В.Н. ВАСЮКОВ

ТЕОРИЯ ЭЛЕКТРИЧЕСКОЙ СВЯЗИ

НОВОСИБИРСК
2005

УДК 621.391(075.8)
В 201

Рецензенты: д-р техн. наук, проф. *А.Г. Вострецов*,
д-р техн. наук, проф. *И.С. Грузман*

Васюков, В. Н.
В 201 Теория электрической связи : учебник / В. Н. Васюков. – Новосибирск : Изд-во НГТУ, 2005. – 392 с. («Учебники НГТУ»).
ISBN 5-7782-0541-4

Учебник содержит изложение основных вопросов теории электрической связи.

Учебник рассчитан на студентов II–III курсов, обучающихся по специальностям «Средства связи с подвижными объектами» и «Многоканальные телекоммуникационные системы», и может быть полезен студентам других направлений и специальностей.

УДК 621.391(075.8)

ISBN 5-7782-0541-4

© В.Н. Васюков, 2005
© Новосибирский государственный технический университет, 2005

ПРЕДИСЛОВИЕ

Системы связи (системы телекоммуникаций, системы передачи информации) в настоящее время переживают этап бурного развития. Достаточно упомянуть такие достижения конца XX века, как глобальная сеть Интернет, спутниковая связь, общедоступная мобильная (сотовая) связь, чтобы оценить уровень и темпы развития техники телекоммуникаций. Этим обусловливается потребность в подготовке высококвалифицированных специалистов в области связи. Есть все основания полагать, что и в ближайшем будущем эта тенденция сохранится.

В подготовке инженеров-связистов и бакалавров по направлению 550400 – «Телекоммуникации» фундаментальную роль играет дисциплина «Теория электрической связи», включенная в Государственный образовательный стандарт в раздел «Общепрофессиональные дисциплины (федеральный компонент)». Предлагаемый учебник является попыткой восполнить недостаток учебной литературы по теории электрической связи.

Теория электрической связи, по мнению автора, представляет собой систему взаимосвязанных положений, взглядов, концепций, составляющих основу мировоззрения специалиста в области телекоммуникаций. Изучение этой теории, в частности, должно дать человеку твердое, научно обоснованное представление о том, что в области связи можно сделать, а чего нельзя ни при каком уровне развития технологии. Кроме того, теоретические положения должны быть подкреплены конкретными образцами их применения на практике.

Эта точка зрения в совокупности с ограниченным объемом определила отбор материала для данного учебника. Сравнительно большое внимание уделено принципиальным вопросам теории сигналов и линейных стационарных цепей, включая концепцию пространства сигналов, ряд и интеграл Фурье, теорему отсчетов,

понятие аналитического сигнала, случайные процессы, основы теории информации и др. Модуляция и демодуляция, экономное (статистическое) и помехоустойчивое кодирование, основы теории помехоустойчивости и некоторые другие вопросы рассмотрены главным образом на уровне идей с подробным изложением частных примеров для усвоения основных понятий, с учетом того, что многочисленные, в том числе технические, подробности этих тем будут изучаться в последующих дисциплинах учебного плана.

Автор выражает глубокую признательность д-ру техн. наук, проф. И.С. Грузману, канд. техн. наук, проф. А.Н. Яковлеву, студентам А.Н. Подовинникову и Д.В. Семенову, взявшим на себя нелегкий труд чтения рукописи и высказавшим много полезных замечаний и предложений по улучшению учебника.

1. ВВЕДЕНИЕ. СИСТЕМЫ СВЯЗИ, СИГНАЛЫ, КАНАЛЫ СВЯЗИ

1.1. ОБЩИЕ СВЕДЕНИЯ О СИСТЕМАХ ЭЛЕКТРИЧЕСКОЙ СВЯЗИ

Системы связи предназначены для передачи *информации*. Информация[1] передается посредством *сообщений*. Таким образом, *сообщение – форма представления информации*. Примерами сообщений могут служить текст телеграммы, фраза в телефонном разговоре, последовательность цифр при передаче данных, изображение в системе фототелеграфии, последовательность изображений (кадров) в системе телевидения и т.п. Сообщение представляет собой совокупность *знаков* (*символов*). Например, текст телеграммы состоит из букв, цифр, пробелов и специальных знаков, а телеграфное сообщение, готовое для передачи по каналу связи, – из канальных символов (например, из «точек», «тире» и пауз при использовании «азбуки Морзе»). В системе черно-белого телевидения сообщением является последовательность кадров, каждый из которых, в свою очередь, представляет собой последовательность значений яркости, упорядоченных согласно схеме телевизионной развертки. В телефонии сообщение – непрерывная последовательность значений напряжения (тока), отображающая изменение во времени звукового давления на мембрану микрофона.

Из приведенных примеров становится ясно, что сообщения могут быть *дискретными* (состоящими из символов, принадлежащих конечному множеству – *алфавиту*) или непрерывными (*континуальными*, *аналоговыми*), описываемыми функциями непрерывного времени.

---

[1] О содержании понятия информации см. разд. 8.

Для передачи сообщения необходим материальный носитель, называемый сигналом. Сигналом может быть свет костра, удар барабана, звук речи или свистка, предмет, находящийся в условленном месте, взмах флажка или шпаги и т.п. В радиотехнике и электрической связи используются электрические сигналы, которые благодаря простоте их генерирования и преобразования наилучшим образом приспособлены для передачи больших объемов данных на большие расстояния. Заметим, что в современных каналах связи и устройствах хранения данных электрические сигналы зачастую преобразуются в оптические или магнитные, но, как правило, предполагается их обратное преобразование.

Естественной формой представления сигнала считается его описание некоторой функцией времени (зависимой переменной чаще всего является напряжение или ток). Сигналы, как и сообщения, могут быть дискретными или континуальными в зависимости от того, рассматриваются ли они как функции дискретного или непрерывного времени. Зависимая переменная также может быть дискретной или непрерывной, в соответствии с чем можно различать четыре типа сигналов (рис. 1.1).

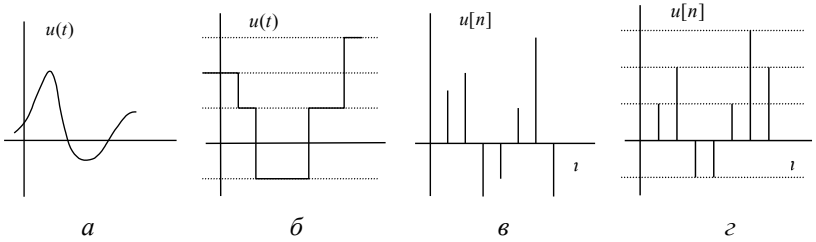


Рис. 1.1. Типы сигналов:

а – аналоговый, *б* – квантованный, *в* – дискретный, *г* – цифровой

Современная *система связи* представляет собой сложную совокупность устройств, выполняющих преобразования сообщений и сигналов с целью наиболее эффективной передачи информации. К показателям эффективности относятся достоверность (верность) и скорость передачи информации, а также некоторые другие величины. Упрощенная схема системы передачи информации (системы связи) показана на рис. 1.2.

Само назначение системы связи предполагает наличие *источника* и *получателя* сообщений. Источник сообщений ИС порождает сообщение a, которое преобразуется преобразователем Пр1 в сигнал $b(t)$, называемый *первичным сигналом*. Например, в систе-

ме телефонии преобразователем служит микрофон, в системе телевидения – передающая телевизионная камера. Первичный сигнал, как правило, непригоден для непосредственной передачи[2], поэтому он поступает на *модулятор*[3] М, где используется для *модуляции* другого колебания $s(t)$, более подходящего для передачи и называемого *переносчиком* или *несущим колебанием*. Модуляция означает изменение одного или нескольких параметров сигнала-переносчика в соответствии с изменением первичного сигнала (или с передаваемым сообщением). Следует отметить, что дискретное сообщение не обязательно должно передаваться дискретным сигналом, а непрерывное – аналоговым. Наоборот, для современных систем связи характерна передача, например, аналоговых сообщений цифровыми сигналами; цифровые первичные сигналы применяют для модуляции аналоговых несущих колебаний и т.д.

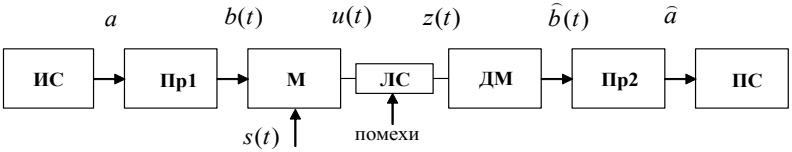


Рис. 1.2. Структурная схема системы связи

Модулированный сигнал $u(t)$ поступает в линию связи ЛС, которая по своим физическим свойствам пригодна для передачи сигнала к приемнику и в которой происходит его искажение под влиянием характеристик линии, а также неизбежное воздействие на сигнал вредных колебаний (*помех*). Вследствие этого колебание $z(t)$, поступающее с выхода линии связи ЛС на *демодулятор*[4] ДМ, отличается от переданного сигнала $u(t)$, поэтому вырабатываемый демодулятором сигнал $\hat{b}(t)$ в общем случае отличается от первичного сигнала $b(t)$. Качество демодулятора (и системы в целом) тем выше, чем меньше это отличие. Сигнал $\hat{b}(t)$ преобразуется преобразователем Пр2 в сообщение $\hat{a}$, передаваемое получателю сооб-

[2] В простых системах проводной телефонии первичный сигнал может передаваться *непосредственно.*

[3] Это устройство часто называют передатчиком; тогда модулятором называют ту часть передатчика, где происходит собственно модуляция, т.е. управление параметрами несущего колебания.

[4] Модулятор и демодулятор часто конструктивно объединяют в одно устройство – *моде́м.*

щения ПС. В радиовещании роль подобного преобразователя играет громкоговоритель, в телевидении – кинескоп и т.д.

Рассмотренная структура системы связи является простейшей и сравнительно редко применяется на практике. В современных системах связи сообщение перед передачей часто *кодируется*, при этом последовательность символов, порождаемая источником (т.е. собственно *сообщение*), преобразуется кодером К (рис. 1.3) в последовательность кодовых символов, которая в виде цифрового сигнала $b_{\text{ц}}(t)$ поступает в модулятор. После прохождения по каналу связи и демодуляции полученная кодовая последовательность *декодируется* декодером[5] ДК, при этом восстанавливается сообщение $\hat{a}$, которое может отличаться от исходного. Кодирование производится для повышения скорости передачи информации (экономное, или энтропийное, кодирование[6]) либо уменьшения вероятности ошибки при приеме сообщения (помехоустойчивое кодирование). Целью кодирования может быть также *согласование* формы передаваемого сообщения с каналом связи. Примером последнего служит кодирование цифробуквенного телеграфного сообщения кодом Морзе. Совокупность всех кодовых символов данного кода называется кодовым *алфавитом*. Количество символов в кодовом алфавите называют *основанием кода*. Например, код Морзе имеет основание 3. Код Бодо, алфавит которого состоит из символов 0 и 1, имеет основание 2.

Обычно один символ исходного сообщения (например, буква в телеграфном сообщении) заменяется последовательностью кодовых символов (кодовой *комбинацией*, кодовым *словом*). *Кодом* называется совокупность всех допустимых кодовых комбинаций. Если каждый символ сообщения заменяется при кодировании одинаковым количеством кодовых символов (т.е. все кодовые слова имеют равную длину), то код называется *равномерным*, иначе – *неравномерным*. Длину кодовой комбинации равномерного кода называют его разрядностью.

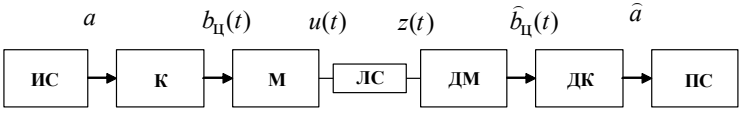


Рис. 1.3. Упрощенная структура дискретной системы связи с кодированием

[5] Часто кодер и декодер конструктивно объединяются в одно устройство, называемое *ко́деком.*

[6] Используются также термины *сжатие* и *статистическое кодирование.*

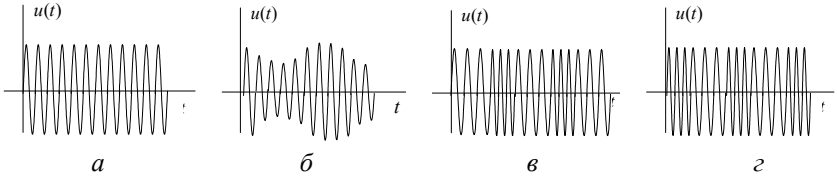
С кодированием не следует путать *шифрование* сообщений, иногда применяемое в современных системах связи. Цель шифрования состоит в предотвращении *несанкционированного извлечения* или *преднамеренного изменения* информации. При *зашифровании* производится замена открытого сообщения *шифрограммой* (шифртекстом), а при *расшифровании* происходит обратное преобразование. Шифрование выполняется до преобразования сообщения в первичный сигнал или в кодовую последовательность.

Таким образом, для модуляции в зависимости от сложности системы применяется первичный сигнал или последовательность кодовых символов. В качестве переносчика часто используют гармоническое колебание $A\cos(\omega t + \varphi)$, которое имеет три параметра: амплитуду A, круговую частоту $\omega = 2\pi f$ и начальную фазу φ. Поэтому возможны три вида модуляции гармонического переносчика аналоговым сигналом: амплитудная модуляция (АМ), частотная модуляция (ЧМ) либо фазовая модуляция (ФМ)[7], рис. 1.4.

Во многих случаях роль переносчика в системах связи играет периодическая последовательность импульсов одинаковой формы (часто импульсы считают в первом приближении прямоугольными[8]). При заданной форме импульсов последовательность характеризуется амплитудным (пиковым) значением, длительностью импульсов и периодом повторения. Поэтому при аналоговом первичном сигнале различают:

– амплитудно-импульсную модуляцию (АИМ), при которой по закону изменения первичного сигнала изменяется амплитуда импульсов;

– широтно-импульсную модуляцию (ШИМ), при которой изменяется длительность («ширина») импульсов[9];



а *б* *в* *г*

Рис. 1.4. Несущее гармоническое колебание (*а*) и получаемые на его основе модулированные сигналы: АМ (*б*), ЧМ (*в*) и ФМ (*г*)

[7] Два последних вида модуляции часто объединяют под общим названием *угловая модуляция.*

[8] Как станет ясно в дальнейшем, строго прямоугольные импульсы на практике получить невозможно.

[9] ШИМ также называют ДИМ, от слова *длительность*.

– времяимпульсную модуляцию (ВИМ), при которой изменяется время задержки импульсов относительно среднего положения;

– частотно-импульсную модуляцию (ЧИМ), когда в такт с первичным сигналом изменяется частота следования импульсов.

Широко применяют также модуляцию гармонического колебания квантованным (цифровым) первичным сигналом. Различают три вида дискретной (цифровой) модуляции (манипуляции): амплитудную (ДАМ, ЦАМ), частотную (ДЧМ, ЦЧМ) и фазовую (ДФМ, ЦФМ), рис. 1.5. Участок манипулированного сигнала, в течение которого модулируемый параметр постоянен, называется *элементарной посылкой*, или просто посылкой.

Колебание при дискретной модуляции характеризуют *технической скоростью* (скоростью *модуляции*, скоростью *телеграфирования*), равной количеству элементарных посылок в секунду. Единицей измерения скорости модуляции является бод[10] (1 бод соответствует одной посылке в секунду).

Наиболее важными показателями качества систем связи являются достоверность и помехоустойчивость. *Достоверность* дискретных систем связи определяется вероятностью безошибочного приема сообщения или отдельной посылки. Достоверность систем передачи непрерывных сообщений часто характеризуется средним квадратом ошибки

$$\varepsilon^2 = \frac{1}{T}\int_0^T \left| b(t) - \hat{b}(t) \right|^2 dt ,$$

где T – время наблюдения сигнала.

Помехоустойчивость системы связи характеризуют отношением средних мощностей сигнала и помехи, при котором обеспечивается заданная достоверность.

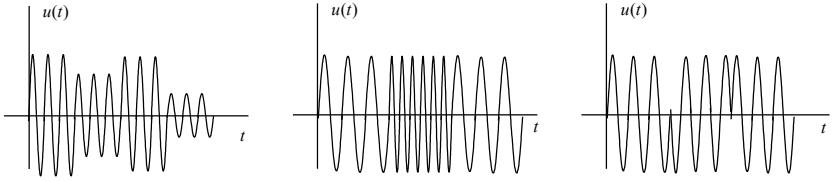


Рис. 1.5. Виды дискретной модуляции (манипуляции) гармонического колебания: ДАМ (*а*), ДЧМ (*б*), ДФМ (*в*)

[10] Назван в честь Ж.М.Э. Бодо (1845 – 1903) – известного французского инженера.

Демодуляция заключается в восстановлении первичного сигнала по принятому искаженному колебанию, а *декодирование* – в восстановлении дискретного сообщения по демодулированному сигналу. Часто перед демодуляцией применяют дополнительное преобразование с целью повышения достоверности (уменьшения вероятности ошибки). Такое преобразование называют *обработкой*. *Оптимальной* называется обработка, обеспечивающая наивысшую достоверность решения. Если оптимальная обработка оказывается слишком сложной и/или дорогостоящей, применяют *квазиоптимальную* (субоптимальную) обработку, которая проще и дешевле и при этом обеспечивает достоверность, близкую к предельной. Часто квазиоптимальная обработка представляет собой *фильтрацию* принятого колебания с целью подавления помех.

1.2. СИГНАЛЫ И ПОМЕХИ

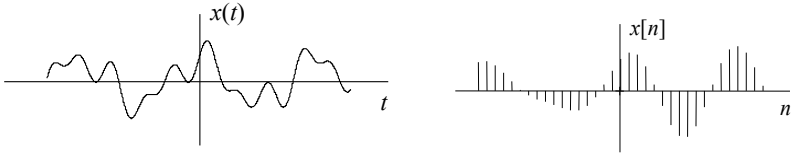
Общий подход к разработке и проектированию современных технических систем, в том числе систем связи, заключается в получении оптимальных или хотя бы субоптимальных технических решений. Такие решения, как правило, не могут быть получены эмпирическим путем, для этого необходимо располагать соответствующими теоретическими, т.е. математическими, методами.

Теория сигналов представляет собой математическую теорию, описывающую с единых позиций все многообразие электрических сигналов, применяемых в проводной и радиосвязи, радио- и телевизионном вещании, радиолокации и радионавигации, автоматике и телемеханике, глобальных и локальных компьютерных сетях и во многих других областях техники.

В настоящее время в технике используется множество различных сигналов, которые классифицируются по различным признакам, связанным со свойствами функций, описывающих сигналы.

Аналоговые (*континуальные*) и *дискретные* сигналы различаются по типу независимой переменной (чаще всего это время). Аналоговый сигнал $x(t)$ описывается функцией *непрерывной* переменной, принимающей значения, например, из множества вещественных чисел $t \in \mathbb{R}$ (хотя сама функция при этом может содержать разрывы – скачки), а дискретный сигнал $x[n]$ – функцией *дискретной* переменной (аргумент, принимающий дискретные значения, принято заключать в квадратные скобки). В качестве дискретного времени обычно рассматривают целочисленную переменную n, принимающую всевозможные целые значения

$n = \overline{-\infty, +\infty}$, а дискретный сигнал называют *последовательностью*. Примеры аналогового и дискретного сигналов представлены графиками на рис. 1.6. Необходимо отметить, что дискретный сигнал обычно изображают графиком со сплошной осью абсцисс, но *существует* этот сигнал лишь в дискретном множестве её точек.



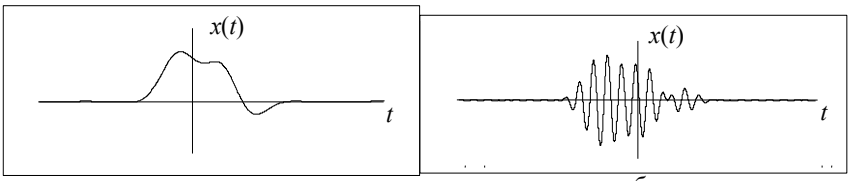
а *б*

Рис. 1.6. Аналоговый сигнал (*а*) и дискретный сигнал (*б*)

Импульсным называется аналоговый сигнал, определенный на непрерывной временно́й оси, но отличный от нуля лишь на ограниченном её участке (носителе сигнальной функции)[11]. Различают *видеоимпульсы,* описываемые функциями, не меняющими знака в пределах носителя, или меняющими его всего несколько раз, а также *радиоимпульсы*, меняющие знак многократно (рис. 1.7). Радиоимпульс можно представить в виде произведения видеоимпульса (называемого в этом случае *огибающей* радиоимпульса) и гармонического *несущего* колебания.

Скалярные и *векторные* сигналы различаются размерностью функций, которые их описывают. В некоторых случаях используются *комплексные* сигналы, принимающие значения из поля $\mathbb{C}$ комплексных чисел. Комплексные числа являются скалярами, хотя иногда их удобнее представлять векторами на так называемой комплексной плоскости.

Многомерные сигналы, в отличие от *одномерных,* описываются функциями многих переменных. Так, черно-белое телевизионное



а *б*

Рис. 1.7. Видеоимпульс (*а*) и радиоимпульс (*б*)

[11] К импульсным также относят сигналы бесконечной длительности, у которых большая часть энергии сосредоточена на конечном участке временной оси (таков, например, экспоненциальный импульс).

или фотографическое изображение описывается функцией двух пространственных переменных, отображающей яркость каждой точки кадра в зависимости от её координат по горизонтали и по вертикали. Цветное изображение можно представить *векторной* (размерности 3) функцией двух переменных, при этом компоненты вектора отображают яркости трех составляющих, например, красного, зеленого и синего цветов. *Пространственно-временные* электромагнитные сигналы описываются векторными функциями четырех переменных, три из которых представляют собой координаты некоторой точки в трехмерном физическом пространстве, а четвертой переменной является время. Размерность векторной функции такого сигнала равна шести, что соответствует представлению в трехмерном пространстве векторов напряженностей электрического и магнитного полей.

Случайные сигналы, в отличие от *детерминированных*, при их наблюдении принимают значения, которые заранее невозможно предсказать точно. Для описания случайных сигналов применяется математический аппарат теории вероятностей (теория случайных процессов), а для построения систем обработки таких сигналов и принятия решений – аппарат математической статистики (теория статистических решений). Строго говоря, *все сигналы являются случайными*, так как если сигнал заранее известен, то нет нужды его принимать (а следовательно, и передавать). Тем не менее часто сигналы при теоретическом рассмотрении описываются детерминированными функциями, например, если случайность сигнала заключается в самом факте его передачи или в его задержке относительно некоторого момента времени и т.п. В таких случаях говорят о *квазидетерминированных* сигналах.

Полезные сигналы отличаются от *мешающих* тем, что полезные сигналы служат для передачи сообщений, в то время как мешающие являются причиной их искажения (потери информации). Часто полезный сигнал называют просто *сигналом,* а мешающий – *помехой.* Сигналы и помехи, рассматриваемые в совокупности, будем называть *колебаниями*. Помехи могут быть *естественными* и *преднамеренными* (*искусственными*)*, шумовыми* (*флюктуационными*) и *импульсными, активными* и *пассивными* и т.д. Необходимо отметить, что одно и то же колебание может быть полезным сигналом по отношению, например, к одной системе связи или радиолокации и помехой – по отношению к другой. Стоит также отметить, что все помехи, как и все сигналы, являются случайными (если помеха детерминированна, то её можно исключить из наблюдаемого колебания и таким образом избавиться от её вредного

воздействия на сообщение). На рис. 1.8 приведены примеры случайного сигнала и случайной (шумовой) помехи. По способу взаимодействия с сигналом помехи подразделяются на *аддитивные* (от английского *add* – складывать), *мультипликативные* (от английского *multiply* – умножать) и смешанные (сюда относятся все взаимодействия, не сводимые к аддитивному или мультипликативному).

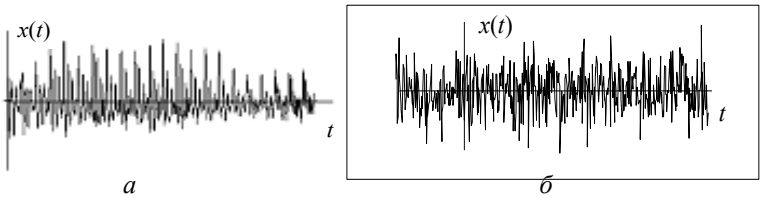


Рис. 1.8. Случайный (речевой) сигнал (*а*) и случайная помеха (шум) (*б*)

Кроме перечисленных, используются и другие признаки классификации сигналов. Иногда различают *информационные* и *управляющие* сигналы (колебания), *модулированные* и *немодулированные*, *узкополосные* и *широкополосные* и т. д. Некоторые из перечисленных типов сигналов будут в дальнейшем рассмотрены подробнее.

В теории электрической связи принято рассматривать сигнал как «объект транспортировки». С этой точки зрения сигнал можно описать тремя «габаритными характеристиками», подобными длине, ширине и высоте груза, перевозимого, скажем, по железной дороге. Первая из таких характеристик – длительность сигнала $T_{\text{с}}$, измеряемая в секундах (с). Любой сигнал можно представить суммой (суперпозицией) гармонических колебаний с определенными частотами, поэтому вторая «габаритная характеристика» – ширина спектра, или полоса частот сигнала $F_{\text{с}}$, равная разности наивысшей и низшей частот его гармонических составляющих и измеряемая в герцах (Гц). Третьей «габаритной» характеристикой служит *динамический диапазон,* измеряемый в децибелах (дБ) и определяемый формулой

$$D_{\text{с}} = 20\lg\left(\frac{X_{\max}}{X_{\min}}\right),$$

где $X_{\max}$ и $X_{\min}$ – соответственно максимальное и минимальное возможные значения сигнала (напряжения или тока). Произведение этих трех величин называется *объемом сигнала*:

$$V_{\text{с}} = T_{\text{с}} F_{\text{с}} D_{\text{с}}.$$

1.3. СИСТЕМЫ И КАНАЛЫ СВЯЗИ

Системы связи подразделяются в соответствии с их назначением на системы телефонии, телеграфии, фототелеграфии, телевидения, телеметрии, телеуправления и передачи данных.

Системы *телефонной связи* предназначены для передачи речевых, а также других звуковых сообщений (например, музыки). Системы телефонной связи подразделяются на профессиональные и вещательные. Таким образом, обычный радиоприемник представляет собой часть вещательной системы телефонной связи.

Системы *телеграфной* связи предназначены для передачи символьных (цифробуквенных) сообщений. В настоящее время в таких системах применяются главным образом печатающие аппараты (телетайпы), хотя изредка еще используют манипуляцию ключом на основе азбуки Морзе.

Системы *фототелеграфной* (*факсимильной*) связи применяют для передачи неподвижных изображений. Изображение (сообщение) путем построчного сканирования «развертывается» в одномерный временно́й (первичный) сигнал, который после передачи по каналу связи подвергается обратному преобразованию в двумерное изображение.

Телевизионные системы также передают неподвижные изображения, но развертка осуществляется многократно (периодически), благодаря чему последовательно сменяющие друг друга изображения (кадры) создают у наблюдателя (получателя сообщения) *иллюзию движения*. Как и телефонные системы, системы телевидения подразделяются на профессиональные и вещательные.

Системы *телеметрии* предназначены для передачи измерительной информации, системы *телеуправления* – для передачи команд (управляющих воздействий).

Для передачи точных цифровых данных (например, для связи между компьютерами в локальных и глобальных сетях) используют системы, которые называются системами *передачи данных*.

Современный уровень цивилизации характеризуется широчайшим использованием систем *записи и воспроизведения* информации, которые также можно считать системами связи, передающими информацию из прошлого в будущее.

Совокупность устройств и линий связи, которые сигнал проходит последовательно между *любыми* двумя точками системы связи, называется *каналом связи*. Таким образом, каналы связи могут соединяться последовательно друг с другом, один канал может входить составной частью в другой канал и т.п.

Если сигнал рассматривается как объект транспортировки, то канал связи можно уподобить транспортному средству, которое характеризуется параметрами, аналогичными параметрам сигнала:

$T_к$ – время действия канала, измеряемое в секундах;

$F_к$ – полоса пропускания канала, измеряемая в герцах;

$D_к$ – динамический диапазон канала в децибелах, определяемый максимальным и минимальным значениями сигнала, которые могут передаваться по данному каналу[12].

Произведение указанных характеристик называется *ёмкостью* (*объёмом*) *канала*:

$$V_к = T_к F_к D_к .$$

Для передачи информации без потерь необходимо выполнение условия

$$V_с \le V_к .$$

Отметим, что при этом возможен «обмен» одних параметров сигнала на другие: например, если время действия канала меньше длительности сигнала, можно «сжать» сигнал во времени путем его записи на магнитную ленту и воспроизведения при передаче с повышенной скоростью. При этом полоса частот сигнала станет во столько же раз шире, во сколько раз сократится время передачи. Ярким примером «обмена» может служить передача информации на сверхбольшие расстояния: например, изображения поверхности Венеры, полученные космической станцией и имеющие большой динамический диапазон, передавались на Землю по каналу связи с малым динамическим диапазоном в течение длительного времени. Можно также «обменять» динамический диапазон на полосу частот, применяя для передачи в условиях сильного шума помехоустойчивый код с короткими широкополосными элементарными сигналами, принимающими всего два значения.

Каналы связи подразделяются:

– *по назначению* – на телеграфные, фототелеграфные, телефонные, телевизионные, телеметрические, каналы звукового вещания, передачи данных и т.д.;

– *по виду используемой среды* – на проводные (воздушные, кабельные, волноводные, световодные) и радиоканалы (радиорелей-

[12] Максимальное значение сигнала ограничивается, в частности, энергетическими характеристиками устройств, входящих в канал, минимальное – шумами (помехами), действующими в канале.

ные, ионосферные, тропосферные, метеорные, спутниковые, космические)[13];

– *по характеру связи* входных и выходных сигналов – на линейные и нелинейные, стационарные и нестационарные, детерминированные и случайные (стохастические);

– *по количеству* независимых переменных в описании сигналов – на временные и пространственно-временные;

– *по характеру* входных и выходных сигналов – на непрерывные (аналоговые), дискретные (цифровые), полунепрерывные (дискретно-аналоговые и аналого-дискретные).

Эта классификация, как и любая другая, является условной и может быть дополнена. В частности, широко известна классификация радиовещательных каналов по длине волны, см. таблицу (названия волн, указанные в скобках, являются нестандартными, но широко используются, при этом волны короче 10 м называют ультракороткими (УКВ)).

Диапазон частот	Диапазон волн	Название частот	Название волн
30…300 Гц	1000…10000 км	Сверхнизкие (СНЧ)	
300…3000 Гц	100…1000 км	Инфранизкие (ИНЧ)	
3…30 кГц	10…100 км	Очень низкие (ОНЧ)	Мириаметровые (сверхдлинные, СДВ)
30…300 кГц	1…10 км	Низкие (НЧ)	Километровые (длинные, ДВ)
300…3000 кГц	100…1000 м	Средние (СЧ)	Гектометровые (средние, СВ)
3…30 МГц	10…100 м	Высокие (ВЧ)	Декаметровые (короткие, КВ)
30…300 МГц	1…10 м	Очень высокие (ОВЧ)	Метровые
300…3000 МГц	10…100 см	Ультравысокие (УВЧ)	Дециметровые
3…30 ГГц	1…10 см	Сверхвысокие (СВЧ)	Сантиметровые
30…300 ГГц	1…10 мм	Крайне высокие (КВЧ)	Миллиметровые
300…3000 ГГц	0,1…1 мм	Гипервысокие (ГВЧ)	Децимиллиметровые

[13] Применяют также *акустические* каналы подводной связи, использующие ультразвуковые колебания.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Дайте определение сообщения, информации, сигнала, помехи, линии связи, искажения.

2. Приведите несколько примеров преобразователей сообщения в первичный сигнал.

3. Зачем нужна модуляция? Назовите виды модуляции при гармоническом переносчике.

4. В чем состоит назначение демодулятора?

5. Что такое оптимальная обработка? квазиоптимальная обработка?

6. Что такое достоверность и помехоустойчивость?

УПРАЖНЕНИЯ

1. Рассчитайте максимальное количество каналов передачи речевых сообщений в диапазонах длинных, средних, коротких, метровых, дециметровых и сантиметровых волн (речевой сигнал по стандарту для телефонной связи занимает полосу частот 300…3400 Гц).

2. В оптоволоконной линии передачи используются волны длиной 0.85...1.8 мкм. Определите максимальное количество речевых сообщений, которые можно передавать одновременно по одному световодному волокну.

3. Рассчитайте количество телевизионных каналов, которые *можно* разместить в диапазонах длинных, средних, коротких, метровых, дециметровых и сантиметровых волн, если полоса частот, отводимая для передачи одной ТВ-программы, составляет 8 МГц. (На практике для ТВ-вещания в метровом диапазоне выделены частоты 48,5...100 МГц (I–V каналы) и 174...230 МГц (VI–XII каналы). В дециметровом диапазоне на частотах 470...1000 МГц располагаются 66 каналов.)

4. Громкость звука часто выражают в децибелах. Уровень громкости определяется выражением $L = 20\lg(p_{\text{эфф}} / p_0)$, где $p_{\text{эфф}}$ – эффективное звуковое давление, а $p_0 = 20$ мкПа – стандартный порог слышимости. Определите звуковое давление, создаваемое шелестом листьев (10 дБ), обычным разговором (60 дБ), громкой музыкой (120 дБ).

5. В децибелах выражают значения величин, имеющих размерность мощности или напряжения (тока). Отношение мощностей, выраженное в децибелах, связано с этой же величиной, выражен-

ной в «разах», соотношением $P\,[\text{дБ}] = 10\lg p\,[\text{раз}]$. Аналогичная формула, связывающая отношения напряжений (токов), имеет вид $U\,[\text{дБ}] = 20\lg u\,[\text{раз}]$. Динамический диапазон речи диктора составляет примерно 30 дБ, симфонического оркестра – 95 дБ. Определите, во сколько раз самый громкий звук речи диктора больше по мощности и по напряжению на выходе микрофона, чем самый слабый звук (то же для оркестра).

6. Телевизионный сигнал изображения занимает полосу частот шириной примерно 6,5 МГц. Изображение передается с частотой 25 кадров в секунду. Считая, что динамический диапазон составляет 48 дБ (уровни яркости от 1 до 256), определите время, необходимое для передачи одного ТВ-кадра по телефонному каналу (полоса частот 300…3400 Гц, динамический диапазон 20 дБ).

2. ОСНОВЫ ТЕОРИИ СИГНАЛОВ

2.1. СИГНАЛЫ И ИХ МАТЕМАТИЧЕСКИЕ МОДЕЛИ

В процессе своего развития любая технология проходит ряд этапов. Вначале устройства и процессы конструируются в большой степени на основе интуиции (эвристическим[14] путем). По мере расширения области применения технологии возрастают цена ошибок, допущенных при проектировании, и суммарные потери вследствие неоптимальности решений. Поэтому параллельно ведутся исследовательские работы по повышению эффективности принимаемых решений (схемных, конструкторских, технологических), а также развиваются теоретические методы анализа и синтеза (построения) систем. Всё сказанное в полной мере относится к электрической связи.

В современных системах связи применяются сложные методы преобразования сигналов, направленные на повышение достоверности передачи информации, помехоустойчивости, надежности связи и т.п. Построение таких систем немыслимо без применения строгих математических методов синтеза и анализа.

Таким образом, естественно возникает вопрос о способах математического описания (*математических моделях*) сигналов и каналов связи и о возможностях преобразования различных моделей друг в друга. В качестве математических моделей сигналов обычно используются подходящие функции или их комбинации (суммы и/или произведения функций, их производных и первообразных и т.п.). Ниже кратко описываются некоторые из таких функций.

А) Гармоническое колебание $A\sin(2\pi ft+\varphi)$, где A – амплитуда, f – частота, φ – начальная фаза колебания. Вместо синуса

[14] От греческого слова *эврика*, произнесенного, согласно легенде, Архимедом в момент озарения.

часто используют косинус. Кроме того, во многих случаях рассматривается комплексное гармоническое колебание $A\exp\ j(2\pi ft+\varphi)$, где $j=\sqrt{-1}$. Это колебание можно представить суммой $A\cos(2\pi ft+\varphi)+\ j\cdot A\sin(2\pi ft+\varphi)$. Иногда в описаниях гармонических колебаний используют круговую частоту $\omega=2\pi f$.

Б) Функция включения Хевисайда (рис. 2.1, *а*), определяемая выражением

$$\sigma(t)=\begin{cases}1 & \text{при } t>0,\\ 0{,}5 & \text{при } t=0,\\ 0 & \text{при } t<0.\end{cases} \qquad (2.1)$$

Функцию Хевисайда, в частности, удобно использовать для представления прямоугольного импульса единичной амплитуды и длительности $\tau_{\text{и}}$ (рис. 2.1, *б*):

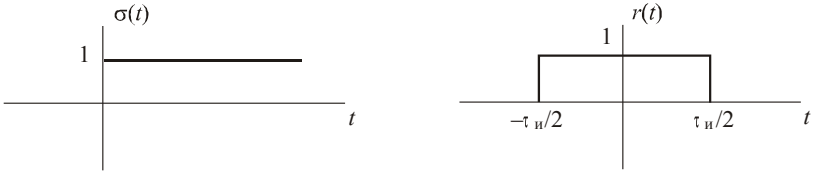
$$r(t)=\sigma(t+\tau_{\text{и}}/2)-\sigma(t-\tau_{\text{и}}/2).$$

В) δ -функция Дирака (читается «дельта-функция») $\delta(t)$, которая на самом деле является *обобщенной* функцией, т.е., строго говоря, *не функцией* в обычном смысле слова [1]. Определяется δ -функция выражением

$$f(t_0)=\int_{-\infty}^{\infty} f(t)\delta(t-t_0)dt\,, \qquad (2.2)$$

которое известно как *стробирующее* (фильтрующее) свойство δ -функции. Оно означает, что δ -функция, входящая в произведение под знаком интеграла, выделяет бесконечно узкий «срез» (*отсчёт*) функции $f(t)$ в точке $t=t_0$. Выражение (2.2) можно понимать как предел

$$f(t_0)=\lim_{\tau_{\text{и}}\to 0}\frac{1}{\tau_{\text{и}}}\int_{-\infty}^{\infty} f(t)r(t-t_0)dt\,.$$



а *б*

Рис. 2.1. Функция Хевисайда (*а*) и прямоугольный импульс (*б*)

С такой точки зрения δ-функцию можно рассматривать как предел последовательности все более коротких прямоугольных импульсов со все большей амплитудой, так что площадь всех импульсов одинакова и равна 1. Тогда (нестрого) можно считать δ-функцию «импульсом» нулевой длительности и бесконечной амплитуды с единичной площадью (рис. 2.2, *а*). Не следует, однако, забывать, что это не обычная, а обобщенная функция, которая имеет особые свойства: так, например, δ-функцию можно дифференцировать[15], но нельзя возводить в квадрат. Поэтому, например, выражение «энергия δ-функции» не имеет смысла. Нужно отметить, что δ-функция играет в теории сигналов совершенно исключительную роль, и в дальнейшем часто будет использоваться.

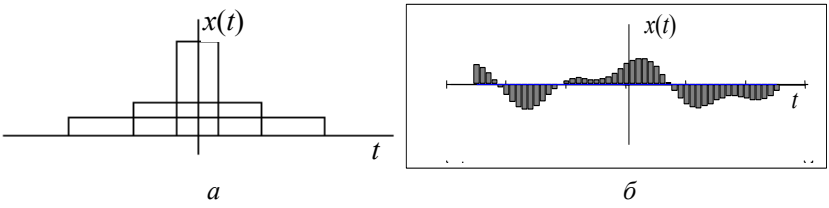
Очевидно, что интеграл

$$\int_{-\infty}^{t} \delta(t)dt = \begin{cases} 0 & \text{при } t < 0, \\ 0{,}5 & \text{при } t = 0, \\ 1 & \text{при } t > 0. \end{cases}$$

Сопоставляя это выражение с формулой (2.1), легко видеть, что функция Хевисайда связана с δ-функцией выражениями

$$\sigma(t) = \int_{-\infty}^{t} \delta(t)dt\,; \quad \delta(t) = \frac{d\sigma(t)}{dt}\,; \tag{2.3}$$

таким образом, δ-функцию можно формально использовать для дифференцирования разрывных функций.



а *б*

Рис. 2.2. δ-функция, как предел последовательности прямоугольных импульсов (*а*) и представление аналогового сигнала «суммой импульсов» (*б*)

---

[15] Производная δ-функции определяется выражением $-f'(t_0) = \int_{-\infty}^{\infty} f(t)\delta'(t-t_0)dt$.

Выражение (2.2), переписанное для сигнала $x(t)$ с учетом четности δ -функции в виде

$$x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau, \qquad (2.4)$$

можно представить, как предел

$$x(t) = \lim_{\tau_и \to 0} \sum_{n=-\infty}^{\infty} x(n\tau_и)\frac{r(t-n\tau_и)}{\tau_и}\tau_и,$$

описывающий сигнал «сплошной суммой бесконечно узких импульсов» (рис. 2.2, *б*). Такое представление часто называют динамическим. Для сигнала, удовлетворяющего условию $x(t) = 0$ при $t < 0$, возможна другая форма динамического представления, основанная на функции Хевисайда

$$x(t) = x(0)\sigma(t) + \int_{0}^{\infty} \frac{dx(\tau)}{d\tau}\sigma(t-\tau)d\tau \qquad (2.5)$$

и получаемая предельным переходом при $\Delta t \to 0$, примененным к выражению

$$x(t) \approx x(0)\sigma(t) + \sum_{n=0}^{\infty} \frac{\left[x\ (n+1)\Delta t\ - x(n\Delta t)\right]}{\Delta t}\sigma(t-n\Delta t)\Delta t,$$

где Δt – временной интервал.

Для представления дискретных сигналов используются функции целого аргумента n, обладающие свойствами, аналогичными свойствам функций (А – В).

а) Гармонические последовательности $x[n] = A\sin(\omega n + \varphi)$ и $x[n] = A\cos(\omega n + \varphi)$ и комплексная экспоненциальная последовательность $x[n] = A\exp[j(\omega n + \varphi)]$.

б) Ступенчатая последовательность (рис. 2.3, *а*), аналогичная функции Хевисайда и определяемая выражением

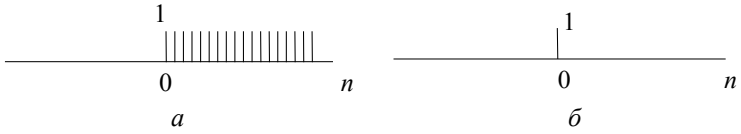
$$u[n] = \begin{cases} 1 & \text{при } n \geq 0, \\ 0 & \text{при } n < 0. \end{cases}$$

в) Функция дискретной переменной, называемая δ -последовательностью и играющая в теории дискретных сигналов роль, ана-

логичную роли δ-функции для аналоговых сигналов, определяется выражением

$$\delta[n] = \begin{cases} 1, & n = 0, \\ 0, & n \neq 0 \end{cases}$$

и является вполне обычной функцией, которую можно представить графиком (рис. 2.3, *б*).



а *б*

Рис. 2.3. Ступенчатая единичная последовательность (*а*) и δ-последовательность (*б*)

Операция дифференцирования для функций дискретного аргумента не имеет смысла и заменяется вычислением разности соседних отсчетов, поэтому выражениям (2.3) соответствуют очевидные соотношения

$$u[n] = \sum_{k=-\infty}^{n} \delta[k] \qquad \text{и} \qquad \delta[n] = u[n] - u[n-1].$$

Дискретный сигнал $x[n]$, $n = \overline{-\infty, \infty}$ можно описать выражением, аналогичным динамическому представлению аналогового сигнала (2.4):

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n-k]. \qquad (2.6)$$

Это очевидное выражение означает, что сигнал $x[n]$ представляется суммой сдвинутых δ-последовательностей при всевозможных целых сдвигах k, при этом каждая δ-последовательность умножается на соответствующий амплитудный коэффициент, равный $x[k]$.

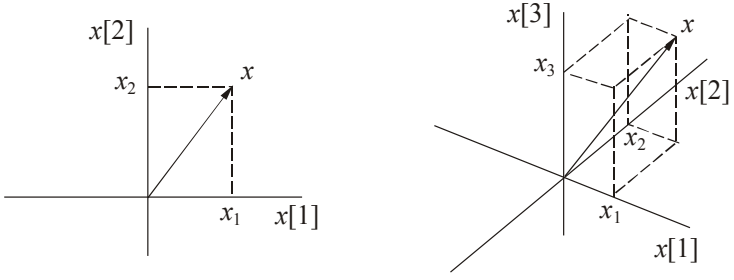
Используя функции (А – В) и (а – в) при различных значениях параметров (амплитуд, частот и начальных фаз для гармонических функций, а также амплитуд и временных сдвигов для остальных), можно получить математические описания (модели) для очень широкого класса сигналов (континуальных и дискретных), фактически для всех сигналов, применяемых на практике. Однако во многих случаях удобнее оказываются иные модели.

Представление сигнала (колебания) в виде *графика* описывающей его функции является наглядным и привычным. В самом деле, большинство сигналов описываются функциями времени, а одним из наиболее распространенных приборов для измерения характеристик электрических сигналов является *осциллограф*, отображающий именно временной график сигнала.

Временно́е представление не является, однако, ни единственным, ни самым лучшим, и на практике при решении конкретных задач следует выбирать наиболее удобные формы описания сигналов.

Основное неудобство, связанное с временным представлением сигналов, заключается в том, что сигналу соответствует *сложный* объект (функция, изображаемая графиком) в *простом* пространстве (на плоскости). В современной теории сигналов используется изображение сигнала *простым* объектом (точкой) в *сложном* пространстве [2]. Это пространство представляет собой множество всевозможных сигналов, рассматриваемых в данной задаче, наделенное соответствующими *структурными* свойствами. При этом свойства сигналов получают наглядное геометрическое истолкование, а для синтеза и анализа сигналов и систем их обработки применяется аппарат современной математики (функциональный анализ).

Основные идеи такого подхода проще изложить для дискретного сигнала. Рассмотрим для примера множество дискретных сигналов, таких, что все значения (отсчеты) этих сигналов равны нулю, за исключением значений, соответствующих $n=1$ и $n=2$. Придавая значениям $x[1]=x_1$ и $x[2]=x_2$ сигнала $x[n]$ смысл абсциссы и ординаты точки (вектора) на плоскости, получаем представление всего множества таких сигналов множеством векторов в двумерном евклидовом пространстве (рис. 2.4, *а*). Множество сигналов,



а *б*

Рис. 2.4. Представление сигнала точкой (вектором) на плоскости (*а*) и представление сигнала вектором в трехмерном пространстве (*б*)

которые могут иметь три ненулевых отсчета (например, при $n=1$, $n=2$ и $n=3$), представляется множеством векторов в трехмерном пространстве (рис. 2.4, *б*).

