $P_1=6$. Исходное положение — точка 0. На основании полученной информации со-

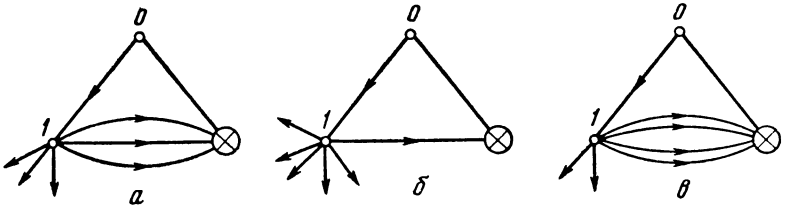


Рис. 1

вершается переход в точку 1. Цель обозначена крестиком. Благоприятные исходы изображены линиями, ведущими к цели. Определим ценность полученной информации во всех трех случаях.

а) Число благоприятных исходов равно трем: $p_0=1/2$, $p_1=3/6=1/2$ и, следовательно,

$$\tilde{J}=\log_2\frac{p_1}{p_0}=\log_2 1=0.$$

б) Имеется один благоприятный исход: $p_0=1/2$, $p_1=1/6$,

$$\tilde{J}=\log_2\frac{1/6}{1/2}=-\log_2 3=-1{,}58.$$

в) Число благоприятных исходов равно четырем: $p_0=1/2$, $p_1=4/6=2/3$,

$$\tilde{J}=\log_2\frac{2/3}{1/2}=\log_2\frac{4}{3}=0{,}42.$$

Как видим, мы совершенно естественно приходим к представлению об отрицательной ценности информации, или просто к отрицательной информации. Отрицательную ценность имеет такая информация, которая, увеличивая исходную неопределенность, уменьшает вероятность достижения цели [1]. Не придумывая новых слов, можно назвать такую отрицательную информацию дезинформацией; таким образом, в примере б) мы получаем дезинфор-

[1] Заметим, что Бриллюэн начисто отрицает возможность появления отрицательной информации в рамках существующей теории (см. [2, стр. 295—296]).

мацию в 1,58 двоичных единиц. В дальнейшем мы будем отождествлять ценность информации с информацией, определяемой согласно (2).

4. Одна простая модель. Очередная задача состоит в том, чтобы составить аналитическую схему рассматриваемой ситуации. Это является пока что очень трудной задачей, общие пути решения которой совершенно не ясны. Мы ограничимся здесь рассмотрением простейшей искусственной схемы, сходной со схемой одномерного случайного блуждания.

Пусть дана прямая с блуждающей точкой, расположенной в исходном положении в начале координат (точка 0). Блуждающая

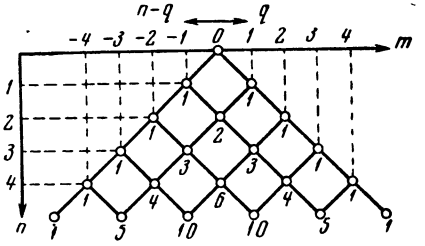


Рис. 2

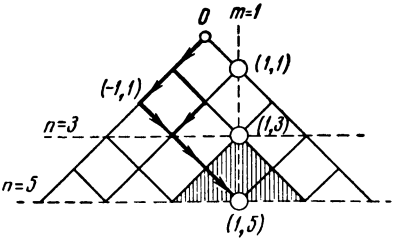


Рис. 3

точка может совершать единичные шаги вправо и влево с равной вероятностью. Цель, которую должна достигнуть движущаяся точка, имеет координату m. В точку с координатой m ведут n/q путей, где n — общее число шагов; q — число шагов вправо. Таким образом, $m=q-(n-q)=2q-n$. Всего же при n шагах имеется 2^n возможных путей. Следовательно, вероятность p достижения цели ровно за n шагов равна

$$p = 2^{-n}\binom{n}{q} = 2^{-n}\frac{n!}{q!\,(n-q)!} = 2^{-n}\frac{n!}{\left(\frac{n+m}{2}\right)!\left(\frac{n-m}{2}\right)!}.$$

Вся схема представляется треугольником Паскаля, изображенным на рис. 2. Цифры около узлов сетки представляют собой биноминальные коэффициенты и выражают число путей, ведущих в данный узел при соответствующем числе шагов. Очевидно, что n и m имеют одинаковую четность.

Движение точки происходит не вполне случайно: каждый очередной ее шаг может быть направлен получаемой информацией; на каждый шаг требуется одна единица информации (так как имеется только два равновероятных исхода: шаг вправо и шаг влево). Будем рассматривать только частный случай $m=1$. В исходном положении для достижения цели достаточно сделать только один шаг вправо; вероятность достижения цели равна $p_0=1/2$. Но пусть на основе полученной информации первый шаг сделан влево, так

что движущаяся точка оказалась в узле 1 (рис. 3). Теперь, чтобы достичь цели, нужно сделать еще не менее двух шагов. Можно поставить задачу двояко:

а) Цель должна быть достигнута ровно за n шагов (включая и первый); $n=3, 5, 7, \ldots$

б) Цель должна быть достигнута числом шагов, не превосходящим n_0; $n_0=3, 5, 7, \ldots$

В принципе возможна еще и такая постановка задачи, когда число шагов не ограничивается; однако эта постановка не представляет интереса, во-первых, потому, что любая практическая задача должна решаться конечной процедурой, а во-вторых, потому, что, как можно показать, вероятность достижения цели стремится к единице при $n_0 \to \infty$.

З а д а ч а а). Вероятности попасть из точки $m=-1$, $n=1$ в точку $m=1$ после ровно n шагов (включая первый) выражаются так:

n	3	5	7	...
p_1	1/4	2/16	5/64	...

Эта табличка составлена следующим образом: знаменатели представляют собой общее число возможных путей (2^{n-1}); числители выражают число путей к точке $m=1$; однако из них исключены пути, ведущие через узлы, соответствующие меньшим значениям n. Будем обозначать узлы значениями их координат (m, n). Тогда из узла (—1,1) в узел (1, 3) ведет только один путь; из (—1,1) в (1,5) ведут всего 4 пути, но два из них проходят через (1,3), остальные два отмечены на рис. 3. Из числа путей, ведущих из (—1,1) в (1,7), сохранены только 5, не проходящих через (1,5) и т. д. Теперь мы можем подсчитать дезынформацию первого шага, учитывая, что $p_0=1/2$. Получаем

n	3	5	7	...
$-J$	1,0	2,0	2,68	...

З а д а ч а б). Комбинаторная задача в этом случае видоизменяется. Число путей определяется следующим образом: при $n_0=3$ имеется только один путь. При $n_0=5$ нужно учесть все пути, проходящие через (1, 3) (заштрихованный треугольник на рис. 3), куда бы они ни шли далее, таких путей 4. К этому нужно добавить пути, ведущие к (1,5), минуя (1,3); таких путей 2 (отмечены на рис. 3). Всего, следовательно, к цели ведет 6 путей из 16 возможных, и $p_1=6/16$. Действуя аналогично, получаем вероятности и дезынформацию:

n_0	3	5	7	...
p_1	1/4	6/16	29/64	...
$-J$	1,0	0,42	0,14	

Описанное рассуждение основано на приведении к ансамблю n_0-шаговых путей. Но можно поступать и иначе, а именно: рассматривать вероятности попадания в точку $m=1$ ровно за

3, 5, 7, . . . шагов как вероятности независимых событий, и попросту суммировать эти вероятности, беря их из таблички задачи а).

Итак, в задаче б) (в отличие от задачи а)) дезынформация убывает с увеличением числа разрешенных шагов. При некотором n_0 J переходит через нуль, так как p_1 все время возрастает, а исходная вероятность $p_0=1/2$.

Эта простая схема не должна рассматриваться как модель какой-либо реальной ситуации; она приведена здесь только для того, чтобы продемонстрировать возможность введения количественных соотношений в вопросе о ценности информации.

Литература

1. *Ф. М. Вудворд.* Теория вероятностей и теория информации с применением к радиолокации. М., «Сов. радио», 128 стр. с черт., 1955.
2. *L. Brillouin.* Science and information theory. N. Y., Acad. press, 1956.

ИНФОРМАЦИЯ И ТЕХНИКА

1. Информация в современных условиях

Первая из черт современной информации состоит в том, что, хотя она и является нематериальным объектом, но получила количественную меру; введено достаточно универсальное определение количества информации. Наука об информации стала точной наукой, применяющей математические методы. Это обусловило возможность необычайно быстрого развития теории информации, свидетелями которого мы являемся.

Теория информации представляет собой молодую (она возникла в послевоенные годы) математическую науку, являющуюся ветвью теории вероятностей. К числу ее достижений, имеющих непосредственное прикладное значение, относится прежде всего установление единого понятия информации вне зависимости от ее источника, смысла, формы и назначения.