Продолжая рассуждения, приходим к представлению множества всех сигналов, определяемых их значениями в *конечном множестве* точек дискретной временно́й оси $n=1, 2, ..., N$ множеством векторов N-мерного *евклидова* пространства. Каждый такой вектор представляет собой *упорядоченный набор* чисел (координат), равных *значениям сигнала* в соответствующие моменты времени. Ясно, что такое представление является взаимно однозначным, а следовательно, не приводит к потере информации.

Несмотря на то, что евклидово пространство размерности выше трёх обычный человек вообразить не в состоянии, N-мерное евклидово пространство является весьма обычным и удобным инструментом исследования, так как свойства евклидова пространства сохраняются при любой его размерности. Кроме того, в большинстве случаев рассматриваются пары сигналов (векторов), а любые два вектора лежат в общем для них двумерном *подпространстве* (плоскости). Таким образом, даже не очень богатого пространственного воображения оказывается вполне достаточно для того, чтобы ориентироваться в сигнальном пространстве любой размерности.

Устремляя N к бесконечности, получаем бесконечномерное евклидово пространство, пригодное для представления *всех дискретных* сигналов, определенных на бесконечной целочисленной временно́й оси $n=-\infty,+\infty$. Это пространство имеет бесконечное, но *счетное* множество «координатных осей». Каждому сигналу взаимно однозначно соответствует бесконечный (счетный) *упорядоченный* набор координат вектора, равных, например, *отсчетам* этого сигнала в соответствующие моменты времени.

Переходя к континуальным сигналам, получаем бесконечномерное пространство с *несчетным* множеством (*континуумом*) «координатных осей», при этом сигналу соответствует бесконечный *несчетный* упорядоченный «набор координат» вектора, равных (нестрого говоря) отсчетам этого сигнала в соответствующие моменты времени, которые теперь следуют друг за другом «бесконечно плотно», т.е. непрерывно. Таким образом, и дискретные, и аналоговые сигналы могут быть представлены векторами в линейных пространствах соответствующих размерностей.

Чтобы использовать преимущества таких моделей, следует вначале убедиться в том, что действиям над элементами линейного пространства (векторами) соответствуют операции, применимые к реальным сигналам.

2.2. СИГНАЛЫ И ДЕЙСТВИЯ НАД НИМИ

В каждой практической задаче, связанной с получением (генерированием), передачей, приемом и обработкой сигналов, рассматриваются сигналы из определенного *множества.* Так, можно, например, рассматривать множество $M(T)$ всех континуальных сигналов, заданных на конечном временно́м интервале $t \in [0, T]$ (интервале наблюдения), или множество всех дискретных сигналов, определенных на конечном участке дискретной временно́й оси $n = \overline{1, N}$. Сигналы из одного множества обладают некоторыми общими свойствами, что и позволяет рассматривать множество как целое.

На практике над сигналами выполняются некоторые действия (операции), такие, например, как *сложение* (суммирование). Для этого применяются устройства, называемые сумматорами. Кроме того, суммирование выполняется естественным путем при распространении различных сигналов в общем канале связи или в пространстве, и в этом случае говорят о взаимных помехах. Суммирование применимо к сигналам, имеющим общую область определения. Например, складывая сигналы $s_1(t)$ и $s_2(t)$, определенные на конечном интервале $[0, T]$, получаем сигнал $s_3(t)$, определенный на этом же интервале (сумма сигналов из множества $M(T)$ снова принадлежит $M(T)$, рис. 2.5. В таких случаях говорят, что множество *замкнуто* относительно сложения.

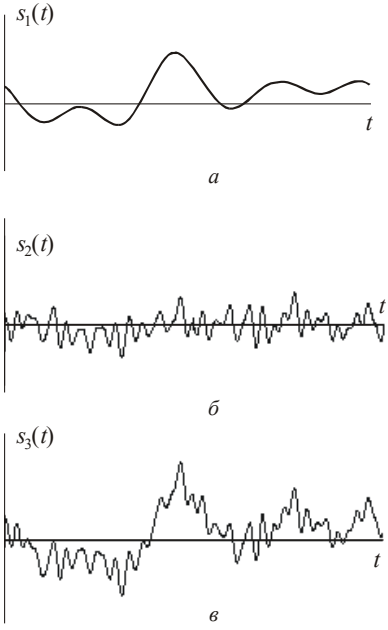


Рис. 2.5. Сигнал (*а*), помеха (*б*) и сумма сигнала и помехи (*в*)

Вторая операция, часто применяемая на практике, – умножение на некоторый постоянный коэффициент. Множитель может быть больше единицы, что соответствует *усилению* сигнала, или меньше единицы, тогда имеет место *ослабление*. Ослабление может быть естественным

(вследствие затухания сигнала в линии передачи или рассеяния энергии в пространстве) или преднамеренным, выполняемым, например, с помощью устройств, называемых *аттенюаторами*. Усиление выполняется при помощи *усилителей*. Множитель может быть и отрицательным, тогда меняется полярность сигнала, а соответствующее устройство называют инвертирующим усилителем, или *инвертором*. На рис. 2.6 сплошной линией показан исходный сигнал, пунктиром тот же сигнал, усиленный вдвое, а штриховой линией – инвертированный сигнал.

Обычно предполагается, что множество сигналов замкнуто относительно умножения на число, таким образом, усиление или ослабление сигнала не нарушает его принадлежности к данному множеству.

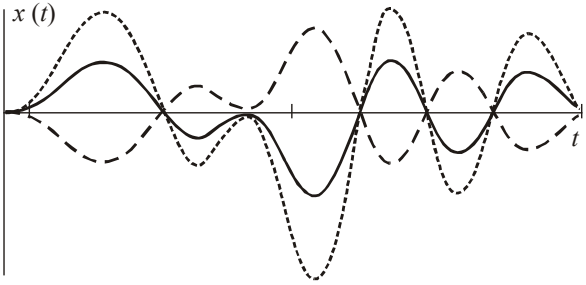


Рис. 2.6. Исходный, усиленный и инвертированный сигналы

Возможность выполнения указанных операций над сигналами обусловливает глубокое сходство множества сигналов с линейным (векторным) пространством. Это позволяет использовать линейное пространство в качестве *модели* для множества сигналов, которое в таком случае становится *пространством* сигналов.

2.3. ЛИНЕЙНОЕ ПРОСТРАНСТВО

Линейным пространством называется множество M объектов (векторов), удовлетворяющее следующим *аксиомам*.

А. Для любых двух векторов из M определена операция сложения, причем сумма вновь принадлежит M (множество M замкнуто относительно сложения), т.е. $\forall x \in M \ \forall y \in M : (x + y) \in M$ ($\forall$ читается «для всех»).

Выполняются следующие *аксиомы сложения*:

1) ассоциативность $\forall x, y, z \in M : x + (y + z) = (x + y) + z$;

2) существование нейтрального элемента (*нулевого вектора*) $\exists \vec{0} \in M : \forall x \in M : x + \vec{0} = x$ ($\exists$ читается «существует»);

3) существование противоположного элемента $\forall x \in M \;\; \exists (-x) \in M : x + (-x) = \vec{0}$;

4) коммутативность $\forall x, y \in M : x + y = y + x$.

Перечисленные аксиомы известны в высшей (абстрактной) алгебре, как аксиомы *коммутативной группы*[16] по сложению.

Б. Для любого вектора из M определена операция умножения на скаляр $\alpha \in \mathbb{F}$ (элемент некоторого *поля* $\mathbb{F}$ – как правило, поля $\mathbb{R}$ вещественных или поля $\mathbb{C}$ комплексных чисел)[17], причем результирующий вектор снова принадлежит M. Иными словами, множество M замкнуто относительно умножения на скаляр: $\forall x \in M \;\; \forall \alpha \in \mathbb{F} : \alpha x \in M$.

Выполняются следующие *аксиомы умножения на скаляр*:

1) ассоциативность $\alpha(\beta x) = (\alpha\beta)x = \alpha\beta x \quad \forall x \in M \;\; \forall \alpha, \beta \in \mathbb{F}$;

2) существование в поле скаляров особого элемента – единицы $\exists 1 \in \mathbb{F} : \forall x \in M : 1x = x$;

3) дистрибутивность сложения векторов и умножения вектора на скаляр $\begin{cases} \alpha(x + y) = \alpha x + \alpha y & \forall x, y \in M \;\; \forall \alpha \in \mathbb{F}, \\ (\alpha + \beta)x = \alpha x + \beta x & \forall x \in M \;\; \forall \alpha, \beta \in \mathbb{F}. \end{cases}$

Нетрудно убедиться непосредственной проверкой, что все эти аксиомы выполняются для сигналов как аналоговых, так и дискретных – вещественных и комплексных. Поэтому *сигналы можно рассматривать как векторы и называть векторами*.

В радиотехнике и связи часто используются комплексные сигналы, принимающие значения из поля $\mathbb{C}$ комплексных чисел. Далее, если явно не сказано обратное, всегда подразумевается, что

---

[16] Коммутативная группа называется также *абелевой* группой в честь Н.Х. Абеля (1802 – 1829), выдающегося норвежского математика.

[17] *Полем* в алгебре называется множество с определенными на нем двумя бинарными операциями, называемыми сложением и умножением, которое является коммутативной группой относительно обеих операций, за исключением существования элемента, противоположного по умножению нейтральному по сложению элементу (запрещено деление на нуль). Кроме полей вещественных и комплексных чисел в теории связи используются конечные *поля Галуа* (см. разд. 8).

сигналы комплексные; вещественные сигналы можно рассматривать как комплексные с нулевой мнимой частью. Таким образом, далее сигналы считаются элементами комплексного пространства (компоненты векторов являются комплексными числами и складываются при сложении векторов, а также умножаются на скаляры по правилам комплексной арифметики).

Очевидно, множество *всех* аналоговых сигналов, заданных на бесконечной временной оси, можно рассматривать как линейное (векторное) пространство (обозначим его L). Большой практический интерес представляет его подмножество – пространство сигналов *ограниченной энергии*, заданных на бесконечной временной оси, которое принято обозначать $L_2(-\infty, \infty)$ или просто L_2. В частных случаях пространство сигналов сужают, например, до *подпространства* $L_2(T)$ сигналов ограниченной энергии, определенных на данном конечном временно́м интервале (сигналов конечной длительности T, тождественно равных нулю вне интервала $[0, T]$), или подпространства $L_2(F)$ сигналов с ограниченной полосой частот F. Линейным пространством является и множество l_2 *всех* дискретных сигналов ограниченной энергии, заданных на всей дискретной временной оси $n = \overline{-\infty, \infty}$. Между двумя последними пространствами, как будет показано ниже, можно установить *взаимно однозначное* соответствие, что делает возможной цифровую обработку сигналов, изначально аналоговых, с последующим преобразованием результата снова в форму аналогового колебания (см. разд. 12).

Поскольку определено сложение векторов и умножение вектора на скаляр, определена и *линейная комбинация* конечной совокупности произвольных векторов $\left\{x_k, k = \overline{1, N}\right\}$:

$$y = \sum_{k=1}^{N} \alpha_k x_k \,, \tag{2.7}$$

где $\left\{\alpha_k, k = \overline{1, N}\right\}$ – произвольный набор скаляров.

Совокупность векторов $\left\{\varphi_k, k = \overline{1, N}\right\}$ *линейно независима*, если равенство $\sum_{k=1}^{N} \alpha_k \varphi_k = \vec{0}$ возможно лишь при условии $\alpha_k = 0$ $\forall k = \overline{1, N}$. Другими словами, линейная независимость означает, что

никакой вектор из данной совокупности нельзя представить линейной комбинацией остальных. Множество всех линейных комбинаций *данной совокупности векторов* при *всевозможных* наборах весовых коэффициентов $\{\alpha_k\}$ образует её *линейную оболочку*. Линейная оболочка совокупности линейно независимых векторов $\{\varphi_k, k=\overline{1,N}\}$ представляет собой линейное пространство; число N называется размерностью этого пространства. Набор векторов $\{\varphi_k, k=\overline{1,N}\}$ в этом случае является *базисом* данного пространства. Для любого пространства существует множество различных базисов, и в каждой задаче можно выбрать наиболее удобный.

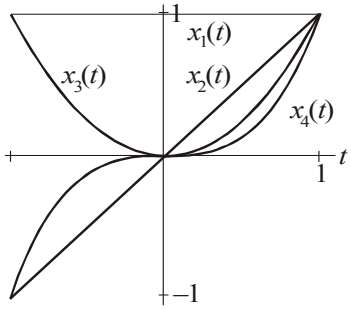


Рис. 2.7. Линейно независимая совокупность функций

Пример 2.1. Множество $S_4 = \{x_1(t)=1, x_2(t)=t, x_3(t)=t^2, x_4(t)=t^3\}$, где $t\in[-1,1]$, линейно независимо (рис. 2.7). Следовательно, оно может служить базисом четырехмерного пространства – пространства всех функций вида $\alpha_1+\alpha_2 t+\alpha_3 t^2+\alpha_4 t^3$ при $t\in[-1,1]$, где коэффициенты α_1, α_2, α_3, α_4 принимают всевозможные комплексные значения. ◄ (Символ ◄ здесь и далее отмечает окончание примера.)

Пример 2.2. Множество $Q_8 = \left\{x_k[n]=\cos\left(\frac{2\pi}{8}kn\right),\ \ k=\overline{0,7}\right\}$ функций целой переменной, определенных на участке дискретной временно́й оси $n=\overline{0,7}$, линейно независимо. Поэтому оно может служить базисом восьмимерного пространства, например, пространства всех дискретных сигналов вида $\sum\limits_{k=0}^{7}\alpha_k\cos\left(\frac{2\pi}{8}kn\right)$, $n=\overline{0,7}$, где $\{\alpha_k, k=\overline{0,7}\}$ – произвольные наборы вещественных чисел. ◄

Пространство всех аналоговых сигналов *бесконечномерно*, поэтому никакая *конечная* совокупность сигналов (функций) не может служить его базисом. Бесконечная совокупность функций

$\{x_k(t), k = \overline{1,\infty}\}$ может быть базисом бесконечномерного пространства L, если множество всех линейных комбинаций вида $\sum_{k=1}^{\infty} \alpha_k x_k(t)$ при *всевозможных* наборах весовых коэффициентов $\{\alpha_k\}$ совпадает с пространством L. Тогда произвольный сигнал из L можно однозначно задать бесконечным набором *коэффициентов разложения* относительно данного базиса, называемого в таком случае *полным* (разумеется, для конкретного сигнала может оказаться, что лишь конечное множество коэффициентов отлично от нуля). Вопрос о полноте базиса бесконечномерного пространства решается в общем случае не просто, однако для базисов, обычно применяемых на практике, полнота доказана [3].

Пространство всех дискретных сигналов, заданных при $n = \overline{-\infty,\infty}$, также бесконечномерно. Один из полных базисов этого пространства определяется выражением (2.6)

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n-k] = \sum_{k=-\infty}^{\infty} \alpha_k \delta[n-k]$$

и представляет собой бесконечный набор δ-последовательностей при всевозможных целочисленных сдвигах $\{\delta[n-k],\ k = \overline{-\infty,\infty}\}$.

Пример 2.3. Множество всех двоичных векторов $B_8 = \{b_k,\ k = \overline{1,8}\}$ при $b_k \in \{0;1\}$ содержит лишь конечное множество элементов (а именно 256). Тем не менее оно может рассматриваться как линейное пространство, если сложение векторов определить через сложение их компонент по модулю 2, а за поле скаляров принять так называемое *поле Галуа* $GF^2 = \{0;1\}$, содержащее всего два числа – 0 и 1. Такие пространства играют очень важную роль, например, в теории кодирования, которая составляет важнейшую часть теории связи. За базис данного пространства можно принять любые 8 линейно независимых ненулевых векторов. ◄

Пример 2.4. В состав декодирующего устройства мобильного телефона стандарта D-AMPS входит устройство памяти, хранящее две «кодовые книги» [4]. Каждая из них содержит по 128 кодовых слов (двоичных векторов), состоящих из 40 компонент и, следовательно, принадлежащих 40-мерному пространству. Однако фактически они принадлежат 7-мерному подпространству, натянутому

на 7 базисных векторов, поэтому для задания кодового слова достаточно указать набор его координат в этом базисе, состоящий из 7 двоичных символов. ◄

2.4. МЕТРИКА, НОРМА И СКАЛЯРНОЕ ПРОИЗВЕДЕНИЕ

Наличие в пространстве полного базиса позволяет описать любой вектор (сигнал) из этого пространства путем задания набора коэффициентов. Во многих случаях нужно не только знать индивидуальное описание сигналов, но и иметь возможность определить количественно отличие сигналов друг от друга. Для этого вводят скалярный функционал[18] $d(x, y)$, определенный для всех пар элементов пространства x и y, называемый расстоянием (*метрикой*), а пространство в таком случае называют *метрическим*.

Метрика должна удовлетворять аксиомам (знак $\rightharpoonup$ читается «только если»):

а) $d(x, y) \geq 0$ и $d(x, y) = 0 \rightharpoonup x = y$;

б) $d(x, y) = d(y, x)$;

в) $d(x, z) \leq d(x, y) + d(y, z)$ (неравенство треугольника).

Отметим, что различные метрики, введенные на одном и том же множестве сигналов, дают различные метрические пространства. Например, на множестве $L(T)$ всех аналоговых сигналов, заданных на интервале $[0, T]$, можно определить следующие метрики [2]:

1) $d_1(x, y) = \int_0^T |\, x(t) - y(t)\,|\, dt$;

2) $d_2(x, y) = \left(\int_0^T |\, x(t) - y(t)\,|^2\, dt \right)^{1/2}$;

3) $d_3(x, y) = \max\limits_{t \in [0, T]} \{ |\, x(t) - y(t)\,| \}$ и т.п.

[18] Функционалом называется *отображение*, ставящее функции (или совокупности функций) в соответствие *число*.

На множестве l всех дискретных сигналов, заданных при $n=\overline{-\infty,\infty}$, можно ввести метрики:

4) $d_4(x,y)=\sum\limits_{n=-\infty}^{\infty}|x[n]-y[n]|$;

5) $d_5(x,y)=\left(\sum\limits_{n=-\infty}^{\infty}|x[n]-y[n]|^2\right)^{1/2}$;

6) $d_6(x,y)=\max\limits_{n=\overline{-\infty,\infty}}\{|x[n]-y[n]|\}$ и т.д.

Иногда при сравнении сигналов нет необходимости знать точный вид сигнала и можно ограничиться лишь его числовой характеристикой (энергией, максимальным значением и т.п.). Обобщением для таких характеристик служит понятие *нормы* вектора. Функционал, исполняющий роль нормы вектора x и обозначаемый $\|x\|$, должен удовлетворять следующим условиям:

а) $\|x\|\ge 0$ и $\|x\|=0\to x=\vec{0}$;

б) $\|x+y\|\le\|x\|+\|y\|$;

в) $\|\alpha x\|=|\alpha|\,\|x\|$.

Норму, как и метрику, можно ввести различными способами. Для аналоговых сигналов чаще всего применяется норма

$$\|x\|_2=\sqrt{\int_0^T|x(t)|^2dt}=\sqrt{E_x},$$

имеющая смысл квадратного корня из *энергии* E_x сигнала. Находят также применение нормы $\|x\|_1=\int\limits_0^T|x(t)|dt$ и $\|x\|_p=\sqrt[p]{\int\limits_0^T|x(t)|^p dt}$. Аналогично вводится норма для дискретных сигналов. Наиболее часто используются нормы $\|x\|_2=\sqrt{\sum\limits_{n=-\infty}^{\infty}|x[n]|^2}$ и $\|x\|_1=\sum\limits_{n=-\infty}^{\infty}|x[n]|$. Пространство с нормой называется *нормированным*. Следует отметить, что, как и в случае метрики, способ задания нормы влияет на свойства пространства.

Ввиду очевидного сходства аксиом нормы и метрики часто (но не всегда!) метрику определяют как норму разности векторов:

$$d(x,y)=\|x-y\|.$$

Например, в пространстве дискретных сигналов норме $\|x\|_2$ можно поставить в соответствие упомянутую выше метрику $d_5(x,y)=\left(\sum_{n=-\infty}^{\infty}|x[n]-y[n]|^2\right)^{1/2}$, которая обобщает на бесконечномерный случай евклидову метрику (расстояние находится «по теореме Пифагора»). Очевидно, метрика $d_2(x,y)=\left(\int_0^T|x(t)-y(t)|^2dt\right)^{1/2}$ является обобщением *евклидовой* метрики на пространство континуальных сигналов.

В большинстве практических задач, связанных с анализом и обработкой сигналов, важную роль играет операция, называемая *скалярным произведением*. Ввести скалярное произведение можно, определив для произвольной пары векторов данного линейного пространства число (скаляр) из соответствующего поля $\mathbb{F}$. Таким образом, скалярное произведение представляет собой функционал. Скалярное произведение векторов x и y, обозначаемое (x,y), должно удовлетворять следующим условиям (аксиомам) [2]:

а) $(x,y)=(y,x)^*$;

б) $(\alpha x+\beta y,z)=\alpha(x,z)+\beta(y,z)$;

в) $(x,x)\geq 0$ и $(x,x)=0\rightarrow x=\vec{0}$.

Знак * в условии а) обозначает комплексное сопряжение величин. Условие б) означает линейность скалярного произведения относительно одного из операндов. Из условия в) следует, что через скалярное произведение можно задать норму, определяемую выражением

$$\|x\|=\sqrt{(x,x)}\,.$$

Таким образом, скалярное произведение *порождает* норму, а через неё – метрику. Если пространство со скалярным произведением и порожденными им нормой и метрикой *полно* (т.е. вместе с любой сходящейся последовательностью векторов содержит и предел этой последовательности [2]), то оно называется *гильбертовым* пространством[19]. Наиболее часто в теории сигналов используются

[19] Названо в честь Д. Гильберта (1862 – 1943) – выдающегося немецкого математика.

именно гильбертовы пространства. Отметим, что в конечномерном случае гильбертово пространство является евклидовым.

Пример 2.5. Множество аналоговых сигналов *ограниченной энергии*, заданных на конечном интервале $[0, T]$, становится гильбертовым пространством, если определить скалярное произведение выражением

$$(x, y) = \int_0^T x(t) y^*(t) dt ,$$

а норму и метрику соответственно выражениями

$$\|x\|_2 = \sqrt{\int_0^T |x(t)|^2 dt} \quad \text{и} \quad d(x, y) = \sqrt{\int_0^T |x(t) - y(t)|^2 dt} .$$

Это пространство принято обозначать $L_2(T)$. Если носитель сигнала – вся вещественная (временная) ось, то пространство сигналов ограниченной энергии обозначается $L_2(-\infty, +\infty)$ или просто L_2. ◄

Пример 2.6. Множество дискретных сигналов (последовательностей) бесконечной протяженности становится гильбертовым пространством, если определить скалярное произведение выражением

$$(x, y) = \sum_{n=-\infty}^{\infty} x[n] y^*[n]$$

и ввести норму и метрику выражениями

$$\|x\|_2 = \sqrt{\sum_{n=-\infty}^{\infty} |x[n]|^2} \text{ и } d(x, y) = \sqrt{\sum_{n=-\infty}^{\infty} |x[n] - y[n]|^2} .$$

Пространство, содержащее все последовательности *конечной нормы* $\|x\|_2$, обозначается l_2 и называется пространством *квадратично суммируемых* последовательностей. ◄

Пример 2.7. Важную роль в теории сигналов и цепей играют нормированные пространства $L_1(T)$ и l_1 аналоговых и дискретных сигналов с нормами, определяемыми соответственно выражениями

$$\|x\|_1 = \int_0^T |x(t)| dt \quad \text{и} \quad \|x\|_1 = \sum_{n=-\infty}^{\infty} |x[n]| .$$

Эти пространства не являются гильбертовыми. ◄

2.5. ГИЛЬБЕРТОВО ПРОСТРАНСТВО

Важная роль гильбертовых пространств, как моделей для пространств сигналов, связана со скалярным произведением и теми преимуществами, которые дает введение этой операции на множестве сигналов.

Скалярное произведение позволяет сравнивать сигналы более полно, чем это возможно в метрическом или нормированном пространстве. Из определения скалярного произведения следует *неравенство Шварца* $|(x, y)|^2 \le (x, x)(y, y)$, которое можно переписать в виде $\dfrac{|(x, y)|}{\|x\|_2 \|y\|_2} \le 1$. Смысл неравенства Шварца в том, что в гильбертовом пространстве, как и в евклидовом, скалярное произведение двух сигналов не может превзойти по модулю произведения их норм, поэтому для пары вещественных[20] сигналов можно определить угол φ между ними выражением $\cos\varphi = \dfrac{(x, y)}{\|x\|_2 \|y\|_2}$.

Рассмотрим два частных случая. В первом случае $(x, y) = \|x\|_2 \|y\|_2$; это означает, что сигналы x и y имеют одинаковую форму и отличаются только нормой. Большой практический интерес представляет второй случай, когда для ненулевых сигналов x и y скалярное произведение $(x, y) = 0$, тогда сигналы называются *ортогональными*. Можно сказать, что первый случай соответствует максимальному сходству сигналов, тогда ортогональность означает их максимальное несходство.

Пример 2.8. Приёмное устройство системы связи с ортогональными сигналами, структура которого иллюстрируется рис. 2.8, содержит m каналов, каждый из которых «настроен» на прием одного сигнала, причем все m сигналов взаимно ортогональны. В приемном устройстве формируются (генерируются) m опорных колебаний, каждое из которых совпадает с одним из ожидаемых сигналов. В каждом канале вычисляется скалярное произведение принимаемого колебания $s(t)$ и одного из опорных колебаний

[20] Для комплексных пространств угол определяется в общем случае неоднозначно [2].

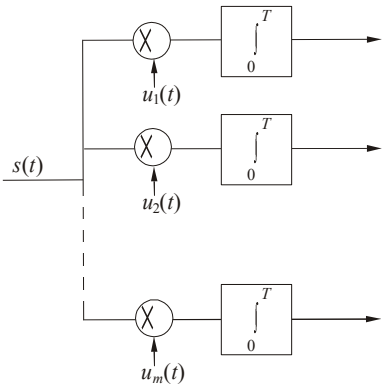


Рис. 2.8. Структура приёмника системы связи с ортогональными сигналами

$u_k(t), k=\overline{1,m}$. Предположим, что принимаемое колебание совпадает по форме с одним из опорных сигналов, но может отличаться по норме (например, вследствие затухания в канале связи). Пусть, например, принимаемое колебание совпадает по форме с опорным сигналом $u_n(t)$, тогда на всех выходах схемы, кроме n-го, по окончании интервала наблюдения будет нулевое значение, а на n-м выходе – произведение норм принимаемого и опорного колебаний, заметно отличное от нуля. Таким образом, измеряя напряжения на выходах схемы, можно определить, какой из m ортогональных сигналов присутствует на входе. Реальное входное колебание всегда содержит смесь сигнала с шумом и/или помехой, поэтому скалярные произведения на выходах каналов отличаются от указанных точных значений; в этом случае ортогональность сигналов гарантирует высокую *помехоустойчивость* системы (подробнее см. разд. 9). ◄

Второе преимущество пространств со скалярным произведением связано с представлением векторов относительно заданного базиса. Скалярное произведение позволяет находить коэффициенты разложения произвольного вектора в данном базисе. Пусть $\{\varphi_k, k=\overline{1,N}\}$ – базис пространства, в котором определено скалярное произведение. Можно построить другой базис $\{\chi_k, k=\overline{1,N}\}$, называемый *сопряженным*, или взаимным, такой, что при любых k и m справедливо выражение $(\varphi_k, \chi_m)=\delta_{km}$, где $\delta_{km}=\begin{cases}1, & k=m\\ 0, & k\neq m\end{cases}$ – символ Кронекера. Это означает, что каждый вектор сопряженного базиса ортогонален всем векторам первого базиса, кроме одного, с которым он имеет скалярное произведение, равное 1. Сопряженный базис является вспомогательным средством для разложения векторов в основном базисе.

Пусть вектор x представляется в виде линейной комбинации базисных векторов $x = \sum_{k=1}^{N} \alpha_k \varphi_k$. Тогда коэффициент α_m находится как скалярное произведение заданного вектора x и вектора χ_m из сопряженного базиса:

$$(x, \chi_m) = \left(\sum_{k=1}^{N} \alpha_k \varphi_k , \chi_m \right) = \sum_{k=1}^{N} \alpha_k (\varphi_k , \chi_m) = \sum_{k=1}^{N} \alpha_k \delta_{km} = \alpha_m .$$

Особенно просто находятся коэффициенты разложения, если базис состоит из взаимно ортогональных векторов, нормы которых равны 1. Такой базис называется *ортонормированным* или *ортонормальным*. Нетрудно убедиться, что ортонормальный базис $\left\{ u_k , k = \overline{1, N} \right\}$ является *самосопряженным*, так как для него выполняется условие $(u_k , u_m) = \delta_{km}$, поэтому коэффициенты разложения для произвольного вектора находятся его скалярным умножением на базисные векторы $\alpha_k = (x, u_k)$, $k = \overline{1, N}$.

Пример 2.9. В пространстве комплексных сигналов конечной длительности T, заданных на интервале $(-T/2, T/2)$, базис $\left\{ \varphi_k (t) = e^{j\omega kt}, \ k = \overline{-\infty, \infty} \right\}$, где $\omega = 2\pi / T$, является ортогональным. В самом деле, для двух произвольно выбранных функций из этого базиса скалярное произведение равно

$$(\varphi_n , \varphi_m) = \int_{-T/2}^{T/2} \varphi_n (t) \varphi_m^{*} (t) dt = \int_{-T/2}^{T/2} e^{j\omega (n-m)t} dt = T \cdot \delta_{m,n} = \begin{cases} 0, & n \neq m \\ T, & n = m \end{cases} .$$

Нормируя базисные функции путем умножения на $1/\sqrt{T}$, можно получить ортонормальный базис $\left\{ \psi_k (t) = \frac{1}{\sqrt{T}} e^{j\omega kt}, k = \overline{-\infty, \infty} \right\}$, для которого справедливо равенство $(\psi_n , \psi_m) = \delta_{m,n}$. Коэффициенты разложения сигнала в данном базисе находятся как скалярные произведения

$$\alpha_k = (x, \psi_k) = \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} x(t) e^{-j\frac{2\pi}{T} kt} dt \quad \forall \ k = \overline{-\infty, \infty} , \qquad (2.8)$$

так что любой сигнал на интервале $(-T/2, T/2)$ можно представить *рядом Фурье*[21]

$$x(t) = \sum_{k=-\infty}^{\infty} \alpha_k \frac{1}{\sqrt{T}} e^{j\frac{2\pi}{T}kt}. \qquad (2.9)$$

Часто используется и представление в ортогональном базисе

$$x(t) = \sum_{k=-\infty}^{\infty} C_k e^{j\frac{2\pi}{T}kt}, \qquad (2.10)$$

где

$$C_k = \frac{\alpha_k}{\sqrt{T}} = \frac{1}{T}\int_{-T/2}^{T/2} x(t) e^{-j\frac{2\pi}{T}kt} dt \quad \forall\ k = \overline{-\infty, \infty}, \qquad (2.11)$$

также называемое рядом Фурье.

Базисы, упомянутые в данном примере, полны в пространстве $L_2(-T/2, T/2)$. Следует, однако, отметить, что, например, в $L_2(-\infty, +\infty)$ они не полны [3]. ◄

Представление сигнала (вектора) относительно *произвольного полного ортонормального* базиса $\{u_k, k = \overline{-\infty, \infty}\}$, $(u_k, u_m) = \delta_{m,k}$, называется *обобщенным рядом Фурье*:

$$x = \sum_{k=-\infty}^{\infty} \alpha_k u_k. \qquad (2.12)$$

Набор $\{\alpha_k, k = \overline{-\infty, \infty}\}$ коэффициентов разложения (2.12) называется *спектром* сигнала x относительно базиса $\{u_k, k = \overline{-\infty, \infty}\}$. Аналогично совокупность всех коэффициентов (2.11) называется спектром сигнала относительно комплексного ряда Фурье (2.10).

Ортонормальные базисы обладают и другим замечательным свойством: зная коэффициенты разложения относительно такого базиса, легко найти нормы и скалярные произведения векторов.

---

[21] Жан Батист Жозеф Фурье (1768 – 1830) – выдающийся французский математик, один из основоположников математической физики.

Действительно, пусть вектор x представлен рядом (2.12). Его норма

$$\|x\|_2 = \sqrt{(x,x)} = \sqrt{\left(\sum_{k=-\infty}^{\infty}\alpha_k u_k, \sum_{m=-\infty}^{\infty}\alpha_m u_m\right)} = \sqrt{\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\alpha_k\alpha_m^*(u_k,u_m)} =$$

$$= \sqrt{\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\alpha_k\alpha_m^*\delta_{km}} = \sqrt{\sum_{m=-\infty}^{\infty}|\alpha_m|^2}\,. \qquad (2.13)$$

Таким образом, доказано *равенство Парсеваля*. В пространствах L_2 и l_2 равенство Парсеваля для сигналов, заданных спектрами $\{\alpha_m, m=\overline{-\infty,\infty}\}$ и $\{\beta_m, m=\overline{-\infty,\infty}\}$ относительно полных ортонормальных базисов, принимает соответственно вид

$$\int_{-\infty}^{\infty}|x(t)|^2\,dt = \sum_{m=-\infty}^{\infty}|\alpha_m|^2 \quad \text{и} \quad \sum_{n=-\infty}^{\infty}|x[n]|^2 = \sum_{m=-\infty}^{\infty}|\beta_m|^2\,.$$

Пусть два вектора представлены в некотором полном ортонормальном базисе выражениями $x=\sum_{k=-\infty}^{\infty}\alpha_k u_k$ и $y=\sum_{k=-\infty}^{\infty}\beta_k u_k$. Тогда их скалярное произведение

$$(x,y) = \left(\sum_{k=-\infty}^{\infty}\alpha_k u_k, \sum_{m=-\infty}^{\infty}\beta_m u_m\right) =$$

$$= \sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\alpha_k\beta_m^*\delta_{km} = \sum_{m=-\infty}^{\infty}\alpha_m\beta_m^*\,. \qquad (2.14)$$

Это выражение носит название *обобщенной формулы Рэлея*. Значение этих равенств состоит в возможности оперировать вместо сигналов коэффициентами их представления в полных ортонормальных базисах (спектрами), даже не интересуясь конкретным видом базиса.

Обобщенный ряд Фурье (ОРФ), представляющий сигнал из бесконечномерного пространства L, содержит в общем случае бесконечно много слагаемых. Часто на практике приходится рассматривать *усеченный* ряд, сумма $\breve{x}$ которого аппроксимирует данный сигнал x:

$$x \approx \breve{x} = \sum_{k=1}^{K}\alpha_k u_k\,.$$

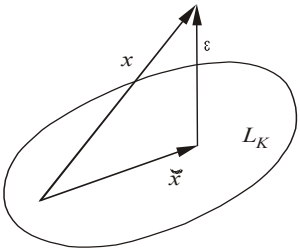


Рис. 2.9. Конечномерная аппроксимация сигнала

Усеченный ОРФ представляет сигнал в виде линейной комбинации K базисных векторов, поэтому x принадлежит K-мерному подпространству L_K пространства L. Поскольку все базисные векторы взаимно ортогональны, ошибка аппроксимации $\varepsilon = x - \breve{x}$ ортогональна по отношению к L_K и принадлежит ортогональному дополнению $L_\perp$, такому, что $L = L_K \oplus L_\perp$ (рис. 2.9). Символ $\oplus$ обозначает прямую сумму пространств (например, трехмерное евклидово пространство можно представить прямой суммой плоскости и прямой, ортогональной этой плоскости). Очевидно,

$$\|\varepsilon\|_2^2 = \|x - \breve{x}\|_2^2 = \|x\|_2^2 - \left\|\sum_{k=1}^{K} \alpha_k u_k\right\|_2^2 = \|x\|_2^2 - \sum_{k=1}^{K} |\alpha_k|^2 \geq 0,$$

откуда следует *неравенство Бесселя*

$$\sum_{k=1}^{K} |\alpha_k|^2 \leq \|x\|_2^2, \qquad (2.15)$$

которое означает, что при аппроксимации сигнала конечной суммой обобщенного ряда Фурье энергия аппроксимирующего сигнала не может превзойти энергию аппроксимируемого сигнала. Равенство возможно только в том случае, если сам сигнал принадлежит подпространству L_K.

С увеличением размерности подпространства L_K, т.е. с увеличением числа слагаемых, входящих в конечную сумму обобщенного ряда Фурье, норма ошибки стремится к нулю (в этом и состоит *практический смысл* требования полноты базиса). Таким образом, располагая полным ортонормальным базисом, можно обеспечить *сколь угодно точную* аппроксимацию сигнала суммой конечного числа наперед заданных функций с соответствующими весовыми коэффициентами; при этом гарантируется, что при заданном числе слагаемых ошибка аппроксимации будет минимальной.

Пример 2.10. Прямоугольный импульс длительности $\tau_и$ и амплитуды A, изображенный на рис. 2.10, на интервале $(-T/2, T/2)$,

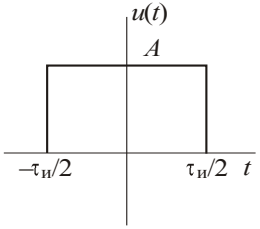


Рис. 2.10. Прямоугольный видеоимпульс

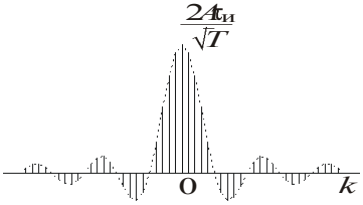


Рис. 2.11. Спектральная диаграмма прямоугольного импульса, заданного на конечном временно́м интервале

$T > \tau_и$, можно представить рядом (2.9) с коэффициентами, найденными согласно выражению

$$\alpha_k = \frac{1}{\sqrt{T}} \int_{-\tau_и/2}^{\tau_и/2} A e^{-j\frac{2\pi}{T}kt}\, dt = \frac{2A\tau_и}{\sqrt{T}} \frac{\sin\left(k\pi\tau_и / T\right)}{k\pi\tau_и / T}.$$

Диаграмма, отображающая спектр прямоугольного импульса относительно ортонормального базиса Фурье, приведена на рис. 2.11. Аппроксимации прямоугольного импульса, полученные как конечные суммы $x(t) = \sum_{k=-K}^{K} \alpha_k \frac{1}{\sqrt{T}} e^{j\frac{2\pi}{T}kt}$ при $K = 5$, $K = 10$ и $K = 20$, показаны различными линиями на рис. 2.12. ◄

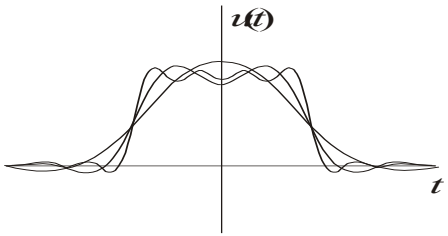


Рис. 2.12. Аппроксимации прямоугольного импульса конечными суммами ряда Фурье

Пример 2.11. Базис, составленный из *функций Уолша,* является ортонормальным полным базисом для $L_2(-1/2, 1/2)$. Графики четырёх первых функций Уолша относительно нормированного времени $\theta = t/T$ показаны на рис. 2.13. Функции Уолша привлекли

внимание благодаря простоте их генерирования при помощи переключательных схем.

Функции Уолша определяются с помощью рекуррентного соотношения

$$\mathrm{wal}(2n+p,\theta)=(-1)^{[n/2]+p}\left\{\mathrm{wal}\left(n,2\theta+\frac{1}{2}\right)+(-1)^{n+p}\,\mathrm{wal}\left(n,2\theta-\frac{1}{2}\right)\right\},$$

$$n=0,1,2,\ldots,\quad p=0,1;\quad \mathrm{wal}(0,\theta)=\begin{cases}1, & \theta\in\left(-\frac{1}{2},\frac{1}{2}\right),\\ 0 & \text{в противном случае.}\end{cases}$$

Здесь $[n/2]$ обозначает целую часть числа $n/2$.

Иногда используют систему функций Уолша, заданную на интервале нормированного времени $(0;1)$. Эта система составляет полный ортонормальный базис для пространства $L_2(0,1)$; такие функции Уолша определяются рекуррентными соотношениями

$$\mathrm{wal}(2n+p,\theta)=$$
$$=\mathrm{wal}(n,2\theta)+(-1)^{n+p}\,\mathrm{wal}(n,2\theta-1),$$
$$n=0,1,2,\ldots,\ p=0,1;$$
$$\mathrm{wal}(0,\theta)=\begin{cases}1, & \theta\in(0,1),\\ 0 & \text{в противном случае.}\end{cases}$$

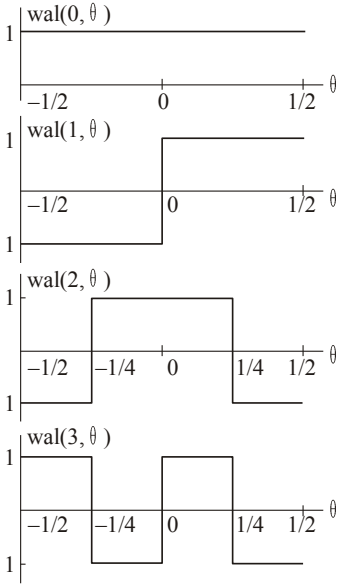


Рис. 2.13. Функции Уолша

Известны также способы определения функций Уолша через матрицы Адамара; при этом получаемые системы функций отличаются нумерацией (способом упорядочения); подробнее см., например, [23]. ◄

Разложение сигналов в различных ортонормальных или ортогональных базисах применяется на практике в тех случаях, когда оперировать спектром сигнала удобнее, чем его временно́й функцией. Устройство, вычисляющее спектральные коэффициенты сигнала, называется *анализатором спектра* (рис. 2.14).

Зная спектральные коэффициенты и базисные функции, можно восстановить сигнал, т.е. выполнить его синтез согласно рис. 2.15. Разложение сигналов относительно неортогонального базиса также возможно, но оно сложнее и его результаты труднее интерпретировать.

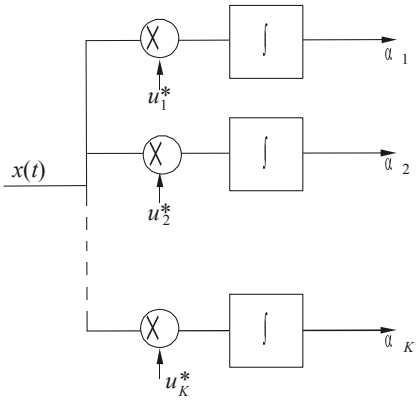


Рис. 2.14. Структура анализатора спектра

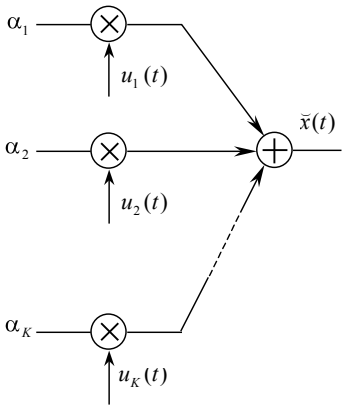


Рис. 2.15. Синтез сигнала по его спектру

Существует алгоритм, называемый *процедурой Грама – Шмидта*[22], позволяющий по имеющемуся набору линейно независимых функций (векторов) построить ортонормальный базис.

Пусть $\left\{v_k, k=\overline{1,\infty}\right\}$ – совокупность линейно независимых векторов, на основе которой требуется построить ортонормальный базис. Введем обозначение $\left\{w_k, k=\overline{1,\infty}\right\}$ для вспомогательной совокупности векторов, а также обозначение $\left\{u_k,\ k=\overline{1,\infty}\right\}$ для ортонормального базиса, который получается в результате выполнения следующих шагов:

1) первый вспомогательный вектор приравнивается первому вектору исходного линейно независимого базиса $w_1 = v_1$; первый

[22] Йорген Грам (1850 – 1916) – датский математик, известен исследованиями в области математической статистики, теории чисел, теории приближения функций рядами; Эрхард Шмидт (1876 – 1959) – немецкий математик, известен результатами исследований в области интегральных уравнений и функционального анализа.

вектор результирующего ортонормального базиса получается нормировкой

$$u_1 = \frac{1}{\|w_1\|_2} w_1;$$

2) второй вспомогательный вектор w_2 получается вычитанием из второго вектора v_2 исходной совокупности его проекции на уже построенный вектор u_1 ортонормального базиса, после чего производится его нормировка и получается второй вектор ортонормального базиса

$$w_2 = v_2 - (v_2, u_1)u_1, \qquad u_2 = \frac{1}{\|w_2\|_2} w_2;$$

3) третий вспомогательный вектор w_3 формируется путем вычитания из очередного вектора v_3 исходной совокупности его проекций на уже построенные векторы u_1 и u_2 ортонормального базиса, после чего этот вектор нормируется

$$w_3 = v_3 - (v_3, u_1)u_1 - (v_3, u_2)u_2, \qquad u_3 = \frac{1}{\|w_3\|_2} w_3 \text{ и т.д.}$$

Продолжая процедуру Грама – Шмидта, можно построить ортонормальный базис любой размерности.

Пример 2.12. Множество $S_4=\{v_0(t)=1, v_1(t)=t, v_2(t)=t^2, v_3(t)=t^3\}$, где $t \in [-1, 1]$, линейно независимо (см. пример 2.1). В результате

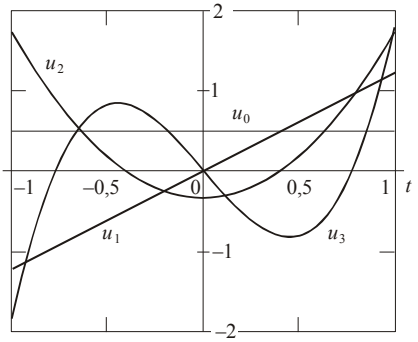


Рис. 2.16. Базис, полученный применением процедуры Грама–Шмидта к совокупности степенных функций, показанной на рис. 2.7

применения процедуры Грама – Шмидта получается ортонормальный базис, состоящий из четырех функций, показанных на рис. 2.16. Это известные полиномы Лежандра, нормированные к единице по норме пространства $L_2(-1,1)$. ◄

Очевидно, в пространствах аналоговых и дискретных сигналов можно построить бесконечно много ортонормальных базисов. Выбор наиболее подходящего базиса определяется конкретной решаемой задачей.

2.6. НЕПРЕРЫВНЫЕ ПРЕДСТАВЛЕНИЯ СИГНАЛОВ

Обобщенный ряд Фурье представляет сигнал в виде взвешенной суммы *счетного*[23] множества базисных функций. Иногда счетный базис неудобен или не годится для описания сигнала. Например, счетный базис Фурье, полный в $L_2(T)$, *не полон* в $L_2(-\infty,\infty)$ и поэтому непригоден для представления сигналов бесконечной длительности. С другой стороны, известные полные в $L_2(-\infty,\infty)$ счетные базисы не обладают теми привлекательными свойствами, которые обусловили широкое применение базиса Фурье в теории и практике и о которых далее будет сказано подробно (см. разд. 2.9).

Гармонические функции, аналогичные функциям базиса Фурье, могут применяться для представления сигналов из $L_2(-\infty,\infty)$, но для этого мощность их множества должна быть *больше мощности счетного множества* (иначе говоря, множество должно быть непрерывным).

Таким образом, понятие обобщенного ряда Фурье подвергается дальнейшему обобщению. Суть этого обобщения заключается в замене суммы бесконечного *счетного* множества базисных функций, умноженных на спектральные коэффициенты, интегралом от функции двух переменных (которая представляет собой «*несчетное* множество» базисных функций), умноженной на функцию одной переменной, называемой *спектральной плотностью*.

Ниже приведены попарно термины и формулы, соответствующие дискретному и непрерывному (интегральному) представлениям аналоговых сигналов.

---

[23] Элементы счетного множества могут быть *пронумерованы*, т.е. поставлены в соответствие элементам множества целых неотрицательных чисел.

Дискретное представление	Интегральное представление
$v_k(t), k = \overline{-\infty, \infty}$ – базис (необязательно ортогональный)	$v(s,t)$ – базисное ядро интегрального представления
$\alpha_k, k = \overline{-\infty, \infty}$ – спектр сигнала относительно выбранного базиса	$\alpha(s)$ – спектральная плотность сигнала относительно выбранного ядра
$x(t) = \sum_{k=-\infty}^{\infty} \alpha_k v_k(t)$ – дискретное представление сигнала	$x(t) = \int_{-\infty}^{\infty} \alpha(s) v(s,t) ds$ – интегральное представление сигнала
$\alpha_k = \langle x, w_k \rangle = \int_{-\infty}^{\infty} x(t) w_k^*(t) dt$ – формула нахождения спектрального коэффициента с использованием сопряженного (взаимного) базиса $w_k(t), k = \overline{-\infty, \infty}$	$\alpha(s) = \int_{-\infty}^{\infty} x(t) w^*(s,t) dt$ – (2.16) формула нахождения спектральной плотности с использованием сопряженного ядра $w(s,t)$
$\langle v_k, w_m \rangle = \delta_{km}$ – условие взаимности (сопряженности) базисов $v_k(t), k = \overline{-\infty, \infty}$ и $w_k(t), k = \overline{-\infty, \infty}$	$\int_{-\infty}^{\infty} v(s,t) w^*(\sigma, t) dt = \delta(s-\sigma)$ – (2.17) условие сопряженности ядер $v(s,t)$ и $w(s,t)$
$\langle u_k, u_m \rangle = \delta_{km}$ – условие ортонормальности (самосопряженности) базиса $u_k(t), k = \overline{-\infty, \infty}$	$\int_{-\infty}^{\infty} u(s,t) u^*(\sigma, t) dt = \delta(s-\sigma)$ и $\int_{-\infty}^{\infty} u(s,t) u^*(s, \tau) ds = \delta(t-\tau)$ – условия самосопряженности базисного ядра $u(s,t)$
$x(t) = \sum_{k=-\infty}^{\infty} \alpha_k u_k(t)$ – обобщенный ряд Фурье (представление сигнала в ортонормальном базисе $\{u_k(t), k = \overline{-\infty, \infty}\}$)	$x(t) = \int_{-\infty}^{\infty} \alpha(s) u(s,t) ds$ – интегральное представление сигнала относительно самосопряженного базисного ядра $u(s,t)$
$\alpha_k = \langle x, u_k \rangle = \int_{-\infty}^{\infty} x(t) u_k^*(t) dt$ – формула нахождения спектрального коэффициента относительно ортонормального базиса	$\alpha(s) = \int_{-\infty}^{\infty} x(t) u^*(s,t) dt$ – формула нахождения спектральной плотности относительно самосопряженного ядра $u(s,t)$

Таким образом, интегральное представление имеет много общего с обобщенным рядом Фурье.

Пример 2.13. Для представления сигналов из пространства $L_2(-\infty,\infty)$ очень часто используется базисное ядро $u(f,t)=e^{j\cdot 2\pi ft}$ (вместо переменной s в обозначении ядра использовано общеупотребительное обозначение частоты буквой f). Ядро является самосопряженным, так как

$$\int_{-\infty}^{\infty} u(f,t)u^*(\varphi,t)dt=\int_{-\infty}^{\infty} e^{j\cdot 2\pi ft}e^{-j\cdot 2\pi\varphi t}dt=$$

$$=\lim_{T\to\infty}\int_{-T}^{T} e^{j\cdot 2\pi(f-\varphi)t}dt=\lim_{T\to\infty}\frac{\sin 2\pi T(f-\varphi)}{\pi(f-\varphi)}=\delta(f-\varphi)$$

(аналогично доказывается и второе условие самосопряженности).

Поэтому спектральная плотность сигнала $x(t)$ относительно данного ядра, которую обозначим $X(f)$, определяется выражением

$$X(f)=\int_{-\infty}^{\infty} x(t)e^{-j\cdot 2\pi ft}dt\,, \tag{2.18}$$

известным как *преобразование Фурье*; формула интегрального представления сигнала

$$x(t)=\int_{-\infty}^{\infty} X(f)e^{j\cdot 2\pi ft}df \tag{2.19}$$

называется *обратным преобразованием Фурье.* ◄

Запишем скалярное произведение двух сигналов $x(t)$ и $y(t)$, выразив сигналы через спектральные плотности при помощи обратного преобразования Фурье:

$$(x,y)=\int_{-\infty}^{\infty} x(t)y^*(t)dt=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(\varphi)e^{j\cdot 2\pi\varphi t}d\varphi\int_{-\infty}^{\infty} Y^*(f)e^{-j\cdot 2\pi ft}df\,dt=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(\varphi)Y^*(f)\int_{-\infty}^{\infty} e^{j\cdot 2\pi(\varphi-f)t}dtdfd\varphi=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X(\varphi)Y^*(f)\delta(\varphi-f)dfd\varphi=\int_{-\infty}^{\infty} X(f)Y^*(f)df\,.$$

Таким образом, получена обобщенная формула Рэлея

$$\int_{-\infty}^{\infty} x(t)y^*(t)dt = \int_{-\infty}^{\infty} X(f)Y^*(f)df \qquad (2.20)$$

для интегрального представления сигналов относительно базисного ядра Фурье $e^{j\cdot 2\pi ft}$. Аналогичное выражение будет справедливо для интегрального представления сигналов относительно *любого самосопряженного* ядра.

Подставляя в (2.20) $y(t) = x(t)$, получаем *равенство Парсеваля*

$$\int_{-\infty}^{\infty} |x(t)|^2\, dt = \int_{-\infty}^{\infty} |X(f)|^2\, df\ . \qquad (2.21)$$

Симметричная форма левых и правых частей выражений (2.20) и (2.21) должна наводить на мысль, что «естественное» временно́е представление сигнала есть на самом деле представление относительно некоторого самосопряженного ядра. Справедливость такого утверждения устанавливается в следующем примере.

Пример 2.14. Для пространства сигналов $L_2(-\infty,\infty)$ примем в качестве базисного ядра сдвинутую (задержанную) δ-функцию $u(t,\tau) = \delta(t-\tau)$ (вместо переменной s использовано обозначение задержки буквой τ). Это ядро является самосопряженным [2]. Поэтому спектральная плотность сигнала $x(t)$ относительно данного ядра определяется выражением

$$x(\tau) = \int_{-\infty}^{\infty} x(t)\delta(t-\tau)dt\ , \qquad (2.22)$$

а интегральное представление сигнала задается формулой

$$x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau\ . \qquad (2.23)$$

Полученное выражение, описывающее стробирующее свойство δ-функции и совпадающее с динамическим представлением сигнала (2.4), явно демонстрирует тот факт, что обычное временно́е представление сигнала можно рассматривать как интегральное (спектральное) представление относительно базисного ядра $u(t,\tau) = \delta(t-\tau)$ со спектральной плотностью $x(\tau)$. Иными словами, временная функция $x(\cdot)$, описывающая сигнал, есть не что

иное, как спектральная плотность. Таким образом, с математической точки зрения *временно́е представление сигнала является не более* (*и не менее*) *естественным, чем частотное* (2.18) *или любое другое представление относительно самосопряженного базисного ядра.* ◄

Пример 2.15. Очень важную роль в теории сигналов играет представление относительно ядра вида $u(t,\tau)=\dfrac{1}{\pi(\tau-t)}$ (вместо переменной s использована переменная τ, имеющая смысл времени). Это ядро является самосопряженным. Поэтому спектральная плотность сигнала $x(t)$ относительно данного ядра определяется выражением

$$\hat{x}(\tau)=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}\frac{x(t)}{\tau-t}dt, \tag{2.24}$$

а интегральное представление сигнала – формулой

$$x(t)=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}\frac{\hat{x}(\tau)}{\tau-t}d\tau. \tag{2.25}$$

Выражения (2.24) и (2.25) называются соответственно прямым и обратным *преобразованиями Гильберта* и используются, в частности, для описания узкополосных детерминированных и случайных колебаний, см. разд. 2.12 и 3.6. ◄

Пример 2.16. Для представления дискретных сигналов из пространства l_2 используется ядро $u(f,n)=e^{j\cdot 2\pi fn}$, зависящее от непрерывной переменной f, имеющей смысл частоты, и от дискретной (целой) переменной n. Спектральную плотность дискретного сигнала $x[n]$ относительно данного ядра можно определить выражением

$$X(f)=\sum_{n=-\infty}^{\infty}x[n]e^{-j\cdot 2\pi fn},\quad -1\le f\le 1, \tag{2.26}$$

а интегральное представление сигнала – выражением

$$x[n]=\int\limits_{-0.5}^{0.5}X(f)e^{j\cdot 2\pi fn}df,\quad n=\overline{-\infty,\infty}. \tag{2.27}$$

Эти выражения называются соответственно прямым и обратным *преобразованиями Фурье* для последовательностей и используются в цифровой обработке сигналов (подробнее см. разд. 12). ◄

Полезно иметь в виду, что дискретное представление сигнала в виде ряда можно истолковать как частный случай интегрального представления. В самом деле, введем для базисных функций $\{v_k(t),\ k=\overline{-\infty,\infty}\}$ обозначение $\{v(t,s_k), k=\overline{-\infty,\infty}\}$, понимая базисные функции, как различные *сечения* некоторой функции двух переменных $v(t,s)$, соответствующие фиксированным значениям переменной s $\{s=s_k, k=\overline{-\infty,\infty}\}$. Подставим выражение для сигнала

$$x(t)=\sum_{k=-\infty}^{\infty}\alpha_k v_k(t)=\sum_{k=-\infty}^{\infty}\alpha_k v(t,s_k)$$

в формулу для нахождения спектральной плотности относительно некоторого ядра с помощью сопряженного ядра (2.16)

$$\alpha(s)=\int_{-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\alpha_k v(t,s_k)w^*(s,t)dt=\sum_{k=-\infty}^{\infty}\alpha_k\int_{-\infty}^{\infty}v(t,s_k)w^*(s,t)dt\ .$$

С учетом условия сопряженности ядер (2.17) получим

$$\alpha(s)=\sum_{k=-\infty}^{\infty}\alpha_k\delta(s-s_k)\ .$$

Таким образом, дискретное представление действительно можно понимать как интегральное представление со спектральной плотностью $\alpha(s)$, сосредоточенной в счетном множестве точек $\left\{s_k, k=\overline{-\infty,\infty}\right\}$.

2.7. ПРЕОБРАЗОВАНИЯ И ОПЕРАТОРЫ

Всюду, где применяются сигналы, они подвергаются преобразованиям. Под преобразованием можно понимать любое изменение сигнала – как целенаправленное, так и непреднамеренное. Целенаправленные преобразования осуществляются в созданных специально для этого устройствах, которые далее будем называть *цепями*. Непреднамеренными являются преобразования, происходящие, например, в *линиях связи*.

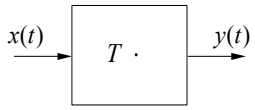


Рис. 2.17. Преобразование сигнала

В наиболее общей форме преобразование изображается схемой рис. 2.17. Обозначая входной сигнал $x(t)$, а выходной

сигнал $y(t)$, можно записать $y(t) = T\{x(t)\}$, где $T\{\cdot\}$ – обозначение преобразования.

С математической точки зрения преобразование представляет собой *отображение* множества входных сигналов $\mathbb{X}$ во множество выходных сигналов $\mathbb{Y}$. Эти множества могут быть одинаковыми, но могут и существенно различаться. Например, в задаче обнаружения[24] полезного сигнала во входном колебании на временно́м интервале T входные колебания принадлежат $L_2(T)$, а множество выходных сигналов состоит из двух значений, условно обозначаемых 0 («сигнала нет») и 1 («сигнал есть»). Здесь будем полагать, что входные и выходные сигналы принадлежат одному и тому же пространству L_2 (или l_2); в этом случае преобразование называется *оператором*. Такая постановка соответствует, в частности, задаче *фильтрации* сигналов. Канал связи представляет собой соединение[25] многих устройств и сред распространения, поэтому осуществляемое им отображение имеет сложный, составной характер. Некоторые из составных частей этого отображения являются операторами, другие, например аналого-цифровые и цифроаналоговые преобразования, описываются отображениями более общего вида.

Рассмотрение преобразований в такой общей постановке не дает каких-либо содержательных результатов именно в силу своей предельной общности. Для того чтобы получить практическую пользу, математическую модель следует конкретизировать (сузить). Очень плодотворный подход состоит в ограничении рассмотрения так называемыми *линейными* операторами[26].

Оператор $\mathbb{L}\{\cdot\}$ называется линейным, если он обладает свойствами аддитивности

$$\mathbb{L}\{x + y\} = \mathbb{L}\{x\} + \mathbb{L}\{y\}$$

и однородности

$$\mathbb{L}\{\alpha x\} = \alpha\mathbb{L}\{x\},$$

---

[24] Подробно эта задача рассматривается в разд. 9.

[25] В простых случаях это каскадное соединение; при *многолучевом* распространении некоторые части канала соединены параллельно.

[26] Некоторые *нелинейные* преобразования колебаний будут рассмотрены в разд. 5, посвященном модуляции и демодуляции.

обычно объединяемыми в одну формулу, выражающую *принцип суперпозиции*:

$$\mathbb{L}\{\alpha x + \beta y\} = \alpha\mathbb{L}\{x\} + \beta\mathbb{L}\{y\}.$$

Таким образом, если оператор, описывающий некоторое устройство (цепь), является линейным, то отклик этой цепи на входной сигнал, представленный обобщенным рядом Фурье $x(t) = \sum_{k=-\infty}^{\infty} \alpha_k u_k(t)$, равен сумме ряда, составленного из откликов на базисные функции с теми же весовыми (спектральными) коэффициентами:

$$y(t) = \mathbb{L}\{x(t)\} = \mathbb{L}\left\{\sum_{k=-\infty}^{\infty} \alpha_k u_k(t)\right\} = \sum_{k=-\infty}^{\infty} \alpha_k \mathbb{L}\{u_k(t)\}. \quad (2.28)$$

Выражение (2.28) описывает *спектральный метод* анализа линейных цепей. Вместо обобщённого ряда Фурье может быть использовано интегральное представление входного сигнала. Чтобы уяснить смысл обсуждаемых понятий, рассмотрим действие линейного оператора в конечномерном линейном пространстве.

Линейные операторы в конечномерных пространствах описываются квадратными матрицами. Рассмотрим пространство дискретных сигналов, каждый из которых представляется N комплексными отсчетами (N-мерное пространство). Результатом воздействия линейного оператора, описываемого матрицей $\Lambda = \left(\lambda_{ij}, i, j = \overline{1, N}\right)$, на вектор-столбец $x = (x_1, ..., x_N)^T$ является вектор-столбец $y = (y_1, ..., y_N)^T$, при этом

$$\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & ... & \lambda_{1N} \\ \lambda_{21} & \lambda_{22} & ... & \lambda_{2N} \\ ... & ... & ... & ... \\ \lambda_{N1} & \lambda_{N2} & ... & \lambda_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_N \end{pmatrix},$$

и значение (отсчет) выходного сигнала описывается выражением $y_k = \sum_{j=1}^{N} \lambda_{kj} x_j$, $k = \overline{1, N}$.

Наглядно представить себе поведение линейного оператора можно на примере его действия на базисные векторы

$e_1 = (1, 0, ..., 0)^T$, $e_2 = (0, 1, ..., 0)^T$, ... , $e_N = (0, 0, ..., 1)^T$. Легко видеть, что вектор e_1 преобразуется в вектор $(\lambda_{11}, \lambda_{21}, ..., \lambda_{N1})^T$, аналогично остальные векторы ортонормального базиса преобразуются в векторы-столбцы, из которых составлена матрица линейного оператора.

Из линейной алгебры известно, что существуют векторы, которые данным оператором преобразуются наиболее простым образом: изменяются лишь их длины (нормы); такие векторы называются *собственными* векторами, а коэффициенты, определяющие изменение длин[27], называются *собственными значениями* оператора. Нетрудно видеть, что если базис пространства составить из собственных векторов данного оператора, то матрица оператора будет *диагональной*

$$\begin{pmatrix} y'_1 \\ y'_2 \\ ... \\ y'_N \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 & ... & 0 \\ 0 & \lambda_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_{NN} \end{pmatrix} \cdot \begin{pmatrix} x'_1 \\ x'_2 \\ ... \\ x'_N \end{pmatrix}, \qquad (2.29)$$

где главная диагональ матрицы составлена из собственных значений, и отсчёты выходного сигнала находятся наиболее просто: $y'_k = \lambda_{kk} x'_k$, $k = \overline{1, N}$ (штрихами отмечены компоненты векторов относительно собственного базиса). Далее будет показано, что аналогичное упрощение может быть достигнуто и для пространства аналоговых сигналов L_2 при соответствующем выборе базисных векторов (функций).

Переход от конечномерного пространства к бесконечномерному пространству дискретных сигналов l_2 приводит к тому, что векторы x и y содержат бесконечно много компонент, соответственно матрица линейного оператора становится бесконечной $\Lambda = \left(\lambda_{ij}, i, j = \overline{-\infty, \infty}\right)$. Значение (отсчет) выходного сигнала определяется выражением $y_k = \sum_{j=-\infty}^{\infty} \lambda_{kj} x_j$, $k = \overline{-\infty, \infty}$, представляющим собой скалярное произведение строки матрицы оператора на вектор-столбец входного сигнала.

---

[27] Сказанное верно для пространства над полем вещественных чисел; операторы, действующие в комплексном пространстве, имеют комплексные собственные значения.

Гильбертово пространство аналоговых сигналов L_2 отличается тем, что множество компонент каждого его вектора *несчетно*, поэтому дискретные индексы заменяются непрерывными переменными, а место матрицы занимает функция $\lambda(\cdot,\cdot)$ двух переменных, называемая *ядром* оператора. Тогда действие линейного оператора на сигнал $x(t)$ описывается интегральным выражением

$$y(t) = \int_{-\infty}^{\infty} \lambda(t, s)x(s)ds \,. \tag{2.30}$$

Здесь переменная s имеет физический смысл и размерность, соответствующие базису, выбранному для описания сигнала x. В частности, это может быть частота, если сигнал задан спектральной плотностью (2.18), или время, если сигнал x задан во временной области (2.22).

2.8. ВРЕМЕННÓЕ ОПИСАНИЕ ЛИНЕЙНЫХ ИНВАРИАНТНЫХ К СДВИГУ (ЛИС) ЦЕПЕЙ

Используя выражение (2.28), найдём отклик цепи на сигнал, представленный выражением (2.23). Очевидно,

$$y(t) = \mathbb{L}\{x(t)\} = \mathbb{L}\left\{\int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)\,d\tau\right\} = \int_{-\infty}^{\infty} x(\tau)\mathbb{L}\{\delta(t-\tau)\}\,d\tau =$$

$$= \int_{-\infty}^{\infty} x(\tau)h(t,\tau)d\tau \,, \tag{2.31}$$

где весовая функция (ядро оператора) $h(t,\tau) = \mathbb{L}\{\delta(t-\tau)\}$ представляет собой отклик (реакцию) цепи в момент t на входной сигнал в виде δ-функции, воздействующий на цепь в момент τ.

Особое значение в анализе цепей имеет случай, когда весовая функция фактически зависит только от разности переменных $h(t,\tau) = h(t-\tau)$, тогда цепь называется *линейной инвариантной к сдвигу* (ЛИС-цепью), или *линейной стационарной*[28], а выражение (2.31) приобретает вид

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \,. \tag{2.32}$$

[28] Нестационарные, или *параметрические*, цепи широко применяются при модуляции и демодуляции сигналов (см. разд. 5).

Выражение (2.32) известно под названием *свёртки*, или *интеграла Дюамеля*[29]. (Иногда используется символическое обозначение свертки выражением $x * h.$)

Если подставить в (2.32) в качестве входного сигнала $x(t) = \delta(t)$, выходной сигнал

$$y(t) = \int_{-\infty}^{\infty} \delta(\tau) h(t-\tau) d\tau = h(t).$$

Таким образом, функция $h(t)$ представляет собой отклик ЛИС-цепи на «бесконечно короткий импульс» (δ-функцию) и называется *импульсной характеристикой* цепи. Зная входной сигнал и импульсную характеристику цепи, всегда можно точно определить выходной сигнал. Поэтому импульсная характеристика (ИХ) составляет *исчерпывающее* описание ЛИС-цепи. Условие $h(t, \tau) = h(t - \tau)$ означает, что, зная реакцию $h(t)$ цепи на воздействие $\delta(t)$, можно определить отклик на сдвинутое воздействие $\delta(t - \tau)$ путем простого сдвига импульсной характеристики на такую же величину τ. Иными словами, поведение такой цепи неизменно во времени.

Пример 2.17. *RC*-фильтр нижних частот, представленный на рис. 2.18, имеет импульсную характеристику $h(t) = \frac{1}{\tau} e^{-t/\tau}$, $\tau = RC$ (рис. 2.19).◀

Для уяснения физического смысла интеграла Дюамеля, играющего важнейшую роль в анализе линейных стационарных цепей, полезно выполнить в (2.32) замену переменных, так что

$$y(t) = \int_{-\infty}^{\infty} x(t-\tau) h(\tau) d\tau. \qquad (2.33)$$

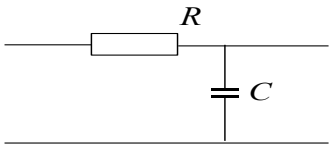


Рис. 2.18. RC-фильтр нижних частот

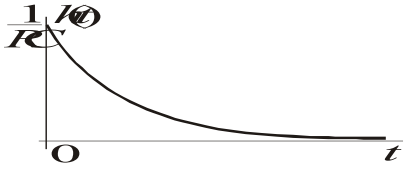


Рис. 2.19. Импульсная характеристика *RC*-фильтра нижних частот

[29] Жан Мари Констан Дюамель (1797 – 1872) – французский математик.

Кроме того, для простоты примем, что импульсная характеристика удовлетворяет условию *каузальности* (причинности)

$$h(t) \equiv 0 \quad \text{при} \quad t < 0 . \tag{2.34}$$

Согласно (2.23) входной сигнал представляется «плотной» последовательностью δ-функций с «амплитудными» коэффициентами, равными значениям сигнала в соответствующие моменты времени. Тогда выражение (2.33) описывает выходной сигнал в момент времени t, как интегральную сумму откликов на все эти δ-функции, воздействовавшие на вход цепи в прошлом. Каждая такая δ-функция отстоит от текущего момента t на величину τ в прошлое, поэтому её вклад в текущее значение выходного сигнала определяется значением импульсной характеристики, соответствующим интервалу τ. Импульсная характеристика любой реальной цепи со временем убывает (затухает), таким образом, цепь постепенно «забывает» значения входного сигнала (рис. 2.20).

Заметим, что ЛИС-цепи представляют собой сравнительно узкий класс цепей (вообще говоря, *никакая* цепь не может быть *строго линейной* хотя бы потому, что любое реальное устройство состоит из веществ, имеющих конечную температуру плавления или возгорания; точно так же реальная цепь не может быть *строго* стационарной уже в силу конечности времени ее существования). Однако очень многие цепи и каналы связи могут считаться *приближенно* линейными инвариантными к сдвигу, а вместе с удобством анализа и синтеза ЛИС-цепей это составляет огромное преимущество линейной стационарной модели и обусловливает ее широкое использование. Нелинейные и/или нестационарные цепи значительно труднее анализировать (не существует, в частности, *общего* метода анализа *всех* нелинейных цепей, аналогичного спектральному методу) и синтезировать, однако некоторые преобразования сигналов, необходимые для практики, *невозможно* осуществить при помощи ЛИС-цепей. Преобразования гармонических

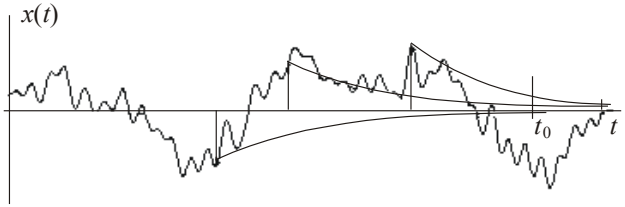


Рис. 2.20. Иллюстрация смысла интеграла Дюамеля

колебаний в нелинейных безынерционных и линейных нестационарных цепях, используемые при модуляции и демодуляции сигналов, будут рассмотрены в разд. 5.

2.9. ЧАСТОТНОЕ ОПИСАНИЕ ЛИС-ЦЕПЕЙ

Интеграл Дюамеля описывает действие оператора ЛИС-цепи на входной сигнал, представленный интегральным выражением (2.23) относительно базисного ядра $\delta(t-\tau)$. Проводя аналогию с конечномерным линейным пространством, можно ожидать, что возможно представление сигнала относительно ядра, аналогичного собственному базису; при этом действие оператора должно описываться более простым выражением. Другими словами, линейному оператору соответствуют векторы (функции), обладающие следующим свойством: действие данного оператора на эти функции сводится к их умножению на скалярные коэффициенты. Обозначим такую *собственную* функцию $\phi(t)$; она должна удовлетворять уравнению

$$\int_{-\infty}^{\infty} \lambda(t,s)\phi(s)ds = \vartheta_{\phi}\phi(t),$$

где ϑ_{ϕ} – некоторый числовой множитель (собственное *значение*, соответствующее данной собственной функции). Различным линейным операторам соответствуют различные наборы собственных функций и собственных значений.

Для линейного инвариантного к сдвигу (стационарного) оператора собственная функция должна удовлетворять уравнению, записываемому с учетом (2.33):

$$\int_{-\infty}^{\infty} \phi(t-\tau)h(\tau)d\tau = \vartheta_{\phi}\phi(t).$$

Легко убедиться, что решением этого интегрального уравнения является комплексная гармоническая функция $e^{j\cdot 2\pi ft}$, где f – её параметр, имеющий смысл частоты:

$$\int_{-\infty}^{\infty} e^{j\cdot 2\pi f(t-\tau)}h(\tau)d\tau = e^{j\cdot 2\pi ft}\int_{-\infty}^{\infty} e^{-j\cdot 2\pi f\tau}h(\tau)d\tau = H(f)e^{j\cdot 2\pi ft}.$$

Итак, если на вход ЛИС-цепи поступает сигнал $e^{j\cdot 2\pi ft}$, то на выходе наблюдается этот же сигнал, умноженный на комплексное

число, зависящее от частоты сигнала. Функция $H(f)$, описывающая эту зависимость, называется *комплексной частотной характеристикой* (КЧХ)[30] цепи и связана с импульсной характеристикой парой преобразований Фурье:

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{-j\cdot 2\pi f t} dt , \qquad (2.35)$$

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{j\cdot 2\pi f t} df . \qquad (2.36)$$

Таким образом, функции времени $\{e^{j\cdot 2\pi f t}\}$ при различных значениях f являются собственными функциями оператора любой ЛИС-цепи, при этом конкретной цепи соответствует определенная КЧХ $H(f)$, определяющая масштабный коэффициент (собственное значение) для каждой функции $e^{j\cdot 2\pi f t}$ при любом значении частоты f.

Запишем входной сигнал в виде интегрального выражения относительно ядра $e^{j\cdot 2\pi f t}$:

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j\cdot 2\pi f t} df . \qquad (2.37)$$

Напомним, что это выражение представляет $x(t)$ «сплошной» суммой базисных функций $e^{j2\pi f t}$ с «амплитудными[31] коэффициентами» $X(f)$. Следовательно, отклик ЛИС-цепи с КЧХ $H(f)$ на этот сигнал представляется интегралом

$$y(t) = \int_{-\infty}^{\infty} H(f) X(f) e^{j\cdot 2\pi f t} df ,$$

так как каждая функция $e^{j\cdot 2\pi f t}$ в разложении (2.37) умножается на $H(f)$. Учитывая, что $y(t) = \int_{-\infty}^{\infty} Y(f) e^{j\cdot 2\pi f t} df$, можно записать вы-

[30] Эту характеристику называют также комплексным коэффициентом передачи, передаточной функцией и т.п.

[31] Ясно, что на самом деле амплитуды гармонических составляющих бесконечно малы.

ражение $Y(f) = H(f)X(f)$, связывающее выходной сигнал ЛИС-цепи с входным сигналом. Заметим, что это выражение соответствует в конечномерном случае умножению вектора на диагональную матрицу (2.29).

Подытоживая, можно сказать, что представление входного сигнала относительно собственного базисного ядра $e^{j\cdot 2\pi ft}$ имеет преимущество перед динамическим представлением, так как вместо *интегрального выражения свертки* связь входного сигнала с выходным описывается *произведением* спектральных плотностей. Уместно еще раз напомнить, что «естественное» временно́е представление сигнала $x(t)$ – это также *спектральная плотность*, только относительно ядра $\delta(t-\tau)$.

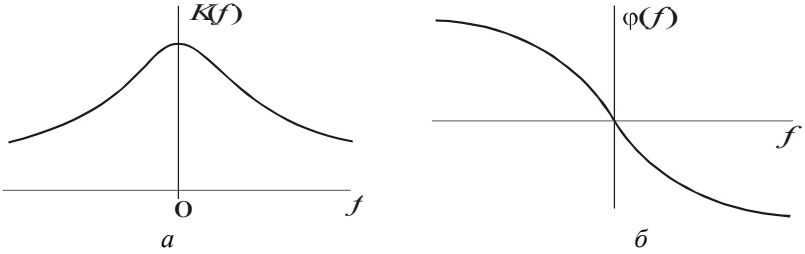
Выражение

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

устанавливающее связь спектральных плотностей сигналов на входе и выходе ЛИС-цепи через её комплексную частотную характеристику, служит основой *спектрального метода анализа* линейных стационарных цепей, широко используемого благодаря своей простоте. Именно этим объясняется *исключительная роль ряда и интеграла Фурье в теории сигналов и цепей*.

Функция $H(f)$ в общем случае является комплексной, $H(f) = K(f)e^{j\varphi(f)}$, что неудобно. Часто рассматривают её модуль и аргумент по отдельности, при этом модуль $K(f) = |H(f)|$ называют *амплитудно-частотной характеристикой* (АЧХ), а аргумент $\varphi(f)$ – *фазочастотной характеристикой* (ФЧХ) цепи.

Пример 2.18. *RC*-фильтр нижних частот, представленный на рис. 2.18, имеет амплитудно-частотную характеристику и фазочастотную характеристику, показанные на рис. 2.21.◀



а *б*

Рис. 2.21. Амплитудно-частотная характеристика (*а*) и фазочастотная характеристика (*б*) *RC*-фильтра нижних частот

Значение комплексной частотной характеристики при заданной частоте f может *в принципе* быть измерено как отношение сигнала на выходе ЛИС-цепи к входному сигналу, если этот входной сигнал – функция $e^{j\cdot 2\pi ft}$. Таким образом, функция $e^{j\cdot 2\pi ft}$ при произвольно задаваемой частоте f может рассматриваться, как *испытательный* сигнал, позволяющий получить описание цепи (КЧХ). Другим испытательным сигналом является δ-функция, которая могла бы быть использована для получения отклика цепи в виде импульсной характеристики. Поскольку КЧХ и импульсная характеристика связаны друг с другом взаимно однозначно (через пару преобразований Фурье), должна существовать связь и между соответствующими им испытательными сигналами. В самом деле, δ-функция может рассматриваться как интегральная сумма *одновременно* воздействующих на вход цепи функций $e^{j\cdot 2\pi ft}$, так как её спектральная плотность

$$\int_{-\infty}^{\infty} \delta(t) e^{-j\cdot 2\pi ft} dt = 1 .$$

Каждая из комплексных гармонических функций умножается цепью на соответствующее значение КЧХ, поэтому импульсная характеристика – отклик на δ-функцию

$$h(t) = \int_{-\infty}^{\infty} H(f)\cdot 1\cdot e^{j\cdot 2\pi ft} df$$

представляет собой, образно говоря, «равнодействующую» откликов на все такие функции.

Заметим, что указанные измерения КЧХ и импульсной характеристики на практике точно выполнить нельзя. Даже если бы существовали абсолютно точные измерительные приборы, потребовалось бы бесконечное время для генерирования функций $e^{j\cdot 2\pi ft}$ (нельзя забывать, что они определены на всей временно́й оси!) и измерения отношений выходных сигналов к входным с бесконечной точностью при всех значениях частоты f. В свою очередь, δ-функция представляет собой «бесконечно короткий импульс бесконечно большой амплитуды», который также не может быть реализован точно. На практике КЧХ и импульсная характеристика ЛИС-цепи могут быть измерены приближенно с помощью отрезков гармонических испытательных сигналов конечной продолжительности и коротких импульсов большой (но конечной) амплитуды.

Часто в выражениях, связанных со спектральным анализом сигналов и ЛИС-цепей, вместо частоты f используется *круговая* частота $\omega = 2\pi f$. Пара (2.18) – (2.19) преобразований Фурье в результате замены переменных принимает вид

$$H(\omega) = \int_{-\infty}^{\infty} h(t) e^{-j\omega t} dt ,$$

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) e^{j\omega t} d\omega .$$

2.10. РЯД ФУРЬЕ И ИНТЕГРАЛ ФУРЬЕ

Как было показано выше, гармонические функции $e^{j\cdot 2\pi ft}$ играют исключительно важную роль в анализе цепей, как собственные функции любого линейного стационарного оператора. Благодаря этому среди всех базисов пространств сигналов, применяемых в теории и практике, базис Фурье получил наибольшее распространение и заслуживает более детального изучения.

2.10.1. РЯД ФУРЬЕ, ЕГО ФОРМЫ, СВОЙСТВА СПЕКТРОВ

Для пространства сигналов конечной длительности и ограниченной энергии $L_2(T)$ ортонормальный базис $\left\{ \frac{1}{\sqrt{T}} e^{j\frac{2\pi}{T}kt}, \; k = \overline{-\infty, \infty} \right\}$ является полным, следовательно, всякий сигнал $x(t) \in L_2(T)$ можно на интервале $(-T/2, T/2)$ представить обобщенным рядом Фурье по ортонормальным функциям

$$x(t) = \sum_{k=-\infty}^{\infty} \alpha_k \frac{1}{\sqrt{T}} e^{j\frac{2\pi}{T}kt} \qquad (2.38)$$

или рядом Фурье по ортогональным функциям

$$x(t) = \sum_{k=-\infty}^{\infty} C_k e^{j\frac{2\pi}{T}kt} . \qquad (2.39)$$

Спектральные коэффициенты для этих рядов определяются выражениями

$$\alpha_k = \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} x(t) e^{-j\frac{2\pi}{T}kt} dt \qquad (2.40)$$

и

$$C_k = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j\frac{2\pi}{T}kt} dt . \qquad (2.41)$$

Для ряда (2.38) справедливо равенство Парсеваля

$$\int_{-T/2}^{T/2} |x(t)|^2 dt = \sum_{k=-\infty}^{\infty} |\alpha_k|^2 .$$

Для ряда (2.39) выполняется равенство

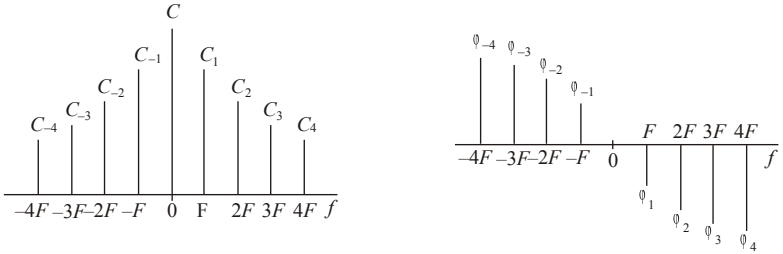
$$\int_{-T/2}^{T/2} |x(t)|^2 dt = T \sum_{k=-\infty}^{\infty} |C_k|^2 .$$

До сих пор базисные функции рассматривались на конечном временно́м интервале $(-T/2, T/2)$. Нетрудно видеть, что эти функции могут рассматриваться и вне этого интервала, т.е. на всей бесконечной временно́й оси. Поскольку все функции $\left\{ e^{j\frac{2\pi}{T}kt}, k = \overline{-\infty, \infty} \right\}$ периодичны, причем для их периодов величина T – наименьшее общее кратное, ряды (2.38) и (2.39), рассматриваемые на всей временно́й оси, определяют *периодическую* функцию, которая представляет собой сигнал $x(t)$, повторяющийся с периодом T.

Таким образом, ряд Фурье одинаково пригоден для представления сигналов конечной длительности и периодических сигналов. Коэффициенты в обоих случаях находятся по формулам (2.40) или (2.41). Далее будет рассматриваться *комплексный ряд Фурье* в форме (2.39).

Коэффициенты ряда Фурье (2.39) даже для вещественного сигнала в общем случае являются комплексными. Для удобства графического представления рассматривают отдельно модули и аргументы коэффициентов $C_k = |C_k| e^{j\varphi_k}$, при этом совокупность $\left\{ |C_k|, k = \overline{-\infty, \infty} \right\}$ называется *амплитудным спектром*, а

$\left\{\varphi_k, k=\overline{-\infty,\infty}\right\}$ – *фазовым спектром* сигнала. Для наглядности амплитудный и фазовый спектр изображают решетчатыми спектральными диаграммами, на которых соответствующие величины показаны длинами отрезков, а сами эти отрезки размещены на частотной оси с шагом, равным в выбранном масштабе частоте повторения сигнала $F=1/T$ (рис. 2.22).



а *б*

Рис. 2.22. Амплитудная и фазовая спектральные диаграммы вещественного сигнала

Если сигнал $x(t)$ принимает вещественные значения, амплитудный спектр обладает свойством четности, а фазовый – свойством нечетности. Действительно, для произвольного спектрального коэффициента

$$C_k=\frac{1}{T}\int\limits_{-T/2}^{T/2}x(t)e^{-j\frac{2\pi}{T}kt}dt$$

с учетом вещественности сигнала $x^*(t)=x(t)$ и

$$C_{-k}=\frac{1}{T}\int\limits_{-T/2}^{T/2}x(t)e^{j\frac{2\pi}{T}kt}dt=\left(\frac{1}{T}\int\limits_{-T/2}^{T/2}x(t)e^{-j\frac{2\pi}{T}kt}dt\right)^*=C_k^*.$$

Таким образом, коэффициенты комплексного ряда Фурье вещественного сигнала попарно комплексно сопряжены. Пользуясь этим свойством, для *вещественных сигналов* можно получить другую форму ряда Фурье, также находящую применение.

Просуммируем пару базисных функций с номерами (индексами) k и $(-k)$ с учетом соответствующих спектральных коэффициентов:

$$C_ke^{j\frac{2\pi}{T}kt}+C_{-k}e^{-j\frac{2\pi}{T}kt}=C_ke^{j\frac{2\pi}{T}kt}+C_k^*e^{-j\frac{2\pi}{T}kt}=$$

$$= |C_k| e^{j\varphi_k} e^{j\frac{2\pi}{T}kt} + |C_k| e^{-j\varphi_k} e^{-j\frac{2\pi}{T}kt} = 2|C_k| \cos\left(\frac{2\pi}{T}kt + \varphi_k\right). \quad (2.42)$$

Тогда ряд Фурье (2.39) можно записать в *тригонометрической* форме

$$x(t) = \sum_{k=0}^{\infty} A_k \cos\left(\frac{2\pi}{T}kt + \varphi_k\right), \quad (2.43)$$

где $A_k = \begin{cases} 2|C_k|, & k \neq 0, \\ C_0, & k = 0; \end{cases}$

все коэффициенты A_k вещественны.

Ещё одна форма ряда Фурье для вещественных сигналов основана на разложении по тригонометрическим функциям, образующим ортогональный базис

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}\left(a_k \cos\frac{2\pi}{T}kt + b_k \sin\frac{2\pi}{T}kt\right),$$

со спектральными коэффициентами

$$a_k = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos\left(\frac{2\pi}{T}kt\right)dt, \quad k = \overline{0,\infty},$$

$$b_k = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\sin\left(\frac{2\pi}{T}kt\right)dt, \quad k = \overline{1,\infty}.$$

Сложим с учетом коэффициентов две функции этого базиса, имеющие одинаковую частоту, и воспользуемся формулой Эйлера:

$$a_k \cos\frac{2\pi}{T}kt + b_k \sin\frac{2\pi}{T}kt = a_k \frac{e^{j\frac{2\pi}{T}kt} + e^{-j\frac{2\pi}{T}kt}}{2} + b_k \frac{e^{j\frac{2\pi}{T}kt} - e^{-j\frac{2\pi}{T}kt}}{2j} =$$

$$= \frac{a_k - jb_k}{2} e^{j\frac{2\pi}{T}kt} + \frac{a_k + jb_k}{2} e^{-j\frac{2\pi}{T}kt}.$$

Сравнивая полученное выражение с выражениями (2.42), видим, что $C_k = \frac{a_k - jb_k}{2}$, а $C_{-k} = \frac{a_k + jb_k}{2}$, откуда следуют связи

между спектральными коэффициентами для различных форм ряда Фурье:

$$|C_k| = \frac{\sqrt{a_k^2 + b_k^2}}{2}, \quad C_0 = \frac{a_0}{2},$$

$$A_k = \sqrt{a_k^2 + b_k^2}, \quad A_0 = \frac{a_0}{2}, \quad \varphi_k = -\operatorname{arctg}\frac{b_k}{a_k}.$$

Очевидно, если сигнал представляет собой четную функцию, то все синусоидальные компоненты ряда равны 0; аналогично, все косинусоидальные компоненты равны нулю, если сигнал – нечетная функция (при этом равна нулю и постоянная составляющая).

Пример 2.19. Периодическая с периодом T последовательность прямоугольных импульсов амплитуды U и длительности $\tau_и$ показана на рис. 2.23.

Спектральные коэффициенты комплексного ряда Фурье находятся как

$$C_k = \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-j\frac{2\pi}{T}kt}dt = \frac{1}{T}\int_{-\tau_и/2}^{\tau_и/2} U\cos\left(\frac{2\pi}{T}kt\right)dt = \frac{U\tau_и}{T}\frac{\sin k\Omega\tau_и/2}{k\Omega\tau_и/2},$$

где введено обозначение круговой частоты $\Omega = \frac{2\pi}{T} = 2\pi F$. Таким образом, диаграмма амплитудного спектра сигнала, показанная на рис. 2.24, имеет огибающую в форме известной функции вида $\frac{\sin x}{x}$. Заметим, что все коэффициенты C_k оказались вещественными, так что фазовый спектр равен нулю для всех k. Значение постоянной составляющей сигнала $C_0 = \frac{U\tau_и}{T} = U/q$, где $q = T/\tau_и$ – параметр импульсной последовательности, называемый *скважностью*.

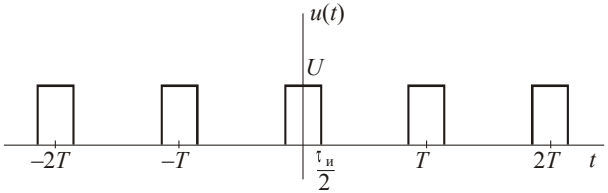


Рис. 2.23. Периодическая последовательность прямоугольных импульсов

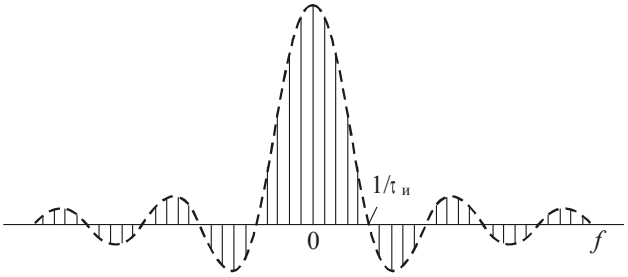


Рис. 2.24. Спектр периодической последовательности прямоугольных импульсов

Последовательности прямоугольных импульсов широко применяются в радиотехнике и связи в качестве моделей реальных сигналов, поэтому спектр данного сигнала достоин более внимательного рассмотрения. Прежде всего, обратим внимание, что огибающая спектра впервые пересекает ось частот при $\frac{\Omega\tau_{и}}{2}=\pi$, т.е. при $\Omega=\frac{2\pi}{\tau_{и}}$ или при $f=1/\tau_{и}$. Таким образом, численное значение скважности прямоугольной импульсной последовательности показывает, во сколько раз полуширина *главного лепестка огибающей* спектра больше шага $F=1/T$ следования по оси частот спектральных составляющих.