В основе определения количества информации лежит представление о том, что всякое сообщение выбирается из некоторого множества возможных сообщений. Чем больше число возможных сообщений, тем больше информации содержит каждое данное сообщение. Так, например, при измерении величины, изменяющейся в определенных пределах, количество информации возрастает с увеличением точности измерения. Затем вводятся вероятности различных возможностей сообщений; считается, что менее вероятные сообщения несут большее количество информации. Так, например, сообщение «завтра, 15 октября, будут заморозки» несет меньше информации, чем сообщение «завтра, 15 июля, будут заморозки». На основании этих представлений и формулируется универсальное математическое определение количества информаций.

К фундаментальным результатам теории информации относятся теоремы о кодировании. Общий смысл этих теорем состоит в том, что при передаче информации по каналу связи возможна безошибочная передача при условии, что скорость передачи не превосходит некоторого предельного значения. Доказано, что возможен способ кодирования, обеспечивающий безошибочную передачу при скорости, сколь угодно близкой к предельной. Эти теоремы имеют очень большое принципиальное значение, так как они устанавливают теоретический предел возможностей в области передачи информации. Но нужно заметить, что в самое последнее время указаны практические методы кодирования, обеспечивающие высокую скорость передачи при очень малой вероятности ошибки. Речь идет о так называемых корректирующих кодах, позволяющих обнаружить и исправить происходящие при передаче ошибки.

Второе обстоятельство заключается в том, что стала ясной всеобъемлющая роль информации не только в сношениях между людьми, но и во взаимодействии человека и машины, а также в жизнедеятельности любого организма. Этим определяется большой интерес к проблемам информации со стороны инженеров всех специальностей, а также представителей естествознания (физиологов, психологов) и общественных наук (в частности экономистов).

Третья характерная черта нынешнего положения: с повышением экономического, технического и культурного уровня общества стремительно растет количество информации, которую нужно собрать, передать и так или иначе использовать для обеспечения всех функций сообщества людей. Эту сторону дела следует обсудить подробнее.

Никакая организованная форма деятельности немыслима без обмена информацией. Без информации невозможно ни планирование, ни управление. Возьмем, к примеру, управление каким-либо промышленным объектом: заводом, цехом или отдельным технологическим процессом. Оно сводится к тому, что на основании сведений о состоянии объекта управления и связанных с ним звеньев принимаются те или иные решения и отдаются соответствующие команды, которые также представляют собой информацию о том, что нужно делать.

Качество информации определяется ее полнотой, достоверностью и своевременностью. Ясно, что, основываясь на неполных, ложных или запоздалых сведениях, нельзя успешно решать поставленные задачи. Нельзя, в частности, принять правильные решения при планировании и управлении. Отсюда и вытекают основные научно-технические задачи в области информации.

Считается, что количество информации растет по меньшей мере пропорционально квадрату промышленного потенциала. Эту закономерность можно пояснить следующим рассуждением. Представим себе, например, объединение нескольких заводов. Между двумя заводами возможна только одна связь, между тремя — уже три, между четырьмя — шесть и т. д. При большом числе заводов число попарных связей между ними равно примерно половине квадрата числа заводов. Если считать промышленный потенциал пропорциональным числу заводов, а количество информации — числу попарных связей между ними, то и получается вышеупомянутая квадратичная зависимость. Результат не изменится, если заводы связаны между собой не непосредственно, а управляются централизованно, например, объединены в трест.

Это, разумеется, не исчерпывает вопроса. Объем информации растет не только с количественным ростом индустрии, т. е. с ростом числа заводов, шахт, электростанций и т. п. Происходит еще и расширение сферы деятельности человека.

Поясним сказанное на примере гидрометеорологической информации. В прежнее время она относилась лишь к ограниченному числу показателей, наблюдаемых в непосредственной бли-

зости от поверхности земли (температура, атмосферное давление, сила и направление ветра, уровень воды и т. п.). С развитием авиации возникла надобность в сведениях, относящихся к более высоким слоям атмосферы, и объем информации сильно возрос (заметим, что одновременно резко возросли и требования к срочности доставки информации и к ее общему количеству; это требуется для составления надежных прогнозов). С развитием коротковолновой радиосвязи возникла потребность в сведениях о состоянии еще более высоких слоев атмосферы; страна покрылась сетью специальных ионосферных станций. С наступлением космической эры нам стали необходимы сведения о физических условиях в зоне, где движутся искусственные спутники и космические корабли. Мы расширяем сферу своей деятельности, и это сопровождается постоянным и быстрым ростом требуемой информации.

Итак, развитие общества и различных форм его деятельности вызывает потребность в огромном количестве информации, на основании которой эта деятельность может быть сделана целесообразной. Сейчас уже видно, что в ряде случаев именно недостаток информации является узким местом, сдерживающим развитие. Так, вскрытые за последнее время просчеты в планировании обусловлены, в частности, неполнотой и неточностью использованной информации.

Рассмотрим теперь несколько подробнее основные проблемы информации, а именно ее передачу, хранение и переработку.

2. Передача информации

Обычно информация возникает в одном месте, а используется в другом. Поэтому самое понятие информации, естественно, связывается с ее передачей на расстояние. С древнейших времен наряду с гонцами, передававшими устные и письменные сообщения, человечество применяло различные сигналы, в частности световые и звуковые. Изобретение электрического телеграфа ознаменовало начало эры электрической связи.

Современные технические задачи в области передачи информации определяются новыми условиями и вытекающими из них новыми, более высокими требованиями. С одной стороны, резко повышаются требования к достоверности передаваемой информации. В процессе передачи информация может быть искажена как вследствие неисправности аппаратуры, так и в результате действия разного рода помех естественного или искусственного происхождения. Последние будут применяться именно тогда, когда потребуется передавать особо важную информацию, от достоверности которой весьма многое зависит (например, в военных условиях).

Искажение текста телеграммы бытового содержания обычно не влечет за собой серьезных последствий, тем более что ошибка чаще всего очевидна. Передача цифровых данных, относящихся

к тем или иным материальным ценностям (например, сведения, передаваемые в системе ЦСУ или Госбанка), требует гораздо более высокой степени достоверности. И, наконец, информация, относящаяся к обороне, имеющая жизненное значение для всей страны, должна передаваться с наивысшей достоверностью, которая может быть обеспечена современными средствами. А так как на достоверность оказывают влияние в первую очередь разного рода помехи, накладывающиеся на передаваемые сигналы, то ясно, что борьба с помехами является одной из актуальнейших задач в области передачи информации.

Наряду с повышением требований к достоверности во многих случаях резко повышаются также требования к скорости передачи информации. Говоря о скорости, нужно различать три различных понятия.

Во-первых, можно говорить о скорости распространения самих электрических сигналов. Эти сигналы распространяются в виде электромагнитных волн вдоль проводов или в открытом пространстве со скоростью, практически равной скорости света (300 тысяч километров в секунду). Эта скорость физически не может быть превзойдена. Для земных условий она достаточно велика, но нужно отметить, что на космических расстояниях — таких, как расстояния до ближайших планет, — время прохождения сигнала составит уже минуты; двусторонняя связь в таких условиях сопряжена со значительными неудобствами.

Во-вторых, можно иметь в виду допустимую задержку между моментом отправления сообщения и моментом его получения в месте назначения. Если эта задержка для письма измеряется сутками, а для телеграммы — часами, то для важных видов информации допустимая задержка измеряется единицами секунд. Большая задержка может совершенно обесценить информацию и повести к самым тяжелым последствиям.

В-третьих, под скоростью передачи информации может пониматься количество информации, передаваемой в единицу времени (в этом смысле термин «скорость передачи» и применяется в дальнейшем). Средствами современной техники можно передать очень большое количество информации за короткое время. Если оценивать скорость передачи таким привычным показателем, как число слов в минуту, то быстродействующий телеграф может передать несколько сотен слов в минуту. Существуют реальные системы, способные передать несколько сотен тысяч слов в минуту, и это еще далеко не предел технических возможностей.

За время после второй мировой войны техника передачи электрических сигналов обогатилась рядом новых способов. К их числу относятся: передача, основанная на использовании рассеивания радиоволн от неоднородностей верхних слоев атмосферы (тропосферы и ионосферы); радиосвязь, применяющая отражение радиоволн от метеорных следов, представляющих собой столбы сильно ионизированного (а следовательно, обладающего большой

электропроводностью) газа. Получили значительное распространение радиорелейные линии, представляющие собой цепочку приемопередающих станций, расположенных друг от друга на расстоянии прямой видимости. Продолжены уже опытные участки волноводных линий. Волновод в простейшем виде представляет собой металлическую трубу; его особенность состоит в том, что частота распространяющихся в нем электромагнитных волн очень велика — порядка миллиардов колебаний в секунду. Это позволяет передавать по волноводу информацию с очень большими скоростями. По сведениям зарубежной печати, ведутся опыты по использованию для целей связи искусственных спутников Земли. В самое недавнее время открылась интереснейшая возможность применения для передачи информации остро направленных пучков световых волн, излучаемых в определенных условиях молекулами вещества. Эта возможность привлекла широкое внимание физиков и инженеров, хотя перспективы здесь еще недостаточно ясны.