Конечная сумма ряда Фурье может служить аппроксимацией сигнала. На рис. 2.25 показаны конечные суммы комплексного ряда Фурье периодической последовательности прямоугольных импульсов при числе слагаемых 5, 11 и 25. Видно, что аппроксимация становится точнее с ростом количества слагаемых. Ошибка аппроксимации при удержании в сумме $2N+1$ слагаемых (от $-N$-го до N-го) может быть найдена на основе равенства Парсеваля как

$$\|\varepsilon\|^2 = T\sum_{k=-\infty}^{-N-1}|C_k|^2 + T\sum_{k=N+1}^{\infty}|C_k|^2 = T\sum_{k=-\infty}^{\infty}|C_k|^2 - T\sum_{k=-N}^{N}|C_k|^2 \,. \blacktriangleleft$$

При увеличении числа слагаемых ряда Фурье ошибка аппроксимации периодического сигнала стремится к нулю по норме пространства $L_2(T)$, т.е.

$$\|\varepsilon\|^2 = \int_{-T/2}^{T/2}\left[x(t)-\breve{x}(t)\right]^2 dt \to 0 \,. \qquad (2.44)$$

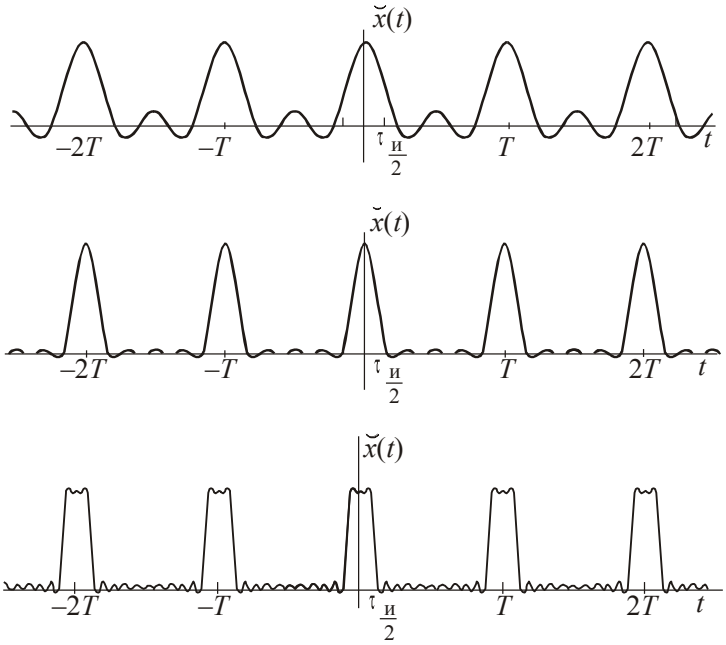


Рис. 2.25. Аппроксимация периодической последовательности, показанной на рис. 2.23, суммой 5, 11 и 25 членов ряда Фурье

Здесь $\breve{x}(t)$ – аппроксимация сигнала $x(t)$. При этом *максимальное* значение разности стремится не к нулю, а к конечной величине (порядка 9 % от амплитуды импульса). Это явление известно как *явление Гиббса*[32]. Причиной явления Гиббса является *неравномерная* сходимость[33] ряда Фурье для разрывных функций. При равномерной сходимости предел последовательности *непрерывных* функций, каковыми являются конечные суммы ряда Фурье, сам должен быть непрерывной функцией; в рассматриваемом же примере пределом является разрывная (скачкообразная) функция. В некоторых практических задачах таких, как синтез цифровых фильтров, явление Гиббса нежелательно; существуют методы уменьшения гиббсовских пульсаций (*осцилляций*), основанные на коррекции коэффициентов ряда Фурье [5].

---

[32] Джосайя Уиллард Гиббс (1839–1903) – выдающийся американский физик, один из основателей статистической физики.

[33] Сходимость, описываемая выражением (2.44), называется среднеквадратической.

2.10.2. СВОЙСТВА ПРЕОБРАЗОВАНИЯ ФУРЬЕ

Ряд Фурье представляет собой удобный инструмент анализа сигналов, заданных на конечном временном интервале, а также периодических колебаний, так как позволяет заменить *несчетное* множество (континуум) значений аналогового сигнала *счетным* множеством спектральных коэффициентов. Базис Фурье $\left\{ e^{j\frac{2\pi}{T}kt}, k = \overline{-\infty, \infty} \right\}$ полон в пространстве $L_2(T)$, поэтому любой сигнал из $L_2(T)$ можно сколь угодно точно аппроксимировать *конечной* суммой ряда Фурье, выбрав достаточно большое число слагаемых. Среди всех полных в $L_2(T)$ базисов базис Фурье имеет то преимущество, что он составлен из функций, собственных для любого ЛИС-оператора. Это максимально упрощает анализ воздействия периодических сигналов на ЛИС-цепи.

Для пространства $L_2(-\infty, +\infty)$ сигналов ограниченной энергии, заданных на всей временной оси, базис $\left\{ e^{j\frac{2\pi}{T}kt}, k = \overline{-\infty, \infty} \right\}$ не является полным ни при каком T и, следовательно, непригоден для представления сигналов, так как ошибку аппроксимации нельзя в общем случае сделать произвольно малой путем учета достаточного числа слагаемых ряда Фурье. В самом деле, если сигнал имеет бесконечную длительность и конечную энергию, т.е. принадлежит пространству $L_2(-\infty,+\infty)$, то он должен убывать при стремлении $t \to \pm\infty$, и притом достаточно быстро. При любом выборе T ряд Фурье для такого сигнала определяет периодическую функцию, которая может совпадать с заданным сигналом только на интервале длительности T, а за его пределами неизбежно будет отличаться от него. Более того, периодическая функция всегда имеет бесконечную энергию, поэтому и ошибка аппроксимации при любом T будет иметь бесконечную норму. Это и означает неполноту счетного базиса Фурье в $L_2(-\infty,+\infty)$[34]. Итак, единственным способом использовать комплексные экспоненты в качестве базисных функ-

[34] Напомним, что в $L_2(-\infty,+\infty)$ существуют полные ортонормальные *счетные* базисы (например, базис, составленный из функций Эрмита [3]), но они, к сожалению, не являются собственными для ЛИС-цепей.

ций является переход к непрерывному представлению сигналов из $L_2(-\infty,+\infty)$ *интегралом* Фурье

$$x(t)=\int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df\,, \tag{2.45}$$

где

$$X(f)=\int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \tag{2.46}$$

– спектральная плотность.

Между рядом и интегралом Фурье имеется тесная связь. Рассмотрим непериодический сигнал $x(t)$ конечной длительности τ_c. (Функция, равная нулю всюду за пределами интервала конечной длины, называемого *носителем* функции, называется *финитной*.) Спектральная плотность $X(f)$ сигнала $x(t)$ определяется выражением прямого преобразования Фурье (2.46). Повторение финитного сигнала $x(t)$ с периодом T, большим, чем длительность τ_c, дает периодический сигнал $\tilde{x}(t)=\sum_{k=-\infty}^{\infty} x(t+kT)$, который в силу своей периодичности может быть представлен рядом Фурье со спектральными коэффициентами, определяемыми выражением (2.41). Сравнивая выражения (2.46) и (2.41) и учитывая, что интеграл в бесконечных пределах от финитной функции равен интегралу по интервалу, содержащему носитель функции, можно записать равенство

$$C_k=\frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-j\frac{2\pi}{T}kt}dt=\frac{1}{T}X\left(\frac{k}{T}\right). \tag{2.47}$$

Таким образом, спектральная плотность импульсного сигнала имеет форму *огибающей* спектральных коэффициентов ряда Фурье периодической последовательности, образованной повторением данного импульсного сигнала с произвольным периодом. Заметим, что с ростом периода повторения спектральные составляющие следуют друг за другом по оси частот все более плотно. Непериодический сигнал представляет собой предельный случай периодического при $T\to\infty$, поэтому можно считать (нестрого!), что спектральная плотность – это «сплошная» совокупность спектральных коэффициентов. Следует, однако, иметь в виду, что «амплитуда» каждой спектральной составляющей при этом также стремится к

нулю, в отличие от спектральной плотности. Кроме того, не следует забывать, что для сигнала $x(t)$, имеющего размерность напряжения, спектральная плотность $X(f)$ имеет размерность [Вольт/Герц], в то время как единицей измерения коэффициентов ряда Фурье (2.41) является Вольт[35]. Поэтому спектр периодического сигнала и спектральная плотность финитного сигнала – два различных объекта. Тем не менее формальное сходство, выражаемое формулой (2.47), можно использовать, например, для расчета спектра последовательности, полученной периодическим повторением финитного сигнала с известной спектральной плотностью.

Рассмотрим основные свойства преобразования Фурье, которые полезно знать при практическом его использовании. Для краткости будем использовать обозначение $x(t) \Leftrightarrow X(f)$ для функций времени и частоты, связанных парой преобразований Фурье.

1. *Линейность*

$$\sum_k \alpha_k x_k(t) \Leftrightarrow \sum_k \alpha_k X_k(f).$$

2. *Дуальность* (*частотно-временная симметрия*)

$$x(f) \Leftrightarrow X(-t), \qquad (2.48)$$

где $x(\cdot)$ понимается как спектральная плотность временной функции $X(\cdot)$.

Читателю предлагается доказать это свойство в качестве упражнения.

3. *Теорема сдвига*

Рассмотрим сигнал $x_\tau(t) = x(t-\tau)$. Его спектральная плотность

$$X_\tau(f) = \int_{-\infty}^{\infty} x(t-\tau)e^{-j\cdot 2\pi ft}dt = \int_{-\infty}^{\infty} x(\theta)e^{-j\cdot 2\pi f(\theta+\tau)}d\theta = e^{-j\cdot 2\pi f\tau}X(f).$$

Таким образом, $x(t-\tau) \Leftrightarrow e^{-j\cdot 2\pi f\tau}X(f)$.

4. *Теорема изменения масштаба*

Рассмотрим сигнал $x_m(t) = x(mt)$, представляющий собой сигнал $x(t)$, сжатый по оси времени в m раз, $m > 0$. Его спектральная плотность

$$X_m(f) = \int_{-\infty}^{\infty} x(mt)e^{-j\cdot 2\pi ft}dt = \int_{-\infty}^{\infty} x(\theta)e^{-j\cdot 2\pi f\frac{\theta}{m}}\frac{d\theta}{m} = \frac{1}{m}X\left(\frac{f}{m}\right). \quad (2.49)$$

[35] Подразумевается, что функция $\exp(j\cdot 2\pi ft)$ физической размерности не имеет.

Теперь положим, что множитель $m = -\mu < 0$. Тогда

$$X_m(f) = \int_{-\infty}^{\infty} x(-\mu t) e^{-j\cdot 2\pi f t} dt = \int_{\infty}^{-\infty} x(\theta) e^{-j\cdot 2\pi f \frac{\theta}{-\mu}} \frac{d\theta}{-\mu} =$$

$$= \frac{1}{\mu} \int_{-\infty}^{\infty} x(\theta) e^{-j\cdot 2\pi \frac{f}{-\mu}\theta} d\theta = \frac{1}{\mu} X\left(\frac{f}{-\mu}\right). \qquad (2.50)$$

Итак, объединяя (2.49) и (2.50), можно окончательно записать

$$x(mt) \Leftrightarrow \left|\frac{1}{m}\right| X\left(\frac{f}{m}\right).$$

5. *Теорема дифференцирования*

Обозначим через $x_d(t) = dx(t)/dt$ производную по времени сигнала $x(t)$. Спектральная плотность производной равна

$$X_d(f) = \int_{-\infty}^{\infty} \frac{dx(t)}{dt} e^{-j\cdot 2\pi f t} dt = x(t) e^{-j\cdot 2\pi f t}\Big|_{-\infty}^{+\infty} + j\cdot 2\pi f \int_{-\infty}^{\infty} x(t) e^{-j\cdot 2\pi f t} dt .$$

Здесь использована формула интегрирования по частям. Первое слагаемое полученного выражения равно нулю, так как сигнал $x(t)$ в силу принадлежности $L_2(-\infty, +\infty)$, т.е. ограниченности энергии, стремится к нулю при $t \to \pm\infty$. Таким образом, $\frac{dx(t)}{dt} \Leftrightarrow j\cdot 2\pi f \cdot X(f)$.

6. *Теорема интегрирования*

Обратной к теореме дифференцирования является теорема интегрирования

$$\int_{-\infty}^{t} x(t) dt \Leftrightarrow \frac{1}{j\cdot 2\pi f} X(f) + \frac{X(0)\delta(f)}{2}. \qquad (2.51)$$

7. *Теорема модуляции*

Под модуляцией здесь подразумевается умножение сигнала $x(t)$ на комплексную экспоненциальную функцию $e^{j\cdot 2\pi f_0 t}$:

$$\int_{-\infty}^{\infty} x(t) e^{j\cdot 2\pi f_0 t} e^{-j\cdot 2\pi f t} dt = \int_{-\infty}^{\infty} x(t) e^{-j\cdot 2\pi (f - f_0) t} dt = X(f - f_0),$$

так что $x(t) e^{j\cdot 2\pi f_0 t} \Leftrightarrow X(f - f_0)$.

8. *Теорема свёртки*

$$x(t) * y(t) \Leftrightarrow X(f)Y(f)$$

была фактически доказана в разд. 2.9.

9. *Теорема умножения*

$$x(t)y(t) \Leftrightarrow X(f) * Y(f)$$

справедлива в силу теоремы свёртки и свойства дуальности преобразования Фурье.

10. *Теорема сопряжения*

Если комплексному сигналу $x(t)$ соответствует спектральная плотность $X(f)$, то для комплексно сопряженного сигнала справедливо соответствие $x^*(t) \Leftrightarrow X^*(-f)$:

$$\int_{-\infty}^{\infty} x^*(t)e^{-j\cdot 2\pi ft}dt = \left(\int_{-\infty}^{\infty} x(t)e^{-j\cdot 2\pi(-f)t}dt \right)^* = X^*(-f).$$

11. *Теорема обращения*

Обращение сигнала означает перемену знака аргумента (времени). Обозначим сигнал $x(t)$, обращенный во времени, $x\_(t) = x(-t)$. Его спектральная плотность:

$$X\_(f) = \int_{-\infty}^{\infty} x(-t)e^{-j\cdot 2\pi ft}dt = \int_{\infty}^{-\infty} x(\theta)e^{j\cdot 2\pi f\theta}(-d\theta) =$$

$$= \int_{-\infty}^{\infty} x(\theta)e^{-j\cdot 2\pi(-f)\theta}d\theta = X(-f),$$

так что обращение временно́й оси в обычном временно́м описании сигнала эквивалентно обращению оси частотной в его спектральном представлении: $x(-t) \Leftrightarrow X(-f)$.

Рассмотренные свойства преобразования Фурье справедливы для произвольных комплексных сигналов. На практике часто имеются дополнительные сведения о сигнале, которые позволяют упростить решение задачи спектрального анализа с учётом частных свойств спектральных плотностей.

Например, предположение о том, что сигнал $x(t)$ является *вещественным*, приводит к свойству *сопряженной симметрии* спектральной плотности:

$$X(f) = X^*(-f),$$

или, что равносильно, $|X(f)| = |X(-f)|$ и $\arg X(f) = -\arg X(-f)$. В самом деле,

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt = \left(\int_{-\infty}^{\infty} x(t)e^{j2\pi ft}dt \right)^* = \left(\int_{-\infty}^{\infty} x(t)e^{-j2\pi(-f)t}dt \right)^*.$$

Это обстоятельство следует учитывать при решении практических задач, так как в большинстве случаев рассматриваются именно вещественные сигналы. В частности, такая симметрия спектра (спектральной плотности) используется в технике связи: для уменьшения требуемой пропускной способности каналов связи применяются амплитудно-модулированные сигналы с одной боковой полосой (ОБП-сигналы).

Если сигнал является *вещественным и четным*, то его спектральная плотность также *вещественна и чётна*:

$$X(f) = X(-f).$$

Это утверждение следует из того, что обращение во времени не изменяет вещественного чётного сигнала, а следовательно, не влияет и на его спектральную плотность, которая должна, таким образом, быть инвариантной к обращению частоты, т.е. вещественной и чётной.

Если сигнал является *вещественным и нечетным*, то его спектральная плотность – *мнимая и нечетная*:

$$X(f) = -X(-f).$$

Действительно, обращение во времени *изменяет знак* нечётного сигнала, следовательно, его спектральная плотность также должна при обращении частоты лишь менять знак, но, поскольку спектральная плотность вещественного сигнала сопряженно-симметрична, отсюда следует, что её вещественная часть равна нулю, т.е. спектральная плотность является мнимой.

Спектральная плотность сигнала $e^{j\cdot 2\pi f_0 t}$ как «обычная функция» не существует, так как $e^{j\cdot 2\pi f_0 t}$ не принадлежит пространству $L_2(-\infty,\infty)$. В то же время решение многих задач упрощается, если все же определить спектральную плотность комплексной экспоненты в терминах теории обобщенных функций. Отыскание спектральной плотности сигнала $e^{j\cdot 2\pi f_0 t}$ сводится к нахождению прямого преобразования Фурье отрезка функции $e^{j\cdot 2\pi f_0 t}$ длительности

τ и предельному переходу при $\tau \to \infty$. В том, что спектральная плотность сигнала $e^{j\cdot 2\pi f_0 t}$ представляет собой δ-функцию (является *сингулярной*)[36], легко убедиться, если найти сигнал, соответствующий спектральной плотности $\delta(f - f_0)$, через обратное преобразование Фурье:

$$\int_{-\infty}^{\infty} \delta(f - f_0) e^{j\cdot 2\pi f t} df = e^{j\cdot 2\pi f_0 t}.$$

Спектральные плотности гармонических сигналов $\cos(2\pi f_0 t)$ и $\sin(2\pi f_0 t)$ легко находятся с учетом формул Эйлера:

$$\cos(2\pi f_0 t) \Leftrightarrow \frac{1}{2}\left[\delta(f - f_0) + \delta(f + f_0)\right],$$

$$\sin(2\pi f_0 t) \Leftrightarrow \frac{1}{2j}\left[\delta(f - f_0) - \delta(f + f_0)\right].$$

Пример 2.20. *Балансно-модулированное* колебание (см. разд. 5) может быть получено путем перемножения модулирующего сигнала $x(t)$ и несущего гармонического колебания $\cos(2\pi f_0 t)$. Спектральную плотность балансно-модулированного сигнала можно найти, воспользовавшись теоремой умножения с учетом вида спектральной плотности косинусоидального колебания:

$$x(t)\cos(2\pi f_0 t) \Leftrightarrow \frac{X(f + f_0)}{2} + \frac{X(f - f_0)}{2}.$$

Тот же результат можно получить на основе теоремы модуляции и свойства линейности преобразования Фурье. ◄

Во многих задачах одновременно присутствуют периодические и непериодические сигналы. Для того чтобы можно было пользоваться общим математическим аппаратом интеграла Фурье, найдем спектральную плотность T-периодического сигнала, который можно записать в виде ряда Фурье

$$x(t) = \sum_{k=-\infty}^{\infty} C_k e^{j\frac{2\pi}{T}kt}.$$

[36] От англ. single – единственный; в названии отражается тот факт, что вся спектральная «масса» сосредоточена в одной точке частотной оси.

Учитывая линейность преобразования Фурье и зная спектральную плотность комплексной экспоненциальной функции, запишем спектральную плотность в виде

$$X(f) = \sum_{k=-\infty}^{\infty} C_k \delta\left(f - k\frac{2\pi}{T} \right). \qquad (2.52)$$

Таким образом, спектральная плотность периодического сигнала сингулярна, т.е. состоит из δ-функций, сосредоточенных на частотах, кратных частоте повторения сигнала.

2.10.3. КОРРЕЛЯЦИОННО-СПЕКТРАЛЬНЫЕ ХАРАКТЕРИСТИКИ ДЕТЕРМИНИРОВАННЫХ СИГНАЛОВ

Согласно обобщенной формуле Рэлея (2.20) скалярное произведение двух детерминированных сигналов может быть найдено как во временной, так и в частотной области:

$$(x, y) = \int_{-\infty}^{\infty} x(t) y^*(t) dt = \int_{-\infty}^{\infty} X(f) Y^*(f) df \ .$$

Подынтегральное выражение правой части этого равенства

$$W_{xy}(f) = X(f) Y^*(f)$$

называется *взаимной спектральной плотностью* сигналов $x(t)$ и $y(t)$.

В частном случае при $x(t) = y(t)$ взаимная спектральная плотность превращается в *энергетический спектр* сигнала

$$W_x(f) = |X(f)|^2 .$$

Смысл энергетического спектра выясняется, если выразить энергию сигнала через скалярное произведение

$$E_x = (x, x) = \int_{-\infty}^{\infty} X(f) X^*(f) df = \int_{-\infty}^{\infty} W_x(f) df \ .$$

Таким образом, функция $W_x(f)$ описывает распределение энергии сигнала по частотной оси (поэтому правильнее было бы называть ее *спектральной плотностью энергии*). Заметим, что

спектральная плотность сигнала $X(f)$ является комплексной функцией, аргумент которой теряется при переходе от $X(f)$ к $W_x(f)$, поэтому в общем случае *сигнал нельзя восстановить по его энергетическому спектру*[37].

Взаимная спектральная плотность характеризует сходство сигналов в том смысле, что интеграл от нее равен их скалярному произведению. В частности, для ортогональных сигналов взаимная спектральная плотность такова, что при интегрировании дает 0, т.е. различные частотные составляющие взаимной спектральной плотности ортогональных сигналов при интегрировании компенсируют друг друга.

Для энергетического спектра и взаимной спектральной плотности, как функций частоты, можно определить при помощи обратного преобразования Фурье (если оно существует) соответствующие временные функции. Запишем вначале обратное преобразование Фурье для взаимной спектральной плотности

$$B_{xy}(\tau) = \int_{-\infty}^{\infty} W_{xy}(f) e^{j\cdot 2\pi f\tau} df = \int_{-\infty}^{\infty} X(f) Y^*(f) e^{j\cdot 2\pi f\tau} df .$$

Обозначим $Y^*(f)e^{j\cdot 2\pi f\tau} = Y_\tau^*(f)$, где $Y_\tau(f) = Y(f)e^{-j\cdot 2\pi f\tau}$ – спектральная плотность сигнала $y_\tau(t) = y(t-\tau)$, равного сигналу $y(t)$, задержанному на величину τ. Тогда

$$B_{xy}(\tau) = \int_{-\infty}^{\infty} X(f) Y_\tau^*(f) df = \int_{-\infty}^{\infty} x(t) y^*(t-\tau) dt = (x, y_\tau) .$$

Полученная функция характеризует сходство сигнала $x(t)$ и сигнала $y_\tau(t) = y(t-\tau)$ в зависимости от значения сдвига, и называется *взаимно корреляционной функцией* (ВКФ) детерминированных сигналов $x(t)$ и $y(t)$.

Аналогично функция

$$B_x(\tau) = \int_{-\infty}^{\infty} W_x(f) e^{j\cdot 2\pi f\tau} df = \int_{-\infty}^{\infty} X(f) X^*(f) e^{j\cdot 2\pi f\tau} df ,$$

[37] Заметим, что в частных случаях это можно сделать, если имеются дополнительные сведения, например о том, что $X(f)$ есть вещественная четная функция.

характеризующая сходство сигнала $x(t)$ и его задержанной копии $x_\tau(t) = x(t-\tau)$

$$B_x(\tau) = \int_{-\infty}^{\infty} X(f)X_\tau^*(f)df = \int_{-\infty}^{\infty} x(t)x^*(t-\tau)dt = (x, x_\tau), \quad (2.53)$$

называется *автокорреляционной* функцией (АКФ) детерминированного сигнала.

Автокорреляционная функция обладает некоторыми свойствами, которые важно знать для ее правильного использования.

1. Автокорреляционная функция достигает максимума при $\tau = 0$ и равна при этом значении аргумента энергии сигнала:

$$B_x(0) = \max_\tau B_x(\tau) = E_x. \quad (2.54)$$

Доказать это свойство легко при помощи неравенства Шварца.

2. Автокорреляционная функция обладает свойством сопряженной симметрии:

$$B_x(\tau) = \int_{-\infty}^{\infty} x(t)x^*(t-\tau)dt = \int_{-\infty}^{\infty} x(\theta+\tau)x^*(\theta)d\theta =$$

$$= \left[\int_{-\infty}^{\infty} x(\theta)x^*(\theta+\tau)d\theta\right]^* = B_x^*(-\tau).$$

В частности, АКФ вещественного сигнала – четная функция.

Взаимно корреляционная функция указанными свойствами не обладает: например, для ортогональных сигналов она равна нулю при нулевом сдвиге; в остальном ее форма определяется формами обоих сигналов.

Введенные функции играют очень важную роль, в частности, при выборе сигналов для *синхронизации* систем связи, что иллюстрируется следующим примером.

Пример 2.21. Многие системы связи нуждаются в синхронизации, т.е. в одновременном начале (с точки зрения устройств передачи и приема) интервалов времени, в течение которых передаются и принимаются сигналы[38]. Для того чтобы синхронизировать приемник, необходимо время от времени передавать некоторый спе-

[38] Говоря здесь об одновременности, мы для простоты не учитываем время передачи сигнала по каналу связи.

циальный сигнал, играющий роль временно́й метки, временно́е положение которой приемник должен измерить, чтобы «сверить часы». Измерение временно́го положения синхронизирующего сигнала производится при помощи многоканального устройства, структурная схема которого показана на рис. 2.26.

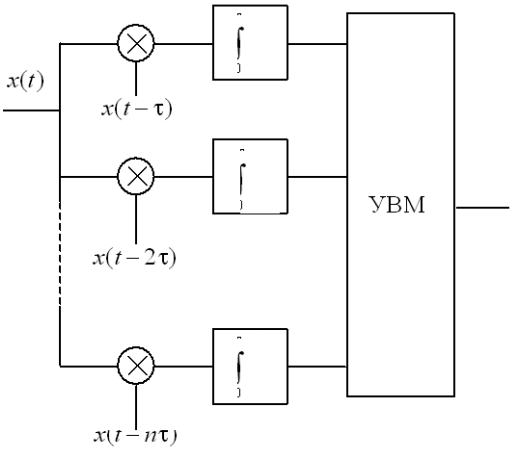


Рис. 2.26. Измерение временно́го положения синхронизирующего сигнала

В каждом канале вырабатывается значение взаимно корреляционной функции принятого сигнала и одного из n опорных сигналов; каждый из опорных сигналов представляет собой копию входного сигнала, задержанную на величину, кратную некоторому шагу τ. Шаг выбирается исходя из требуемой точности измерения задержки, при этом для достижения большей точности шаг необходимо уменьшать. Число каналов n определяется диапазоном измеряемых задержек и величиной шага τ. Легко видеть, что величины на выходах интеграторов представляют собой отсчеты АКФ принимаемого сигнала, взятые с интервалом τ.

С учетом 1-го свойства АКФ, если задержка входного сигнала составляет $k\tau$, то на выходе k-го канала значение ВКФ достигнет максимума и будет равно энергии сигнала. Устройство выбора максимума УВМ вырабатывает оценку $\tilde{\tau}$ по номеру канала, на выходе которого имеет место максимальное значение интеграла

$$\tilde{\tau} = \tau \cdot \arg\max_{k}(x, x_{k\tau}) \quad \blacktriangleleft$$

Пример 2.22. Автокорреляционная функция прямоугольного импульса длительности $\tau_и$ имеет вид треугольника (рис. 2.27).

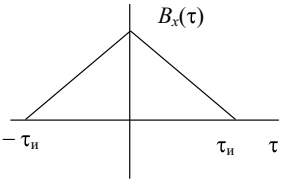
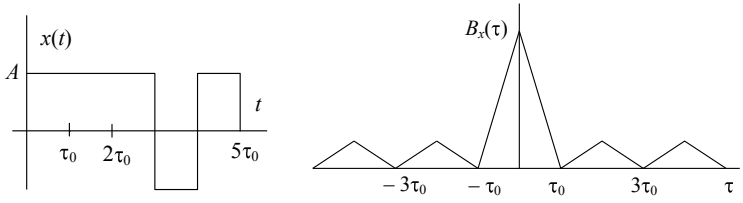


Рис. 2.27. АКФ прямоугольного импульса длительности $\tau_и$

Максимальное значение АКФ равно $A^2\tau_и$, где A – амплитудное (максимальное) значение импульса. ◄

Пример 2.23. С точки зрения точности синхронизации выгодно использовать сигнал, который имеет «острую» (игольчатую) АКФ, близкую по форме к δ-функции. Реальные сигналы с конечной шириной спектра к этому идеалу могут только приближаться. Одним из хороших приближений является *сигнал Баркера,* состоящий из N разнополярных прямоугольных элементарных импульсов; пример такого сигнала при $N = 5$ показан на рис. 2.28, *а*.

Отличительное свойство сигнала Баркера состоит в том, что его АКФ (рис. 2.28, *б*) имеет лепестковый вид, причем ширина каждого лепестка равна удвоенной длительности элементарного импульса, а уровни боковых лепестков в N раз меньше, чем уровень главного лепестка (равный, очевидно, $N\tau_0 A^2$). К сожалению, сигналы Баркера существуют только при $N \in \{2, 3, 4, 5, 7, 11, 13\}$. Таким образом, максимальное превышение главного лепестка над боковыми, которое определяет эффективность (помехоустойчивость) синхронизации[39], не может быть для сигналов Баркера больше, чем 13. Бóльшее превышение достигается для длинных последовательностей разнополярных прямоугольных импульсов, называемых *m*-последовательностями (для них, однако, уровни боковых лепестков могут быть лишь в $\sqrt{N}$ раз меньше главного). ◄



а *б*

Рис. 2.28. Сигнал Баркера ($N = 5$) и его АКФ

[39] Чем больше указанное превышение, тем меньше вероятность принять боковой лепесток за главный из-за шумовых выбросов.

2.11. ДИСКРЕТИЗАЦИЯ СИГНАЛОВ. ТЕОРЕМА ОТСЧЁТОВ

Коэффициенты обобщенного ряда Фурье, составляющие спектр некоторого сигнала относительно полного ортонормального базиса, вычисляются путем скалярного умножения этого сигнала на базисные функции. Существует счетный базис, для которого эта операция эквивалентна взятию *отсчётов мгновенных значений* аналогового сигнала через равные промежутки времени. Таким образом, аналоговый сигнал однозначно представляется дискретной последовательностью своих отсчётов. Такая *дискретизация* аналоговых сигналов имеет огромное значение для современной техники, так как является основой цифровой обработки сигналов.

Следует сразу же отметить, что, как и базис Фурье, упомянутый базис не обладает свойством полноты в пространстве $L_2(-\infty, +\infty)$, поэтому он непригоден для представления с заданной точностью *любых* аналоговых сигналов ограниченной энергии, однако для подпространства сигналов, спектральная плотность которых сосредоточена на конечном интервале $(-F_в, F_в)$ частотной оси[40], этот счетный базис полон.

Условие финитности спектра не является слишком обременительным, так как спектральные плотности всех сигналов из $L_2(-\infty, +\infty)$ при $f \to \pm\infty$ быстро убывают[41], поэтому их можно *с любой точностью* аппроксимировать финитными функциями. При выборе конкретного базиса в качестве конечного интервала $(-F_в, F_в)$ принимается так называемая эффективная ширина спектра сигнала. *Эффективной* шириной спектра можно считать ширину частотного интервала, в котором сосредоточена заданная доля (например, 99 %) всей энергии сигнала. Обычно на практике перед дискретизацией сознательно ограничивают ширину спектра сигнала путем его предварительной фильтрации, так как это уменьшает ошибку восстановления аналогового сигнала.

Возможность представления аналогового сигнала последовательностью его отсчетов и условия применимости такого представления устанавливаются *теоремой отсчетов.* Приведенное

[40] Напомним, что такие функции называются финитными.
[41] См. по этому поводу п. 2.10.2.

ниже доказательство теоремы отсчетов принадлежит В.А. Котельникову[42].

Рассмотрим произвольный сигнал $x(t)$, спектральная плотность которого $X(f)$ равна нулю вне конечного интервала $(-F_{\text{в}}, F_{\text{в}})$ частотной оси. Выразим функцию спектральной плотности $X(f)$ в виде ряда Фурье

$$X(f) = \sum_{k=-\infty}^{\infty} C_k e^{j\frac{2\pi}{2F_{\text{в}}}kf}, \qquad (2.55)$$

вполне аналогичного комплексному ряду Фурье (2.39), представляющему временну́ю функцию на интервале $(-T/2, T/2)$, с той очевидной разницей, что базисные функции здесь зависят не от t, а от f. Очевидно, коэффициенты ряда находятся как

$$C_k = \frac{1}{2F_{\text{в}}} \int_{-F_{\text{в}}}^{F_{\text{в}}} X(f) e^{-j\frac{2\pi}{2F_{\text{в}}}kf} df, \quad k = \overline{-\infty, \infty}.$$

Выразим сигнал $x(t)$ через его обратное преобразование Фурье, подставляя в качестве спектральной плотности её представление рядом Фурье (2.55)

$$x(t) = \int_{-F_{\text{в}}}^{F_{\text{в}}} X(f) e^{j\cdot 2\pi ft} df = \sum_{k=-\infty}^{\infty} C_k \int_{-F_{\text{в}}}^{F_{\text{в}}} e^{j\cdot 2\pi f\left(t + k\frac{1}{2F_{\text{в}}}\right)} df =$$

$$= \sum_{k=-\infty}^{\infty} C_k \int_{-F_{\text{в}}}^{F_{\text{в}}} \cos\left[2\pi f\left(t + k\frac{1}{2F_{\text{в}}}\right)\right] df =$$

$$= \sum_{k=-\infty}^{\infty} C_k \cdot 2F_{\text{в}} \frac{\sin\left[2\pi F_{\text{в}}\left(t + k\frac{1}{2F_{\text{в}}}\right)\right]}{2\pi F_{\text{в}}\left(t + k\frac{1}{2F_{\text{в}}}\right)}. \qquad (2.56)$$

[42] Владимир Александрович Котельников (1908–2005) – выдающийся русский инженер и математик; доказал теорему отсчетов в 1933 г. Известны также варианты доказательства теоремы отсчётов, связанные с именами Э. Уиттекера (1916), Х. Найквиста (1928), Д. Габора (1946), К. Шеннона (1948).

Заметим, что

$$C_k = \frac{1}{2F_\text{в}} \int\limits_{-F_\text{в}}^{F_\text{в}} X(f) e^{-j\frac{2\pi}{2F_\text{в}}kf} df =$$

$$= \frac{1}{2F_\text{в}} \int\limits_{-F_\text{в}}^{F_\text{в}} X(f) e^{j\cdot 2\pi ft} df \Big|_{t=-k\frac{1}{2F_\text{в}}} = \frac{x\left(-k\dfrac{1}{2F_\text{в}}\right)}{2F_\text{в}}. \qquad (2.57)$$

Подставив (2.57) в (2.56) и введя обозначение для интервала (шага) дискретизации $T_d = \dfrac{1}{2F_\text{в}}$, запишем сигнал в виде ряда

$$x(t) = \sum_{k=-\infty}^{\infty} x(-kT_d) \frac{\sin\left[\dfrac{\pi}{T_d}(t+kT_d)\right]}{\dfrac{\pi}{T_d}(t+kT_d)} = \sum_{n=-\infty}^{\infty} x(nT_d) \frac{\sin\left[\dfrac{\pi}{T_d}(t-nT_d)\right]}{\dfrac{\pi}{T_d}(t-nT_d)}, \qquad (2.58)$$

известного под названием *ряда Котельникова*. Коэффициенты $x(nT_d)$ этого ряда представляют собой *отсчеты* (мгновенные значения) аналогового сигнала $x(t)$, взятые через равные промежутки времени $T_d = \dfrac{1}{2F_\text{в}}$. Базисные функции ряда Котельникова получаются сдвигами на такие же промежутки времени единственной функции. Обозначим эту исходную функцию $\kappa_0(t) = \sin\left(\dfrac{\pi}{T_d}t\right) \Big/ \left(\dfrac{\pi}{T_d}t\right)$, тогда базис будет совокупностью $\left\{\kappa_n(t),\ n = \overline{-\infty, \infty}\right\}$, $\kappa_n(t) = \kappa_0(t - nT_d)$. Несколько базисных функций показаны на рис. 2.29. Этот ряд даёт *точное* представление (интерполяцию) значений сигнала $x(t)$ в любой точке временно́й оси по известным значениям сигнала в дискретном множестве её точек (узлов интерполяции). Следовательно, нет необходимости передавать или хранить *всё* непрерывное множество (континуум) значений аналогового сигнала с финитным спектром, достаточно передать (или зафиксировать на некото-

ром носителе) счетную последовательность его дискретных отсчетов $\left\{x[n]=x\left(nT_d\right), n=\overline{-\infty,\infty}\right\}$, по которым при необходимости сигнал может быть *точно* восстановлен.

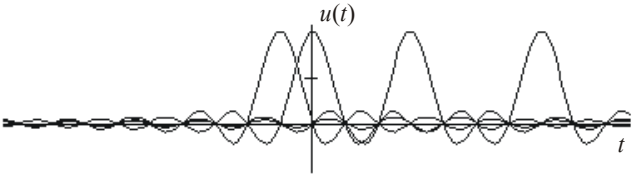


Рис. 2.29. Базисные функции ряда Котельникова при $n = -1, 0, 3, 7$

Чтобы выяснить свойства полученного базиса, найдем скалярное произведение пары базисных функций. Сравним выражения (2.56) и (2.55) для сигнала и его спектральной плотности. Очевидно, функция

$$2F_{\text{в}}\frac{\sin\left\lfloor 2\pi F_{\text{в}}\left(t+k\dfrac{1}{2F_{\text{в}}}\right)\right\rfloor}{2\pi F_{\text{в}}\left(t+k\dfrac{1}{2F_{\text{в}}}\right)}$$

имеет спектральную плотность, равную $e^{j\frac{2\pi}{2F_{\text{в}}}kf}$ на частотном интервале $\left(-F_{\text{в}},F_{\text{в}}\right)$, и нулю вне его. Тогда согласно обобщенной формуле Рэлея скалярное произведение k-й и m-й функций базиса Котельникова равно

$$(\kappa_k,\kappa_m)=\frac{1}{4F_{\text{в}}^{\,2}}\int_{-F_{\text{в}}}^{F_{\text{в}}}e^{j\frac{2\pi}{2F_{\text{в}}}(k-m)f}\,df=\frac{1}{2F_{\text{в}}}\delta_{km}\,.$$

Таким образом, базис Котельникова ортогонален, но не нормирован. По существу, *базис Котельникова во временной области – это базис Фурье в частотной области* [ср. (2.55) и (2.56)]. Поэтому свойства ортогональности и полноты одинаково справедливы для этих базисов.

Чтобы восстановить (интерполировать) аналоговый сигнал по последовательности его отсчетов, необходимо просуммировать все

базисные функции Котельникова при $n=\overline{-\infty,\infty}$ с весовыми коэффициентами, равными отсчётам $x(nT_d)$. Технически эту операцию можно в принципе осуществить, располагая ЛИС-цепью, имеющей импульсную характеристику, совпадающую с функцией Котельникова $\kappa_0(t)$, и подавая на вход этой цепи в моменты nT_d, $n=\overline{-\infty,\infty}$ воздействия в виде δ-функций с амплитудными множителями $x(nT_d)$, $n=\overline{-\infty,\infty}$. Следовательно, на вход такой цепи должен подаваться возбуждающий сигнал в виде T_d-периодической последовательности δ-функций, умноженных на отсчеты аналогового сигнала $v(t)=\sum\limits_{n=-\infty}^{\infty} x(nT_d)\delta(t-nT_d)$. Откликом цепи на воздействие $\delta(t-nT_d)$ является сдвинутая на nT_d импульсная характеристика $\kappa_n(t)=\kappa_0(t-nT_d)$. С учетом линейности и стационарности цепи очевидно, что отклик на воздействие $v(t)$ представляет собой правую часть выражения (2.58), поэтому на выходе цепи наблюдается восстановленный сигнал $x(t)$.

Учитывая, что δ-функция имеет нулевую длительность, можно представить возбуждающий сигнал в виде

$$v(t)=\sum_{n=-\infty}^{\infty} x(nT_d)\delta(t-nT_d)=x(t)\sum_{n=-\infty}^{\infty}\delta(t-nT_d)=x(t)\tilde{\delta}(t), \qquad (2.59)$$

где $\tilde{\delta}(t)=\sum\limits_{n=-\infty}^{\infty}\delta(t-nT_d)$ – периодическая последовательность δ-функций. Считая (нестрого!) δ-функцию «коротким импульсом», можно назвать сигнал $v(t)$ идеализированным сигналом с амплитудно-импульсной модуляцией (иАИМ-сигналом). Поскольку иАИМ-сигнал равен произведению (2.59), его спектральная плотность равна свертке спектральных плотностей сигналов $x(t)$ и $\tilde{\delta}(t)$. Найдем спектральную плотность последовательности $\tilde{\delta}(t)$. Для этого вначале запишем T_d-периодическую последовательность в виде ряда Фурье

$$\tilde{\delta}(t)=\sum_{n=-\infty}^{\infty} S_n e^{j\frac{2\pi}{T_d}nt},$$

коэффициенты которого, определяемые согласно (2.41), равны:

$$S_n = \frac{1}{T_d}\int\limits_{-T_d/2}^{T_d/2} \delta(t) e^{-j\frac{2\pi}{T_d}nt}\, dt = \frac{1}{T_d}.$$

Поэтому, учитывая выражение (2.52), запишем спектральную плотность последовательности $\tilde{\delta}(t)$ в виде

$$\tilde{\Delta}(f) = \frac{1}{T_d}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T_d}\right).$$

Теперь найдем спектральную плотность иАИМ-сигнала как свертку:

$$V(f) = \int\limits_{-\infty}^{\infty} X(\phi)\tilde{\Delta}(f-\phi)d\phi = \int\limits_{-\infty}^{\infty} X(\phi)\frac{1}{T_d}\sum_{n=-\infty}^{\infty}\delta\left(f-\phi-\frac{n}{T_d}\right)d\phi =$$

$$= \frac{1}{T_d}\sum_{n=-\infty}^{\infty}\int\limits_{-\infty}^{\infty} X(\phi)\delta\left(f-\phi-\frac{n}{T_d}\right)d\phi = \frac{1}{T_d}\sum_{n=-\infty}^{\infty} X\left(f-\frac{n}{T_d}\right) =$$

$$= \frac{1}{T_d}\sum_{n=-\infty}^{\infty} X(f - nF_d).$$

Таким образом, спектральная плотность иАИМ-сигнала представляет собой с учетом масштабного коэффициента $1/T_d$ сумму (суперпозицию) бесконечного множества копий спектральной плотности аналогового сигнала $x(t)$, отличающихся друг от друга сдвигами по оси частот, кратными частоте дискретизации (рис. 2.30).

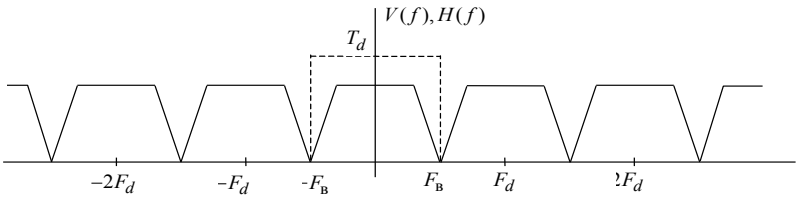


Рис. 2.30. Спектральная трактовка восстановления сигнала

Следовательно, восстановление аналогового сигнала интерполирующей цепью равносильно подавлению в спектре сигнала $v(t)$ всех спектральных составляющих, не принадлежащих интервалу $(-F_{\text{в}}, F_{\text{в}})$, поэтому интерполирующий фильтр должен иметь П-образную (прямоугольную) комплексную частотную характеристику (на рис. 2.29 показана штриховой линией):

$$K(f) = \begin{cases} T_d, & -F_{\text{в}} < f < F_{\text{в}}, \\ 0 & \text{в противном случае.} \end{cases} \qquad (2.60)$$

Легко убедиться, что идеальная интерполирующая цепь должна в таком случае иметь импульсную характеристику $h(t) = \sin\left[\frac{\pi}{T_d}t\right] \Big/ \left(\frac{\pi}{T_d}t\right)$, совпадающую с базисной функцией $\kappa_0(t)$. Таким образом, спектральный подход к восстановлению аналогового сигнала по его отсчетам приводит к тому же выводу, что и временной.

Идеальная интерполирующая цепь, строго говоря, нереализуема, так как её импульсная характеристика имеет бесконечно большую протяженность в области отрицательных времен. Однако в принципе можно построить цепь, сколь угодно точно её аппроксимирующую, правда, при этом восстановленный сигнал будет получаться с задержкой (тем бóльшей, чем выше требуемая точность аппроксимации)[43]. В самом деле, физическая реализуемость предполагает каузальность импульсной характеристики, т.е. выполнение условия (2.34). На рис. 2.31 показаны импульсная характеристика идеальной интерполирующей цепи $\kappa_0(t)$ (штриховая линия)

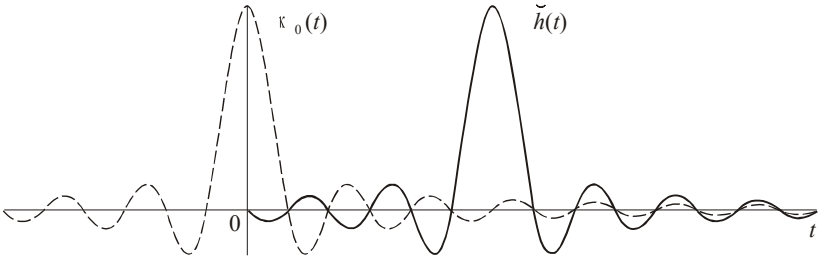


Рис. 2.31. Аппроксимация импульсной характеристики интерполирующей цепи

[43] Сложность цепи при этом также возрастает.

и аппроксимирующая ее функция $\breve{h}(t)$ (сплошная линия). Очевидно, при соблюдении условия каузальности повышение точности аппроксимации неизбежно приводит к сдвигу функции вправо, а значит, увеличивает задержку восстановленного сигнала.

Нереализуемым является и сигнал, описываемый выражением (2.59), так как в него входят δ-функции. На практике вместо них используются короткие[44] импульсы.

Необходимо отметить, что выражение (2.59), описывающее процесс *восстановления* аналогового сигнала по его отсчетам, иногда неправильно связывают с процессом *дискретизации* сигнала. На самом деле взятие (одиночного) отсчета аналогового сигнала в произвольный момент времени t_0 представляет собой *стробирование* и описывается выражением типа *свертки*

$$x(t_0) = \int_{-\infty}^{\infty} x(t)\delta(t - t_0)dt = \int_{-\infty}^{\infty} x(t)\delta(t_0 - t)dt\,, \qquad (2.61)$$

а не умножения, как в (2.59).