Как уже говорилось, техника передачи информации при помощи электрических сигналов достигла высокого уровня. Передача информации осуществляется с требуемыми скоростью и достоверностью многими способами на любые расстояния земного масштаба. Однако современность выдвигает новые задачи. Уже при посылке космических лабораторий-автоматов на ближайшие планеты требуется осуществить двустороннюю связь на расстоянии десятков миллионов километров. Трудность этой задачи усугубляется тем, что на космических объектах жестко ограничиваются вес и габариты аппаратуры. Это означает ограничение энергетических ресурсов, а следовательно, мощности передатчика; затрудняет применение антенн больших размеров.

При осуществлении дальней космической связи возникает и ряд совершенно новых проблем: укажем лишь одно обстоятельство, не играющее роли в земных условиях, но приобретающее решающее значение в условиях космической связи. Речь идет о дискретной природе электромагнитного излучения, т. е. о том, что энергия испускается излучателем не непрерывно, а отдельными порциями — квантами. При высоких частотах и больших расстояниях число квантов, достигающих приемника, настолько уменьшается, что прием сигнала затрудняется, а ниже определенного предела становится и вовсе невозможным. Это не означает, конечно, что мы не сумеем осуществить космическую связь на любых расстояниях; приведенный пример показывает лишь, что системы космической связи должны строиться с учетом ряда новых факторов.

3. Хранение информации

Существует информация оперативная: она используется немедленно и по использовании не сохраняется. Другой вид информации представляют собой факты, предназначенные для долговременного хранения. Примером оперативной информации могут

служить данные о траектории ракеты, получаемые в процессе запуска и служащие для корректирования ее полета. Примером информации долговременного хранения являются всякого рода учетные данные, могущие быть использованными неоднократно в любое время, например, для изучения тех или иных статистических закономерностей.

Трудность проблемы долговременного хранения информации определяется тем, что количество фактических данных, добываемых человечеством, растет во все убыстряющемся темпе. Ярким примером может служить информация, заключенная в научных публикациях. По данным Всесоюзного института научной и технической информации, мировой книжный фонд составляет 30 млн. названий, число патентов — 12 млн.; ежеминутно в мире выпускается около 2 тыс. страниц печатных публикаций. Считается, что число научных публикаций растет в геометрической прогрессии, а именно удваивается за каждые 10—15 лет. Ясно, что при таких условиях проблема состоит уже не только в том, чтобы зафиксировать и сохранить всю эту информацию, но и в том, чтобы при надобности разыскать и извлечь ее из хранилищ. На такого рода розыски тратится колоссальное количество времени и труда. Доходит до того, что иногда проще и быстрее выполнить какое-либо исследование заново, чем разыскивать результаты аналогичных исследований в литературе.

Проблемы, возникающие в связи с хранением научной и другой долговременной информации и с организацией информационно-справочной службы, совершенно специфичны и очень трудны. В самых общих чертах они состоят в следующем. Прежде всего нужно создать систему классификации, которая позволила бы упорядочить всю массу хранящихся и вновь поступающих сведений. Эта система не должна изменяться при неограниченном росте фонда данных. Необходимо разработать специальный условный язык, при помощи которого можно было бы кратко обозначать (индексировать) каждый источник данных. При разработке такого языка следует предусматривать возможность автоматизации поиска.

Ценность данных, находящихся в любом хранилище, существенно зависит от того, насколько быстро ими можно воспользоваться. Поэтому необходимо стремиться не только к автоматизации поиска, но и к автоматизации самого извлечения справки. Ясно, что печатная продукция, являющаяся обычным средством хранения данных, совершенно для этого не приспособлена. К тому же книги и журналы занимают очень много места. Поэтому стоит вопрос об изыскании иных носителей информации. Применение микрофильма позволяет резко сократить физический объем хранилищ. Но предпочтительны такие виды носителей, с которых информация может быть непосредственно считана в форме электрических сигналов. Примером является всем известная запись на магнитную ленту. Однако в современном своем

виде магнитная запись, конечно, не удовлетворяет требованиям информационной службы, и предстоят еще длительные поиски приемлемого решения.

Поставленное выше требование автоматического получения данных из хранилища в форме электрических сигналов важно потому, что при его выполнении становится возможным быстрое получение требуемых справок на любом расстоянии от хранилища с использованием современных средств передачи информации.

Самая трудная задача — это автоматизация сортировки поступающих данных и аннотирование содержания источника. Решение этой задачи было бы чрезвычайно упрощено, если бы в результате специального международного соглашения каждое печатное издание заранее снабжалось едиными по форме индексом и аннотацией на всеобщем информационном языке и в форме знаков, приспособленных для прочтения их машиной.

В будущем можно представить себе центральное хранилище информации, полностью автоматизированное, включенное в сеть связи и позволяющее по запросу немедленно или с минимальной задержкой получить в любом месте любую справку (библиографическую, статистическую и т. п.).

О больших проблемах хранения информации и информационной службы здесь говорится преимущественно в будущем времени. Действительно, они далеки от окончательного решения: ведь их значение понято лишь совсем недавно. Ныне они стоят во весь рост. А это в свою очередь обусловлено тем, что быстрый рост запаса фактических данных привел к качественно новой ситуации.

4. Переработка информации

Всякая информация собирается с определенной целью и для достижения этой цели подвергается соответствующей переработке. Так, первичная информация, получаемая от контрольных приборов, подвергается соответствующей переработке для управления ходом технологического процесса; первичные данные народнохозяйственного учета могут подвергаться переработке с целью выявления статистических закономерностей; иная переработка тех же данных дает основу для планирования.

Всякое решение есть результат переработки имеющейся информации; целесообразность решения, как уже говорилось, находится в прямой зависимости от полноты, своевременности и достоверности информации. Очень важно, что переработка информации сводится к формально-логическим операциям в гораздо большем числе случаев, чем можно было бы думать. Творческие или волевые акты здесь чаще всего не требуются, а иногда могут только испортить дело. Не нужно творческого начала для составления прогноза погоды или для составления плана железнодорожных перевозок. Это значит, что во всех подобных случаях переработка информации вплоть до принятия решения может быть возложена

на машину. Такое заключение имеет само по себе огромное революционизирующее значение: колоссальное количество людей, занятых формальными видами умственного труда, может быть высвобождено и использовано для более высоких форм человеческой деятельности. А их прежняя работа может быть выполнена машинами с не меньшим, а иногда и с большим успехом. Это уже подтверждено фактами. Имеется немалое число планов (например, в области перевозок), для которых машинами найдены более выгодные варианты, нежели предложенные людьми.

Но есть и еще одно обстоятельство первостепенной важности. Оно состоит в том, что человеческие возможности ограничены. Во-первых, человек выполняет арифметические и логические действия сравнительно медленно; машина работает в тысячи раз быстрее. Во-вторых, — и это самое главное — человек не может оперировать одновременно очень большим количеством исходных данных. А ведь именно это и требуется при решении сколько-нибудь сложной задачи, например, планирования. На современном уровне развития хозяйства имеется огромное число отраслей, относящихся к сырьевым ресурсам, энергетике, транспорту, обрабатывающей промышленности, торговле, культуре и быту. Все эти отрасли связаны множеством взаимодействий, переплетающихся самым сложным образом. Только полный учет этих взаимных связей может привести к правильному решению плановых задач, начиная от таких сравнительно мелких, как определение объема производства швейных машин или готовой одежды для нужд населения, и кончая таким важнейшим делом, как составление топливного баланса страны.

Человеческая голова просто не вмещает всего необходимого количества исходных данных. Это заставляет поручать планирование многим людям; планирующие органы делятся на управления, управления — на отделы. Но в итоге вместо того, чтобы объединять исходные данные с сохранением действительно существующих взаимосвязей, мы вынуждены их разобщать. Однако дальнейшее развитие общества и усложнение взаимосвязей между различными сферами его деятельности приведут к более тяжелому положению, если своевременно и радикально не изменить техническую основу плановой работы.

Необходимо широчайшим образом использовать для переработки больших количеств информации электронные вычислительно-логические машины. Они уже достигли высокого совершенства. Но нужно развивать их и далее: повышать их надежность и быстродействие, увеличивать емкость запоминающих устройств, организовать их взаимодействие с информационными машинами, обслуживающими хранилища долговременной информации. Следует изыскивать наиболее удобные методы ввода в машину первичной информации и вывода из машины результатов ее переработки. И, наконец, требует самого усиленного внимания проблема взаимодействия машины и человека.

Итак, основную массу работы по переработке информации не только можно, но и необходимо переложить на машины. Так как при этом единым процессом охватывается огромное количество разнообразной первичной информации, то использование машин, естественно, представляется как организация крупных вычислительных и управляющих центров, к которым стекается с периферии первичная информация, так сказать, «информационное сырье». Эти центры могут быть в какой-то степени (пока не вполне ясно, в какой) специализированы; однако между ними должны существовать постоянные связи, чтобы результаты работы одного центра полностью учитывались другими.