Реальное взятие отсчета производится устройством, в котором выполняется свертка аналогового сигнала не с δ-функцией, как в выражении (2.61), а с некоторым реальным импульсом $d(t)$. Этот импульс должен быть «похож» на δ-функцию, в частности, он должен быть коротким и интеграл от него должен быть равен 1. Для простоты примем в качестве $d(t)$ прямоугольный импульс длительности Δ и амплитуды $1/\Delta$. Свертке сигнала $x(t)$ с таким импульсом соответствует умножение спектральной плотности $X(f)$ на спектральную плотность прямоугольного импульса, имеющую, как известно, форму функции вида $\sin x / x$, поэтому при стробировании реальным импульсом конечной длины *всегда* происходит искажение спектра сигнала. Для уменьшения такого искажения необходимо стремиться к уменьшению длительности импульса $d(t)$, при этом форма импульса не играет заметной роли.

Все реальные сигналы имеют конечную длительность, поэтому спектральная плотность реального сигнала *не может быть финитной*. Нефинитность спектра сигнала приводит к тому, что «хвосты» копий спектральной плотности $X(f)$ при периодическом

[44] Здесь импульс считается коротким, если его длительность много меньше величины $1/F_{\text{в}}$.

повторении накладываются друг на друга и суммируются, приводя к *необратимому* искажению сигнала[45]. Применяемая *до дискретизации* фильтрация сигнала при помощи фильтра нижних частот с характеристикой, близкой к прямоугольной, подавляет эти «хвосты», уменьшая погрешность интерполяции.

Итак, *точному* восстановлению аналогового сигнала по последовательности его отсчётов препятствуют:

1) конечная длительность любого реального сигнала и, как следствие, бесконечная ширина его спектра;

2) конечная длительность реального стробирующего импульса и, как следствие, искажение формы спектра сигнала при дискретизации;

3) невозможность точно реализовать интерполирующий фильтр.

Несмотря на эти ограничения, дискретизация широко применяется на практике, в частности, она является необходимой частью *цифровой обработки сигналов*.

2.12. АНАЛИТИЧЕСКИЙ СИГНАЛ

В теории электрических цепей, как известно, широко используется метод представления гармонических колебаний комплексными векторами (метод комплексных амплитуд), состоящий в том, что гармоническое колебание $U_m \cos(2\pi f t + \phi)$ рассматривается как вещественная часть комплексной функции $U_m e^{j(2\pi f t + \phi)}$, которая изображается вектором в комплексной плоскости, вращающимся с постоянной угловой скоростью $\omega = 2\pi f$; при этом вектор имеет длину U_m и при $t = 0$ составляет с вещественной осью комплексной плоскости угол ϕ. Аналогичное представление можно ввести для сигнала произвольной формы

$$x(t) = \mathrm{Re}\{z(t)\},$$

где $z(t) = x(t) + j \cdot \hat{x}(t)$ – комплексное колебание (*аналитический сигнал*), мнимая часть которого $\hat{x}(t)$ должна однозначно определяться исходным сигналом $x(t)$.

[45] Это явление называется *подменой частот*; в англоязычной литературе используется название aliasing.

Заметим, что для гармонического колебания справедливо равенство

$$\cos(2\pi ft) = \frac{e^{j\cdot 2\pi ft} + e^{-j\cdot 2\pi ft}}{2},$$

т.е. переход от гармонического колебания $U_m \cos(2\pi ft + \phi)$ к его комплексному представлению $U_m e^{j(2\pi ft+\phi)}$ сводится к отбрасыванию составляющей с отрицательной частотой и умножению оставшегося слагаемого на 2. Поскольку колебание произвольной формы можно представить *суперпозицией гармонических* колебаний (в форме ряда или интеграла Фурье), для *любого* сигнала $x(t)$ переход к его комплексному представлению $z(t)$ должен сводиться к тем же операциям – подавлению спектральных составляющих с отрицательными частотами и удвоению остальных.

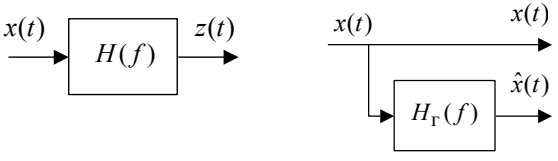
Таким образом, преобразование произвольного сигнала $x(t)$ в аналитический сигнал $z(t)$ эквивалентно его прохождению через ЛИС-цепь с комплексной частотной характеристикой

$$H(f) = \begin{cases} 2, & f \geq 0, \\ 0, & f < 0 \end{cases} \qquad (2.62)$$

(рис. 2.32, *а*). Поскольку вещественная часть аналитического сигнала есть исходный сигнал $x(t)$, это преобразование можно также представить схемой рис. 2.32, *б*.

Рассматривая спектральные представления исходного и аналитического сигналов $X(f)$ и $Z(f) = X(f) + j\hat{X}(f)$ совместно с выражением (2.62), видим, что для спектральных плотностей должны выполняться условия

$$\begin{cases} j\hat{X}(f) = X(f) & \text{при } f \geq 0, \\ j\hat{X}(f) = -X(f) & \text{при } f < 0, \end{cases} \qquad (2.63)$$



а *б*

Рис. 2.32. Преобразование вещественного сигнала в аналитический сигнал

которые можно переписать в форме

$$\begin{cases} \hat{X}(f) = -jX(f) & \text{при } f \geq 0, \\ \hat{X}(f) = jX(f) & \text{при } f < 0. \end{cases} \tag{2.64}$$

Таким образом, мнимая часть $\hat{x}(t)$ может быть получена воздействием сигнала $x(t)$ на фильтр–преобразователь *Гильберта* с характеристикой

$$H_{\text{г}}(f) = \begin{cases} -j, & f \geq 0, \\ j, & f < 0. \end{cases}$$

АЧХ такого фильтра постоянна и равна 1 при всех значениях частоты, а его ФЧХ равна $-\pi/2$ в области положительных частот и $\pi/2$ при отрицательных частотах (рис. 2.33).

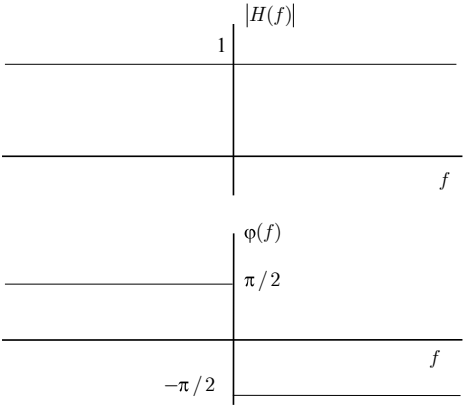


Рис. 2.33. АЧХ и ФЧХ преобразователя Гильберта

Найдем импульсную характеристику преобразователя Гильберта. Непосредственное определение обратного преобразования Фурье невозможно, так как КЧХ является неинтегрируемой функцией. Представим АЧХ пределом $|H(f)| = \lim\limits_{\varepsilon \to 0} e^{-\varepsilon |f|}$ функции, спадающей при увеличении модуля частоты экспоненциально с параметром ε, тогда

$$h_{\text{г}}(t) = \int\limits_{-\infty}^{\infty} H(f) e^{j \cdot 2\pi f t} df =$$

$$= \lim_{\varepsilon \to 0} \int_{-\infty}^{0} j e^{\varepsilon f} e^{j \cdot 2\pi f t} df + \lim_{\varepsilon \to 0} \int_{0}^{\infty} -j e^{-\varepsilon f} e^{j \cdot 2\pi f t} df =$$

$$= \lim_{\varepsilon \to 0} \left[j \int_{-\infty}^{0} e^{(\varepsilon + j \cdot 2\pi t) f} df - j \int_{0}^{\infty} e^{(-\varepsilon + j \cdot 2\pi t) f} df \right] =$$

$$= \lim_{\varepsilon \to 0} \left\{ j \left[\frac{1}{\varepsilon + j 2\pi t} e^{(\varepsilon + j \cdot 2\pi t) f} \Big|_{-\infty}^{0} - \frac{1}{-\varepsilon + j \cdot 2\pi t} e^{(-\varepsilon + j \cdot 2\pi t) f} \Big|_{0}^{\infty} \right] \right\} =$$

$$= \lim_{\varepsilon \to 0} \left\{ j \left[\frac{1}{\varepsilon + j \cdot 2\pi t} + \frac{1}{-\varepsilon + j \cdot 2\pi t} \right] \right\} = \frac{1}{\pi t}.$$

Полученная импульсная характеристика показана на рис. 2.34.

Поскольку $x(t)$ преобразуется в $\hat{x}(t)$ ЛИС-цепью, выходной сигнал можно записать как свертку

$$\hat{x}(t) = x(t) * h_{\Gamma}(t) = \int_{-\infty}^{\infty} x(\tau) h_{\Gamma}(t - \tau) d\tau,$$

или

$$\hat{x}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau. \qquad (2.65)$$

Сравнивая (2.63) и (2.64), можно видеть, что $x(t) = -\hat{x}(t) * h_{\Gamma}(t)$, или

$$x(t) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{x}(\tau)}{t - \tau} d\tau = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{x}(\tau)}{\tau - t} d\tau. \qquad (2.66)$$

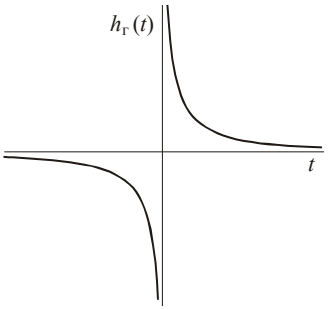


Рис. 2.34. Импульсная характеристика преобразователя Гильберта

Выражения (2.65) и (2.66) представляют собой прямое и обратное *преобразования Гильберта* (см. пример 2.15). Очевидно, эти преобразования линейны, поэтому, заменив τ на s, можно (2.66) рассматривать как интегральное представление сигнала

$x(t)$ спектральной плотностью $\hat{x}(\cdot)$ относительно ядра $\dfrac{1}{\pi(s-t)}$ и записать в виде

$$x(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{x}(s)}{s-t} ds \, .$$

Прямое преобразование (2.65) можно переписать в виде

$$\hat{x}(s) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(t)}{s-t} dt \, ,$$

откуда видно, что ядро является самосопряженным (так как ядра прямого и обратного преобразований Гильберта являются комплексно сопряженными[46]).

Ввиду важности преобразования Гильберта для техники связи приведем его основные свойства. Напомним, что самосопряженное ядро является непрерывным аналогом ортонормального базиса, поэтому выполняется обобщенная формула Рэлея

1) $(\hat{x}, \hat{y}) = (x, y)$

и, в частности, равенство Парсеваля

2) $(\hat{x}, \hat{x}) = (x, x)$.

Преобразование Гильберта, таким образом, сохраняет энергию сигнала (что естественно, поскольку фильтр Гильберта имеет АЧХ, тождественно равную 1). Более того, сохраняется энергетический спектр сигнала, а значит, и АКФ:

3) $W_{\hat{x}}(f) = W_x(f)$;

4) $R_{\hat{x}}(\tau) = R_x(\tau)$;

5) если $x(t) =$ const, то $\hat{x}(t) = 0$ в силу нечетности функции $h_{\text{Г}}(t)$;

6) если $x(t)$ – вещественный сигнал, то $x(t)$ *и* $\hat{x}(t)$ ортогональны. Действительно,

$$(x, \hat{x}) = \int_{-\infty}^{\infty} X(f)\hat{X}^*(f)df = \quad \text{[в соответствии с (2.64)]}$$

$$= -j \int_{-\infty}^{0} X(f)X^*(f)df + j\int_{0}^{\infty} X(f)X^*(f)df = 0 \, ,$$

[46] В данном случае они просто совпадают, так как являются вещественными.

поскольку подынтегральное выражение представляет собой чётную функцию – энергетический спектр вещественного сигнала.

Аналитический сигнал можно записать в форме $z(t) = A(t)e^{j\Phi(t)}$, где функция $\Phi(t)$ представляет угловое положение вектора на комплексной плоскости, т.е. *фазу*. Производная фазы по времени называется (круговой) *мгновенной частотой* и определяется выражением

$$\omega = \frac{d\Phi(t)}{dt} = \frac{d}{dt}\operatorname{Im}\{\ln z(t)\}.$$

Тогда мгновенная частота

$$f(t) = \frac{1}{2\pi}\operatorname{Im}\left[\frac{d}{dt}\ln z(t)\right] = \frac{1}{2\pi}\operatorname{Im}\left[\frac{z'(t)}{z(t)}\right] =$$

$$= \frac{1}{2\pi}\operatorname{Im}\left[\frac{[x'(t) + j\cdot\hat{x}'(t)][x(t) - j\cdot\hat{x}(t)]}{x^2(t) + \hat{x}^2(t)}\right] =$$

$$= \frac{1}{2\pi}\frac{\hat{x}'(t)x(t) - x'(t)\hat{x}(t)}{x^2(t) + \hat{x}^2(t)}.$$

Понятие аналитического сигнала оказывается полезным при описании *узкополосных сигналов*, для которых спектральная плотность аналитического сигнала $Z(f)$ в основном сосредоточена около некоторой центральной частоты F_0 (рис. 2.35). В частности, узкополосными являются сигналы, полученные путем модуляции гармонических несущих колебаний. Для узкополосного сигнала функция $A(t)$ имеет смысл *огибающей*, а фаза $\Phi(t) = 2\pi F_0 t + \psi(t)$ складывается из линейно растущего со временем слагаемого и медленно[47] меняющейся *начальной фазы* $\psi(t)$.

Согласно теореме модуляции (см. п. 2.10.2) умножение произвольного сигнала на комплексную экспоненту $e^{j\cdot 2\pi f_0 t}$ эквивалентно сдвигу его спектральной плотности *вправо* на величину f_0.

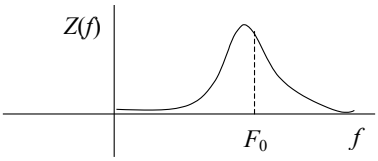


Рис. 2.35. К понятию узкополосного сигнала

[47] В сравнении с колебанием частоты F_0.

Сдвинем спектральную плотность $Z(f)$ с «центром тяжести» F_0, показанную на рис. 2.35, *влево* на величину F_0, тогда получится колебание

$$\gamma(t) = z(t)e^{-j\cdot 2\pi F_0 t}$$

со спектральной плотностью $\Gamma(f) = Z(f + F_0)$, рис. 2.36. Колебание $\gamma(t)$, очевидно, является комплексным и низкочастотным (в том смысле, что его спектральная плотность сосредоточена около нулевой частоты). Можно считать, что аналитический сигнал $z(t)$ получен *модуляцией* несущего гармонического колебания с частотой F_0 комплексным колебанием $\gamma(t)$, которое поэтому называется *комплексной огибающей*. Комплексную огибающую $\gamma(t) = A(t)e^{j\psi(t)}$ можно представить в виде векторной диаграммы, показанной на рис. 2.37.

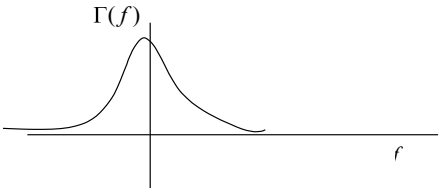


Рис. 2.36. Спектральная плотность комплексной огибающей узкополосного сигнала

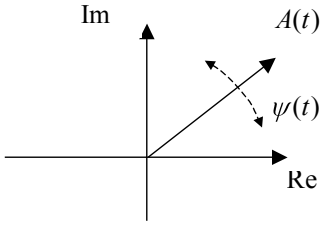


Рис. 2.37. Векторная диаграмма комплексной огибающей узкополосного сигнала

Длина вектора, изображающего комплексную огибающую, и угол между ним и вещественной осью комплексной плоскости медленно меняются в соответствии с функциями $A(t)$ и $\psi(t)$. Аналогичное представление аналитического сигнала отличается тем, что вектор дополнительно вращается против часовой стрелки с угловой скоростью (круговой частотой) $2\pi F_0$. Таким образом, узкополосный сигнал $x(t)$ можно рассматривать как гармоническое колебание, модулированное по амплитуде и фазе соответственно «медленными» функциями $A(t)$ и $\psi(t)$:

$$x(t) = A(t)\cos\left[2\pi F_0 t + \psi(t)\right] = A(t)\cos\Phi(t). \qquad (2.67)$$

Очевидно, $A(t) = \sqrt{x^2(t) + \hat{x}^2(t)}$, $\Phi(t) = \arg\left[x(t) + j\cdot\hat{x}(t)\right]$.

Комплексная огибающая, как медленная комплексная функция, представляет собой сумму двух медленно меняющихся слагаемых $\gamma(t) = u(t) + jv(t)$, поэтому исходный сигнал можно представить в виде

$$x(t) = \mathrm{Re}\left\{\gamma(t)e^{j\cdot 2\pi F_0 t}\right\} =$$

$$= \mathrm{Re}\left\{\left[u(t) + jv(t)\right]\left[\cos(2\pi F_0 t) + j\sin(2\pi F_0 t)\right]\right\} =$$

$$= u(t)\cos(2\pi F_0 t) - v(t)\sin(2\pi F_0 t). \qquad (2.68)$$

Колебание $u(t)$ называется *синфазной*, а колебание $v(t)$ – *квадратурной* составляющей (компонентой) узкополосного сигнала $x(t)$. Вместе две эти функции называются *квадратурными компонентами*. Отметим, что по известным квадратурным компонентам и частоте F_0 формула (2.68) позволяет точно восстановить исходный сигнал $x(t)$, поэтому вся информация, содержащаяся в сигнале, сохраняется в паре его квадратурных составляющих. Именно это дает основание заменять узкополосный сигнал его комплексной огибающей (или, что эквивалентно, парой квадратурных компонент) тогда, когда это удобно с точки зрения решаемой задачи.

В частности, такая замена может быть очень эффективной при *дискретизации* узкополосного сигнала. С информационной точки зрения вместо узкополосного сигнала можно передавать его квадратурные компоненты (предполагается, что частота F_0 известна). Поэтому частота дискретизации должна выбираться так, чтобы по отсчетам можно было восстановить низкочастотные квадратурные компоненты, т.е. дискретизация узкополосного сигнала может производиться с частотой, вдвое превышающей верхнюю частоту в спектре комплексной огибающей, а не самого сигнала. Например, если сигнал занимает полосу частот от 990 кГц до 1 МГц, то частоту дискретизации достаточно выбрать равной 10 кГц, в то время как без учета рассмотренных в этом разделе понятий частота дискретизации сигнала должна быть не менее 2 МГц. Требования к частоте дискретизации имеют жизненно важное значение при разработке цифровых систем связи, так как этим определяются сложность и стоимость системы, а иногда и ее принципиальная реализуемость.

Синфазная компонента может быть найдена следующим образом:

$$u(t) = \mathrm{Re}\left\{\gamma(t)\right\} = \mathrm{Re}\left\{z(t)e^{-j\cdot 2\pi F_0 t}\right\} =$$

$$= \mathrm{Re}\left\{\left[x(t) + j\cdot\hat{x}(t)\right]\left[\cos(2\pi F_0 t) - j\sin(2\pi F_0 t)\right]\right\} =$$

$$= x(t)\cos(2\pi F_0 t) + \hat{x}(t)\sin(2\pi F_0 t)\,.$$

Аналогично можно найти квадратурную компоненту

$$v(t) = \mathrm{Im}\left\{\gamma(t)\right\} = \hat{x}(t)\cos(2\pi F_0 t) - x(t)\sin(2\pi F_0 t)\,.$$

Таким образом, для сигнала $x(t)$ нужно вначале определить сопряженный по Гильберту сигнал $\hat{x}(t)$, после чего легко найти квадратурные компоненты. Есть и другой способ, более простой с практической точки зрения и реализуемый схемой, показанной на рис. 2.38. Покажем, что приведенная схема действительно выделяет квадратурные компоненты (с точностью до амплитудного множителя). В соответствии с выражением (2.68)

$$x(t)\cos(2\pi F_0 t) = u(t)\cos^2(2\pi F_0 t) - v(t)\sin(2\pi F_0 t)\cos(2\pi F_0 t) =$$

$$= u(t)\frac{(1+\cos(2\cdot 2\pi F_0 t))}{2} - v(t)\frac{\sin(2\cdot 2\pi F_0 t)}{2} \xrightarrow{\text{ФНЧ}} \frac{u(t)}{2}\,.$$

(Здесь символ $\xrightarrow{\text{ФНЧ}}$ означает подавление составляющих высоких частот, имеющих порядок F_0 и выше).

Аналогично

$$-x(t)\sin(2\pi F_0 t) = -u(t)\cos(2\pi F_0 t)\sin(2\pi F_0 t) + v(t)\sin^2(2\pi F_0 t) =$$

$$= -u(t)\frac{\sin(2\cdot 2\pi F_0 t)}{2} + v(t)\frac{1-\cos(2\cdot 2\pi F_0 t)}{2} \xrightarrow{\text{ФНЧ}} \frac{v(t)}{2}\,.$$

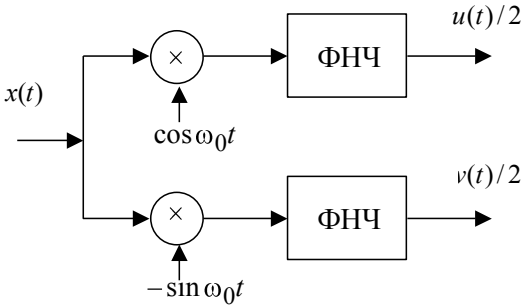


Рис. 2.38. Выделение квадратурных компонент узкополосного сигнала

Отметим, что опорные колебания отличаются лишь фазовым сдвигом $\pi/2$, поэтому на практике обычно вырабатываются одним генератором с использованием фазовращателя. Значение квадратурных компонент для практики иллюстрируется следующими примерами.

Пример 2.24. Предположим, что рассматриваемый узкополосный сигнал $x(t)$ представляет собой *амплитудно-модулированное колебание*, тогда вектор, изображающий комплексную огибающую, изменяется только по длине (норме), угловое же его положение определяется начальной фазой несущего колебания и постоянно. Начало отсчета начальной фазы определяется на практике начальной фазой опорного колебания в канале выделения синфазной компоненты, поэтому если начальная фаза несущего колебания известна, можно обеспечить синфазность (когерентность) несущего и опорного колебаний. Тогда вектор, изображающий комплексную огибающую, направлен вдоль вещественной оси комплексной плоскости, и квадратурная компонента всегда равна нулю. Квадратурный канал схемы, показанной на рис. 2.36, оказывается ненужным. На выходе синфазного канала наблюдается колебание $u(t)/2$, пропорциональное закону изменения огибающей, т.е. закону модуляции. Таким образом, синфазная часть схемы может использоваться при указанных условиях для демодуляции (детектирования) АМ-колебаний и называется в таких случаях *синхронным (когерентным) детектором*. ◄

Пример 2.25. Предположим теперь, что узкополосный сигнал $x(t)$ представляет собой *колебание с фазовой модуляцией* (ФМ-сигнал), тогда вектор, изображающий комплексную огибающую, имеет постоянную длину, а его угловое положение медленно меняется по закону модуляции фазы. Если изменения угла невелики (индекс модуляции мал), синфазная компонента меняется в небольших пределах и может считаться приближенно постоянной, а синфазный канал можно исключить из схемы. Квадратурная составляющая, напротив, меняется заметно и при условии малости индекса приближенно пропорциональна изменениям угла, т.е. закону фазовой модуляции. Квадратурная часть схемы представляет собой синхронный детектор, в котором опорное колебание сдвинуто по фазе на 90º относительно несущего колебания; такое устройство применяется для детектирования ФМ-колебаний (подробнее см. разд. 5). ◄

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие преимущества дает представление сигналов как элементов векторного пространства?

2. Какие сигналы называются ортогональными?

3. Что можно сказать о сигналах x и y, если $(x, y) = -\|x\|_2 \|y\|_2$?

4. Что такое ортонормальный базис?

5. В чем состоит практический смысл требования полноты базиса?

6. Что такое явление Гиббса и в чем его причина?

7. Сформулируйте принцип суперпозиции.

8. Запишите выражение, описывающее произвольный линейный оператор, действующий в пространстве $L_2(-\infty, +\infty)$.

9. Запишите выражение, описывающее линейный инвариантный к сдвигу (стационарный) оператор, действующий в пространстве $L_2(-\infty, +\infty)$.

10. Чем объясняется особая роль ряда и интеграла Фурье в анализе сигналов и цепей?

11. Что такое импульсная характеристика ЛИС-цепи? Можно ли её измерить точно? приближенно? Как это сделать?

12. Что такое комплексная частотная характеристика ЛИС-цепи? Можно ли её измерить (точно, приближенно)? Как это сделать?

13. Что такое автокорреляционная функция детерминированного сигнала? Что она характеризует?

14. Что такое взаимно корреляционная функция детерминированных сигналов? Что она характеризует?

15. Как связаны ВКФ и скалярное произведение детерминированных сигналов?

16. Как технически можно восстановить аналоговый сигнал по последовательности его отсчетов?

17. Почему при дискретизации аналогового сигнала, производимой путем стробирования, форма стробирующего импульса не играет заметной роли, если импульс достаточно короток?

18. Что препятствует на практике точному восстановлению аналогового сигнала по последовательности его отсчетов?

19. В чем причина погрешности при реальном стробировании аналогового сигнала в устройстве выборки-хранения?

20. Для чего частоту дискретизации на практике выбирают больше удвоенной верхней частоты спектра сигнала?

21. Для чего перед дискретизацией аналоговый сигнал подвергают НЧ-фильтрации?

22. Что такое аналитический сигнал? Как связаны вещественный сигнал и соответствующий ему аналитический сигнал?

УПРАЖНЕНИЯ

1. Выведите из (2.5) представление сигнала, отличного от нуля на всей вещественной оси.

2. Докажите, что в линейном пространстве должно выполняться условие

$$\exists 0 \in \mathbb{F} : \forall x \in M : 0x = \vec{0}.$$

3. Докажите линейную независимость функций, упомянутых в примере 2.1.

4. Рассчитайте первые 8 функций Уолша согласно рекуррентным выражениям примера 2.11 и постройте их графики (выполните это задание для функций Уолша, заданных на интервалах (–0,5; 0,5) и (0; 1)). Сравните полученные результаты.

5. Запишите обобщенную формулу Рэлея для пространств L_2 и l_2.

6. Докажите самосопряженность ядра Гильберта $\dfrac{1}{\pi(s-t)}$.

7. Докажите свойство дуальности (2.48) преобразования Фурье.

8. Объясните качественно присутствие δ-функции в выражении (2.51).

9. Докажите свойство (2.54).

10. Выведите формулу, приведенную в примере 2.20.

11. Докажите, что АКФ вещественного сигнала обладает свойством четности.

12. Найдите выражение АКФ прямоугольного видеоимпульса (пример 2.22) на основании общего выражения (2.53)

13. Убедитесь, что импульсная характеристика идеального интерполирующего фильтра с КЧХ (2.60) имеет вид $h(t) = \sin\left[\dfrac{\pi}{T_d}t\right] \Big/ \left(\dfrac{\pi}{T_d}t\right)$.

14. Часто в качестве моделей импульсных сигналов используют нефинитные функции, имеющие бесконечную длительность.

В таких случаях вводят *эффективную* длительность сигнала – временной интервал, на котором сосредоточена бóльшая часть его энергии. Определите эффективную длительность экспоненциального импульса

$$s(t) = \begin{cases} Ae^{-\alpha t}, & t \geq 0 \\ 0, & t < 0 \end{cases},$$

как интервал, на котором сосредоточено 95 % энергии.

15. Многие модели сигналов, удобные в аналитическом отношении, характеризуются нефинитным спектром (или спектральной плотностью). Для них вводят понятие *эффективной ширины* спектра – частотного интервала, содержащего заданную долю $k_э$ энергии сигнала. Определите эффективную ширину спектра прямоугольного импульса при $k_э = 0.9664$.

3. СЛУЧАЙНЫЕ ПРОЦЕССЫ

Как отмечалось в разд. 1, все сигналы и помехи являются случайными, т. е. непредсказуемыми. Математическими моделями случайных сигналов и помех служат *случайные процессы*. В терминах современной теории вероятностей всякий случайный процесс (СП) связан с некоторым воображаемым множеством, или *пространством элементарных событий* $\Omega = \{\omega\}$. Выбор одного из элементарных событий происходит некоторым способом, который наблюдателю не известен и представляется *случайным*, поэтому исход такого эксперимента заранее предсказать нельзя. Выбор конкретного элемента ω приводит к осуществлению вполне определенной *реализации* случайного процесса. Важно подчеркнуть, что в основе теории вероятностей лежит допущение о возможности *многократного повторения случайного эксперимента в одинаковых условиях* и получения любого количества реализаций случайного события, *случайной величины* или случайного процесса. С пространством Ω связана функция $P : \Omega \to \mathbb{R}$, называемая *вероятностной мерой* (распределением вероятностей), в соответствии с которой элементарные события имеют бóльшие или меньшие шансы быть выбранными в конкретном опыте. Более точно, мера P ставит в соответствие любому подмножеству A множества Ω неотрицательное вещественное число, не превышающее единицы, называемое вероятностью $P(A)$ случайного события A. При этом $P(\Omega) = 1$, т. е. вероятность *достоверного* события равна 1. *Невозможное* событие, обозначаемое $\varnothing$, имеет вероятность $P(\varnothing) = 0$. Строго говоря, в современной теории вероятностей для полного задания вероятностного описания необходимо также определить систему измеримых подмножеств (подмножеств множества Ω, для которых определена мера), замкнутую относительно операций объединения и пересечения и называемую σ-алгеброй или σ-полем.

Теория случайных процессов представляет собой большой раздел теории вероятностей, который изучается в курсе математики, поэтому здесь приводится лишь краткое изложение основных понятий, необходимых для понимания изучаемых разделов теории электрической связи. Для более полного изучения теории вероятностей и теории случайных процессов следует обратиться к специальным учебникам, например [6, 7].

3.1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ И ИХ ХАРАКТЕРИСТИКИ

Случайной величиной (СВ) называется любая функция $\mathbf{x}(\cdot)$, определенная на множестве Ω и принимающая вещественные значения[48]. При фиксированном $\omega \in \Omega$ значение $x = \mathbf{x}(\omega)$ также фиксировано; оно называется реализацией случайной величины $\mathbf{x}$. Далее для краткости случайные величины обозначаются так же, как и их реализации, если это не приводит к двусмысленности.

Полное описание случайной величины составляет кумулятивная *функция распределения,* определяемая выражением

$$F(a) = \mathbf{P}\{x = \mathbf{x}(\omega) \le a\},$$

где $\mathbf{P}\{\cdot\}$ обозначает вероятность события, состоящего в том, что случайная величина принимает значение, не превосходящее заданного значения a. Случайная величина, принимающая значения из дискретного множества, называется дискретной. Функция распределения такой СВ имеет ступенчатый вид. Если функция распределения является непрерывной и дифференцируемой, то можно определить *плотность распределения вероятностей* (ПРВ), называемую также для краткости плотностью вероятности (а иногда просто плотностью)

$$w(x) = \frac{dF(x)}{dx}, \text{ при этом } F(x) = \int_{-\infty}^{x} w(x)dx.$$

Очевидно, функция распределения по определению должна быть неотрицательной неубывающей функцией со свойствами

[48] *Комплексная* случайная величина определяется парой вещественных случайных величин, рассматриваемых совместно и играющих роль вещественной и мнимой частей комплексной СВ.

$F(-\infty)=0$, $F(\infty)=1$. Следовательно, плотность распределения должна быть неотрицательной функцией, удовлетворяющей *условию нормировки* $\int_{-\infty}^{\infty} w(x)dx=1$.

Иногда используется описание случайной величины *характеристической функцией*, которая определяется выражением

$$\varphi(u)=\int_{-\infty}^{\infty} w(x)e^{jux}dx,$$

совпадающим с преобразованием Фурье плотности распределения вероятностей (с точностью до знака показателя экспоненты).

Иногда нет необходимости использовать полное описание случайной величины и можно ограничиться ее числовыми характеристиками. Чаще всего этими характеристиками служат так называемые *моменты,* определяемые следующими выражениями. *Начальный* момент k -го порядка (k -й начальный момент)

$$m_k=\int_{-\infty}^{\infty} x^k w(x)dx=\overline{x^k}=\mathbf{E}\{x^k\},$$

где горизонтальная черта и $\mathbf{E}\{\cdot\}$ – символические обозначения интегрального оператора *усреднения по ансамблю*[49]. Наиболее часто используется первый начальный момент

$$m=m_1=\int_{-\infty}^{\infty} xw(x)dx=\bar{x}, \qquad (3.1)$$

называемый *математическим ожиданием*, или центром распределения[50]. Смысл этого понятия становится яснее из физической аналогии: если плотность распределения вероятностей рассматривать как линейную плотность бесконечно тонкого стержня единичной массы, расположенного вдоль оси абсцисс, то математическое ожидание будет равно координате центра масс этого стержня.

---

[49] Под ансамблем понимается множество реализаций случайной величины или процесса вместе с вероятностной мерой, заданной на этом множестве.

[50] Часто употребляется также термин «среднее», например, типично выражение «случайный процесс с нулевым средним».

Отметим, что усреднение по ансамблю можно применять к любой функции случайной величины; так, характеристическую функцию можно трактовать как результат усреднения комплексной экспоненты $\varphi(u) = \overline{e^{jux}}$.

Центральный момент k-го порядка (k-й центральный момент) равен k-му начальному моменту *центрированной* случайной величины $(x - m)$:

$$M_k = \int_{-\infty}^{\infty} (x-m)^k w(x)dx = \overline{(x-m)^k} = \mathbf{E}\{(x-m)^k\}.$$

Наиболее употребительным из центральных моментов является второй центральный момент, или *дисперсия*

$$D = M_2 = \int_{-\infty}^{\infty} (x-m)^2 w(x)dx = \overline{(x-m)^2} = \mathbf{E}\{(x-m)^2\}. \quad (3.2)$$

В упомянутом выше механическом примере дисперсии соответствует момент инерции стержня при вращении его вокруг центра масс. Дисперсия тем меньше, чем более сосредоточена «масса вероятности» около среднего значения случайной величины (математического ожидания). Вместо дисперсии часто оперируют величиной, равной $\sigma = \sqrt{D}$ и называемой *среднеквадратическим отклонением* (СКО) случайной величины.

Еще одной числовой характеристикой СВ является *средний квадрат*, или второй начальный момент $m_2 = \int_{-\infty}^{\infty} x^2 w(x)dx = \overline{x^2} =$ $= \mathbf{E}\{x^2\}$. Нетрудно видеть, что он связан с дисперсией и математическим ожиданием:

$$D = \int_{-\infty}^{\infty} (x-m)^2 w(x)dx = \int_{-\infty}^{\infty} (x^2 - 2mx + m^2) w(x)dx =$$

$$= m_2 - 2m^2 + m^2 = m_2 - m^2.$$

Пример 3.1. В математике и технике часто используется нормальное, или гауссово (гауссовское), распределение с ПРВ

$$w(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

где m и σ – параметры распределения. Подставив эту плотность в (3.1) и (3.2), можно убедиться, что математическое ожидание гауссовской случайной величины равно m, а дисперсия равна σ^2. Таким образом, параметр σ имеет смысл СКО и характеризует степень «размазанности» распределения (рис. 3.1). ◄

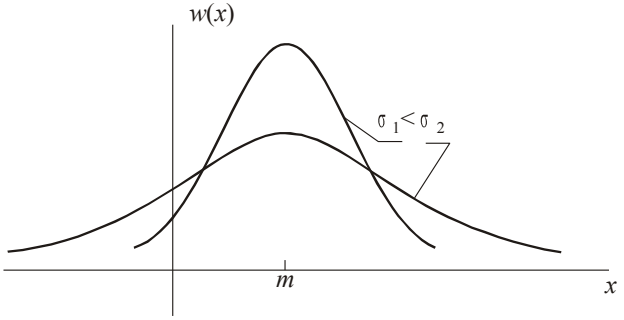


Рис. 3.1. Гауссовские ПРВ при различных значениях СКО

В теории информации используется числовая характеристика распределения случайной величины, называемая (дифференциальной) *энтропией* и определяемая выражением

$$H(x) = \overline{\log_2 \frac{1}{w(x)}} = -\int_{-\infty}^{\infty} w(x) \log_2 w(x) dx .$$

Две случайные величины $x = \mathbf{x}(\omega)$ и $y = \mathbf{y}(\omega)$, заданные на общем пространстве Ω, характеризуются совместной плотностью распределения $w(x, y)$. Числовыми характеристиками совместной плотности служат начальные и центральные смешанные моменты

$$m_{kn} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^n w(x, y) dx dy ,$$

$$M_{kn} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)^k (y - m_y)^n w(x, y) dx dy ,$$

где k и n – произвольные целые положительные числа.

Наиболее часто используются смешанные моменты второго порядка – начальный (*корреляционный* момент)

$$m_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyw(x, y)\,dx\,dy = k_{xy} \tag{3.3}$$

и центральный (*ковариационный* момент, или *ковариация*[51])

$$M_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - m_x)(y - m_y)w(x, y)\,dx\,dy = R_{xy}\,. \tag{3.4}$$

Ковариация представляет собой простейшую характеристику степени статистической (вероятностной) связи случайных величин x и y.

Пример 3.2. Для пары гауссовских случайных величин двумерная совместная ПРВ имеет вид

$$w(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - 2r\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]},$$

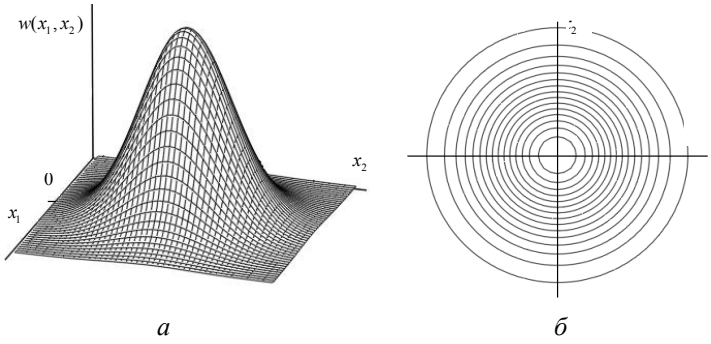
где σ_1, σ_2 – среднеквадратические отклонения, m_1, m_2 – математические ожидания, r – *коэффициент корреляции*, представляющий собой нормированный ковариационный момент

$$r = \frac{\overline{(x_1 - m_1)(x_2 - m_2)}}{\sigma_1\sigma_2} = \frac{R_{x_1x_2}}{\sigma_1\sigma_2}\,. \blacktriangleleft$$

На рис. 3.2, *а* показана двумерная гауссовская ПРВ при $r = 0$, $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2$, а на рис. 3.2, *б* – ее отображение линиями уровня. Для сравнения на рис. 3.3 показаны линии уровня гауссовской ПРВ при $r = 0.9$ (*а*) и $r = –0.9$ (*б*). Из рисунков видно, что при положительном коэффициенте корреляции двух СВ более вероятны их реализации с близкими значениями, а при отрицательном – с близкими по модулю, но различающимися по знаку. При нулевом коэффициенте корреляции очевидно, что

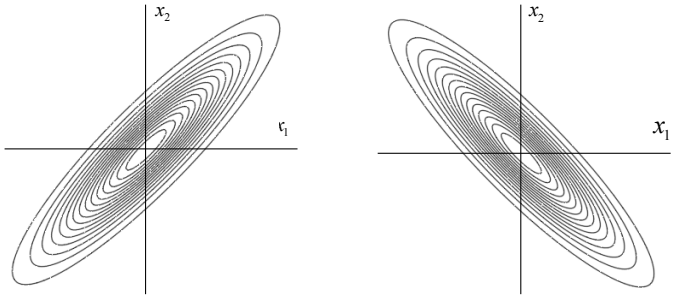
$$w(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}\, \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} = w(x_1)w(x_2)\,,$$

[51] В литературе иногда момент, определяемый формулой (3.3), называют ковариационным, тогда корреляционным называют момент (3.4).



а *б*

Рис. 3.2. Двумерная гауссовская ПРВ



а *б*

Рис. 3.3. Линии уровня двумерной гауссовской ПРВ при $r = 0,9$ (*а*) и $r = -0,9$ (*б*)

т.е. *некоррелированные* гауссовские случайные величины *независимы* (напомним, что совместная ПРВ независимых случайных величин равна произведению их одномерных ПРВ). Для негауссовских случайных величин некоррелированность не означает независимости, хотя из их независимости следует некоррелированность. Таким образом, вообще говоря, независимость – более *сильное* свойство, чем некоррелированность.

Поскольку случайная величина является функцией, на множестве случайных величин можно определить структуру гильбертова пространства. Действительно, случайные величины (как функции на пространстве Ω) можно складывать, при этом сумма снова будет случайной величиной. Случайные величины можно умножать на скалярные коэффициенты, причем множество случайных величин замкнуто относительно такого умножения. Справедливость

аксиом линейного пространства (разд. 2.3) легко проверяется непосредственно. Таким образом, множество всех вещественных случайных величин, заданных на общем множестве элементарных событий Ω, можно рассматривать как линейное пространство над полем $\mathbb{R}$ вещественных чисел (аналогично можно ввести пространство комплексных случайных величин над полем $\mathbb{C}$ комплексных чисел и т.д.). Дальнейшее усовершенствование структуры пространства связано с введением нормы, метрики и скалярного произведения. Для того чтобы пространство было гильбертовым, необходимо, чтобы норма порождалась скалярным произведением, а метрика – нормой [2]. Операцию скалярного умножения определим для *вещественных* случайных величин x и y как смешанный момент второго порядка (корреляционный момент)

$$(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyw(x, y)dxdy = \overline{xy}. \qquad (3.5)$$

В частности, если две величины имеют нулевой корреляционный момент, то они являются ортогональными. Проверим выполнение аксиом скалярного произведения.

Из (3.5) очевидно выполнение равенства $(x, y) = (y, x)$.

Проверка выполнения условия $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ может быть произведена непосредственно:

$$(\alpha x + \beta y, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\alpha x + \beta y)zw(x, y, z)dxdydz =$$

$$= \alpha\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xzw(x, z)dxdz + \beta\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yzw(y, z)dydz = \alpha\overline{xz} + \beta\overline{yz}.$$

Здесь $w(x, y, z)$ – совместная плотность распределения вероятностей трех случайных величин, которая при интегрировании по одному из аргументов дает совместную плотность оставшихся двух случайных величин[52], например $\int_{-\infty}^{\infty} w(x, y, z)dx = w(y, z)$.

Третье условие, очевидно, выполняется: $(x, x) \geq 0$, поскольку (x, x) – не что иное, как средний квадрат, неотрицательный по определению. Равенство нулю среднего квадрата (как второго на-

[52] ПРВ, которая получается интегрированием ПРВ большей размерности по одной или нескольким переменным, называется *маргинальной.*

чального момента) возможно только в том случае, если вся «вероятностная масса» сосредоточена в точке $x = 0$. Таким образом, роль нулевого вектора в рассматриваемом пространстве играет случайная величина, которая принимает значение 0 с вероятностью 1 (ПРВ такой случайной величины равна $\delta(x)$).

Норма случайной величины определяется через скалярное произведение, как $\|x\| = \sqrt{(x, x)} = \sqrt{\overline{x^2}}$, а метрика задается через норму

$$d(x, y) = \|x - y\| = \sqrt{\overline{(x - y)^2}} .$$

Итак, множество случайных величин, определенных на общем пространстве элементарных событий, становится гильбертовым пространством. К нему применимы все ранее введенные для гильбертова пространства понятия, такие, как базис, ортонормальный базис, ортогонализация Грама – Шмидта, равенство Парсеваля и т.п.

В следующем примере предполагается, что математическое ожидание случайных величин равно нулю, тогда средний квадрат совпадает с дисперсией, а корреляционный момент – с ковариационным (вторым смешанным *центральным* моментом).

Пример 3.3. Задача *оптимальной фильтрации* состоит в том, чтобы по наблюдаемому колебанию $z(t)$ наилучшим образом оценить полезный (случайный) сигнал $x(t)$. И наблюдаемый, и полезный сигналы здесь будем понимать, как *наборы случайных величин* – отсчетов сигнала (их множество может быть *несчетным,* т.е. «сплошным»). Оптимальный *линейный* фильтр – это линейный оператор $\mathbb{L}\{\cdot\}$, вырабатывающий на основе колебания $z(t)$ оценку, такую, что дисперсия ошибки оценивания $\overline{[x(t) - \mathbb{L}\{z(t)\}]^2}$ минимальна.

Результат воздействия на сигнал $z(t)$ линейного оператора – это, нестрого говоря, линейная комбинация *всех отсчетов* сигнала. Поэтому оценка принадлежит линейной оболочке отсчетов сигнала $z(t)$, или *подпространству, натянутому* на эти отсчеты (которые представляют собой случайные величины, т.е. векторы). Полезный сигнал $x(t)$ в общем случае лежит вне этого подпространства (рис. 3.4) .

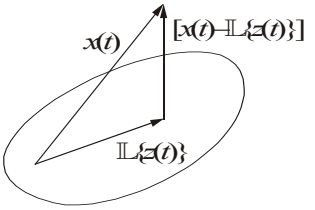


Рис. 3.4. Геометрическая интерпретация принципа оптимального линейного оценивания

Из геометрических соображений ясно, что СКО ошибки оценивания (норма ошибки) будет минимально в том случае, если вектор ошибки будет *ортогонален* этому подпространству (ошибка не коррелирована с наблюдаемым сигналом в произвольный момент времени θ), отсюда условие оптимальности оператора

$$\overline{[x(t)-\mathbb{L}\{z(t)\}]z(\theta)}=0\,.$$

Учитывая, что линейный оператор выражается интегралом, для оптимальной линейной оценки получаем

$$\overline{\left[x(t)-\int_{-\infty}^{\infty}h(t,\tau)z(\tau)d\tau\right]z(\theta)}=0\,,$$

где $h(t,\tau)$ – весовая функция (ядро оператора), имеющая смысл отклика фильтра в момент времени t на значение наблюдаемого сигнала в момент τ; θ – переменная, имеющая размерность времени. Раскрывая скобки и выполняя усреднение, получаем

$$\overline{x(t)z(\theta)}-\int_{-\infty}^{\infty}h(t,\tau)\overline{z(\tau)z(\theta)}\,d\tau=0\,,$$

откуда следует *уравнение Винера – Хопфа*

$$\int_{-\infty}^{\infty}h(t,\tau)k_{zz}(\tau,\theta)d\tau=k_{xz}(t,\theta)\,,$$

где $k_{zz}(\tau,\theta)=\overline{z(\tau)z(\theta)}$ – второй смешанный момент отсчетов случайного процесса $z(t)$ в моменты времени τ и θ, называемый функцией *автокорреляции* процесса $z(t)$, а $k_{xz}(t,\theta)=\overline{x(t)z(\theta)}$ – второй смешанный момент отсчетов *различных* процессов, называемый функцией *взаимной* корреляции процессов $x(t)$ и $z(t)$ (см. п. 3.2).

Характеристика $h(t,\tau)$ оптимального линейного устройства оценивания находится, как решение уравнения Винера – Хопфа (подробнее см. разд. 10).◀

3.2. СЛУЧАЙНЫЕ ПРОЦЕССЫ И ИХ ОПИСАНИЕ

В теории связи большую роль играют *случайные процессы,* являющиеся математическими моделями как сигналов, так и помех. Случайный процесс – это колебание, принимающее в любой заданный момент времени значение, которое невозможно точно

предсказать. Таким образом, можно понимать случайный процесс как упорядоченную последовательность случайных величин, следующих друг за другом в порядке возрастания некоторой переменной (чаще всего времени). Перейти от описания случайной величины к описанию случайного процесса можно, рассматривая совместные распределения двух значений процесса в различные моменты времени, трех значений и т.д. В частности, рассматривая процесс в n временных *сечениях* (при $t = t_1,...,t_n$), получаем n-мерные совместные функцию распределения и плотность распределения вероятностей случайных величин $x(t_1), \dots\ x(t_n)$, определяемые выражением

$$\mathbf{P}\{\xi_1 \le x_1,...,\xi_n \le x_n\} = F(x_1,...,x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} w(x_1,...,x_n)dx_1...dx_n .$$

Здесь и далее зависимость от времени явно не указана для упрощения записи. Для n-мерной ПРВ выполняется условие нормировки

$$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} w(x_1,...,x_n)dx_1...dx_n = 1 .$$

Случайный процесс считается полностью определенным, *если для любого n можно записать его совместную ПРВ при любом выборе моментов времени* $t_1,...,t_n$. Следует отметить, что на практике это удается сделать крайне редко[53].

Часто при описании случайного процесса можно ограничиться совокупностью его смешанных начальных моментов (если они существуют, т.е. сходятся соответствующие интегралы)

$$m_{k_1...k_n} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} x_1^{k_1}...x_n^{k_n} w(x_1,...,x_n)dx_1...dx_n \qquad (3.6)$$

и смешанных центральных моментов

$$M_{k_1...k_n} = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_1 - m_{x_1})^{k_1}...(x_n - m_{x_n})^{k_n} w(x_1,...,x_n)dx_1..dx_n \qquad (3.7)$$

при целых неотрицательных $k_1,...,k_n$ и целом n.

---

[53] Исключение составляют *гауссовские* и *марковские* процессы, а также процессы с *распределением Гиббса.*

В частности, полагая в (3.6) $k_1 = 1$, $k_2 = ... = k_n = 0$, получаем первый начальный момент (математическое ожидание) при $t = t_1$; при $k_1 = 2$, $k_2 = ... = k_n = 0$ выражение (3.6) определяет средний квадрат, а выражение (3.7) – дисперсию в соответствующем сечении. Если под случайным процессом подразумевается сигнал в форме напряжения, то математическое ожидание имеет смысл его среднего значения («медленной» составляющей), средний квадрат – полной мощности, а дисперсия – мощности флюктуационной («быстрой») составляющей.

В общем случае моменты совместной ПРВ зависят от расположения сечений на оси времени и называются *моментными функциями*. Чаще всего используют второй смешанный центральный момент

$$M_{11} = R_x(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - m_{x_1})(x_2 - m_{x_2})w(x_1, x_2)dx_1 dx_2 =$$

$$= \overline{(x_1 - m_{x_1})(x_2 - m_{x_2})},$$

называемый функцией автокорреляции или автокорреляционной функцией[54] (АКФ). Напомним, что здесь и далее явно не указана зависимость от времени, в частности, функциями времени являются $m_{x_1} = m_{x_1}(t_1)$ и $m_{x_2} = m_{x_2}(t_2)$.

Можно рассматривать совместно два случайных процесса $x(t)$ и $y(t)$, которые в общем случае не являются независимыми в вероятностном смысле; такое рассмотрение предполагает их совместное описание в виде совместной многомерной ПРВ, а также в виде совокупности всех моментов, в том числе смешанных. Наиболее часто при этом используют второй смешанный центральный момент

$$R_{xy}(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - m_{x_1})(y_2 - m_{y_2})w(x_1, y_2)dx_1 dy_2 =$$

$$= \overline{(x_1 - m_{x_1})(y_2 - m_{y_2})},$$

[54] Ни в коем случае не следует путать понятия АКФ для случайных процессов и для детерминированных сигналов, которые имеют *совершенно* разный смысл!

называемый взаимно корреляционной функцией $R_{xy}(t_1, t_2)$. Как и АКФ, взаимно корреляционная функция (ВКФ) является функцией двух переменных.

Среди всех случайных процессов выделяют СП, для которых совместная n-мерная ПРВ не изменяется при одновременном изменении (сдвиге) всех моментов времени на одну и ту же величину. Такие процессы называются *стационарными в узком смысле*, или *строго стационарными*.

Чаще всего на практике ограничивают рассмотрение случайными процессами с ослабленным условием стационарности. СП называется *стационарным в широком смысле*, если при одновременном сдвиге сечений не изменяются лишь его моменты не выше второго порядка. Практически это означает, что СП $x(t)$ стационарен в широком смысле, если он имеет постоянные *среднее* (математическое ожидание m_x) и *дисперсию* D_x, а АКФ зависит только от разности моментов времени, но не от их положения на временно́й оси:

1) $m_x(t) = m_x$,

2) $R_x(t_1, t_2) = R_x(t_2 - t_1) = R_x(\tau)$.

Заметим, что $R_x(0) = D_x$, откуда и следует постоянство дисперсии.

Нетрудно убедиться, что процесс, стационарный в узком смысле, стационарен и в широком смысле. Обратное вообще неверно, хотя существуют процессы, для которых стационарность в широком смысле означает и стационарность в узком смысле.

Пример 3.4. Совместная n-мерная ПРВ отсчетов $x_1,...,x_n$ гауссовского процесса, взятых в моменты времени $t_1,...,t_n$, имеет вид

$$w(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2....\sigma_n \,|\,\mathbf{r}\,|^{1/2}} e^{-\frac{1}{2|\mathbf{r}|}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{A}_{ij}\frac{(x_i - m_i)}{\sigma_i}\frac{(x_j - m_j)}{\sigma_j}}, \tag{3.8}$$

где $|\mathbf{r}|$ – определитель квадратной матрицы, составленной из попарных коэффициентов корреляции отсчетов; $\mathbf{A}_{ij}$ – алгебраическое дополнение элемента r_{ij} этой матрицы.

Как видно из выражения (3.8), совместная ПРВ полностью определяется математическими ожиданиями, дисперсиями и коэффи-

циентами корреляции отсчетов. Таким образом, зная моментные функции не выше второго порядка, при любом n можно записать совместную ПРВ. Если процесс стационарен в широком смысле, то все математические ожидания одинаковы, все дисперсии (а значит, и СКО) равны друг другу, а коэффициенты корреляции определяются только тем, насколько моменты времени $t_1,...,t_n$ отстоят друг от друга. Тогда, очевидно, ПРВ (3.8) не изменится, если все моменты $t_1,...,t_n$ сдвинуть влево или вправо на одну и ту же величину. Отсюда следует, что *гауссовский процесс, стационарный в широком смысле, стационарен и в узком смысле* (строго стационарен). ◄

Среди стационарных случайных процессов часто выделяют более узкий класс *эргодических* случайных процессов. Для эргодических процессов моменты, найденные усреднением по ансамблю, равны соответствующим моментам, найденным усреднением по времени[55] единственной реализации $\xi(t)$:

$$m_k = \left\langle \left[x(t)\right]^k \right\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T \left[\xi(t)\right]^k dt\,,$$

$$M_k = \left\langle \left[x(t)-m\right]^k \right\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T \left[\xi(t)-m\right]^k dt$$

(здесь $\langle\cdot\rangle$ – символическое обозначение оператора усреднения по времени).

В частности, для эргодического процесса математическое ожидание, дисперсия и АКФ равны соответственно:

$$m = \left\langle x(t)\right\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T \xi(t)dt\,,$$

$$D = \left\langle \left[x(t)-m\right]^2 \right\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T \left[\xi(t)-m\right]^2 dt\,,$$

$$R(\tau) = \left\langle \left[x(t)-m\right]\left[x(t+\tau)-m\right] \right\rangle = \left\langle x(t)x(t+\tau)\right\rangle - m^2 =$$

$$= \lim_{T\to\infty} \frac{1}{T}\int_0^T \xi(t)\xi(t+\tau)dt - m^2.$$

[55] Строгое определение эргодичности выходит за рамки учебника (см., напр., [7]).

Эргодичность представляет собой весьма желательное свойство, так как дает возможность практически измерять числовые характеристики случайного процесса. Дело в том, что обычно наблюдателю доступна лишь одна (хотя, возможно, достаточно длинная) реализация случайного процесса. Эргодичность означает, по существу, то, что эта единственная реализация является полноправным представителем всего ансамбля.

Пример 3.5. Измерение характеристик эргодического процесса может быть выполнено при помощи простых измерительных устройств; так, если процесс представляет собой напряжение, зависящее от времени, то вольтметр *магнитоэлектрической* системы измеряет его математическое ожидание (постоянную составляющую), вольтметр электромагнитной или термоэлектрической системы, подключенный через разделительную емкость (для исключения постоянной составляющей), – его среднеквадратическое значение (СКО). Устройство, структурная схема которого показана на рис. 3.5, позволяет измерить значения функции автокорреляции при различных значениях задержки τ.

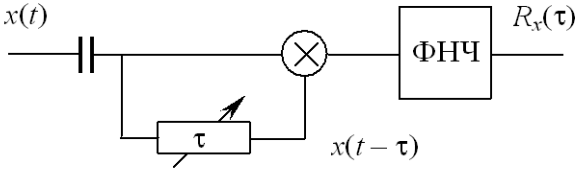


Рис. 3.5. Измерение автокорреляционной функции эргодического процесса

Фильтр нижних частот играет здесь роль интегратора, конденсатор выполняет центрирование процесса, так как не пропускает постоянную составляющую тока. Это устройство называется *коррелометром*.◀

Достаточными условиями эргодичности стационарного случайного процесса служат условие стремления АКФ к нулю

$$\lim_{\tau\to\infty} R(\tau) = 0\,,$$

а также менее сильное *условие Слуцкого*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T R(\tau)d\tau = 0\,.$$

3.3. КОРРЕЛЯЦИОННО-СПЕКТРАЛЬНАЯ ТЕОРИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ

Точное решение задач, связанных с анализом случайных процессов и их воздействия на ЛИС-цепи, сопряжено с большими трудностями, так как предполагает отыскание совместной n-мерной ПРВ для выходного процесса. Значительно проще решается задача анализа, если интересоваться только моментными характеристиками первого и второго порядка, которые определяют свойство стационарности в широком смысле. Учитывая, что большинство реально наблюдаемых процессов удовлетворительно описываются гауссовской моделью, а гауссовские процессы полностью определяются моментными характеристиками не выше второго порядка, во многих случаях ограничиваются анализом на уровне математических ожиданий (средних) и корреляционных функций.

Рассмотрим *вещественный* стационарный случайный процесс $x(t)$ с нулевым средним. Его реализация, как детерминированная функция, может быть представлена обратным преобразованием Фурье

$$\xi(t) = \int_{-\infty}^{\infty} \Xi(f) e^{j\cdot 2\pi f t} df ,$$

где $\Xi(f)$ – спектральная плотность реализации. (Следует иметь в виду, что почти все[56] реализации стационарного СП не принадлежат пространству сигналов конечной энергии L_2, поэтому их спектральные плотности можно рассматривать лишь в терминах *обобщенных* функций, так как соответствующие интегралы в классическом смысле расходятся. Однако, поскольку нас будут интересовать лишь усредненные величины и функции, важна лишь интегрируемость соответствующих математических ожиданий.) Случайный процесс $x(t)$ можно записать в виде

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j\cdot 2\pi f t} df ,$$

где $X(f)$ – также случайный процесс (в соответствии с природой преобразования Фурье – это *тот же* процесс, представленный в другом «базисе»). Выясним, что это за процесс.

[56] Заметим, что здесь выражение «почти все» понимается в строгом вероятностном смысле и означает «с вероятностью единица».

Поскольку $x(t)$ – случайный процесс с нулевым средним, $X(f)$ также имеет нулевое среднее:

$$\overline{x(t)} = \int_{-\infty}^{\infty} \overline{X(f)} e^{j\cdot 2\pi ft} df = 0 \Rightarrow \overline{X(f)} = 0 .$$

Автокорреляционная функция вещественного процесса $x(t)$

$$R_x(\tau) = \overline{x(t)x(t+\tau)} = \overline{x^*(t)x(t+\tau)} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{X^*(f)X(\phi)} e^{-j\cdot 2\pi ft} e^{j\cdot 2\pi\phi t} e^{j\cdot 2\pi\phi\tau} df d\phi =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_x(f)\delta(f-\phi) e^{j\cdot 2\pi\phi\tau} df d\phi .$$

Последнее равенство записано на том основании, что АКФ не зависит от переменной t, а это может быть только при том условии, что $f \equiv \phi$, т.е.

$$\overline{X^*(f)X(\phi)} = W_x(f)\delta(f-\phi) . \qquad (3.9)$$

С учетом стробирующего свойства δ-функции можно записать

$$R_x(\tau) = \int_{-\infty}^{\infty} W_x(f) e^{j\cdot 2\pi f\tau} df , \qquad (3.10)$$

а следовательно,

$$W_x(f) = \int_{-\infty}^{\infty} R_x(\tau) e^{-j\cdot 2\pi f\tau} d\tau . \qquad (3.11)$$

Выражения (3.10) – (3.11) составляют запись *теоремы Винера – Хинчина*[57].

При $\tau = 0$ из выражения (3.10) следует

$$R_x(0) = \int_{-\infty}^{\infty} W_x(f) df ,$$

[57] Александр Яковлевич Хинчин (1894 – 1959) – русский математик, известен своими трудами в области теории вероятностей, теории функций, теории массового обслуживания и др.

а поскольку $R_x(0) = D_x$ – мощность случайного процесса (с нулевым средним), функция $W_x(f)$ называется *спектральной плотностью мощности* (СПМ). Очевидно, СПМ – *неотрицательная* функция. Если процесс имеет ненулевое математическое ожидание m, то к СПМ добавляется слагаемое $m^2\delta(f)$.

Для вещественного процесса АКФ – четная вещественная функция, тогда СПМ – тоже четная вещественная. Поэтому иногда используется *односторонняя* СПМ:

$$R_x(\tau) = \int_0^\infty N_x(f)\cos(2\pi f\tau)df \ .$$

Очевидно, $N_x(f) = 2W_x(f), \ f \geq 0.$

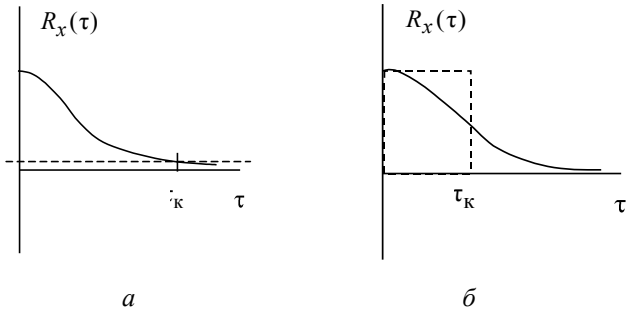
Иногда нет необходимости знать точный вид АКФ и СПМ и можно ограничиться числовыми характеристиками, в роли которых выступают интервал корреляции и эффективная ширина спектра. *Интервал корреляции* определяют по-разному, в частности, известны следующие определения.

1. Интервал корреляции – такое значение τ, при котором АКФ спадает до заданного уровня, например до 1/10 максимального значения (рис. 3.6, *а*).

2. Интервал корреляции – основание прямоугольника, имеющего площадь, равную площади под графиком АКФ (рис. 3.6, *б*).

Эффективную ширину спектра определяют по спектральной плотности мощности способами, аналогичными показанным на рис. рис. 3.6, *а* и 3.6, *б*.

Очень часто используют следующие две модели стационарных случайных процессов.



а *б*

Рис. 3.6. К определению интервала корреляции

Пример 3.6. *Белый шум*. Так называется стационарный случайный процесс с нулевым средним, имеющий АКФ вида

$$R_x(\tau) = \frac{N_0}{2}\delta(\tau).$$

Очевидно, что в этом случае СПМ постоянна на всех частотах от $-\infty$ до ∞:

$$W_x(f) = N_0/2.$$

(Принято использовать обозначение $N_0/2$ для двусторонней СПМ; односторонняя обозначается N_0.)

Легко видеть, что никакой реальный случайный процесс не может быть белым шумом, так как белый шум имеет бесконечную дисперсию (мощность). Кроме того, для белого шума теряет смысл понятие распределения. Однако эта модель чрезвычайно удобна в анализе вследствие δ-образности АКФ, поэтому она широко используется.◀

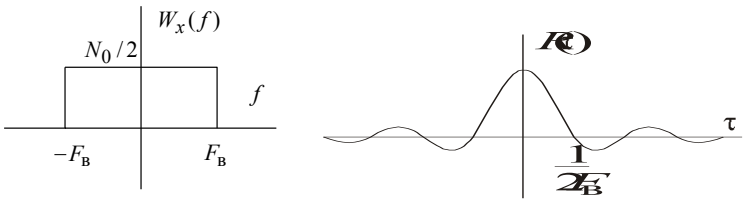
Пример 3.7. *Квазибелый шум* (шум, белый в ограниченной полосе частот от $-F_в$ до $F_в$), рис. 3.7, *а*. Такой процесс имеет СПМ вида

$$W_x(f) = \begin{cases} N_0/2 & \text{при } |f| < F_в, \\ 0 \text{ в противном случае.} \end{cases}$$

АКФ квазибелого шума согласно теореме Винера – Хинчина имеет вид

$$R_x(\tau) = \frac{N_0}{2}\int_{-F_в}^{F_в}\cos(2\pi f\tau)df = F_в N_0 \frac{\sin(2\pi F_в\tau)}{2\pi F_в t},$$

показанный на рис. 3.7, *б*. Особенность такого процесса заключается в том, что график его АКФ пересекает ось времени в точках, кратных $1/(2F_в)$. Таким образом, дискретизация квазибелого шума



а *б*

Рис. 3.7. СПМ (*а*) и АКФ (*б*) квазибелого шума

с шагом $T_d = 1/(2F_в)$ дает последовательность некоррелированных случайных величин. В частности, если квазибелый шум является гауссовским процессом, то его отсчеты, взятые с шагом T_d, оказываются независимыми, что упрощает анализ (см., например, разд. 9.3).◀

3.4. ВОЗДЕЙСТВИЕ СТАЦИОНАРНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ НА ЛИС-ЦЕПИ

Рассматривая воздействие стационарного (здесь и далее – в широком смысле) случайного процесса на ЛИС-цепь в рамках корреляционно-спектральной теории, достаточно интересоваться только моментами не выше второго порядка. Отсюда следует, что при воздействии ССП $x(t)$ на ЛИС-цепь с КЧХ $H(f)$ и импульсной характеристикой $h(t)$ можно ставить задачу найти среднее значение (математическое ожидание) и АКФ выходного процесса $y(t)$, а также взаимно корреляционные функции $R_{xy}(\tau)$ и $R_{yx}(\tau)$ процессов $x(t)$ и $y(t)$.

В задаче анализа ЛИС-цепи при стационарном случайном воздействии в качестве входного процесса обычно рассматривается процесс с нулевым средним. Если математическое ожидание входного процесса (постоянное вследствие стационарности) отлично от нуля, $m_x \neq 0$, всегда можно рассмотреть прохождение через ЛИС-цепь постоянной и флюктуационной составляющих отдельно. Очевидно, математическое ожидание выходного процесса $m_y = H(0)m_x$. Далее полагаем, что на вход цепи с КЧХ $H(f)$ воздействует стационарный процесс $x(t)$ с нулевым средним и АКФ

$$R_x(\tau) = \int_{-\infty}^{\infty} W_x(f) e^{j\cdot 2\pi f\tau} df \ .$$

Каждая реализация $\eta(t)$ процесса $y(t)$ получается сверткой реализации $\xi(t)$ процесса $x(t)$ с импульсной характеристикой ЛИС-цепи, или, что то же самое, обратным преобразованием Фурье произведения КЧХ цепи $H(f)$ на спектральную плотность $\Xi(f)$ входной реализации $\xi(t)$:

$$\eta(t) = \int_{-\infty}^{\infty} H(f) \left\{ \int_{-\infty}^{\infty} \xi(t_1) e^{-j\cdot 2\pi f t_1} dt_1 \right\} e^{j\cdot 2\pi f t} df =$$

$$= \int_{-\infty}^{\infty} H(f)\Xi(f)e^{j\cdot 2\pi ft}df \ .$$

(Здесь, как и в разд. 3.4, следует иметь в виду, что интеграл в фигурных скобках в классическом смысле расходится.)

Переходя от реализаций к процессам, можно записать

$$y(t) = \int_{-\infty}^{\infty} H(f)X(f)e^{j\cdot 2\pi ft}df = \int_{-\infty}^{\infty} Y(f)e^{j\cdot 2\pi ft}df \ ,$$

где $Y(f)$ – случайная функция частоты (тот же случайный процесс, представленный в другом «базисе»). Заметим, что из $\overline{x(t)} = 0$ следует $\overline{X(f)} = 0$ и далее $\overline{Y(f)} = 0$.

Автокорреляционная функция процесса $y(t)$

$$R_y(\tau) = \overline{y^*(t)y(t+\tau)} =$$

$$= \overline{\int_{-\infty}^{\infty} H^*(f)X^*(f)e^{-j\cdot 2\pi ft}df \int_{-\infty}^{\infty} H(f_1)X(f_1)e^{j\cdot 2\pi f_1 t}e^{j\cdot 2\pi f_1\tau}df_1} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |H(f)|^2 \overline{X(f_1)X^*(f)}e^{j\cdot 2\pi(f_1-f)t}e^{j\cdot 2\pi f_1\tau}df df_1 \ .$$

Учитывая (3.9), запишем

$$R_y(\tau) = \int_{-\infty}^{\infty} |H(f)|^2 W_x(f)e^{j\cdot 2\pi f\tau}df = \int_{-\infty}^{\infty} W_y(f)e^{j\cdot 2\pi f\tau}df \ .$$

Из последнего выражения следует, что отклик ЛИС-цепи на стационарный случайный процесс имеет спектральную плотность мощности, равную входной СПМ, умноженной *на квадрат модуля* КЧХ (т. е. на квадрат АЧХ) цепи:

$$W_y(f) = |H(f)|^2 W_x(f) \ . \qquad (3.12)$$

Это выражение описывает спектральный метод анализа ЛИС-цепей при случайных стационарных воздействиях.

Поскольку частотные функции в (3.12) умножаются, соответствующие временные функции взаимодействуют путем свертки

$$R_y(\tau) = R_h(\tau) * R_x(\tau) \ .$$

Здесь временна́я функция $R_x(\tau)$ соответствует спектральной плотности мощности $W_x(f)$ входного процесса, а

$$R_h(\tau) = \int_{-\infty}^{\infty} |H(f)|^2 e^{j\cdot 2\pi f\tau} df = \int_{-\infty}^{\infty} H(f)H^*(f)e^{j\cdot 2\pi f\tau} df . \quad (3.13)$$

Заметим, что $H(f)$ – преобразование Фурье импульсной характеристики $h(t)$, вещественной по предположению. Тогда $H^*(f)$ соответствует функции $h(-t)$. Действительно, согласно теореме обращения (см. п. 2.10.2) $h(-t) \Leftrightarrow H(-f)$, а для вещественных функций $H(-f) = H^*(f)$. Умножению частотных функций в правой части выражения (3.13) соответствует свертка их временных прообразов, поэтому

$$R_h(\tau) = h(\tau) * h(-\tau) = \int_{-\infty}^{\infty} h(t)h(t+\tau)d\tau ,$$

и функцию $R_h(\tau)$ можно назвать автокорреляционной функцией импульсной характеристики[58]. АКФ импульсной характеристики может быть измерена при помощи коррелометра, подключенного к выходу цепи, если на ее вход подать белый шум с единичной спектральной плотностью мощности. Действительно, при этом $R_x(\tau) = \delta(\tau)$, следовательно, $R_y(\tau) = R_h(\tau)$.

Взаимно корреляционная функция входного и выходного процессов

$$R_{xy}(\tau) = \overline{x^*(t)y(t+\tau)} =$$

$$= \overline{\int_{-\infty}^{\infty} X^*(f)e^{-j\cdot 2\pi ft} df \int_{-\infty}^{\infty} H(f_1)X(f_1)e^{j\cdot 2\pi f_1 t} e^{j\cdot 2\pi f_1\tau} df_1} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(f_1)\overline{X(f_1)X^*(f)}e^{j\cdot 2\pi(f_1-f)t} e^{j\cdot 2\pi f_1\tau} df df_1 =$$

[58] Заметим, что здесь имеется в виду АКФ детерминированной функции.

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(f_1)W_x(f)\delta(f-f_1)e^{j\cdot 2\pi(f_1-f)t}e^{j\cdot 2\pi f_1\tau}dfdf_1 =$$

$$= \int_{-\infty}^{\infty} H(f)W_x(f)e^{j\cdot 2\pi f\tau}df = \int_{-\infty}^{\infty} h(t)R_x(\tau-t)dt\,.$$

Для ВКФ входного и выходного процессов выполняется свойство

$$R_{xy}(\tau) = R_{yx}(-\tau)\,.$$

Анализ распределения шума на выходе цепи в общем случае весьма сложен, однако во многих практически важных случаях выходной процесс можно считать гауссовским. Это предположение оправдано, когда:

1) эффективная ширина спектра входного процесса намного шире, чем полоса пропускания цепи (при этом происходит *нормализация* процесса, так как интеграл Дюамеля можно приближенно представить суммой независимых «прошлых» отсчетов входного процесса с весовыми коэффициентами, равными соответствующим отсчетам импульсной характеристики, причем количество этих независимых отсчетов равно отношению длины ИХ к интервалу корреляции входного процесса, или, что то же самое, отношению полосы частот входного процесса к полосе пропускания цепи; согласно центральной предельной теореме Ляпунова распределение суммы независимых случайных величин стремится к нормальному с увеличением числа слагаемых);

2) на входе ЛИС-цепи гауссовский процесс, причем необязательно широкополосный (при этом значение выходного процесса равно сумме гауссовских случайных величин, которая имеет гауссово распределение независимо от числа слагаемых) .

3.5. БЕЗЫНЕРЦИОННЫЕ НЕЛИНЕЙНЫЕ ПРЕОБРАЗОВАНИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ

Ранее рассматривались линейные преобразования детерминированных колебаний и случайных процессов. Напомним, что линейными называются преобразования, подчиняющиеся принципу суперпозиции (см. разд. 2). В технике связи часто используются цепи, не удовлетворяющие этому принципу, т.е. *нелинейные*.

К сожалению, общей теории нелинейных цепей и их взаимодействия с сигналами, столь же простой, как теория ЛИС-цепей, не существует. Некоторые задачи, связанные с воздействием детерминированных колебаний на нелинейные цепи, будут рассмотрены в разд. 5. Анализ многомерного распределения шума на выходе *нелинейной* цепи в общем случае представляет собой крайне сложную задачу, однако иногда рассматриваемая нелинейная цепь является безынерционной[59], т. е. значение выходного процесса зависит только от значения входного процесса *в этот же момент времени*. Кроме того, во многих практически важных задачах достаточно знать распределение процесса лишь в одном временно́м сечении, т. е. его одномерное распределение (распределение *мгновенного значения* случайного процесса). Таким образом, фактически при этом рассматривается нелинейное преобразование *случайной величины,* анализ которого представляет собой сравнительно простую задачу.

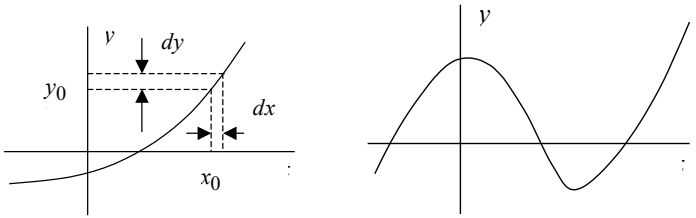
Нелинейная безынерционная цепь описывается характеристикой $y = f(x)$, указывающей зависимость мгновенного значения y выходного процесса $y(t)$ от мгновенного значения x входного процесса $x(t)$. Предположим, что эта зависимость монотонна, (рис. 3.8, *а*). Обозначим одномерную ПРВ мгновенного значения входного процесса $w_x(x)$, а аналогичную функцию для выходного процесса – $w_y(y)$ (индексы указывают, к какой случайной величине относится соответствующая ПРВ). Поскольку вероятность попадания случайной величины x с плотностью $w_x(x)$ в бесконечно узкий интервал $(x_0, x_0 + dx)$ равна $w_x(x_0)dx$ и при этом с такой же вероятностью случайная величина y попадает в интервал $(y_0, y_0 + dy)$, при любом x и соответствующем y выполняется равенство

$$w_x(x)dx = w_y(y)dy\,.$$

Формально отсюда следует равенство

$$w_y(y) = w_x(x)\frac{dx}{dy}\,,$$

[59] Можно сказать, что безынерционная цепь не имеет памяти.



а *б*

Рис. 3.8. К формулам (3.14), (3.15)

но нужно учесть, во-первых, что правая часть, как и левая, должна зависеть только от y, и, во-вторых, что ПРВ не может быть отрицательной. С учетом сказанного можно записать

$$w_y(y) = w_x\left[\varphi(y)\right]\left|\frac{d\varphi(y)}{dy}\right|, \qquad (3.14)$$

где $x = \varphi(y)$ – функция, обратная по отношению к характеристике $f(\cdot)$ нелинейной безынерционной цепи.

Если характеристика цепи не является монотонной и содержит N участков монотонности (рис. 3.8, *б*, $N = 3$), то формула (3.14) приобретает вид

$$w_y(y) = \sum_{k=1}^{N} w_x\left[\varphi_k(y)\right]\left|\frac{d\varphi_k(y)}{dy}\right|, \qquad (3.15)$$

где $\varphi_k(y)$ – функция, обратная к характеристике нелинейной цепи на k-м участке монотонности.

3.6. УЗКОПОЛОСНЫЕ СЛУЧАЙНЫЕ ПРОЦЕССЫ

Представление, аналогичное аналитическому сигналу, существует и для случайных процессов. Оно особенно удобно для описания узкополосных случайных процессов. Узкополосные СП играют чрезвычайно важную роль в теории связи, так как они описывают модулированные сигналы, переносящие информацию, а также узкополосные (сосредоточенные по спектру) помехи, часто имеющие место в каналах связи.

Предположим, что на вход цепи, изображенной на рис. 2.32, *б*, поступает стационарный в широком смысле случайный процесс

$x(t)$ со спектральной плотностью мощности $W_x(f)$. Поскольку АЧХ фильтра Гильберта тождественно равна 1, СПМ сопряженного процесса $\hat{x}(t)$, которую обозначим $W_{\hat{x}}(f)$, равна $W_x(f)$. Значит, равны и автокорреляционные функции этих процессов:

$$R_{\hat{x}}(\tau) = R_x(\tau). \tag{3.16}$$

Рассмотрим комплексный СП

$$z(t) = x(t) + j \cdot \hat{x}(t). \tag{3.17}$$

Поскольку он формируется фильтром с КЧХ (2.62), его СПМ должна быть равна

$$W_z(f) = \begin{cases} 4W_x(f), & f \geq 0, \\ 0, & f < 0. \end{cases}$$

Согласно теореме Винера – Хинчина АКФ комплексного СП

$$R_z(\tau) = 4\int_0^\infty W_x(f) e^{j \cdot 2\pi f \tau} df =$$

$$= 4\int_0^\infty W_x(f)\cos(2\pi f\tau)df + j \cdot 4\int_0^\infty W_x(f)\sin(2\pi f\tau)df. \tag{3.18}$$

Заметим, что в силу четности СПМ вещественного процесса $x(t)$ его АКФ

$$R_x(\tau) = 2\int_0^\infty W_x(f)\cos(2\pi f\tau)df.$$

Тогда

$$R_z(\tau) = 2R_x(\tau) + j \cdot 2R_*(\tau), \tag{3.19}$$

где для мнимой части (3.18) введено обозначение $2R_*(\tau)$.

Вспомним, что спектральная плотность аналитического сигнала является правосторонней (равна 0 при отрицательных частотах). Так как АКФ и СПМ комплексного случайного процесса $z(t)$ также связаны парой преобразований Фурье, то очевидно, что правосторонней СПМ должна соответствовать АКФ, имеющая *вид аналитического сигнала*, причем ее вещественная и мнимая части связаны парой преобразований Гильберта:

$$R_z(\tau) = R_{re}(\tau) + jR_{im}(\tau). \tag{3.20}$$

Сравнивая (3.20) и (3.19), видим, что функция $R_{re}(\tau) = 2R_x(\tau)$ является четной вещественной (это видно и из выражения (3.18), определяющего $R_{re}(\tau)$ суммой косинусоид). Аналогично $R_{im}(\tau)$ – вещественная нечетная функция, как сумма синусоид, каждая из которых сопряжена по Гильберту соответствующей косинусоиде в $R_{re}(\tau)$.

Запишем АКФ комплексного СП $z(t)$:

$$R_z(\tau) = \overline{z(t+\tau)z^*(t)} = \overline{[x(t+\tau) + j\cdot\hat{x}(t+\tau)][x(t) - j\cdot\hat{x}(t)]} =$$

$$= \overline{x(t+\tau)x(t)} + \overline{\hat{x}(t+\tau)\hat{x}(t)} + j\cdot\overline{\hat{x}(t+\tau)x(t)} - j\cdot\overline{x(t+\tau)\hat{x}(t)} =$$

$$= R_x(\tau) + R_{\hat{x}}(\tau) + j[R_{\hat{x}x}(\tau) - R_{x\hat{x}}(\tau)]. \qquad (3.21)$$

Найдем слагаемые мнимой части:

$$R_{x\hat{x}}(\tau) = \overline{x(t+\tau)\hat{x}(t)} = \overline{x(t+\tau)\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(s)}{t-s}ds} =$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\overline{x(t+\tau)x(s)}}{t-s}ds = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R_x(t+\tau-s)}{t-s}ds =$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R_x(\xi)}{\xi-\tau}d\xi = \hat{R}_x(\tau).$$

Полученное выражение представляет функцию, сопряженную по Гильберту автокорреляционной функции исходного процесса.

$$R_{\hat{x}x}(\tau) = \overline{\hat{x}(t+\tau)x(t)} = \overline{\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(s)}{t+\tau-s}ds\cdot x(t)} =$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\overline{x(s)x(t)}}{t+\tau-s}ds = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R_x(s-t)}{\tau-(s-t)}ds =$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R_x(\xi)}{\tau-\xi}d\xi = -\hat{R}_x(\tau).$$

Таким образом, слагаемые мнимой части отличаются только знаком; учитывая это, а также тот факт, что $R_{\hat{x}}(\tau) = R_x(\tau)$ [см. (3.16)], запишем на основании (3.21)

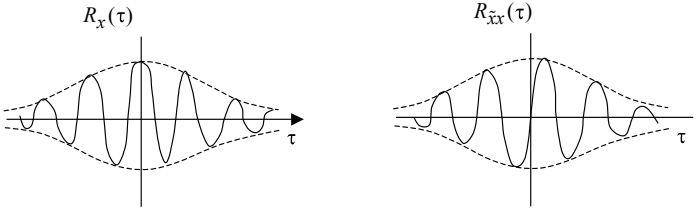
$$R_z(\tau) = 2R_x(\tau) + j \cdot 2R_{\hat{x}x}(\tau).$$

Еще раз отметим, что вещественная часть АКФ комплексного СП $z(t)$ является четной, а мнимая – нечетной функциями. В частности, отсюда следует, что случайные процессы, сопряженные по Гильберту, некоррелированны в совпадающие моменты времени (при $\tau = 0$).

Все сказанное справедливо для *любого* комплексного случайного процесса $z(t)$, определенного выражением (3.17) (необязательно узкополосного). Если же процесс является узкополосным, то для него характерно наличие некоторой средней частоты и медленно меняющейся огибающей (комплексной). Типичный вид мнимой и вещественной частей АКФ комплексного *узкополосного* СП приведен на рис. 3.9. Характерными особенностями являются их колебательный характер, симметрия огибающих, а также одинаковая частота квазигармонического заполнения.

Модель комплексного случайного процесса используется при нахождении плотности распределения вероятностей огибающей узкополосного гауссовского случайного процесса, которая необходима, в частности, для анализа качества приема (демодуляции) амплитудно-модулированных сигналов на фоне шума. Подробно задача оптимального приема сигналов в присутствии помех рассматривается в разд. 9.

Предположим, что $x(t)$ – гауссовский узкополосный СП с нулевым средним (это предположение на практике обычно выполняется, так как случайный процесс на входе демодулятора представляет собой результат полосовой фильтрации входного широкополосного шума, а при этом происходит его нормализация [8]).



а *б*

Рис. 3.9. Вещественная и мнимая части АКФ комплексного СП

В совпадающие моменты времени значения процессов $x(t)$ и $\hat{x}(t)$ некоррелированны, а следовательно, в силу гауссовости, и независимы. Кроме того, они имеют одинаковую дисперсию, поэтому можно записать совместную плотность распределения вероятностей отсчетов x и y процессов $x(t)$ и $\hat{x}(t)$ в некоторый момент времени t_0:

$$w(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Рассматривая x и y как декартовы координаты точки на плоскости, введем элементарную площадку $dxdy$, которая в полярных координатах выражается, как $dA \cdot Ad\Phi$, где A – длина радиуса, Φ – угол. Вероятность того, что точка с координатами x и y попадает в элементарную площадку $dxdy$, равна

$$\mathbf{P}\{x < \xi \le x + dx,\ y < \eta \le y + dy\} = w(x, y)dxdy = W(A, \Phi)dAd\Phi,$$

где $W(A,\Phi)$ – совместная плотность распределения вероятностей огибающей A и начальной фазы Φ комплексного случайного процесса в момент времени t_0. Исходя из этого очевидного равенства и выражая x и y через A и Φ, можно записать

$$W(A, \Phi) = \frac{w\left[x(A, \Phi),\ y(A, \Phi)\right] AdAd\Phi}{dAd\Phi} =$$

(здесь $x(A, \Phi)$, $y(A, \Phi)$ – обратные функции, описывающие преобразование полярных координат в декартовы)

$$= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} A = \frac{A}{2\pi\sigma^2} e^{-\frac{A^2}{2\sigma^2}} = \frac{A}{\sigma^2} e^{-\frac{A^2}{2\sigma^2}} \frac{1}{2\pi} = W_A(A)W_\Phi(\Phi).$$

Из полученного выражения видно, что огибающая и начальная фазы в некоторый фиксированный момент времени представляют собой независимые случайные величины с плотностями распределения вероятностей

$$W_A(A) = \frac{A}{\sigma^2} e^{-\frac{A^2}{2\sigma^2}} \qquad (3.22)$$

и

$$W_{\Phi}(\Phi) = \frac{1}{2\pi}.$$

Плотность (3.22) называется *рэлеевской*[60] (рис. 3.10, кривая *1*). Начальная фаза имеет равномерное в интервале $(0, 2\pi)$ распределение. Если случайный процесс имеет ненулевое математическое ожидание (в задаче анализа помехоустойчивости приема сигнала на фоне шума это соответствует присутствию полезного сигнала в принимаемом колебании), то распределение вероятностей огибающей становится более сложным и принимает вид *обобщенного распределения Рэлея*, или распределения Рэлея – Райса с плотностью

$$W_A(A) = \frac{A}{\sigma^2} e^{-\frac{A^2+U^2}{2\sigma^2}} I_0\left(\frac{AU}{\sigma^2}\right),$$

где U – амплитуда сигнала, $I_0(\cdot)$ – модифицированная функция Бесселя нулевого порядка. Плотность обобщенного распределения Рэлея показана на рис. 3.10, кривые *2*, *3* ($A = 1, \sigma = 1$).

Распределение начальной фазы для этого случая не является равномерным; его точный вид достаточно сложен и здесь не рассматривается.

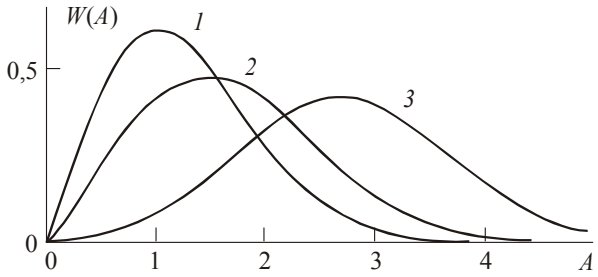


Рис. 3.10. Распределения Рэлея (кривая *1*) и Рэлея – Райса при $U = 1,2$ (кривая *2*) и при $U = 2,5$ (кривая *3*)

[60] Джон Уильям Стретт, лорд Рэлей (1842 – 1919) – знаменитый английский физик, известный трудами в области теории колебаний и др.; нобелевский лауреат 1904 г.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какое из утверждений правильно: 1) случайная величина – это число; 2) случайная величина – это функция? Почему?
2. Перечислите свойства функции распределения и плотности распределения вероятностей случайной величины.
3. Что такое моменты распределения СВ?
4. Как определить энтропию распределения СВ?
5. Может ли дисперсия СВ равняться нулю?
6. Может ли математическое ожидание СВ равняться нулю?
7. Что такое корреляция и как она связана с ковариацией?
8. Что является полным описанием случайного процесса?
9. Что получится, если N-мерную совместную плотность распределения вероятностей случайного процесса проинтегрировать по $(N-1)$ переменным?
10. Какой процесс называется стационарным в узком смысле?
11. Какой процесс называется стационарным в широком смысле?
12. Утверждение: «Если процесс стационарен в узком смысле, то он стационарен и в широком смысле» – верно? неверно? верно для некоторых процессов (каких именно)? неверно для некоторых процессов (каких именно)?
13. Утверждение: «Если процесс стационарен в широком смысле, то он стационарен и в узком смысле» – верно? неверно? верно для некоторых процессов (каких именно)? неверно для некоторых процессов (каких именно)?
14. Утверждение: «Если случайные величины независимы, то они и некоррелированны» – верно? неверно? верно для некоторых распределений (каких именно)? неверно для некоторых распределений (каких именно)?
15. Утверждение: «Если случайные величины некоррелированны, то они и независимы» – верно? неверно? верно для некоторых распределений (каких именно)? неверно для некоторых распределений (каких именно)?
16. Какие процессы называются эргодическими?
17. Как связаны АКФ и СПМ стационарного случайного процесса?
18. Что такое белый шум? Какова его АКФ? Какова его СПМ?
19. Как связана СПМ процесса на выходе ЛИС-цепи с СПМ процесса на входе?
20. Какую ПРВ имеют огибающая и фаза гауссовского узкополосного случайного процесса?

УПРАЖНЕНИЯ

1. Докажите, что процесс, стационарный в узком смысле, стационарен и в широком смысле.

2. Реализации случайного процесса представляют собой функции времени $\cos(\omega t + \phi)$, $t \in (-\infty, \infty)$ при некотором фиксированном значении ω и случайной начальной фазе ϕ, имеющей распределение, равномерное в интервале $(0, 2\pi)$. Проверьте, является ли этот процесс стационарным и эргодическим.

3. Мгновенное значение случайного процесса имеет распределение вероятностей с плотностью вида $w(x) = C \cdot \exp(-0,5|x|)$. Найдите константу C, математическое ожидание и дисперсию. Постройте графики плотности распределения вероятностей и функции распределения вероятностей (друг под другом в одном масштабе).

4. Мгновенное значение случайного процесса описывается функцией распределения $F(x) = 1 - \exp(-2x),\ x \geq 0$. Найдите плотность распределения вероятностей, математическое ожидание, дисперсию. Постройте графики плотности распределения вероятностей и функции распределения вероятностей (друг под другом в одном масштабе).

5. Функция автокорреляции стационарного случайного процесса имеет вид $B_x(\tau) = D_x e^{-a|\tau|}$, где a – некоторая постоянная. Найдите спектральную плотность мощности случайного процесса.

6. Спектральная плотность мощности окрашенного шума $x(t)$ имеет вид $G_x(f) = Ce^{-a|f|}$, где C и a – некоторые постоянные. Найдите АКФ процесса.

4. МЕТОДЫ АНАЛИЗА ЛИС-ЦЕПЕЙ

Задача анализа состоит в нахождении выходного сигнала по заданному входному сигналу и известному описанию цепи. Не существует метода решения этой задачи, подходящего для всех цепей – линейных и нелинейных, стационарных и нестационарных. Наиболее хорошо развиты в настоящее время методы анализа линейных инвариантных к сдвигу цепей. Невыполнение требований линейности и стационарности (или хотя бы одного из них) приводит к большим трудностям в решении задачи анализа.

Как было показано в разд. 2, сигнал на выходе ЛИС-цепи может быть найден точно, если известно описание цепи в форме импульсной или комплексной частотной характеристики (функциональное описание [14]). Часто цепь описывается другими (структурными) способами (принципиальной схемой, дифференциальным уравнением и т.п.). Поскольку любая форма исчерпывающего описания определяет *одну и ту же* цепь, существует внутренняя связь между этими формами, и выяснение этой связи входит в задачу анализа[61]. В этом смысле все точные методы анализа эквивалентны друг другу. В основе всякого приближенного метода лежит какое-нибудь упрощающее предположение, поэтому такие методы всегда приводят к *различным* решениям, однако они должны мало[62] отличаться от точного решения.

Точными методами анализа ЛИС-цепей являются метод, основанный на решении дифференциального уравнения цепи, операторный и спектральный методы, метод комплексной огибающей.

---

[61] Согласно [14] анализ состоит в нахождении функционального описания цепи по известному ее структурному описанию; нахождение структурного описания по заданному функциональному составляет существо задачи *синтеза.*

[62] В каждом конкретном случае нужно определить количественную меру отличия и количественный критерий малости.