Одним из следствий такой организации явится резкое, скачкообразное увеличение количества информации, передаваемой в различных направлениях. Если сейчас информацией обмениваются главным образом люди, то в предвидимом будущем основными потребителями информации станут машины, в частности вычислительные, справочные и управляющие центры. Общее количество передаваемой информации возрастет в десятки раз, а может быть, и еще больше. Эта перспектива приводит к постановке ряда новых проблем.

5. Единая система связи

Нам привычно представление о том, что деятельность промышленности и сельского хозяйства тесно связана с транспортировкой грузов: сырья, топлива, полуфабрикатов и изделий. Так же привычно и представление о транспортировке электрической энергии. Но за последнее время, как говорилось в начале статьи, на передний план выступает информация, являющаяся организующим началом любой формы человеческой деятельности.

Информацию также нужно транспортировать, т. е. передавать из одного места в другое. Сама информация нематериальна, но она возникает в результате материальных процессов и может быть передана только при помощи материальных процессов. Именно поэтому между проблемами транспортировки грузов и энергии, с одной стороны, и проблемой транспортировки информации — с другой, существует определенное сходство.

Железные дороги СССР образуют железнодорожную сеть, единую в организационном и техническом отношениях. Мы никогда не говорим «единая железнодорожная сеть» потому, что указанное единство представляется (по крайней мере нам в условиях социалистического государства) совершенно естественным и необходимым. Применительно же к электроэнергетике мы пользуемся термином «единая энергетическая система» (ЕЭС); обусловлено это тем, что слияние систем различных энергетических районов в одно целое произошло сравнительно недавно. Преимущества такого объединения настолько очевидны, что нет надобности здесь о них говорить.

Совершенно естественно возникает идея о необходимости построения единой общегосударственной системы передачи информации, обеспечивающей все нужды страны с учетом всех тех новых обстоятельств, о которых говорилось выше. Для краткости будем в дальнейшем называть эту систему единой системой связи (ЕСС), так как передача информации и связь — это по существу одно и то же. Здесь даем некоторые подробности. В стране, разумеется, существует уже система связи, и притом весьма обширная. Но эта система в будущем (а отчасти уже и сейчас) нас не может удовлетворить. Во-первых, она территориально не полна, т. е. не обеспечивает передачу информации из любого места страны в любое другое (уточним: речь идет об электрической связи). Во-вторых, пропускная способность по магистральным направлениям недостаточна.

Конечно, оба недостатка относятся только к количественным характеристикам сети; они устраняются путем наращивания сети, которое происходит все время. Важно другое, а именно: существующая сеть не обладает ни организационным, ни техническим единством. Представляется естественным, чтобы сеть связи имела одного хозяина в лице Министерства связи СССР. Между тем у нас до сих пор очень много ведомственных линий и сетей, разумеется, далеко не полностью загруженных. На сооружение и эксплуатацию этих сетей затрачиваются большие средства. Не будем здесь анализировать причины, приведшие к такому положению. Нельзя, однако, не заметить, что нечто подобное в железнодорожной сети (т. е. наличие ведомственных железных дорог) показалось бы нам совершенно противоестественным.

Что касается технического единства, то речь идет о том, что по существующей сети ведется передача посредством телефона, телеграфа, фототелеграфа, вещания, телевидения и (пока в очень небольших количествах) так называемой цифровой информации. Все эти различные виды сообщений предъявляют совершенно разные требования к оконечной аппаратуре, к линиям и коммутационным устройствам. Дело усложняется еще и тем, что сеть связи состоит из линий с существенно различными характеристиками и эксплуатационными свойствами. Самую развитую часть существующей сети образуют телефонные каналы со своими коммутационными узлами.

Техническое единство является необходимой предпосылкой единства вообще. Чтобы пояснить это грубым примером, отметим, что железные дороги нельзя было бы объединить в сеть, если бы они имели разную ширину колеи; электрические станции нельзя было бы объединить в систему, если бы они работали на разной частоте. Применительно к системе связи речь идет о более тонких вещах, а именно, о том, чтобы различные виды сообщений (перечисленные выше) передавать посредством сигналов одного и того же типа. При выборе единого типа сигнала следует, очевидно,

ориентироваться на тот вид сообщений, который будет играть преобладающую роль.

Из сказанного ранее следует, что основными потребителями информации будут вычислительные и управляющие центры. Информация к ним должна поступать в естественной для них форме — в виде последовательности некоторых чисел. Это и есть «цифровая информация», упомянутая выше. Передача информации в цифровой форме ведется точно так же, как передача текста по телеграфу: каждой букве или цифре на основании принятого кода присваивается сигнал в виде определенного сочетания электрических импульсов. Такие сигналы называют кодированными, а соответствующие методы передачи — кодовыми. Оказывается (это уже специальный вопрос, который здесь обсуждать нет возможности), что кодовыми методами могут пользоваться и телефон, и фототелеграф, и телевидение. Таким образом, возможно унифицировать метод передачи вне зависимости от характера передаваемого сообщения (различные виды сообщений будут отличаться друг от друга только по скорости передачи информации). Этим закладывается основа технического единства ЕСС.

Обсудим теперь некоторые вопросы построения ЕСС. Общие требования к системе легко могут быть сформулированы, если стать на точку зрения потребителя (отправителя и получателя информации). ЕСС предоставляет ему свои услуги и должна дать определенные гарантии. Потребителю должно быть гарантировано, что:

— сообщение будет доставлено по назначению;
— время, затраченное на доставку сообщения, не превзойдет заранее обусловленного;
— искажение сообщения не превзойдет допустимого;
— будет обеспечена передача с требуемой для данного сообщения скоростью.

Потребителя совершенно не интересует, какими техническими средствами обеспечивается выполнение этих естественных и, казалось бы, простых требований. Но их реализация ставит целый комплекс труднейших технических задач.

Попытаемся теперь взглянуть на ЕСС с технической точки зрения. Придется, конечно, сделать это в самых общих чертах, не вдаваясь в специальные вопросы. Одна из первых задач — выбор рациональной структуры сети. Должна быть обеспечена связь между любыми двумя абонентами, охваченными сетью. Для этого, разумеется, не требуется соединять всех абонентов попарно отдельными линиями; обычно применяется радиальная система, в которой абоненты данной территориальной группы соединены линиями с коммутационным узлом, а узлы связаны между собой магистральными линиями с должным образом выбранной пропускной способностью. Говоря о структуре, имеют в виду расположение узлов и строение системы магистралей. Оптимальная структура сети обеспечивает все требуемые связи с наименьшими затра-

тами. При этом учитывается географическое расположение объектов, их взаимное тяготение. Но этого мало: структура сети должна обеспечивать достаточный запас живучести. Это значит, что в случае перегрузки или выхода из строя тех или иных участков сети должна предусматриваться возможность направления потока информации по обходным путям.

Требования в отношении достоверности и допустимой задержки, очевидно, неодинаковы для разных видов сообщений. Здесь следует установить несколько категорий с соответствующими количественными нормативами по обоим указанным показателям. Это сильно усложняет работу коммутационного узла. Во-первых, он должен предоставлять соединение в зависимости от категории по допустимой задержке, а во-вторых, включать дополнительные кодирующие устройства в зависимости от категории по достоверности. Вообще функции коммутационного узла в ЕСС много сложнее, чем, скажем, функции АТС. К тому же и техника коммутации будет совершенно иной: на узлах ЕСС не будет ни механических реле, ни искателей; соединения будут осуществляться с помощью чисто электронной аппаратуры и, по-видимому, с использованием особых свойств кодированных сигналов. На узле будет решаться также задача обеспечения требуемой скорости передачи. Для получения нужной пропускной способности узел будет выбирать для передачи данного сообщения подходящий канал или подключать несколько каналов одновременно.

Нелегкой задачей является увязывание в единый комплекс различных видов линий: кабельных, волноводных, радиорелейных и всех видов радиолиний, в том числе тропосферных, метеорных, использующих спутники, и др. Каждый из этих видов имеет свою специфику; нужно не только унифицировать входы и выходы оконченных устройств, но и выработать согласованный режим эксплуатации.

Сказанное дает некоторое представление о сложности функций ЕСС. Четкое выполнение этих функций возможно лишь при условии, если вся сеть будет работать под централизованным автоматическим управлением. Управляющий центр ЕСС должен располагать исчерпывающей информацией о состоянии сети в каждый данный момент (о технической исправности и нагрузке всех звеньев сети, а также об изменениях в потоках информации). Центр должен быть в состоянии прогнозировать эти изменения на некоторое время вперед и создавать необходимые оперативные резервы. Центр назначает маршруты прохождения потоков информации в зависимости от общей обстановки на сети; в случае надобности центр может сосредоточить все ресурсы сети для выполнения особых заданий по передаче информации. Короче говоря, управляющий центр ЕСС выполняет функции диспетчерского управления сетью. По причинам, изложенным выше, невозможно поручить эти функции людям. Центр будет представлять собой большую группу специализированных вычислительно-логических

машин, предназначенных для решения в непрерывно изменяющихся условиях одной и той же задачи: обеспечение наиболее благоприятных условий для прохождения по назначению всех наличных потоков информации.