Они являются точными и универсальными в том смысле, что позволяют *в принципе* решить задачу анализа точно при любой ЛИС-цепи и любом воздействии. Однако во многих практически важных случаях точное решение оказывается слишком трудоемким. В то же время некоторые дополнительные сведения о сигнале и цепи могут существенно упростить анализ и привести к результату хотя и приближенному, но достаточно близкому к точному решению с практической точки зрения. К приближенным относятся метод мгновенной частоты (см. п. 5.5.2) и некоторые другие. В этом разделе кратко рассматриваются точные методы анализа ЛИС-цепей.

4.1. МЕТОД, ОСНОВАННЫЙ НА РЕШЕНИИ ДИФФЕРЕНЦИАЛЬНОГО УРАВНЕНИЯ

Дифференциальные уравнения вообще связывают значения некоторых физических величин со скоростями их изменения, скоростями изменения скоростей (ускорениями) и т.д. Эти связи, выраженные в форме дифференциальных уравнений, отражают объективные физические законы, которым подчиняется реальный мир. Линейные стационарные цепи с сосредоточенными параметрами описываются наиболее простыми дифференциальными уравнениями – обыкновенными линейными дифференциальными уравнениями с постоянными коэффициентами[63] вида

$$a_n \frac{d^n y(t)}{dt^n} + ... + a_0 y(t) = b_m \frac{d^m x(t)}{dt^m} + ... + b_0 x(t), \qquad (4.1)$$

где $x(t)$ – входной сигнал, $y(t)$ – выходной сигнал, а целые числа n и m определяются сложностью цепи. Если входной сигнал задан, то тем самым задана вся правая часть уравнения, которую можно обозначить $f(t)$. Тогда уравнение можно записать в виде

$$a_n \frac{d^n y(t)}{dt^n} + ... + a_0 y(t) = f(t), \qquad (4.2)$$

при этом число n определяет порядок дифференциального уравнения. Знание уравнения, описывающего цепь, а также состояния цепи[64] в начальный момент времени (начальных условий) позволя-

[63] Цепи *с распределенными параметрами* описываются дифференциальными уравнениями в частных производных.

[64] Состояние цепи определяется набором величин $y(t)$, $dy(t)/dt$, ..., $dy^n(t)/dt^n$.

ет найти состояние цепи в любой будущий момент времени (цепь считается каузальной, т.е. ее поведение не зависит от будущих значений входного и выходного сигналов). Уравнение (4.2) имеет ненулевую правую часть и называется неоднородным; его решение представляет собой сумму некоторого *частного* решения неоднородного уравнения и *общего* решения однородного уравнения

$$a_n \frac{d^n y(t)}{dt^n} + ... + a_0 y(t) = 0 . \qquad (4.3)$$

Для решения однородного дифференциального уравнения (4.3) нужно решить алгебраическое характеристическое уравнение цепи

$$a_n \gamma^n + a_{n-1}\gamma^{n-1} + ... + a_1\gamma + a_0 = 0 . \qquad (4.4)$$

Если коэффициенты уравнения вещественны (а это практически всегда так), то корни либо вещественны, либо образуют комплексно-сопряженные пары. При этом некоторые корни могут совпадать (быть кратными). Для случая, когда все корни $\gamma_1, \gamma_2, ..., \gamma_n$ являются различными (простыми), общее решение однородного дифференциального уравнения (4.3) описывает собственные колебания цепи и имеет вид

$$y(t) = C_1 e^{\gamma_1 t} + C_2 e^{\gamma_2 t} + ... + C_n e^{\gamma_n t} ,$$

где постоянные $C_1, C_2, ..., C_n$ определяются начальными условиями. В случае кратных корней в решении присутствуют с соответствующими весовыми коэффициентами слагаемые вида $e^{\gamma_k t}$, $te^{\gamma_k t}, ..., t^{m-1}e^{\gamma_k t}$, где γ_k – корень уравнения (4.4) кратности m. Заметим, что для *устойчивости* цепи свободные колебания должны затухать со временем, а отсюда следует, что все корни характеристического уравнения должны иметь отрицательные вещественные части (лежать в левой половине комплексной плоскости).

Метод анализа, основанный на решении дифференциального уравнения, может использоваться для нахождения откликов несложных ЛИС-цепей на простые воздействия (например, функцию включения, и т.п.). В более сложных случаях этот метод применяется редко.

Пример 4.1. Для *RC*-фильтра нижних частот (см. пример 2.17), обозначая ток, протекающий через емкость, $i(t)$, входное и выходное

напряжения $u_{\text{вх}}(t)$ и $u_{\text{вых}}(t)$ и учитывая, что $i(t) = Cdu_{\text{вых}}(t)/dt$, запишем неоднородное дифференциальное уравнение

$$u_{\text{вх}}(t) = RCdu_{\text{вых}}(t)/dt + u_{\text{вых}}(t).$$

Соответствующее однородное уравнение имеет вид $\tau du_{\text{вых}}(t)/dt + u_{\text{вых}}(t) = 0$, где $\tau = RC$ – постоянная времени, и характеристическое уравнение $\tau\gamma + 1 = 0$. Единственный корень характеристического уравнения $\gamma = -1/\tau$, и общее решение однородного уравнения, описывающее свободные колебания на выходе цепи, имеет вид $u_{\text{вых}}(t) = Ce^{-t/\tau}$, где C определяется некоторым частным решением неоднородного уравнения и начальными условиями. Если входное напряжение описывается функцией включения, частным решением неоднородного уравнения является установившаяся реакция цепи, равная, очевидно, функции включения. Тогда решение неоднородного уравнения имеет вид

$$u_{\text{вых}}(t) = Ce^{-t/\tau} + \sigma(t).$$

Начальным условием для уравнения является условие $u_{\text{вых}}(0) = 0$, откуда $C = -1$, таким образом, отклик цепи на функцию включения, называемый *переходной характеристикой*, равен при $t \geq 0$ $g(t) = 1 - e^{-t/\tau}$. ◄

4.2. СПЕКТРАЛЬНЫЙ МЕТОД

Предположим, что на вход ЛИС-цепи, описываемой дифференциальным уравнением (4.1), воздействует комплексное гармоническое колебание вида $x(t) = e^{j\omega t}$, где ω – круговая частота колебания. Как было показано в разд. 2, сигнал на выходе ЛИС-цепи при таком воздействии равен $y(t) = H(\omega)e^{j\omega t}$, где $H(\omega)$ – значение КЧХ на частоте ω. Подставляя $x(t)$ и $y(t)$ в уравнение (4.1) и решая полученное алгебраическое уравнение относительно $H(\omega)$, получаем комплексную частотную характеристику, выраженную через коэффициенты дифференциального уравнения цепи:

$$H(\omega) = \frac{b_m (j\omega)^m + b_{m-1}(j\omega)^{m-1} + \ldots + b_1 j\omega + b_0}{a_n (j\omega)^n + a_{n-1}(j\omega)^{n-1} + \ldots + a_1 j\omega + a_0}.$$

Таким образом, зная дифференциальное уравнение, описывающее ЛИС-цепь, можно для нахождения выходного сигнала воспользоваться *спектральным* методом, для чего найти спектральную плотность $X(\omega)$ входного сигнала, умножить ее на КЧХ, а от полученной спектральной плотности $Y(\omega)$ выходного сигнала перейти к его временному представлению путем обратного преобразования Фурье.

Пример 4.2. КЧХ фильтра нижних частот, рассмотренного в примере 4.1, равна, очевидно, $H(\omega)=1/(1+j\omega RC)$. Рассматривая в качестве входного сигнала δ-функцию со спектральной плотностью, тождественно равной 1, имеем спектральную плотность выходного сигнала $Y(\omega)=1/(1+j\omega RC)$, откуда обратным преобразованием Фурье находится импульсная характеристка $h(t)=\frac{1}{\tau}e^{-t/\tau}$, $t\geq 0$, равная производной переходной характеристики. ◄

4.3. ОПЕРАТОРНЫЙ МЕТОД

Операторный метод[65] основан на символической замене оператора дифференцирования множителем, который принято обозначать буквой p. При этом функции времени $x(t)$ и $y(t)$ должны быть заменены их изображениями $X(p)$ и $Y(p)$, определяемыми преобразованием[66] Лапласа[67]

$$X(p)=\int_0^\infty x(t)e^{-pt}dt, \quad Y(p)=\int_0^\infty y(t)e^{-pt}dt.$$

При таком допущении дифференциальное уравнение (4.1) преобразуется в алгебраическое уравнение

$$a_n p^n Y(p)+...+a_1 pY(p)+a_0 Y(p)=b_m p^m X(p)+...+b_1 pX(p)+b_0 X(p).$$

---

[65] Операторный метод был в основном разработан Оливером Хевисайдом (1850 – 1925), выдающимся английским инженером и математиком (подробнее см., например, [14]).

[66] Это *одностороннее* преобразование Лапласа определено для функций, равных нулю при $t<0$ и растущих при $t\to\infty$ не быстрее чем $e^{\sigma_0 t}$, при $\sigma>\sigma_0$ [14].

[67] Пьер Симон Лаплас (1749 – 1827) – знаменитый французский математик, физик и астроном, один из создателей теории вероятностей.

Вводя операторную передаточную функцию $K(p)$ формулой

$$K(p)=\frac{Y(p)}{X(p)},$$

получаем

$$K(p)=\frac{b_m p^m + b_{m-1}p^{m-1} + ... + b_1 p + b_0}{a_n p^n + a_{n-1}p^{n-1} + ... + a_1 p + a_0}.$$

Таким образом, нахождение выходного сигнала ЛИС-цепи *операторным* методом сводится к следующим шагам:

- переход от входного сигнала $x(t)$ к его лапласовскому изображению $X(p)$;
- нахождение изображения выходного сигнала $Y(p)=K(p)X(p)$;
- переход от изображения $Y(p)$ к выходному сигналу $y(t)$ путем обратного преобразования Лапласа

$$y(t)=\frac{1}{2\pi j}\int_{c-j\infty}^{c-j\infty} Y(p)e^{pt}dt,$$

где c – константа, которая должна быть больше абсциссы абсолютной сходимости [8].

Легко видеть, что передаточная функция ЛИС-цепи получается из ее КЧХ путем замены $j\omega \to p=\sigma+j\omega$ (что соответствует *аналитическому продолжению* функции, заданной на мнимой оси комплексной плоскости, на всю комплексную плоскость). Верно и обратное: зная передаточную функцию ЛИС-цепи, путем замены $p \to j\omega$ (*сужения* функции комплексного переменного на мнимую ось комплексной p-плоскости) можно получить КЧХ: $H(\omega)=K(j\omega)$. Учитывая, что для многих функций лапласовские изображения давно найдены и сведены в таблицы, операторный метод анализа ЛИС-цепей оказывается во многих случаях самым простым. Подробное обсуждение операторного метода с многочисленными примерами можно найти в [14].

Пример 4.3. Передаточная функция *RC*-фильтра нижних частот равна $K(p)=\frac{1}{1+p\tau}=\frac{1/\tau}{p+1/\tau}$, а изображение δ-функции равно 1.

Перемножая, получим $Y(p)=\dfrac{1/\tau}{p+1/\tau}$. В таблице преобразований Лапласа [14] есть изображение $1/(p+\alpha)\Leftrightarrow e^{-\alpha t}$. Полагая $\alpha=1/\tau$, получаем отклик на δ-функцию (импульсную характеристику) вида $h(t)=\dfrac{1}{\tau}e^{-t/\tau}$. ◄

4.4. МЕТОД КОМПЛЕКСНОЙ ОГИБАЮЩЕЙ

Метод комплексной огибающей обычно применяется для анализа частотно-избирательных цепей (ЧИЦ) при узкополосных воздействиях. Именно эта ситуация имеет место в приемных устройствах, где модулированные узкополосные колебания воздействуют на частотно-избирательные ЛИС-цепи (фильтры).

Узкополосный сигнал $x(t)$ со средней частотой F_0 можно выразить через его комплексную огибающую $\gamma(t)$ следующим образом (см. разд. 2):

$$x(t)=\mathrm{Re}\left\{\gamma(t)e^{j\cdot 2\pi F_0 t}\right\}=\frac{1}{2}\left[\gamma(t)e^{j\cdot 2\pi F_0 t}+\gamma^*(t)e^{-j\cdot 2\pi F_0 t}\right],$$

поэтому спектральная плотность $X(f)$ сигнала $x(t)$ может быть записана в виде

$$X(f)=\frac{1}{2}\left[\Gamma\left(f-F_0\right)+\Gamma^*\left(-f-F_0\right)\right], \qquad (4.5)$$

где $\Gamma(f)$ – спектральная плотность комплексной огибающей $\gamma(t)$.

Импульсная характеристика $h(t)$ частотно-избирательной цепи (полосового фильтра с центральной частотой F_0), рассматриваемая как сигнал, также имеет узкополосный характер и может быть представлена в форме

$$h(t)=\mathrm{Re}\left\{\lambda(t)e^{j\cdot 2\pi F_0 t}\right\}=\frac{1}{2}\left[\lambda(t)e^{j\cdot 2\pi F_0 t}+\lambda^*(t)e^{-j\cdot 2\pi F_0 t}\right],$$

поэтому КЧХ такой цепи можно записать в виде

$$H(f)=\frac{1}{2}\left[\Lambda\left(f-F_0\right)+\Lambda^*\left(-f-F_0\right)\right]. \qquad (4.6)$$

Тогда спектральная плотность $Y(f)$ сигнала $y(t)$ на выходе фильтра равна

$$Y(f) = H(f)X(f) =$$

$$= \frac{1}{4}\left[\Lambda(f - F_0)\Gamma(f - F_0) + \Lambda^*(-f - F_0)\Gamma^*(-f - F_0)\right]. \quad (4.7)$$

Здесь учтено то обстоятельство, что функции $\Gamma(f - F_0)$ и $\Lambda(f - F_0)$ равны нулю при $f < 0$, а функции $\Lambda^*(-f - F_0)$ и $\Gamma^*(-f - F_0)$ при $f > 0$.

Для приведения (4.7) к виду, аналогичному выражениям (4.5), (4.6), запишем

$$Y(f) = \frac{1}{2}\left[\mathrm{B}(f - F_0) + \mathrm{B}^*(-f - F_0)\right],$$

где $\mathrm{B}(f)$ – спектральная плотность комплексной огибающей $\beta(t)$ выходного сигнала

$$y(t) = \mathrm{Re}\left\{\beta(t)e^{j2\pi F_0 t}\right\} = \frac{1}{2}\left[\beta(t)e^{j2\pi F_0 t} + \beta^*(t)e^{-j2\pi F_0 t}\right]. \quad (4.8)$$

Заметим, что при этом $\mathrm{B}(f) = \frac{1}{2}\Lambda(f)\Gamma(f)$, откуда следует, что комплексная огибающая выходного сигнала может быть найдена как свертка

$$\beta(t) = \gamma(t) * \frac{1}{2}\lambda(t)$$

комплексной огибающей $\gamma(t)$ входного сигнала и (комплексной) импульсной характеристики $\frac{1}{2}\lambda(t)$. Цепь с такой ИХ называется низкочастотным эквивалентом частотно-избирательной цепи. Заметим, что с точностью до множителя $\frac{1}{2}$ импульсная характеристика НЧ эквивалента совпадает с комплексной огибающей импульсной характеристики ЧИЦ.

Таким образом, для нахождения отклика ЧИЦ на узкополосный сигнал достаточно вычислить свертку (интеграл Дюамеля) комплексных огибающих входного сигнала и импульсной характеристики фильтра, умножить результат на $\frac{1}{2}$ и затем найти выходной сигнал согласно выражению (4.8).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Где должны располагаться корни характеристического уравнения устойчивой цепи?
2. Как связана комплексная частотная характеристика ЛИС-цепи с ее передаточной функцией?
3. Как связаны характеристики частотно-избирательной цепи и ее низкочастотного эквивалента?

УПРАЖНЕНИЯ

1. Составьте дифференциальное уравнение RC-фильтра нижних частот (интегрирующей цепочки). Выведите КЧХ и передаточную функцию цепи. Найдите импульсную характеристику.
2. Составьте дифференциальное уравнение RC-фильтра верхних частот (дифференцирующей цепочки). Выведите КЧХ и передаточную функцию цепи. Найдите импульсную характеристику цепи.
3. Найдите низкочастотный эквивалент колебательного контура, если его комплексное сопротивление описывается выражением $Z(j\omega) = \dfrac{R_0}{1 + j\left(2Q\dfrac{\omega - \omega_0}{\omega_0}\right)}$, где R_0 – сопротивление на резонансной частоте ω_0, Q – добротность контура.

5. ПРИНЦИПЫ МОДУЛЯЦИИ И ДЕМОДУЛЯЦИИ

Модуляция – это изменение *одного или нескольких* параметров колебания, называемого несущим колебанием (переносчиком), в соответствии с изменениями первичного (информационного) сигнала. При этом спектр (спектральная плотность) получаемого модулированного сигнала отличается от спектров как первичного сигнала, так и переносчика. Можно сказать, что такое изменение спектра является *целью* модуляции: например, речевой сигнал в системах телефонии занимает полосу частот от 300 до 3400 Гц и его непосредственная передача по каналу радиосвязи невозможна, так как размеры антенны, эффективно излучающей радиоволны столь низких частот, были бы слишком велики для практического применения. В результате амплитудной модуляции таким сигналом гармонического несущего колебания с частотой, например, 1 МГц получается амплитудно-модулированный (АМ) сигнал, занимающий полосу частот от 996 600 до 1 003 400 Гц, излучение которого не составляет проблемы.

Важно отметить, что при модуляции (а также демодуляции) происходят такие преобразования первичного сигнала, которые сопровождаются появлением *новых* частотных составляющих, отсутствовавших в спектре исходного сигнала. Практически во всех случаях после такого обогащения спектра (ОС) производится частотная фильтрация (ЧФ) при помощи подходящей ЛИС-цепи для подавления ненужных или вредных спектральных составляющих, (рис. 5.1). При помощи одних только ЛИС-цепей модуляцию осуществить невозможно. То же относится к демодуляции[68].

[68] Специальный случай демодуляции ЛИС-цепью будет особо рассмотрен позднее (разд. 5.7).

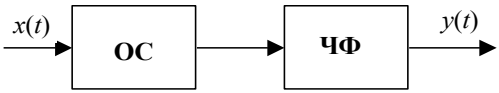
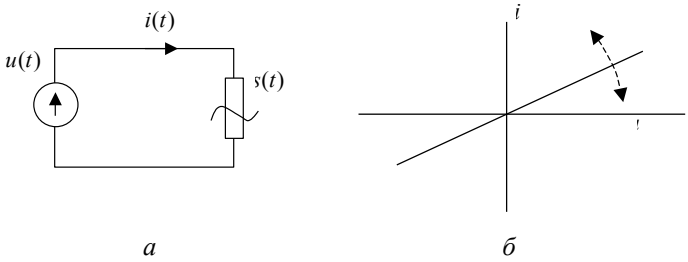


Рис. 5.1. Принцип преобразования спектра колебания

Действительно, предположим, что на вход ЛИС-цепи поступает периодический сигнал $x(t) = \sum_{k=-\infty}^{\infty} C_k e^{j\frac{2\pi}{T}kt}$. Действие ЛИС-цепи на комплексную экспоненту сводится к ее умножению на комплексное число, равное значению КЧХ цепи на частоте данной экспоненты. Ясно, что спектральные составляющие исходного сигнала могут *исчезать* в результате такой фильтрации, но не появляться, если их изначально не было. Очевидно, что и для непериодических сигналов справедливо то же самое. Обогащение спектра сигнала новыми частотами возможно при использовании нелинейных или линейных нестационарных (*параметрических*) цепей. Напомним, что нелинейными называются цепи, для которых не выполняется принцип суперпозиции. Для линейных нестационарных цепей указанный принцип выполняется, однако оператор цепи зависит от времени, вследствие чего в спектре сигнала могут появляться новые частоты.

5.1. ВОЗДЕЙСТВИЕ ГАРМОНИЧЕСКОГО КОЛЕБАНИЯ НА ПАРАМЕТРИЧЕСКУЮ ЦЕПЬ

Простейшая линейная параметрическая цепь, которую можно использовать для обогащения спектра, представляет собой активное сопротивление, меняющееся во времени по некоторому периодическому закону. Удобнее в качестве изменяющегося параметра рассмотреть переменную проводимость $s(t)$, так что под воздействием напряжения $u(t)$ через параметрический элемент протекает ток $i(t) = u(t)s(t)$ (рис. 5.2, *а*). Проводимость $s(t)$ можно трактовать, как переменную крутизну линейной вольт-амперной характеристики параметрического элемента (рис. 5.2, *б*).



а *б*

Рис. 5.2. Воздействие переменного напряжения на линейный параметрический элемент

Рассмотрим простейшую ситуацию, когда напряжение и крутизна изменяются по гармоническому закону с разными частотами

$$u(t) = U\cos(\omega_1 t + \phi_1),$$

$$s(t) = S_0 + S_1\cos(\omega_2 t + \phi_2).$$

Очевидно, спектр напряжения содержит одну гармоническую составляющую (гармонику) с частотой ω_1, а спектр функции $s(t)$ – две составляющие с частотами 0 и ω_2.

Ток, протекающий через параметрический элемент, равен сумме гармонических составляющих (гармоник), амплитуды и начальные фазы которых образуют соответственно амплитудный и фазовый спектры:

$$i(t) = US_0\cos(\omega_1 t + \phi_1) + US_1\cos(\omega_1 t + \phi_1)\cos(\omega_2 t + \phi_2) =$$

$$= US_0\cos(\omega_1 t + \phi_1) + \frac{US_1}{2}\cos\left[(\omega_1 + \omega_2)t + \phi_1 + \phi_2\right] +$$

$$+ \frac{US_1}{2}\cos\left[(\omega_1 - \omega_2)t + \phi_1 - \phi_2\right]. \qquad (5.1)$$

Как видно из полученного выражения, в спектре тока присутствуют гармонические составляющие с частотой ω_1, а также с суммарной $\omega_1 + \omega_2$ и разностной $\omega_1 - \omega_2$ частотами. Очевидно, что при более сложном спектральном составе напряжения и/или крутизны количество новых частот будет больше (в любом случае спектральные коэффициенты можно найти по тригонометрическим формулам, раскрывая произведения косинусов и/или синусов).

Наличие в спектре колебания составляющих с суммарной и разностной частотами позволяет использовать параметрические цепи для *переноса спектра*. В самом деле, подавая ток, описываемый выражением (5.1), на частотно-избирательную нагрузку (полосовой фильтр), получим напряжение частоты $\omega_1 + \omega_2$ или $\omega_1 - \omega_2$, в зависимости от настройки фильтра. Таким образом, получаем перенос частоты ω_1 на величину ω_2 вправо или влево по частотной оси.

На практике сигнал, подлежащий преобразованию, имеет спектр конечной ширины; после умножения сигнала на $s(t)$ при помощи фильтра выделяется спектр такой же формы, но сдвинутый по частоте на ω_2 вверх или вниз. Частными случаями переноса спектра являются преобразование частоты, применяемое при супергетеродинном приеме (см. пример 5.1), а также амплитудная модуляция и синхронное детектирование АМ-сигналов.

5.2. НЕЛИНЕЙНЫЕ ЭЛЕМЕНТЫ И ИХ АППРОКСИМАЦИИ

К нелинейным элементам, наиболее широко применяемым в технике генерирования и обработки сигналов, относятся в первую очередь полупроводниковые приборы (диоды, транзисторы и т.п.), которые описываются характеристиками (чаще всего рассматриваются вольт-амперные характеристики – ВАХ), имеющими весьма сложный вид. Для целей анализа эти характеристики аппроксимируют математическими зависимостями, которые должны быть достаточно простыми и в то же время сохранять существенные черты аппроксимируемых характеристик. Рассмотрим наиболее часто применяемые аппроксимации.

5.2.1. ПОЛИНОМИАЛЬНАЯ (СТЕПЕННАЯ) АППРОКСИМАЦИЯ

Характеристика нелинейного элемента (НЭ) представляется полиномом [13] некоторой степени N

$$i = f(u) = \sum_{k=0}^{N} a_k u^k = a_0 + a_1 u + a_2 u^2 \ldots + a_N u^N . \qquad (5.2)$$

Во всех практических случаях функция $f(u)$ аппроксимирует истинную ВАХ (заданную графически или таблично) не на всей

числовой оси, а только на некотором ограниченном интервале значений независимой переменной, или рабочем участке. Выберем на этом участке $N+1$ точек, обозначив их $u_1, u_2, ..., u_{N+1}$. Для каждого из этих значений напряжения обозначим соответствующие значения тока (взятые из таблицы или найденные по заданному графику ВАХ), как $i_1, i_2, ..., i_{N+1}$. Тогда согласно (5.2) можно составить $N+1$ уравнение

$$\left.\begin{array}{c} a_0 + a_1 u_1 + a_2 u_1^2 ... + a_N u_1^N = i_1, \\ a_0 + a_1 u_2 + a_2 u_2^2 ... + a_N u_2^N = i_2, \\ \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots, \\ a_0 + a_1 u_{N+1} + a_2 u_{N+1}^2 ... + a_N u_{N+1}^N = i_{N+1} \end{array}\right\}$$

с $N+1$ неизвестными $a_0, a_1, ..., a_N$. Решив эту линейную систему уравнений, можно найти полиномиальную функцию (5.2).

Во многих реальных задачах удобно рассматривать четную и нечетную части характеристики. Любую нелинейную характеристику можно, очевидно, представить в виде суммы четной и нечетной функций

$$f(u) = f_{\text{ч}}(u) + f_{\text{н}}(u), \tag{5.3}$$

где четная и нечетная части удовлетворяют следующим выражениям:

$$f_{\text{ч}}(u) = f_{\text{ч}}(-u), \tag{5.4}$$

$$f_{\text{н}}(u) = -f_{\text{н}}(-u). \tag{5.5}$$

Из выражения (5.3) с учетом (5.4), (5.5) следует

$$f(-u) = f_{\text{ч}}(u) - f_{\text{н}}(u), \tag{5.6}$$

откуда, складывая и вычитая выражения (5.3) и (5.6), получаем

$$f_{\text{ч}}(u) = \frac{f(u) + f(-u)}{2}, \tag{5.7}$$

$$f_{\text{н}}(u) = \frac{f(u) - f(-u)}{2}. \tag{5.8}$$

Пример функции $F(u)$ и ее четной и нечетной частей приведен на рис. 5.3.

Очевидно, при полиномиальной аппроксимации ВАХ ее четная и нечетная части складываются из четных и нечетных степеней:

$$f_{\text{ч}}(u) = a_0 + a_2u^2 + a_4u^4 \ldots,$$

$$f_{\text{н}}(u) = a_1u + a_3u^3 + a_5u^5 \ldots.$$

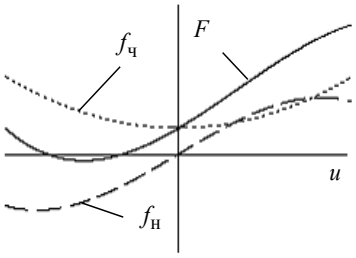


Рис. 5.3. Нелинейная функция и ее четная и нечетная части

Выделение четной и нечетной частей ВАХ полезно, когда работа рассматриваемого устройства определяется либо только четной, либо только нечетной частью. Например, при амплитудной модуляции полезная составляющая колебания определяется только четной частью ВАХ (см. разд. 5.4.2). Тогда аппроксимацию можно проводить только для четной части, предварительно выделив ее графически или по таблице значений, при этом расчет коэффициентов аппроксимации требует решения системы уравнений меньшего порядка. Кроме того, вид нужной (четной или нечетной) части ВАХ позволяет оценить «на глаз» пригодность данного НЭ для реализации требуемого преобразования сигналов.

Важность выделения четной и нечетной частей ВАХ не ограничивается только теоретическим рассмотрением, так как существуют схемные решения, позволяющие получать четные или нечетные характеристики нелинейных цепей путем согласного или встречного включения одинаковых нелинейных элементов. Так строятся, например, *балансные*, мостовые и *кольцевые* схемы.

Для реализации четной характеристики необходимо использовать два НЭ с одинаковыми ВАХ. Пусть ВАХ каждого элемента

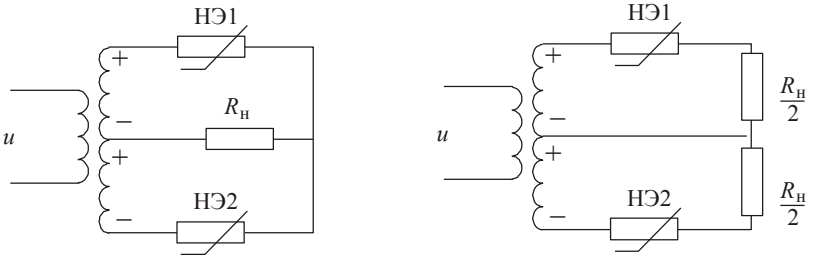


Рис. 5.4. Балансные схемы

имеет вид $i = F(u)$. Четную суммарную характеристику можно получить в соответствии с (5.7) в виде

$$i = F(u) + F(-u).$$

Это означает, что необходимо сложить токи двух НЭ, на которые входное напряжение подается с противоположными знаками (противофазно), что реализуется в схеме, показанной на рис. 5.4, *а*, путем соответствующего подключения вторичной обмотки трансформатора к нелинейным элементам. Сложение токов происходит в общем для них сопротивлении нагрузки.

Реализация нечетной ВАХ осуществляется схемой рис. 5.4, *б* за счет вычитания (противофазного сложения) на нагрузке напряжений, создаваемых токами различных НЭ.

$$i = F(u) - F(-u),$$

что соответствует выражению (5.8).

Если на нелинейный элемент подается сигнал в сумме с постоянным напряжением смещения U_0, определяющим рабочую точку ВАХ, удобно аппроксимирующий полином (5.2) представить в форме

$$i = a'_0 + a'_1(u - U_0) + a'_2(u - U_0)^2 ... + a'_N(u - U_0)^N, \qquad (5.9)$$

где коэффициенты, разумеется, отличаются от коэффициентов $a_0, ..., a_N$; при этом упрощаются математические операции при нахождении спектра тока.

5.2.2. ЭКСПОНЕНЦИАЛЬНАЯ АППРОКСИМАЦИЯ

Критерием выбора аппроксимирующей функции является простота аналитического выражения при приемлемой точности аппроксимации. Вольт-амперные характеристики некоторых нелинейных элементов с хорошей точностью представляются экспоненциальными функциями вида

$$i = f(u) = Ae^{\alpha u}, \qquad (5.10)$$

где A и α – константы.

Очевидно, при $u = 0$ значение тока равно $f(0) = Ae^0 = A$. Таким образом, значение константы A определяется непосредственно по заданной ВАХ, как значение тока при нулевом напряжении. Для нахождения константы α воспользуемся методом *приведения к линейному виду*. Прологарифмировав отношение i / A, получим

$$\ln \frac{i}{A} = \alpha u . \qquad (5.11)$$

По имеющейся заданной ВАХ можно построить график зависимости левой части выражения (5.11) от напряжения при различных u. Если экспоненциальная аппроксимация является подходящей для данной ВАХ, полученный график оказывается практически линейным, а константа α представляет собой тангенс угла наклона графика (или хотя бы касательной к нему в рабочей точке).

Для полупроводниковых диодов характерно нулевое значение тока при нулевом напряжении. Тогда аппроксимация (5.10) неприемлема, и взамен применяют аппроксимацию вида

$$i = f(u) = I_0(e^{\alpha u} - 1),$$

где $I_0 = -f(u)\big|_{u=-\infty}$ – обратный ток диода. Константа α находится аналогично описанному выше случаю.

5.2.3. КУСОЧНО-ЛИНЕЙНАЯ АППРОКСИМАЦИЯ

Эта аппроксимация заключается в замене реальной характеристики набором отрезков прямых линий, которые в совокупности приближенно повторяют форму ВАХ. Наибольшее распространение получила аппроксимация вида

$$i = \begin{cases} 0, & u < U_{\text{н}}, \\ S(u - U_{\text{н}}), & u \geq U_{\text{н}}, \end{cases}$$

где $U_{\text{н}}$ – значение напряжения, соответствующее началу линейно растущего участка, S – крутизна линейной части ВАХ.

5.3. ВОЗДЕЙСТВИЕ ГАРМОНИЧЕСКИХ КОЛЕБАНИЙ НА НЭ

5.3.1. ВОЗДЕЙСТВИЕ ГАРМОНИЧЕСКОГО НАПРЯЖЕНИЯ НА НЭ С ПОЛИНОМИАЛЬНОЙ ХАРАКТЕРИСТИКОЙ

Рассмотрим НЭ с вольт-амперной характеристикой, аппроксимируемой степенным полиномом

$$i = a_0 + a_1(u - U_0) + a_2(u - U_0)^2 ... + a_N(u - U_0)^N ,$$

на который воздействует напряжение вида

$$u(t) = U_0 + U_m \cos\omega t ,$$

гармоническое относительно рабочей точки, определяемой постоянным напряжением U_0.

Подставляя выражение напряжения в ВАХ и раскрывая степени и произведения тригонометрических функций, получим

$$i(t) = I_0 + I_1 \cos\omega t + I_2 \cos 2\omega t + ... + I_N \cos N\omega t ,$$

где

$$I_0 = a_0 + \frac{1}{2} a_2 U_m^2 + \frac{3}{8} a_4 U_m^4 + ... ,$$

$$I_1 = a_1 U_m + \frac{3}{4} a_3 U_m^3 + \frac{5}{8} a_5 U_m^5 + ... , \qquad (5.12)$$

$$I_2 = \frac{1}{2} a_2 U_m^2 + \frac{1}{8} a_4 U_m^4 + ... ,$$

$$I_3 = \frac{1}{4} a_3 U_m^3 + \frac{5}{16} a_5 U_m^5 + ... \text{ и т.д.}$$

Таким образом, спектр тока НЭ при гармоническом воздействии содержит *кратные* гармоники (гармонические составляющие с частотами $n\omega$ при $n = 0, 1, 2, ..., N$).

Если ток НЭ протекает через частотно-избирательную нагрузку (фильтр), то напряжение на нагрузке определяется теми составляющими, для которых нагрузка представляет значительное сопротивление. Например, включая последовательно с НЭ параллельный

колебательный контур, настроенный на вторую (третью, четвертую и т.д.) гармонику, получаем на нагрузке гармоническое напряжение кратной (двойной, тройной и т.д.) частоты. Таким образом реализуется *умножение* частоты на 2 (3, 4 …). Нередко применяют нелинейный активный элемент (транзистор) с нагрузкой в виде параллельного колебательного контура, настроенного на *основную* гармонику ω; таким образом осуществляется *нелинейное* (*резонансное*) *усиление* узкополосных сигналов (достоинством таких нелинейных усилителей является их высокий коэффициент полезного действия). Зависимость $I_n(U_m)$ амплитуды полезной гармоники тока от амплитуды входного гармонического напряжения называется *колебательной характеристикой*.

В таких случаях рассматривают ВАХ как *линеаризованную* зависимость, т. е. как линейную функцию, крутизна которой определяется относительно соответствующей гармоники и называется *средней крутизной*. Ввиду того, что функция линейна, ее крутизна равна просто отношению амплитуды выбранной гармоники тока к амплитуде воздействующего гармонического напряжения. Средняя крутизна в общем случае зависит от входного напряжения нелинейным образом. Очевидно, что средняя крутизна, например, по 1-й гармонике определяется из (5.12) как

$$S_{\text{ср1}} = a_1 + \frac{3}{4}a_3 U_m^2 + \frac{5}{8}a_5 U_m^4 + \ldots .$$

5.3.2. ВОЗДЕЙСТВИЕ ГАРМОНИЧЕСКОГО НАПРЯЖЕНИЯ НА НЭ С КУСОЧНО-ЛИНЕЙНОЙ ВАХ

При кусочно-линейной аппроксимации ВАХ нелинейного элемента протекающий через него ток представляется импульсами, описываемыми отрезками гармонической функции (рис. 5.5). Примем для определенности, что приложенное напряжение описывается выражением

$$u(t) = U_0 + U_m \cos\omega t ,$$

тогда ток будет протекать через НЭ только в пределах временны́х интервалов, определяемых неравенством

$$U_0 + U_m \cos\omega t > U_{\text{н}} .$$

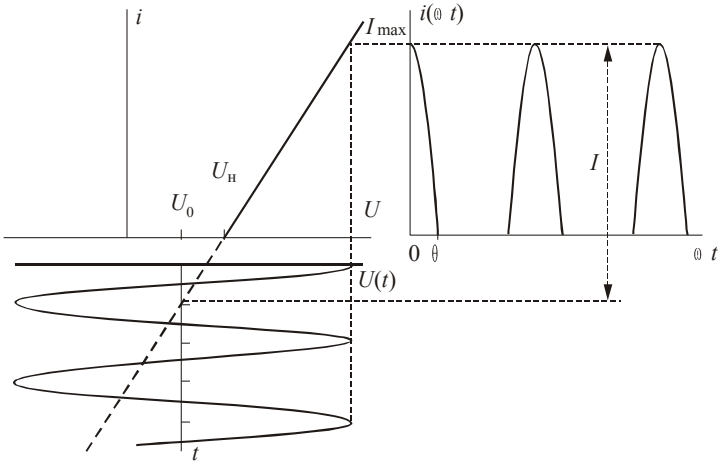


Рис. 5.5. К определению угла отсечки

Величину интервала протекания тока принято характеризовать углом θ (*углом отсечки*), определяемым следующим из этого условия уравнением

$$U_0 + U_m \cos\theta = U_{\text{н}},$$

или

$$\cos\theta = \frac{U_{\text{н}} - U_0}{U_m}. \qquad (5.13)$$

Обозначим максимальное значение тока $I_{\max}$, а буквой I обозначим амплитудное значение тока, который протекал бы через нелинейный элемент, если бы его характеристика была линейной и описывалась функцией $i = S(u - U_{\text{н}})$. Тогда, очевидно,

$$I_{\max} = I(1 - \cos\theta), \qquad (5.14)$$

а ток в пределах угла отсечки

$$i(t) = I_{\max} \frac{\cos\omega t - \cos\theta}{1 - \cos\theta}. \qquad (5.15)$$

Разлагая его (на интервале от $-\pi$ до π), как четную функцию, в ряд Фурье

$$i(t) = I_0 + I_1 \cos\omega t + I_2 \cos 2\omega t + I_3 \cos 3\omega t + \ldots,$$

имеем

$$I_0 = \frac{1}{2\pi}\int_{-\theta}^{\theta} I_{\max}\frac{\cos\omega t - \cos\theta}{1-\cos\theta}\,d(\omega t) = I_{\max}\alpha_0(\theta),$$

$$I_1 = \frac{1}{\pi}\int_{-\theta}^{\theta} I_{\max}\frac{\cos\omega t - \cos\theta}{1-\cos\theta}\cos\omega t\,d(\omega t) = I_{\max}\alpha_1(\theta),$$

$$I_n = \frac{1}{\pi}\int_{-\theta}^{\theta} I_{\max}\frac{\cos\omega t - \cos\theta}{1-\cos\theta}\cos n\omega t\,d(\omega t) = I_{\max}\alpha_n(\theta).$$

Коэффициенты пропорциональности, связывающие максимальное значение импульса тока с амплитудами гармоник тока, зависят от угла отсечки и называются *функциями Берга* (коэффициентами Берга)[69]. Функции Берга можно рассчитать по формулам

$$a_0(\theta) = \frac{\sin\theta - \theta\cos\theta}{\pi(1-\cos\theta)},$$

$$a_1(\theta) = \frac{\theta - \sin\theta\cos\theta}{\pi(1-\cos\theta)},$$

$$a_n(\theta) = \frac{2\left[\sin n\theta\cos\theta - n\cos n\theta\sin\theta\right]}{\pi n(n^2-1)(1-\cos\theta)}.$$

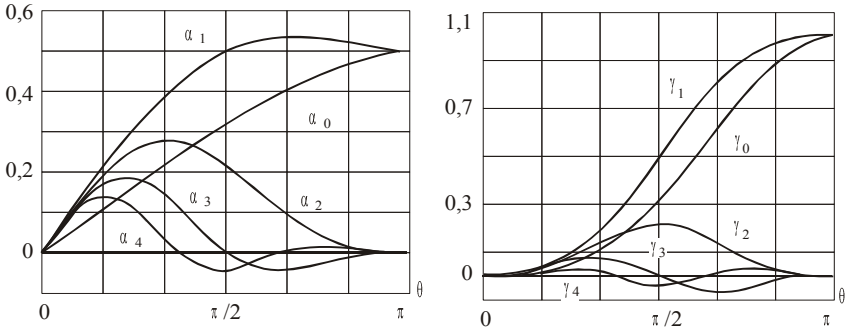
Несколько первых функций Берга показаны на рис. 5.6, *а*. Видно, что, выбирая значение угла отсечки (путем соответствующего задания U_m и U_0), можно добиться максимума нужной гармоники в спектре тока. Такой оптимальный угол отсечки для *n*-й гармоники определяется выражением

$$\theta_{\text{opt}} = \frac{120^\circ}{n}. \qquad (5.16)$$

Учитывая (5.14) и (5.15), можно записать ток в пределах угла отсечки в виде

$$i(t) = I(\cos\omega t - \cos\theta) = SU_m(\cos\omega t - \cos\theta),$$

[69] Аксель Иванович Берг (1892–1979) – известный русский ученый в области радиотехники, академик.



а *б*

Рис. 5.6. Функции Берга

тогда амплитуда *n*-й гармонической составляющей тока

$$I_n = SU_m\gamma_n(\theta),$$

где

$$\gamma_n(\theta) = (1-\cos\theta)\alpha_n(\theta),$$

представляет собой коэффициент пропорциональности между I и I_n. Функции $\gamma_n(\theta)$, показанные на (рис. 5.6, *б*), также носят название функций Берга. Максимума n-я функция достигает при

$$\theta_{\text{opt}} = \frac{180^\circ}{n}. \qquad (5.17)$$

Выбор формулы для нахождения *оптимального* угла отсечки определяется решаемой задачей. Если проектируется мощный усилитель или умножитель частоты, когда задан максимальный ток, обусловленный требованиями максимальной рассеиваемой мощности или электрической прочности устройства, и изменять угол отсечки можно изменением U_0 и U_m, поддерживая заданное значение $I_{\max}$, следует пользоваться формулой (5.16); если задано амплитудное значение входного напряжения, что характерно для маломощных каскадов, и можно для выбора угла отсечки оперировать только смещением U_0, справедлива формула (5.17).

5.3.3. БИГАРМОНИЧЕСКОЕ ВОЗДЕЙСТВИЕ НА НЭ

Рассмотрим теперь воздействие на НЭ с вольт-амперной характеристикой, аппроксимируемой степенным полиномом, *бигармонического* напряжения вида

$$u(t) = U_0 + U_{m1}\cos(\omega_1 t + \phi_1) + U_{m2}\cos(\omega_2 t + \phi_2).$$

Подставляя это выражение в ВАХ и раскрывая степени суммы гармонических функций по формуле бинома Ньютона

$$(a+b)^k = \sum_{m=0}^{k} C_k^m a^{k-m} b^m,\ C_k^m = \binom{k}{m} = \frac{k!}{m!(k-m)!},$$

находим, что в спектре тока будут присутствовать *комбинационные частоты* вида

$$\left|n_1\omega_1 + n_2\omega_2\right|, \tag{5.18}$$

где n_1 и n_2 – целые числа (положительные, отрицательные или 0), причем *порядок* комбинационной частоты $|n_1| + |n_2|$ ограничивается порядком полинома, аппроксимирующего ВАХ: $|n_1| + |n_2| \le N$.

Наличие комбинационных частот позволяет использовать НЭ для переноса спектра колебаний.

5.3.4. НЕЛИНЕЙНЫЙ ЭЛЕМЕНТ В КАЧЕСТВЕ ПАРАМЕТРИЧЕСКОГО

В подавляющем большинстве случаев на практике в качестве параметрических элементов используются элементы нелинейные, при этом должны выполняться определенные условия.

Предположим, что на НЭ подается, кроме постоянного напряжения для выбора рабочей точки, сумма двух гармонических сигналов $u(t) = U_1\cos\omega_1 t + U_2\cos\omega_2 t$, причем один из них имеет настолько малую амплитуду (например, U_1), что изменение напряжения за счет первого слагаемого происходит на участке ВАХ, который можно считать приближенно линейным, а второй сигнал большой амплитуды смещает рабочую точку и изменяет крутизну этого линейного участка.

Таким образом, при воздействии на НЭ сильного и слабого сигналов элемент ведет себя по отношению к слабому сигналу как

линейный параметрический элемент, управляемый по крутизне сильным сигналом.

Пример 5.1. Принцип действия *супергетеродинного приемника* основан на преобразовании частоты, т.е. переносе спектра модулированного колебания из окрестности несущей частоты в окрестность так называемой *промежуточной* частоты без изменения формы модулирующего сигнала. Модулированный сигнал можно рассматривать как узкополосный, не конкретизируя вид модуляции (см. разд. 2):

$$x(t) = A(t)\cos\left[\omega_0 t + \phi(t)\right] = A(t)\cos\Phi(t)\,.$$

Полагая модулированный сигнал слабым, рассмотрим его воздействие на нелинейный элемент с квадратичной ВАХ, управляемый по крутизне сильным опорным колебанием, так что крутизна меняется по закону

$$S(t) = S_0 + S_m \cos\omega_г t\,,$$

где $\omega_г$ – частота *гетеродина* (генератора опорного колебания), S_m – максимальное отклонение крутизны от среднего значения S_0. Ток, протекающий через нелинейный элемент

$$i(t) = A(t)\cos\left[\omega_0 t + \phi(t)\right]\left[S_0 + S_m\cos\omega_г t\right] =$$

$$= A(t)S_0\cos\left[\omega_0 t + \phi(t)\right] + A(t)S_m\cos\left[\omega_0 t + \phi(t)\right]\cos\omega_г t =$$

$$= A(t)S_0\cos\left[\omega_0 t + \phi(t)\right] +$$

$$+\frac{A(t)S_m}{2}\cos\left[\left(\omega_0 + \omega_г\right)t + \phi(t)\right] + \frac{A(t)S_m}{2}\cos\left[\left(\omega_0 - \omega_г\right)t + \phi(t)\right],$$

содержит две составляющие, совпадающие по форме с исходным модулированным сигналом с точностью до константы $S_m/2$ и несущих частот, равных соответственно $\omega_0 + \omega_г$ и $\omega_0 - \omega_г$. Выделив одну из составляющих с помощью полосового фильтра, получим сигнал, перенесенный без изменения закона модуляции на *промежуточную* частоту[70] $\omega_{пр} = \omega_0 + \omega_г$ (или $\omega_{пр} = \omega_0 - \omega_г$). Преимущество супергетеродинного приемника состоит в том, что на-

[70] В зависимости от выбора промежуточной частоты (выше или ниже несущей частоты сигнала) различают преобразование частоты *вверх* и преобразование частоты *вниз* (во втором случае супергетеродинный приемник называют также инфрадинным).