ЕСС будет представлять собой крупнейший инженерный комплекс, который должен вобрать в себя всю существующую сеть связи и развиваться путем планомерного ее наращивания в органическом взаимодействии с системой вычислительных, управляющих и справочных центров.

Ряд общих принципов, обсужденных в этой статье, нашел уже применение за рубежом. Но там эти принципы реализуются в виде специализированных систем узковедомственного назначения. Такова, например, крупная система «Сэйдж» (США), управляющая разветвленным комплексом противовоздушной обороны; имеется ряд систем, обслуживающих отдельные объединения промышленных предприятий, и т. п.

Такой грандиозный замысел, как создание общегосударственной единой системы связи, обеспечивающей в современных условиях и современными средствами все нужды страны, осуществим в полной мере только в социалистическом государстве, в условиях планового хозяйства и централизованного руководства.

НЕКОТОРЫЕ ВОЗЗРЕНИЯ НА МЕХАНИЗМ ТВОРЧЕСКОГО ПРОЦЕССА

Попытки выяснения механизма такого тонкого процесса, как процесс творчества, приобретают в наше время особый интерес в связи с успехами кибернетики, для которой вопрос о принципиальных границах воспроизведения машиной функций человеческого интеллекта очень важен, но далеко не решен.

Определение продукта творчества. Результатом творческой деятельности в сфере искусства является произведение; в сфере науки — метод или теория. Продукт творчества можно представить как некоторое сочетание первичных элементов, уже известных и употребительных в данной области. Так, музыкальное произведение составлено из элементарных звуков, произведение архитектуры — из элементарных архитектурных форм, литературное произведение — из слов. Математические или физические теории также представляют собой сочетания некоторых первичных символов и понятий.

Творчество, как выбор. Творческим произведением является не всякое сочетание элементов, а лишь такое, которое в той или иной мере обладает общим качеством, называемым красотой, или плодотворностью, или целесообразностью в зависимости от сферы, к которой относится произведение. Творческая функция состоит в том, чтобы из необозримого множества возможных сочетаний отобрать те, которые обладают указанным свойством. Таким образом, процесс творчества сводится к выбору.

Высказывания Эшби и Пуанкаре. Приведенная выше точка зрения уже высказывалась с полной определенностью. В качестве примера интересно привести мнения современного ученого, известного специалиста по кибернетике У. Росс Эшби и крупного французского математика А. Пуанкаре.

Эшби пишет [1]: «. . . наше восхищение продуктивностью гения направлено неверно. Ничего нет легче, чем создание новых идей: при соответствующем истолковании калейдоскоп, внутренности овцы или шумовая лампа будут создавать их в изобилии. В гении замечательно умение отсеивать возможности. . .» Пуанкаре пишет [2]: «Творчество состоит именно в том, чтобы не строить бесполезные комбинации, а строить те, которые полезны и которых ничтожное меньшинство. Творить — это распознавать, это выбирать». Заметим, что Пуанкаре говорит здесь, в частности, о математическом творчестве.

Взгляды Пуанкаре на роль подсознания. Число возможных сочетаний сравнительно небольшого количества элементов необозримо велико. Например [3], число возможных русских стихотворений из 400 букв имеет порядок 10^{100}. Этот факт да и прямое наблюдение творческой деятельности показывают, что никакой автор в действительности не делает (и не может делать) прямого сознательного перебора возможных комбинаций. Спрашивается, как же все-таки совершается творческий процесс? Пуанкаре отводит главную роль подсознательной сфере. Он опирается при построении своей концепции на всем известные факты, в частности, на то, что идеи внезапно «приходят нам в голову» в почти законченной форме; что иногда не дающееся в результате сознательной деятельности решение проблемы является нам во сне. Отсюда непосредственно следует, что некоторая часть творческого процесса протекает в подсознании. Пуанкаре отмечает и подчеркивает, что работе подсознания обязательно предшествует некоторый период сознательной деятельности; продукт работы подсознания подвергается последующей сознательной обработке. Длинную цепь своих содержательных и блестящих по форме рассуждений Пуанкаре заканчивает следующей моделью (которую он приводит лишь в качестве грубой аналогии). Представим себе некоторое помещение, на стенах которого развешаны различные первичные элементы. Приступая к творческой работе в определенном направлении, мы снимаем со стен элементы, входящие в игру, и пытаемся сочетать их различным образом. Если нам не удалось найти то, что искали, то сознательная деятельность прекращается; однако приведенные в движение элементы не возвращаются на свои места — они продолжают носиться в помещении, сочетаясь во всевозможных комбинациях. Если образовалась «хорошая» комбинация (Пуанкаре так и выражается), то она всплывает на поверхность сознания.

Проблема критерия. Если стать на изложенную выше точку зрения, то трудность объяснения механизма творческого процесса не снимается, а перемещается; возникает проблема критерия, на основании которого выбирается «хорошая» (красивая, плодотворная, целесообразная) комбинация. Качество творения в области точных наук и техники может быть оценено. Так, например, новая теория хороша, поскольку она обобщает старые, удовлетворительно объясняет все известные экспериментальные факты и позволяет правильно предсказывать новые. Но в сфере искусства критерий должен иметь эстетическую природу, и его точная формулировка (на языке точных наук) даже на низшем уровне наталкивается на большие трудности. Правда, всякое произведение искусства должно удовлетворять некоторому канону, соответствие которому входит в критерий. Но каноны не имеют всеобщего характера; кроме того, они непрерывно меняются.

По-видимому, отбор производится, так сказать, в несколько туров; критерий имеет несколько степеней, причем часть ступеней

отбора происходит еще в подсознании, что может резко сократить поиск «хорошей» комбинации.

Опыты по сочетанию музыки с помощью ЭЦВМ. За последнее время были сделаны многочисленные опыты по решению творческих задач на электронных вычислительных машинах; многие результаты были опубликованы.

На одном из заседаний Совета по кибернетике АН СССР демонстрировались музыкальные произведения (короткие одноголосные мелодии), сочиненные машиной. В программу, заданную машине, были заложены общего характера требования, относящиеся к звукоразряду, ладу, ритму и некоторые условия, относящиеся к архитектонике музыкального произведения. Были показаны образцы разных жанров; все они звучали приятно. Из доклада выяснилось, что показана была лишь небольшая часть продукции машины, специально отобранная. Тотчас был задан вопрос: каким критерием руководствовались при этом последнем туре отбора? Вопрос остался без ответа.

Одно простое устройство. В качестве иллюстрации изложенных выше представлений рассмотрим возможную схему устройства, выполняющего простейшую творческую функцию. Пусть требуется составить повторяющийся орнамент из одной непрерывной линии. Случайным источником является шумовой генератор с линией задержки, позволяющий получить два практически независимых случайных процесса. Эти процессы используются для отклонения в двух взаимно перпендикулярных направлениях, т. е. представляют абсциссу и ординату будущей кривой. Канон, т. е. первая ступень критерия, состоит в требовании плавности кривой. Это требование осуществляется либо фильтром нижних частот на выходе генератора, либо ограничителями по производным. Получаемые образцы записываются (и все время стираются) на двухдорожном барабане, один оборот которого соответствует одному периоду орнамента. Вторая ступень критерия состоит в требовании, чтобы на стыке двух периодов не было разрыва и излома линий, т. е. чтобы значения функции и ее первой производной в начале и в конце периода были одинаковы. Выполнение этого требования проверяется специальным узлом. Если требование выполнено, полученная «хорошая» (т. е. удовлетворяющая критерию) комбинация выдается на выход.

В этом устройстве налицо все основные части механизма, обсужденного ранее. Никакого практического значения описанная модель, конечно, не имеет.

Машина, выполняющая творческие функции, в принципе возможна. Она состоит, если следовать изложенной концепции, из случайного источника и контрольных звеньев, производящих отбор на основании заложенного в машину критерия. Проблема состоит в формулировке критерия. Следовательно, машинное творчество возможно лишь на том уровне, на котором еще возможна формулировка критерия (на приемлемом для машины языке).

Весьма возможно, что машинное «творчество» может найти практическое применение, например, в области прикладного искусства. Так, абстрактные рисунки для набивных тканей могла бы, вероятно, с успехом создавать машина. Опыты в этом направлении были бы очень интересны, если пока и не для промышленности, то для кибернетической науки.

Однако наибольший интерес представили бы попытки точного формулирования различных общих критериев. Для этого необходимо сближение методов и точек зрения представителей гуманитарных и точных наук. Это потребует больших усилий обеих сторон и, в частности, преодоления известных предупреждений. Но результат таких усилий может быть только один — общий прогресс науки.

Литература

1. *Эшби У. Росс.* Схема усилителя мыслительных способностей. Сб. «Автоматы». ИЛ, 1956.
2. *Н. Poincare.* Science et methode. Paris, Flammarion, 1908.
3. *А. Н. Колмогоров.* Теория передачи информации. Изд-во АН СССР, 1965.

НЕКОТОРЫЕ МЕТОДИЧЕСКИЕ ВОПРОСЫ В ПРОБЛЕМЕ ОПОЗНАВАНИЯ[1]

Проблема опознавания (узнавания) как кибернетическая проблема возникла сперва в форме нескольких частных задач (опознавание фигур, звуков речи и т. п.). В настоящее время целесообразно подчеркнуть общий характер проблемы.

Опознавание есть первая ступень переработки информации, поступающей из окружающего мира; первоначальный акт познания, определяющий наше поведение.

Мы наблюдаем окружающий мир при помощи органов чувств и находящихся в нашем распоряжении приборов. Обработка результатов наблюдений производится в несколько этапов. Прежде всего мы опознаем предметы. Затем, наблюдая отношения между предметами, мы опознаем то, что называется обычно ситуацией (сюда входит и отношение окружающих предметов к нам). Наблюдая за предметами и ситуациями в течение некоторого времени, мы опознаем их изменения, т. е. явления. На более высоком уровне обработка наблюдений приводит к обнаружению закономерностей, а следовательно, к возможности прогноза дальнейшего хода событий и т. д.

Переработку поступающей извне информации, приводящую к опознаванию, производит мозг (участие сознания не обязательно). Проблема, которая нас интересует, состоит в том, чтобы переложить функцию опознавания на машину. Сейчас всем уже ясно, как велико общенаучное и прикладное значение этой проблемы.

Опознавание и систематика. Один специальный вид объектов опознавания заслуживает отдельного упоминания. Речь идет о знаковых системах, служащих для общения, т. е. для передачи информации. К таким системам относятся устная речь, все виды письменности, все системы видимых и слышимых сигналов, язык формул, язык графиков и чертежей и т. д.

Процесс общения между живыми существами возможен, разумеется, лишь при условии, что общающиеся особи владеют данной знаковой системой, т. е. прежде всего умеют опознавать знаки, из которых система построена.

---

[1] Статья составляет содержание доклада на методологическом семинаре Института проблем передачи информации АН СССР. Семинар состоялся 24 марта 1965 г.; 30 марта А. А. Харкевич скончался.

В наше время возникла новая ситуация, в которой требуется общение человека с машиной в процессе совместного выполнения весьма сложных функций. Это влечет за собой возникновение ряда новых проблем, среди которых большой интерес представляет проблема создания как символов, так и языков, приемлемых для машины и предназначенных для общения между человеком и машиной.

Из сказанного следует, что без опознавания невозможно никакое общение ни людей между собой, ни человека с машиной: это еще раз говорит о большом значении опознавания.

Задачи опознавания. Попытаемся охарактеризовать задачи, которые возникают при опознавании. Специальную терминологию будем вводить по мере надобности.

З а д а ч а α. Основная задача опознавания может быть сформулирована так. Отнести предъявляемый объект к одному из заранее установленных классов.

Здесь под классом понимается некоторое подмножество (множества всех объектов), члены которого обладают определенной общностью свойств или, проще говоря, сходством.

Характерным для основной задачи является то, что классификация (которая всегда предшествует опознаванию) производится на основании наблюдений над конечной выборкой представителей, тогда как класс может иметь бесконечное (или хотя и конечное, но очень большое и практически необозримое) число членов. Для этого с самого начала вырабатывается обобщенная характеристика класса, содержащая то общее, что объединяет объекты в класс и, наоборот, не содержащая индивидуальных различий реальных представителей одного и того же класса. Эту обобщенную характеристику мы и назовем образом [1].

Здесь надо отметить два существенно различных случая. В одном случае классы характеризуются вполне определенными и заранее известными признаками, представляемыми измеримыми величинами; общность свойств в этом случае есть просто нахождение в одном и том же многомерном интервале (т. е. в некоторой области пространства признаков), заранее определенном для каждого класса самим процессом классификации. Ясно, что задача опознавания в этом случае относительно проста и доступна машине с жесткой программой. Так, например, машина может отличить треугольники от четырехугольников, красные круги от синих, большие шарики от малых. Обозначим задачу опознавания по совокупности заранее известных признаков через α_1.

Второй, значительно более трудный и интересный случай характеризуется тем, что мы производим классификацию, сами не

---

[1] Таким образом, термин «образ» означает в сущности то же, что и термин «понятие». Но во избежание всякого рода недоразумений термин «понятие» лучше оставить в распоряжении психологии и теории познания, а термин «образ» применять в рассматриваемых нами проблемах кибернетики.

зная как, т. е. не будучи в состоянии указать, в чем состоит объективная общность свойств членов каждого класса, хотя наличие такой общности не вызывает сомнений. Так, например, при опознавании букв (т. е. при чтении) мы относим в один класс все возможные начертания буквы А, хотя и не умеем пока сформулировать то, что является общим для всех этих начертаний. Обозначим задачу опознавания в такой постановке через α_2.

В задаче α_2 общность свойств выражена только через наименование или условное обозначение класса. Можно надеяться, что различие между задачами α_1 и α_2 имеет преходящий характер. Иными словами, дело здесь не в принципиальной невозможности формирования образа, а в недостатке нашего опыта и развития: мы пока не умеем строить некоторые нужные нам абстракции.

При таких обстоятельствах приходится искать признаки на ощупь, по наитию, проверяя свои догадки кропотливым экспериментом. Именно этим путем получены приемлемые решения многих практически важных задач. Однако можно с самого начала избрать другой путь: решать задачу α_2 путем обучения, т. е. путем предварительного показа ряда объектов с указанием классов, к которым они относятся. Ясно, что по организации и способу действия машина с обучением существенно отличается от машины, работающей по жесткой программе.

З а д а ч а β. Перейдем теперь к некоторым типичным вырожденным задачам опознавания. К ним относится, в частности, случай, когда для каждого класса известен некоторый эталон, т. е. идеальный представитель; члены класса получаются путем наложения на эталон случайных возмущений. Примером такой ситуации может служить чтение знаков определенного шрифта; реальные знаки могут иметь существенный разброс относительно эталонов вследствие дефектов печати, в особенности в машинописных копиях. Обозначим такого рода задачу через β_1. Примером из несколько иной области является прием нескольких различных, но наперед известных сигналов при наличии помех.

Естественный путь решения задачи β_1 состоит в прямом сличении предъявленного объекта со всеми эталонами. Более того, можно показать, что оптимальный метод решения задачи есть корреляционный метод [1].

К вырожденным задачам отнесем также случай (задача β_2), когда эталоны не заданы, но все члены классов могут быть получены один из другого детерминированным наперед известным преобразованием. В частности, эти преобразования могут представлять собой группу. В качестве наглядного примера приведем изображения букв, предъявляемых в различных параллельных проек-

---

[1] Часто производят подмену задачи α задачей β_1, вводя в качестве эталона результат того или иного усреднения по исходной выборке.

циях; эти изображения образованы группой аффинных преобразований [1].

Естественно избрать такой путь решения задачи β_2: выбрать для каждого класса произвольный член в качестве эталона; предъявленный объект подвергать преобразованию, варьируя параметры преобразования и все время сличая результат с эталонами, пока преобразованный объект не совпадет с одним из эталонов (или окажется к нему значительно ближе, чем к другим). Контроль осуществляется тем же корреляционным методом.

Возможен благоприятный случай, когда в качестве признаков, по которым классы надежно различимы, могут быть взяты величины, инвариантные по отношению к данному преобразованию. Это возвращает нас к основной задаче α_1.

З а д а ч а γ. Рассмотрим теперь случай, когда никакого объективного сходства между членами класса нет и когда состав класса определен произвольным или чисто случайным выбором. Например, мы можем считать классом пассажиров данного вагона метро, хотя никакой общности свойств, кроме нахождения в данном вагоне, между ними может не существовать. Задачу опознавания в этом случае обозначим буквой γ. Ясно, что эта задача специфична, так как при отсутствии устойчивой общности свойств не существует образа. Однако при некотором уточнении задача γ становится осмысленной, а путь ее решения — совершенно ясным.

Воспользуемся в качестве примера тиражной таблицей. Будем считать классом множество номеров, выигравших в данном тираже. Набор чисел, образующих таблицу выигрышей, совершенно случаен (принимаются специальные меры к тому, чтобы это было именно так). Задача γ в данном примере состоит в том, чтобы определить, выиграл ли наш билет. Известно, как решается задача: номер данного билета сличается поочередно со всеми номерами таблицы; мы выиграли, если наш номер совпадет (окажется тождественным) с одним из номеров таблицы выигрышей.

Пример раскрывает особенности задачи γ. Первая состоит в том, что опознавание сводится к отождествлению индивидуального объекта $\mathfrak{C}$ самим собой. Из этого вытекает и вторая особенность, состоящая в том, что для сличения должен быть предъявлен весь класс (а не образ и не эталон, которых в задаче γ не существует). Эти особенности и оправдывают отнесение задачи γ к числу вырожденных.

З а д а ч а δ. Еще один особый вид задачи опознавания — назовем ее задачей δ — отличается тем, что классы различаются по вполне определенному заранее известному признаку (или признакам); однако признак этот по тем или иным причинам ненаблюдаем. В этом случае задача может быть решена, если в результате

---

[1] В этом примере речь идет о линейных преобразованиях координат; могут встречаться произвольные преобразования как координат, так и самих функций.

предварительного изучения ранее накопленных данных установлено, что интересующему нас признаку (или свойству) с достаточно высокой вероятностью сопутствует некоторая комбинация других признаков, которые могут быть непосредственно наблюдаемы. Таким образом, опознавание и производится по сопутствующим признакам. Для применения этого метода необходимо предварительно собрать и обработать «статистику сопутствия». В принципе совершенно безразлично, кто это делает — человек или машина; уже имеющийся опыт показывает, что машина, действующая по специальной программе с обучением, справляется с делом во всяком случае не хуже человека.

Примером задачи δ может служить постановка диагноза на основании совокупности симптомов заболевания. Другой пример — опознавание нефтеносных слоев по данным геофизической разведки. Непосредственно определить наличие нефти при разведке путем пробного бурения затруднительно. Однако сравнительно легко измерить некоторые физические характеристики слов, как-то: электрическое сопротивление, собственный потенциал, радиоактивность и др. Это и есть в данном случае сопутствующие признаки.

Итак, в основе предлагаемой систематики задач опознавания лежит способ образования класса, как подчеркнуто в нижеследующей сводке:

α — класс определен своим образом, образ составлен по исходной выборке;
α_1 — признаки класса заранее известны;
α_2 — признаки класса заранее не известны;
β_1 — класс порожден случайными уклонениями от известного эталона;
β_2 — класс порожден известными преобразованиями своих членов;
γ — класс образован случайным выбором членов;
δ — класс определен заранее известным, но непосредственно не наблюдаемым признаком.

Возможно, что описанная систематика будет в дальнейшем усовершенствована. Но уже сейчас достаточно ясно, что разные задачи требуют разных подходов и что прежде чем обсуждать сравнительные достоинства методов опознавания, нужно отчетливо представить себе характер задачи.

Разные замечания. С х о д с т в о, б л и з о с т ь, к о м п а к т н о с т ь. В настоящее время широко распространены геометрические представления, в основе которых лежит отображение множеств совокупностями точек в пространстве с подходящими свойствами. С точки зрения этих представлений сходство объектов естественно отождествляется с метрической близостью отображающих объекты точек (и соответственно различие между объектами с расстоянием между отображающими точками). Столь же естественно представить себе, что точки, отображающие объекты одного класса (т. е. объекты, сходные между собой), располагаются кучно.

Если классы легко различаются, то и соответствующие скопления легко отграничиваются друг от друга. Это и есть «гипотеза компактности», высказанная Э. М. Бравермaном. Мы не будем уточнять смысл терминов «ложатся кучно», «легко отграничиваются» — уточнения эти занимают много места, но мало что добавляют к интуитивно ясным представлениям. Однако одна сторона дела все же требует обсуждения. Речь идет о том, к какому именно пространству относится гипотеза компактности.

Очевидно, что имеется в виду не пространство признаков. Если определить признак как измеримую величину, значения которой для данного класса лежат в определенном интервале, то в пространстве признаков компактность — уже не гипотеза, а основное свойство, непосредственно вытекающее из определения.

Таким образом, гипотеза компактности подразумевает пространство наблюдений [1].

Но для того, чтобы гипотеза компактности осуществлялась в пространстве наблюдений, это пространство должно быть соответствующим образом организовано; нельзя рассчитывать на то, что все уладится само собой. Относящиеся сюда соображения, хотя и вполне элементарные, изложены ниже довольно подробно.

Прежде всего отметим, что как бы мы ни определили сходство, определение будет практически пригодно лишь при условии, что сходство определено с точностью до разрешенного преобразования и допустимого случайного уклонения. Иными словами, должно быть указано, какие изменения объекта считаются ненарушающими его сходства с другими объектами того же класса. Рис. 1 поясняет сказанное; он показывает, что, например, изображения букв не теряют сходства между собой при аффинных преобразованиях [2].

Из этого вовсе не следует, что точки, соответствующие этим изображениям, ложатся в пространстве наблюдений компактно.

Построим огрубленные изображения цифр, вписывая их в крупную решетку, как показано на рис. 2 (зачерняются клетки, через которые проходит контур цифры). На рис. 3 показаны черно-белые изображения цифр. Принимаем для черных и белых клеток значения яркости соответственно единица и нуль. Значение яркости в каждой клетке — это одна из координат пространства наблюдений, которое в данном случае представляет собой совокупность вершин (9×13) — мерного единичного куба. Выбраны изображения с одинаковым числом единиц (весом), равным 10. Двойки (рис. 3, *а* и *б*) тождественны и отличаются только параллельным

---

[1] Применяются и другие (на мой взгляд менее удобные) наименования, например, пространство входов, пространство рецепторов.

[2] В действительности сходство не теряется при преобразованиях значительно более широкого класса, в том числе и неточечных. Однако по поводу аффинных преобразований следует заметить, что для письменных знаков допускается лишь ограниченный поворот, так как поворот на больший угол может привести к замене одного знака другим (например, Е и Ш, 6 и 9).

переносом. Между тем расстояние между ними (отсчитанное в метрике Хэмминга и вычисленное суммированием по модулю 2 двоичных чисел, выражающих результат наблюдения) составляет $d(a, b)=10$, т. е. ровно столько же, сколько расстояние между первой двойкой и пятеркой $d(a, c)$. В то же время расстояние между второй двойкой и пятеркой — $d(b, c)=6$. Понятно, что для получения компактности в данном случае нужно исключить параллельный перенос, т. е. предъявить изображения цифр так, чтобы занятые ими решетки (3×5) совпадали. Если разрешены аффинные преобразования, то они должны быть предварительно устранены (т. е. должны быть проделаны обратные преобразования). После устранения разрешенных преобразований все объекты одного класса должны были бы изобразиться в пространстве наблюдений одной точкой. В реальных условиях всегда останется некоторый разброс, обусловленный случайными уклонениями. Но наилучшая возможная компактность будет обеспечена.

Естественно возникает мысль: в целях «компактизации» представлять класс как порождение некоторого преобразования (конечно, не обязательно группового) независимо от того, как дело обстоит в действительности. Располагая данной выборкой, т. е. конечным числом представителей класса, можно подыскать преобразование, позволяющее с достаточно малой погрешностью по-

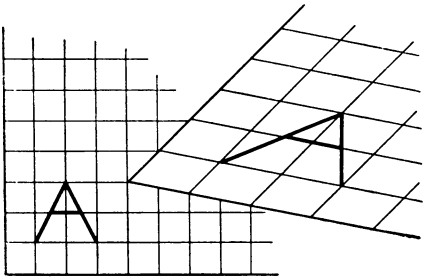


Рис. 1

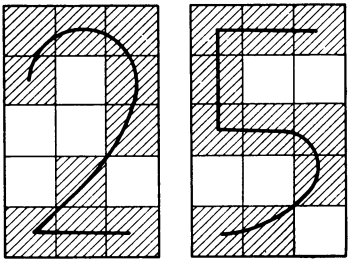


Рис. 2

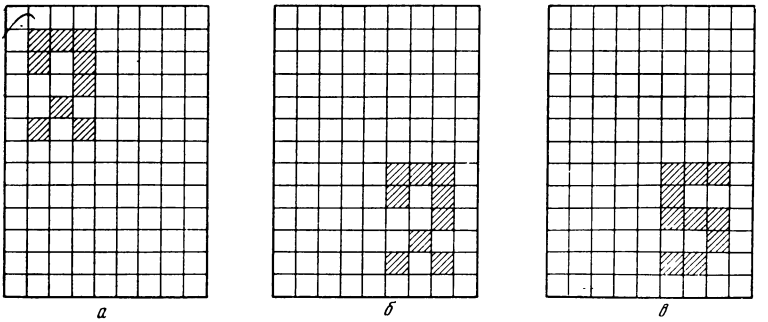


Рис. 3

лучать один объект выборки из другого. Эта возможность заслуживает отдельного рассмотрения.

По поводу наблюдаемости признаков. Если классы различаются по какому-либо признаку, то для того чтобы машина могла осуществить опознавание, она должна быть в состоянии наблюдать этот признак (исключение составляет случай опознавания по сопутствующим признакам, который мы здесь не рассматриваем).

Пусть требуется рассортировать шарики; в зависимости от обстоятельств они могут различаться по размеру, весу, цвету, прозрачности, магнитной проницаемости, степени шероховатости поверхности, радиоактивности и т. д.

Воспринимающее (входное) устройство машины, т. е. та часть машины, которая осуществляет наблюдения, должно иметь соответствующие датчики. Машина, снабженная термометрами того или иного устройства, отличит теплые шарики от холодных, но не красные от синих; для различения по цвету нужно оборудовать машину фотоэлементами с цветными фильтрами [1].

В этом смысле постановка опознавания всегда конкретна; ее формулировка диктуется реальной потребностью и содержит в себе такие ограничения, которые делают задачу технически разрешимой.

Однако кое-что еще недосказано. Ведь в условиях задачи α_2 (см. выше) признаки не сформулированы и неизвестно, что следует наблюдать [2]. В этом случае приходится прибегать к весьма расточительному способу: строить наблюдения по значительным избыткам, но так, чтобы в их результатах заведомо содержались нужные нам признаки. Так, например, если наперед известно, что различение предметов может быть установлено по их внешнему виду, то результатом наблюдения должно быть изображение предмета. Если известно, в частности, что предметы различаются по форме, то в чем бы ни состояли эти различия, они обязательно будут содержаться в монохроматических изображениях. Для различения звуков речи достаточно снять осциллограмму звукового давления в одной точке звукового поля. Эта осциллограмма заведомо содержит в себе интересующие нас признаки отдельных фонем или целых слов.

Следует, быть может, пояснить, что иметь дело непосредственно с фотографией или с осциллограммой решающая часть узнающей машины практически не может — слишком велика размерность пространства наблюдений.

---

[1] В связи с этим представляется неправомерной постановка для машины задачи о «классификации по не заданному заранее признаку». Не говоря уже о том, что непонятно, какая практическая надобность может привести к такой постановке, признаки, которыми оперирует машина, всегда заданы: они содержатся в результатах тех наблюдений, которые ей доступны.

[2] Размерность пространства наблюдений зависит от разрешающей способности средств наблюдения. При обычной технике для звуков число измерений имеет порядок 10^3, а для изображений 10^6.

К тому же на фотографии и на осциллограмме признаки, т. е. элементы образа, неизменные для данного класса и именно поэтому нужные нам для опознавания, буквально завалены бесполезными индивидуальными особенностями данного объекта. Отыскание признаков и сводится к тому, чтобы, отбросив частное, сохранить общее.

Перейдем к особому случаю, когда важные для опознавания признаки могут оказаться наблюдаемыми. Начнем с примера: пусть среди сосудов требуется опознавать пепельницы. Но разнообразие внешних форм пепельниц очень велико; в то же время ясно, что пепельница отличается от других сосудов своим назначением. Этот признак должен быть наблюдаем машиной. Для этого достаточно предъявлять ей различные сосуды с тем наполнением, которое они получают в процессе применения. Если в этих условиях предъявить машине чайную чашку, заполненную пеплом и окурками, то машина признает в этом сосуде пепельницу и будет совершенно права!

В несколько более общей формулировке речь идет о случае, когда признаки объекта заключены не в нем самом, а в его отношениях к другим предметам. В этом случае эти характерные отношения должны быть наблюдаемы. Так, например, X имеет определенный круг знакомых. Установить, принадлежит ли Y к их числу нельзя ни по фотографии Y, ни даже по его паспорту. Но фотография, на которой X и Y запечатлены беседующими в домашней обстановке, позволяет утверждать, что с высокой вероятностью X и Y знакомы.

Заметим, что ситуация, которую мы обсуждаем, может представить затруднения не только для машины, но и для человека. Полезно вспомнить, что для того чтобы облегчить задачу опознавания объектов, выполняющих определенные функции или наделенные особыми свойствами, введены специальные внешние меты, как-то: форменная одежда, знаки отличия; знак депутата; знак, свидетельствующий об окончании вуза; пропуска; служебные удостоверения и жетоны; клетчатые полоски на такси; красный крест на медицинском инвентаре и т. д. Во всех этих примерах непосредственно ненаблюдаемые функции или свойства сделаны наблюдаемыми посредством разного рода условных видимых знаков. Развивая эту тему, мы должны были бы признать, что той же цели служат вообще все надписи и наименования; так, чтобы отличить молочный магазин от парикмахерской, не нужно входить внутрь и наблюдать происходящее. Но тут мы уже удаляемся от проблемы опознавания.

О б о б у ч е н и и ч е л о в е к а м а ш и н о й. В проблеме опознавания применительно к организации общения машины и человека есть одна заслуживающая внимания подробность.

Одна из важных задач в этой области состоит в том, чтобы обучить машину человеческому языку; это облегчает работу оператора, так как отпадает надобность в переводе вводимой в ма-

шину информации с человеческого языка на язык, доступный машине.

В качестве примера можно привести работу с ЭВМ — желательно вводить в машину инструкции и исходные данные не путем набивки перфолент или перфокарт, а непосредственно в обычной письменной или устной форме, как если бы оператор имел дело с человеком, а не с машиной. Для этого входное устройство ЭВМ должно представлять собой читающее устройство или устройство, распознающее команды, подаваемые голосом. При этом обычное требование состоит в том, чтобы машина правильно действовала при произвольном операторе, т. е. не была «настроена» на определенный почерк или определенный голос [1].

Трудность заключается в том, чтобы обеспечить достаточно высокую надежность, т. е. малую вероятность ошибки при опознавании машиной получаемых ею письменных или устных команд. Конечно, возможно и даже необходимо применение различных методов контроля. Но контроль имеет смысл лишь при условии, что исходная надежность достаточно высока. И вот тут возникает важный и интересный вопрос: каков разумный уровень требований, которые следует предъявить машине? Ясно, что нельзя потребовать, чтобы при любом начертании или при любом произнесении машина воспринимала все команды безошибочно. Такое требование — просто бессмыслица. Столь же ясно, что тем больше произвол, допускаемый при начертании тех или иных знаков, тем меньше вероятность правильного их опознавания данной машиной или тем труднее построить машину, осуществляющую опознавание с заданной вероятностью правильного опознавания.

Уместно напомнить, что и люди не могут сговориться, если один из собеседников говорит невнятно по небрежности или вследствие каких-либо дефектов речи; и люди, возможно, окажутся не в состоянии разобрать небрежные каракули. Уместно напомнить, что мы считаем естественным, например, требование аккуратного написания цифр в денежных документах. Из этого следует, что то, что названо выше «разумным уровнем требований к машине», должно быть определено как результат некоторого компромисса между стремлением получить как можно более высокую надежность машины, с одной стороны, с желанием обеспечить этот результат простейшими техническими средствами (т. е. с минимальными затратами) — с другой.

Но надежность зависит не только от машины; надежность в значительной мере зависит и от ее партнера — от управляющего машиной оператора. Машина может лишь обеспечить определенную надежность при заданном допуске на начертание или произнесение команд; оператор должен уложиться в этот допуск. Таково требование, предъявляемое, так сказать, машиной человеку.

---

[1] Надо заметить, что за последнее время поставлена и успешно решается другая задача — опознавания индивидуума по почерку или голосу.

Для выполнения этого требования нужен этап тренировки оператора [1].

Решающее устройство машины снабжается нулевой зоной (зоной неопределенности), так что машина отказывается от опознавания недостаточно отчетливой команды и выдает соответствующий сигнал отказа. Оператор повторяет команду более тщательно, пока она не будет принята машиной. Так происходит обучение человека машиной.

Конечно, увеличивая зону неопределенности, можно повысить надежность. Но это происходит за счет перекладывания трудностей с машины на человека, и в конце концов вместо облегчения условий работы оператора мы придем к противоположному результату. Таким образом, и здесь нужен компромисс, и вся проблема организации взаимодействия человека и машины требует весьма осторожного и тактичного подхода. В расширенной постановке проблема состоит в обеспечении заданной высокой надежности при минимальных затратах на технику и максимально комфортабельных условиях работы человека-оператора.

---

[1] Интересно заметить, что все находят естественным то, что люди, имеющие дело с обычными машинами (станками, автомобилями, строительными или дорожными механизмами), проходят специальное обучение, иногда довольно длительное. Когда же дело доходит до кибернетических машин, выполняющих бесконечно более сложные функции, то необходимость приспособиться к возможностям этих машин почему-то вызывает иногда удивление и даже негодование!

СОДЕРЖАНИЕ

Александр Александрович Харкевич

ИЗБРАННЫЕ ТРУДЫ В ТРЕХ ТОМАХ. Том 3

ТЕОРИЯ ИНФОРМАЦИИ. ОПОЗНАНИЕ ОБРАЗОВ

Утверждено к печати Институтом проблем передачи информации

Редактор издательства Н. Н. Соколова
Художник В. Д. Димитриади. Технические редакторы О. М. Гуськова, Р. М. Денисова

Сдано в набор 21/XI 1972 г. Подписано к печати 9/IV 1973 г. Формат $60\times90^{1}/_{16}$. Бумага № 1.
Усл. печ. л. 32,75. Уч.-изд. л. 30,3. Т-05340. Тираж 4100 экз. Тип. зак. 1538.
Цена 2 руб. 35 коп.

Издательство «Наука», 103717 ГСП, Москва К-62, Подсосенский пер., 21.
1-я типография издательства «Наука», 199034, Ленинград, 9-я линия, д. 12

ИЗДАТЕЛЬСТВО «ХИМИЯ»