стройка на нужный частотный канал осуществляется путем изменения частоты гетеродина, а это значительно проще, чем перестраивать полосовой фильтр – входную цепь *приемника прямого усиления*. Основная частотная селекция, обеспечивающая *избирательность по соседнему каналу*, осуществляется полосовым *фильтром сосредоточенной селекции*, для которого сравнительно просто обеспечиваются хорошие частотно-избирательные свойства благодаря тому, что он не нуждается в перестройке. ◄

Если условие слабости модулированного сигнала нарушено, то преобразовательный элемент следует рассматривать как нелинейный, при этом комбинационные частоты второго порядка $\omega_0 + \omega_г$ и $\omega_0 - \omega_г$ определяются четной частью ВАХ. Если ВАХ аппроксимируется полиномом четвертой степени (или более высокой четной), преобразование частоты сопровождается искажениями закона модуляции [13].

5.4. АМПЛИТУДНАЯ МОДУЛЯЦИЯ ГАРМОНИЧЕСКОГО ПЕРЕНОСЧИКА

5.4.1. ВРЕМЕННОЕ И СПЕКТРАЛЬНОЕ ОПИСАНИЕ АМ-КОЛЕБАНИЙ

Гармонические переносчики часто используются в радиотехнике и связи по ряду причин, среди которых главная заключается в том, что только гармонические колебания не изменяют формы при прохождении через линейные стационарные цепи и каналы связи. Амплитудная модуляция заключается в изменении амплитуды несущего гармонического колебания

$$u_н(t) = U_m \cos(\omega_0 t + \varphi)$$

в соответствии с изменениями первичного (информационного) сигнала. Для простоты анализа примем, что первичный сигнал представляет собой гармоническое колебание низкой (в сравнении с частотой несущего колебания ω_0) частоты Ω. Это случай так называемой *тональной* (однотональной) модуляции. Тогда амплитудно-модулированное колебание (АМК) имеет вид

$$u_{АМ}(t) = U(t)\cos(\omega_0 t + \varphi) = U_m \left[1 + M\cos(\Omega t + \psi)\right]\cos(\omega_0 t + \varphi), \quad (5.19)$$

где M – *коэффициент модуляции*. На рис. 5.7 показано АМ-колебание с коэффициентом модуляции 0,5.

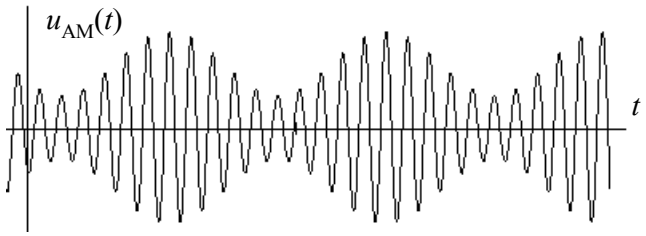


Рис. 5.7. Амплитудно-модулированное колебание

Нетрудно получить выражение для определения коэффициента модуляции по временной диаграмме. Очевидно, огибающая АМК достигает максимума при условии $\cos(\Omega t + \psi) = 1$ и минимума при $\cos(\Omega t + \psi) = -1$, поэтому

$$u_{\max} = U_m(1 + M),$$
$$u_{\min} = U_m(1 - M).$$

Складывая и вычитая эти равенства, получаем систему

$$\left.\begin{aligned} u_{\max} + u_{\min} &= 2U_m, \\ u_{\max} - u_{\min} &= 2U_m M, \end{aligned}\right\}$$

откуда

$$M = \frac{u_{\max} - u_{\min}}{u_{\max} + u_{\min}}.$$

Очевидно, величина M должна лежать в интервале от 0 до 1. При $M > 1$ имеет место искажение огибающей, называемое *перемодуляцией* (рис. 5.8).

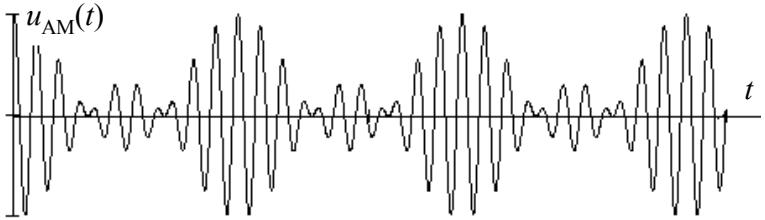


Рис. 5.8. Амплитудно-модулированное колебание с перемодуляцией

Найдем спектр тонального АМК. Напомним, что спектр – это совокупность коэффициентов, определяющих амплитуды гармонических колебаний, составляющих рассматриваемое колебание. Тогда «найти спектр» означает представить АМК в виде суммы гармонических колебаний и определить их амплитуды. Раскроем выражение (5.19) и получим

$$u_{\text{АМ}}(t) = U_m \cos(\omega_0 t + \varphi) + \frac{U_m M}{2} \cos\left[(\omega_0 + \Omega)t + \varphi + \psi\right] +$$

$$+ \frac{U_m M}{2} \cos\left[(\omega_0 - \Omega)t + \varphi - \psi\right]. \qquad (5.20)$$

Следовательно, амплитудная спектральная диаграмма тонального АМК имеет вид, показанный на рис. 5.9, *а*, а фазовая – на рис. 5.9, *б*.

Другой формой наглядного представления АМ-колебаний служит векторная диаграмма (рис. 5.10). Здесь принято, что комплексная плоскость вращается по часовой стрелке с угловой скоростью ω_0, тогда вектор несущего колебания длиной U_m неподвижен, а векторы боковых колебаний вращаются в противоположных направлениях с одинаковыми угловыми скоростями Ω и $-\Omega$, так что

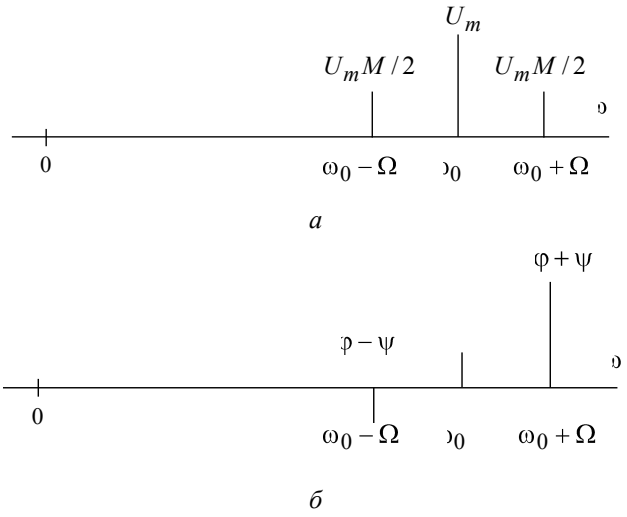


Рис. 5.9. Амплитудная (*а*) и фазовая (*б*) спектральные диаграммы тонального амплитудно-модулированного колебания

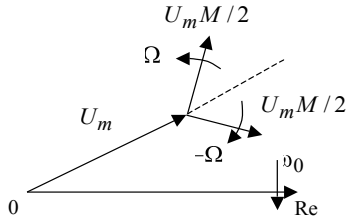


Рис. 5.10. Векторная диаграмма тонального амплитудно-модулированного колебания

их сумма всегда лежит на линии, вдоль которой направлен вектор несущего колебания. Таким образом, сумма всех трех векторов со временем изменяет только длину, оставаясь на той же линии, т.е. имеет место модуляция (изменение) только амплитуды. Очевидно, то же справедливо и при любом первичном сигнале, который всегда можно представить суммой гармонических колебаний. При этом векторная диаграмма приобретает сложный вид и на практике не используется. Спектральная диаграмма может быть найдена с использованием теоремы умножения (см. п. 2.10.2).

Предположим, что первичный сигнал $b(t)$ имеет спектральную плотность $B(\omega)$. При модуляции происходит умножение колебания $\left[1+Mb(t)\right]$ на колебание $U_m \cos\omega_0 t$ (для простоты начальную фазу положим равной 0). Согласно теореме умножения спектральная плотность результата равна свертке спектральных плотностей сомножителей.

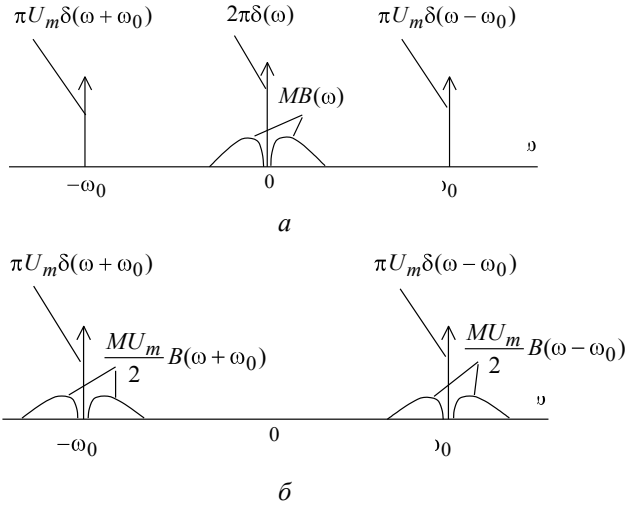


Рис. 5.11. Спектральные плотности несущего и модулирующего сигналов (*а*) и амплитудно-модулированного колебания (*б*)

Спектральная плотность первого сомножителя, очевидно, равна $2\pi\delta(\omega) + MB(\omega)$, а несущего колебания

$$2\pi U_m \frac{\delta(\omega-\omega_0)+\delta(\omega+\omega_0)}{2} = \pi U_m\left[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)\right].$$

Поэтому результирующее модулированное колебание имеет спектральную плотность (рис. 5.11)

$$U_{\text{AM}}(\omega) = \pi U_m\delta(\omega-\omega_0) + \frac{MU_mB(\omega-\omega_0)}{2} +$$

$$+\pi U_m\delta(\omega+\omega_0) + \frac{MU_mB(\omega+\omega_0)}{2}. \qquad (5.21)$$

Таким образом, спектральная плотность АМК имеет вид двух масштабированных (в $MU_m/2$ раз) копий спектральной плотности первичного сигнала, сдвинутых вправо и влево по оси частот на величину несущей частоты. Несущее колебание представлено в спектральной плотности АМК двумя δ-функциями с весом πU_m.

5.4.2. ПОЛУЧЕНИЕ АМ-КОЛЕБАНИЙ

Реализовать амплитудную модуляцию можно при помощи структурной схемы, показанной на рис. 5.12, где полосовой фильтр ПФ предназначен для подавления ненужных составляющих спектра колебания. (Если перемножитель идеальный, фильтр не нужен, однако на практике умножение осуществляется при помощи реальных нелинейных устройств, что приводит к возникновению комбинационных частот, в том числе ненужных.)

Другой способ получения АМ-колебания можно реализовать, подавая на НЭ сумму несущего и модулирующего колебаний и выделяя на частотно-зависимой нагрузке (фильтре) полезные составляющие. В схеме на рис. 5.13 нелинейным элементом является диод, а полосовым фильтром – параллельный контур.

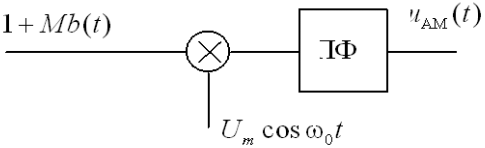


Рис. 5.12. Структура амплитудного модулятора

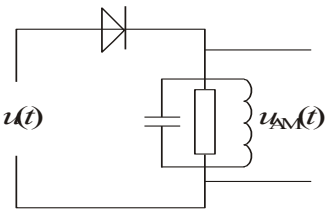


Рис. 5.13. Схема модулятора на диоде

Пусть ВАХ нелинейного элемента имеет квадратичный вид

$$i = a_0 + a_1 u + a_2 u^2, \qquad (5.22)$$

а сумма напряжений

$$u(t) = U_{m1}\cos\omega_0 t + U_{m2}\cos\Omega t. \qquad (5.23)$$

Подставим (5.23) в (5.22) и после преобразований получим (пренебрегая всеми составляющими, лежащими вне полосы пропускания фильтра)

$$i_{\text{AM}}(t) = a_1 U_{m1}\cos\omega_0 t + a_2 U_{m1} U_{m2}\cos(\omega_0 + \Omega)t + $$
$$+ a_2 U_{m1} U_{m2}\cos(\omega_0 - \Omega)t = $$
$$= a_1 U_{m1}\left[1 + M\cos\Omega t\right]\cos\omega_0 t,$$

где $M = 2\dfrac{a_2}{a_1}U_{m2}$. Видно, что, во-первых, глубина модуляции тока пропорциональна амплитуде первичного сигнала, так что модуляция на квадратичном элементе является линейной[71]. Во-вторых, модуляция тем глубже, чем больше отношение коэффициентов аппроксимации ВАХ. Для выбора рабочей точки ВАХ используется постоянное напряжение смещения, прибавляемое к входному напряжению. Заметим, что параллельный контур оказывает максимальное сопротивление току на резонансной частоте, а при отклонении частоты от резонансной сопротивление убывает тем быстрее, чем выше добротность контура. Поэтому коэффициент модуляции напряжения на контуре всегда будет меньше коэффициента модуляции тока.

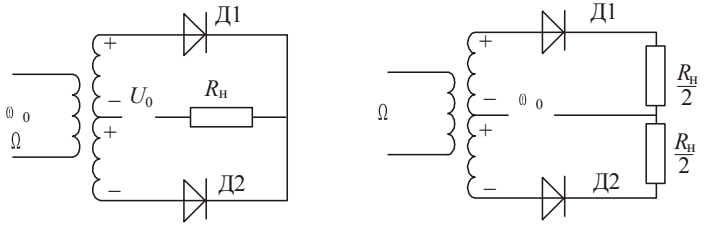
При любом законе амплитудной модуляции вектор несущего колебания не меняется (по модулю) во времени и поэтому не может нести информацию. Мощность несущего колебания равна $U_m^2/2$, в то время как суммарная мощность боковых составляющих при тональной модуляции равна $2\dfrac{M^2U_m^2}{4\cdot 2} = \dfrac{M^2U_m^2}{4}$ и при $M = 1$ составляет лишь половину мощности несущего колебания.

[71] Имеется в виду линейная зависимость огибающей от первичного сигнала; операция АМ является, конечно, нелинейной.

Таким образом, 2/3 мощности передатчика затрачивается впустую[72]. Поэтому на практике часто применяют частичное или полное подавление несущего колебания (во втором случае амплитудная модуляция называется *балансной*, поскольку реализуется в балансных схемах).

Сущность балансной модуляции легко понять, если подойти к построению схемы модулятора с точки зрения обеспечения четности ВАХ. В самом деле, частоты боковых составляющих АМ-колебания являются комбинационными частотами (5.18) второго порядка, т.е. $n_1 = n_2 = 1$, а частота несущего колебания представляет собой комбинационную частоту первого порядка ($n_1 = 1$, $n_2 = 0$). Поэтому если ВАХ нелинейного элемента имеет четный характер, в спектре тока будут отсутствовать все нечетные комбинационные частоты, включая несущую. Схема, показанная на рис. 5.14, *а*, реализует четную характеристику и применяется для балансной модуляции.

Другая разновидность балансной схемы, показанная на рис. 5.14, *б*, применяется для подавления вредных комбинационных составляющих. Например, если нелинейные элементы имеют кубическую ВАХ, то появляются нежелательные комбинационные частоты, самые вредные из которых – частоты $\omega_0 \pm 2\Omega$, так как их трудно подавить (они слишком близки к несущей частоте). Выход заключается в том, чтобы сделать схему модулятора нечетной по отношению к модулирующему колебанию. При этом относительно несущей частоты схема является четной, вследствие чего подавляется и несущее колебание.



а *б*

Рис. 5.14. Балансные модуляторы

[72] В случае амплитудной модуляции гармонического несущего колебания *реальным речевым* сигналом доля полезной мощности составляет всего около 0,1 % полной мощности АМК [14].

Выходное напряжение в такой схеме равно $U_{\text{вых}} = R(i_1 - i_2)$, где i_1 и i_2 – токи, протекающие через верхний и нижний НЭ, $R = \frac{R_{\text{н}}}{2}$.

Обозначим напряжение на половине вторичной обмотки через u_2, а напряжение несущей частоты через u_1. Полагая R малым, можно записать напряжение, приложенное к первому диоду, в виде $u_{\text{д1}} = u_1 + u_2$, а напряжение, приложенное ко второму диоду, в виде $u_{\text{д2}} = u_1 - u_2$.

Полагая, что

$$u_1(t) = U_{m1}\cos\omega_0 t, \tag{5.24}$$

$$u_2(t) = U_{m2}\cos\Omega t, \tag{5.25}$$

а ВАХ диода имеет вид кубического полинома, запишем токи диодов:

$$i_1 = a_0 + a_1(u_1 + u_2) + a_2(u_1 + u_2)^2 + a_3(u_1 + u_2)^3, \tag{5.26}$$

$$i_2 = a_0 + a_1(u_1 - u_2) + a_2(u_1 - u_2)^2 + a_3(u_1 - u_2)^3. \tag{5.27}$$

Разность токов равна

$$i_1 - i_2 = 2a_1u_2 + 4a_2u_1u_2 + 6a_3u_1^2u_2 + 2a_3u_2^3. \tag{5.28}$$

Тогда напряжение на выходе схемы

$$u_{\text{вых}} = 2R\left(a_1u_2 + 2a_2u_1u_2 + 3a_3u_1^2u_2 + a_3u_2^3\right). \tag{5.29}$$

Учитывая выражения (5.24) и (5.25), замечаем, что в это напряжение входят спектральные составляющие с частотами Ω, $\omega_0 \pm \Omega$, $2\omega_0 \pm \Omega$ и 3Ω. Отсутствуют четные гармоники частоты Ω и соответствующие комбинационные частоты (в том числе $\omega_0 \pm 2\Omega$), а также нечетные гармоники частоты ω_0 (в том числе сама частота ω_0). Все составляющие, кроме полезных (с частотами $\omega_0 \pm \Omega$), легко подавляются резонансной нагрузкой в виде параллельного контура, включаемой вместо $R_{\text{н}}$.

Предельная степень компенсации ненужных составляющих достигается в *кольцевом* модуляторе, показанном на рис. 5.15.

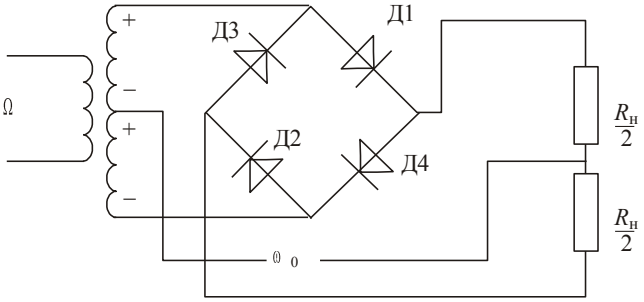


Рис. 5.15. Схема кольцевого модулятора

Считая по-прежнему, что сопротивление нагрузки пренебрежимо мало, и обозначая u_1 – напряжение несущей частоты (с частотой ω_0), а u_2 – напряжение на половине вторичной обмотки, запишем для токов, протекающих через диоды Д1 – Д4:

$$i_1 = F(u_1 + u_2), \tag{5.30}$$

$$i_2 = F(u_1 - u_2), \tag{5.31}$$

$$i_3 = F(-u_1 - u_2), \tag{5.32}$$

$$i_4 = F(-u_1 + u_2), \tag{5.33}$$

где $F(\cdot)$ – ВАХ одного диода (диоды считаются одинаковыми), а направления токов соответствуют прямому включению диодов. Тогда выходное напряжение равно

$$u_{\text{вых}} = R(i_1 + i_3) - R(i_2 + i_4) = R(i_1 - i_2) + R(i_3 - i_4).$$

Учитывая, что токи i_1 и i_2 по-прежнему определяются выражениями (5.26), (5.27), а i_3 и i_4 отличаются от них знаком аргумента [см. (5.30) – (5.33)], видим, что

$$i_3 = a_0 - a_1(u_1 + u_2) + a_2(u_1 + u_2)^2 - a_3(u_1 + u_2)^3,$$

$$i_4 = a_0 - a_1(u_1 - u_2) + a_2(u_1 - u_2)^2 - a_3(u_1 - u_2)^3,$$

откуда

$$i_3 - i_4 = -2a_1u_2 + 4a_2u_1u_2 - 6a_3u_1^2u_2 - 2a_3u_2^3. \tag{5.34}$$

Складывая выражения (5.28) и (5.34) и умножая на R, получаем выходное напряжение

$$u_{\text{вых}} = 8Ra_2u_1u_2 .$$

Таким образом, если ВАХ диодов имеют кубический вид, то кольцевая схема является точным перемножителем сигналов. Балансно-модулированное колебание имеет вид (при тональной модуляции)

$$u_{\text{БМ}}(t) = U_m \cos\Omega t \cos\omega_0 t ;$$

его график показан на рис. 5.16.

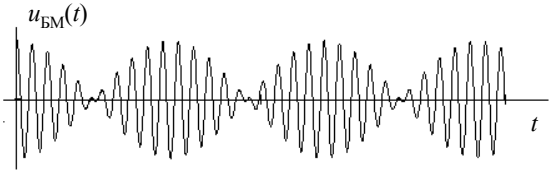


Рис. 5.16. Балансно-модулированное колебание

Учитывая симметрию спектра вещественного первичного сигнала, вследствие которой спектр АМ-сигнала также оказывается симметричным, можно сократить вдвое требуемую полосу частот канала, если сформировать АМ-сигнал, имеющий одну боковую полосу (*ОБП-сигнал*). Это возможно с использованием сигнала, сопряженного по Гильберту.

Аналитический сигнал, соответствующий первичному сигналу $b(t)$, имеет вид $\dot{b}(t) = b(t) + j\hat{b}(t)$ и спектральную плотность $2B(\omega)$ на положительных частотах. Сдвиг спектральной плотности вправо по оси частот на ω_0 достигается умножением сигнала на $e^{j\omega_0 t}$. При этом получается комплексное колебание, вещественная часть которого представляет собой действительный сигнал с одной (верхней) боковой полосой

$$u_{\text{ОБПверх}}(t) = \text{Re}\left\{\dot{b}(t)e^{j\omega_0 t}\right\} = b(t)\cos\omega_0 t - \hat{b}(t)\sin\omega_0 t .$$

Колебание $\dot{b}'(t) = b(t) - j\hat{b}(t)$ имеет спектральную плотность, сосредоточенную на отрицательных частотах, поэтому умножение его на $e^{j\omega_0 t}$ формирует комплексное колебание с одной боковой (нижней) полосой

$$u_{\text{ОБПнижн}}(t) = \text{Re}\left\{\dot{b}'(t)e^{j\omega_0 t}\right\} = b(t)\cos\omega_0 t + \hat{b}(t)\sin\omega_0 t .$$

5.4.3. ДЕТЕКТИРОВАНИЕ АМ-КОЛЕБАНИЙ

Детектирование, или демодуляция, представляет собой процесс, обратный модуляции. Результатом детектирования, следовательно, должен быть первичный (модулирующий) сигнал. В случае АМ это означает, что при детектировании должен получиться сигнал, повторяющий по форме огибающую АМ-колебания.

Известны методы детектирования АМ-колебаний на основе нелинейных и параметрических устройств.

Детектирование АМК параметрической цепью имеет много общего с модуляцией. В самом деле, амплитудная модуляция представляет собой фактически перенос спектра первичного сигнала на несущую частоту без изменения его формы (см. рис. 5.1). Демодуляции соответствует, очевидно, обратный перенос спектра на такую же величину. Поэтому умножение АМ-колебания на опорное гармоническое колебание несущей частоты с последующей фильтрацией нижних частот позволяет получить первичный сигнал, т.е. выполнить детектирование, при помощи параметрического устройства, показанного на рис. 5.17.

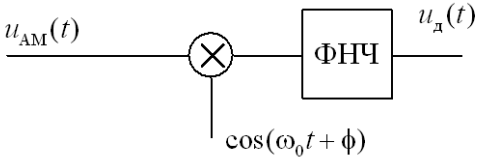


Рис. 5.17. Структура синхронного детектора

Умножив АМ-сигнал (5.19) на опорное колебание, получим

$$u_{\text{АМ}}(t)\cos(\omega_0 t + \phi) = U(t)\cos(\omega_0 t + \varphi)\cos(\omega_0 t + \phi) =$$

$$= \frac{U(t)}{2}\cos(2\omega_0 t + \varphi + \phi) + \frac{U(t)}{2}\cos(\varphi - \phi).$$

Очевидно, первое слагаемое не представляет интереса и должно быть подавлено фильтром нижних частот. Второе слагаемое является полезным сигналом, при этом полезный эффект будет максимальным, если опорное колебание будет иметь не только ту же частоту, что и несущее колебание АМ-сигнала, но и ту же начальную фазу, т. е. для наилучшего детектирования должно выполняться условие $\varphi - \phi = 0$. Поэтому такое детектирование называется *синхронным* или *когерентным* (см. пример 2.24).

Другой способ детектирования АМ-сигналов основан на использовании нелинейного элемента с последующей фильтрацией нижних частот. Пусть ВАХ нелинейного элемента аппроксимируется полиномом второго порядка

$$i = a_0 + a_1 u + a_2 u^2 .$$

Ток, протекающий через НЭ при воздействии на него тонального АМ-сигнала, равен (здесь и далее $\varphi = \psi = 0$)

$$i(t) = a_0 + a_1 U_m \left[1 + M\cos\Omega t\right]\cos\omega_0 t + a_2 U_m^2 \left[1 + M\cos\Omega t\right]^2 \cos^2\omega_0 t =$$

$$= a_0 + a_1 U_m \left[1 + M\cos\Omega t\right]\cos\omega_0 t +$$

$$+ a_2 U_m^2 \left[1 + 2M\cos\Omega t + \frac{M^2}{2} + \frac{M^2}{2}\cos 2\Omega t\right]\left(\frac{1}{2} + \frac{1}{2}\cos 2\omega_0 t\right) =$$

$$= a_0 + \frac{a_2 U_m^2}{2} + \frac{a_2 U_m^2 M^2}{4} + a_2 U_m^2 M\cos\Omega t + \frac{a_2 U_m^2 M^2}{4}\cos 2\Omega t + \text{ВЧ},$$

где символом ВЧ для краткости обозначены высокочастотные составляющие, подавляемые фильтром нижних частот. Из полученного выражения видно, что помимо постоянной составляющей и полезного сигнала (в данном случае это колебание с частотой Ω), а также высокочастотных составляющих в спектре тока содержится также колебание с частотой 2Ω, которое и является наиболее вредным, так как его трудно подавить фильтром нижних частот[73] (частота 2Ω находится слишком близко от частоты полезного колебания Ω). Если модулирующий сигнал отличается от гармонического тонального колебания (а на практике это всегда так) и имеет спектр конечной ширины, то полезные и вредные составляющие разделить практически невозможно[74]. Таким образом, нелинейное (квадратичное) детектирование сопровождается *нелинейными* ис-

[73] Здесь под трудностью подавления вредных составляющих подразумевается то, что для достижения нужной степени подавления может потребоваться очень сложный (многозвенный) фильтр.

[74] Например, для полосы частот телефонного канала 300…3400 Гц вторая гармоника нижней частоты равна 600 Гц и не может быть отделена от полезной составляющей сигнала с такой же частотой.

кажениями. Нелинейные искажения принято характеризовать *коэффициентом нелинейных искажений*

$$k_{\text{н}} = \frac{\sqrt{I_2^2 + I_3^2 + I_4^2 + \ldots}}{I_1},$$

где I_n – амплитуда n-й гармоники тока. В данном случае $k_{\text{н}} = \frac{I_2}{I_1} = \frac{M}{4} \le 25\ \%$. Такой уровень нелинейных искажений в большинстве практических приложений (в частности, в радиовещании) совершенно недопустим. Заметим, что реальные ВАХ полупроводниковых приборов могут быть аппроксимированы квадратичным полиномом лишь в пределах небольшого участка (при малом сигнале).

При сильном сигнале более подходящей является кусочно-линейная аппроксимация (см. п. 5.2.3). Рассмотрим детектор на биполярном транзисторе, принципиальная схема которого показана на рис. 5.18.

Полагая, что зависимость тока коллектора от напряжения база – эмиттер аппроксимируется кусочно-линейной функцией

$$i_{\text{к}}(u) = \begin{cases} 0 & \text{при} \quad u < U_{\text{н}}, \\ S(u - U_{\text{н}}) & \text{при} \quad u \ge U_{\text{н}}, \end{cases}$$

и что напряжение смещения U_0 равно напряжению начала линейного участка характеристики $U_{\text{н}}$, видим, что угол отсечки θ согласно (5.13) равен 90°, тогда при воздействии на вход схемы напряжения АМ-сигнала (5.19) ток коллектора имеет вид импульсов

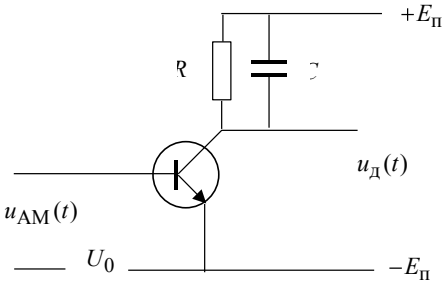


Рис. 5.18. Детектор на биполярном транзисторе

гармонической формы с частотой следования ω_0, скважностью 2 и амплитудой

$$I_m = SU_m\left(1 + M\cos\Omega t\right),$$

меняющейся медленно (по закону модулирующего сигнала).

Постоянная составляющая[75] импульсного тока также медленно меняется

$$I_0(t) = \gamma_0(\theta)SU_m\left(1 + M\cos\Omega t\right) = 0.318SU_m\left(1 + M\cos\Omega t\right).$$

Таким образом, низкочастотная составляющая выходного напряжения транзисторного детектора

$$u_{\text{н}}(t) = 0.318SU_mR\left(1 + M\cos\Omega t\right). \qquad (5.35)$$

Детекторы характеризуются *коэффициентом детектирования*

$$k_{\text{д}} = \frac{U_{m\text{вых}}}{MU_{m\text{вх}}},$$

где $U_{m\text{вых}}$ и $U_{m\text{вх}}$ – амплитуды низкочастотной составляющей выходного напряжения и несущего колебания соответственно. В данном случае

$$k_{\text{д}} = \frac{0.318SU_mRM}{MU_m} = 0.318SR.$$

Низкочастотная составляющая выходного напряжения прямо пропорциональна модулирующему колебанию, т.е. нелинейные искажения отсутствуют (если пренебречь отличиями реальной характеристики от кусочно-линейной). Отметим, что если $U_0 \neq U_{\text{н}}$, угол отсечки будет зависеть от времени и появятся нелинейные искажения.

Наиболее простое устройство для детектирования АМ-колебаний – диодный детектор (рис. 5.19).

В этой схеме в отличие от транзисторного детектора угол отсечки определяется не внешним источником напряжения смещения, а выходным постоянным[76] напряжением, приложенным к диоду в обратном направлении. В самом деле, ток, протекающий через

---

[75] На самом деле это не постоянная, а низкочастотная составляющая сигнала, но для каждого отдельного импульса тока она находится как постоянная составляющая тока на протяжении одного периода.

[76] См. предыдущую сноску.

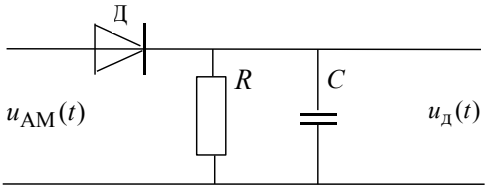


Рис. 5.19. Диодный детектор

диод в прямом направлении, заряжает конденсатор до некоторого напряжения, полярность которого такова, что оно стремится «запереть» диод. В результате открытое или запертое состояние диода в каждый момент времени определяется разностью $u(t)$ входного напряжения $u_{\text{АМ}}(t)$ и выходного напряжения $u_{\text{д}}(t)$, показанной сплошной линией на графике рис. 5.20. Медленно меняющееся выходное напряжение показано пунктирной линией.

Согласно формуле (5.13)

$$\cos\theta = \frac{U_{\text{н}} - U_0}{U_m},$$

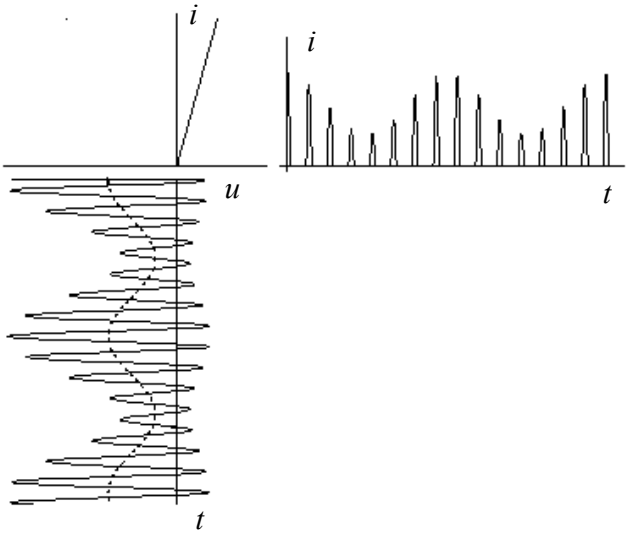


Рис. 5.20. Диаграммы напряжений и тока в схеме диодного детектора

а параметр аппроксимации $U_{\text{н}}$ равен нулю, поэтому $\cos\theta = \dfrac{-U_0}{U_m}$, но смещение в данном случае – это напряжение на нагрузке детектора, равное $U_0 = -I_0 R$, откуда

$$\cos\theta = \frac{I_0 R}{U_m}. \tag{5.36}$$

Учитывая это уравнение и выражая I_0 через функцию Берга, запишем

$$I_0 R = S U_m \gamma_0(\theta) R = U_m \cos\theta.$$

Раскрывая функцию $\gamma_0(\theta)$ и сокращая U_m, получим

$$\frac{SR}{\pi}\left(\sin\theta - \theta\cos\theta\right) = \cos\theta,$$

откуда, поделив обе части уравнения на $\cos\theta$, будем иметь уравнение

$$\frac{SR}{\pi}(\operatorname{tg}\theta - \theta) = 1. \tag{5.37}$$

Заметим, что в это уравнение не входит U_m. Это означает, что в линейном детекторе угол отсечки есть величина постоянная, зависящая только от параметров схемы. Используем разложение тангенса в степенной ряд, ограниченное двумя слагаемыми [8]

$$\operatorname{tg}\theta \approx \theta + \frac{1}{3}\theta^3.$$

Тогда из выражения (5.37) получается уравнение $\dfrac{SR}{\pi}\dfrac{\theta^3}{3} = 1$, откуда $\theta \approx \sqrt[3]{\dfrac{3\pi}{SR}}$.

Полезная составляющая тока нагрузки, как следует из формулы (5.36),

$$I_0 = \frac{U_m}{R}\cos\theta,$$

пропорциональна U_m, что и означает линейность детекторной характеристики[77]. Коэффициент детектирования, очевидно, равен

$$k_{\text{д}} = \cos\theta = \cos\left(\sqrt[3]{3\pi/(SR)}\right).$$

Практическое применение диодного детектора предполагает правильный выбор параметров ФНЧ. Необходимо, во-первых, чтобы сопротивление нагрузки было много больше внутреннего сопротивления диода (в прямом направлении). Это обеспечивает при быстром заряде конденсатора сравнительно медленный его разряд, что приводит к выделению огибающей АМ-сигнала. Во-вторых, емкость конденсатора должна выбираться из того условия, чтобы граничная частота ФНЧ была больше верхней частоты полезного сигнала и в то же время меньше несущей частоты. Поскольку используемый ФНЧ первого порядка имеет очень пологую АЧХ, постоянная времени RC -цепи должна удовлетворять двойному неравенству

$$\frac{1}{\omega_0} \ll RC \ll \frac{1}{\Omega}.$$

Нарушение левой части неравенства приводит к слишком быстрому разряду конденсатора (напряжение на нагрузке пульсирует с частотой ω_0), нарушение правой части – к слишком медленному разряду, вследствие чего напряжение на нагрузке может «не успевать» за более быстрыми изменениями огибающей АМ-сигнала. При этом форма выделяемой огибающей сильно отличается от модулирующего сигнала, что соответствует *нелинейным* искажениям.

5.5. УГЛОВАЯ МОДУЛЯЦИЯ

5.5.1. ОПИСАНИЕ УМ-КОЛЕБАНИЙ

При угловой модуляции гармонического колебания результирующий сигнал имеет постоянную амплитуду и зависящую от первичного (информационного) сигнала фазу, поэтому его можно записать в общем виде, как

$$u_{\text{УМ}}(t) = U_m \cos\Phi(t) = U_m \cos\left[\omega_0 t + \phi(t)\right], \qquad (5.38)$$

где $\Phi(t)$ – фаза колебания, а $\phi(t)$ – его *начальная* фаза.

---

[77] Диодный детектор можно считать линейным только при сильном сигнале, когда ВАХ диода удовлетворительно аппроксимируется кусочно-линейной функцией.

Для описания УМ-колебаний полезно ввести понятие *мгновенной* частоты

$$\omega(t) = \frac{d\Phi(t)}{dt} = \omega_0 + \frac{d\phi(t)}{dt}. \qquad (5.39)$$

Очевидно, при неизменной начальной фазе мгновенная частота равна частоте несущего колебания, однако изменение начальной фазы приводит к изменению мгновенной частоты. При *фазовой модуляции* (ФМ) начальная фаза меняется по закону первичного сигнала, следовательно, мгновенная частота меняется по закону его производной. При *частотной модуляции* (ЧМ) в соответствии с первичным сигналом меняется мгновенная частота, значит, начальная фаза меняется как интеграл первичного сигнала. В любом случае при угловой модуляции по виду модулированного сигнала невозможно определить вид модуляции (ЧМ или ФМ), *если не известен закон модуляции*.

Пусть частота модулируется по гармоническому закону

$$\omega(t) = \omega_0 + \omega_д \cos\Omega t \qquad (5.40)$$

(максимальное отклонение мгновенной частоты от среднего значения называется *девиацией*[78] частоты $\omega_д$). Тогда фаза модулированного колебания

$$\Phi(t) = \int_0^t \omega(t)dt + \phi_0 = \omega_0 t + \frac{\omega_д}{\Omega}\sin\Omega t + \phi_0 \qquad (5.41)$$

состоит из линейно меняющегося слагаемого $\omega_0 t$, постоянной ϕ_0 и гармонического слагаемого, амплитудное значение которого называется *индексом модуляции* (или девиацией фазы), численно[79] равным при тональной модуляции

$$m = \frac{\omega_д}{\Omega}.$$

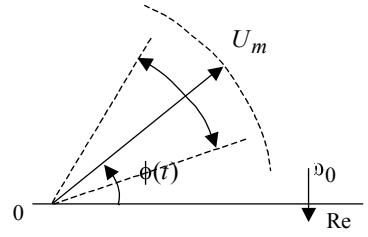


Рис. 5.21. Векторная диаграмма колебания с угловой модуляцией

На рис. 5.21 показана векторная диаграмма УМ-колебания. Вектор, изображающий колебание, не изменяет своей длины, но

[78] От англ. deviation – отклонение.
[79] Индекс модуляции, очевидно, имеет размерность *радиан.*

с течением времени он изменяет свое положение между двумя штриховыми линиями, отстоящими от среднего положения на величину индекса модуляции, при этом конец вектора перемещается по окружности.

Рассмотрим спектр колебания с угловой модуляцией по гармоническому закону. Для простоты будем считать $\phi_0 = 0$. Примем, что фаза модулируется по синусоидальному закону:

$$u_{\text{УМ}}(t) = U_m \cos\left(\omega_0 t + m \sin \Omega t\right) =$$

$$= U_m \cos\left(m \sin \Omega t\right) \cos \omega_0 t - U_m \sin\left(m \sin \Omega t\right) \sin \omega_0 t . \quad (5.42)$$

Заметим, что закон модуляции подвергается нелинейным преобразованиям $\cos(\cdot)$ и $\sin(\cdot)$, а это должно приводить к обогащению спектра.

Вначале рассмотрим частный случай малого индекса модуляции $m \ll 1$. Тогда

$$\cos\left(m \sin \Omega t\right) \approx 1 ; \quad (5.43)$$

$$\sin\left(m \sin \Omega t\right) \approx m \sin \Omega t . \quad (5.44)$$

С учетом этих приближенных равенств перепишем (5.42) в виде

$$u_{\text{УМ}}(t) = U_m \cos \omega_0 t - U_m m \sin \Omega t \sin \omega_0 t =$$

$$= U_m \cos \omega_0 t + \frac{U_m m}{2} \cos\left(\omega_0 + \Omega\right) t - \frac{U_m m}{2} \cos\left(\omega_0 - \Omega\right) t .$$

Полученное выражение похоже на выражение (5.20) для АМ-колебания. Однако отличие в знаке последнего слагаемого приводит к тому, что суммарный вектор колебания с течением времени изменяет свое угловое положение (рис. 5.22). При этом, очевидно, конец вектора суммарного колебания движется по прямой. Это отличие от идеального УМ-колебания является следствием использования приближенных выражений (5.43), (5.44). Если $m \ll 1$, то прямая мало отличается от окружности (тем меньше, чем меньше m).

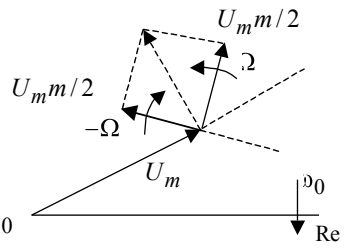


Рис. 5.22. Векторная диаграмма колебания при тональной УМ с малым индексом

Рассмотренный пример представляет лишь иллюстративный интерес, так как на практике используются УМ-колебания с большим индексом. Запишем УМ-колебание в виде

$$u_{\text{УМ}}(t) = U_m \cos\left(\omega_0 t + m \sin \Omega t\right) = U_m \operatorname{Re}\left\{e^{j\omega_0 t} e^{jm \sin \Omega t}\right\}.$$

Известна формула

$$e^{jm \sin x} = \sum_{k=-\infty}^{\infty} J_k(m) e^{jkx},$$

где $J_k(\cdot)$ – функция Бесселя первого рода k -го порядка. С учетом этого равенства

$$u_{\text{УМ}}(t) = U_m \operatorname{Re}\left\{ e^{j\omega_0 t} \sum_{k=-\infty}^{\infty} J_k(m) e^{jk\Omega t} \right\} =$$

$$= U_m \sum_{k=-\infty}^{\infty} J_k(m) \cos\left(\omega_0 + k\Omega\right) t .$$

Таким образом, даже при тональной модуляции спектр УМ-колебания имеет бесконечно много составляющих, амплитуды которых определяются значениями функции Бесселя $J_k(m)$, рассматриваемой как функция номера гармоники k при заданном значении m.

На рис. 5.23 изображены значения $J_k(m)$ при $m = 40$. Видно, что при $k > m$ они быстро убывают. Благодаря такому поведению функции Бесселя можно считать, что УМ-колебание имеет спектр с эффективной шириной, равной

$$2(m+1)\Omega \approx 2m\Omega = 2\omega_{\text{д}}.$$

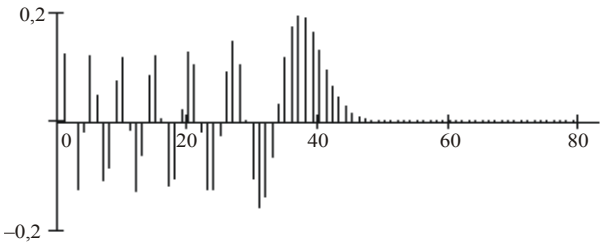


Рис. 5.23. Значения функций Бесселя при различных k и при $m = 40$

Таким образом, можно в первом приближении считать, что ширина спектра УМ-сигнала равна диапазону изменения частоты при модуляции, равному удвоенной девиации частоты.

5.5.2. ПРИБЛИЖЕННЫЙ АНАЛИЗ ВОЗДЕЙСТВИЯ УМ-КОЛЕБАНИЙ НА ЛИС-ЦЕПИ

Сигналы с угловой модуляцией применяются в технике связи очень широко. При этом часто представляет большой практический интерес задача анализа колебаний на выходе ЛИС-цепи (фильтра, линейного стационарного канала связи) при воздействии УМ-колебания. При негармоническом первичном сигнале и большом индексе модуляции точный анализ методами, рассмотренными в разд. 4, практически невозможен. Поэтому для приближенного анализа прохождения сигналов с угловой модуляцией через частотно-избирательные цепи применяется метод мгновенной частоты.

При угловой модуляции несущего колебания с частотой $\omega_0 = 2\pi f_0$ и амплитудой U_m низкочастотным колебанием (первичным сигналом) $b(t)$ получается сигнал, который можно приближенно рассматривать как «гармоническое колебание с медленно меняющейся частотой»[80]. Под частотой здесь понимается *мгновенная* частота УМ-колебания, а ее изменения могут считаться медленными, если мгновенные частоты УМ-колебания и отклика на него частотно-избирательной цепи практически совпадают. Для этого, очевидно, требуется, чтобы скорость протекания переходных процессов в ЧИЦ была велика в сравнении со скоростью изменения модулирующего сигнала. Отсюда вытекает требование, чтобы верхняя частота спектра модулирующего сигнала Ω была намного меньше ширины $\Delta\omega$ полосы пропускания цепи $\Omega << \Delta\omega$. Однако скорость изменения мгновенной частоты УМ-сигнала зависит также от амплитуды модулирующего сигнала, которая определяет девиацию частоты $\omega_д$ – максимальное отклонение мгновенной частоты от среднего значения [см. (5.40)]. Принято считать [23], что для применения метода мгновенной частоты достаточно, чтобы выполнялось условие $\omega_д < \Delta\omega$.

[80] Разумеется, в строгом смысле такое колебание *не является гармоническим* и имеет сложный спектр, рассмотренный в разд. 5.5.1.

Перепишем (5.38) в виде

$$u_{\text{УМ}}(t) = U_m \cos\Phi(t) = U_m \operatorname{Re}\left\{e^{j\Phi(t)}\right\} = U_m \operatorname{Re}\left\{e^{j\left[\omega_0 t + \phi(t)\right]}\right\}.$$

Тогда колебание на выходе ЛИС-цепи с комплексной частотной характеристикой $H(\omega) = K(\omega)e^{j\varphi(\omega)}$ имеет вид

$$u_{\text{вых}}(t) = U_m K(\omega(t)) \operatorname{Re}\left\{e^{j\left[\omega_0 t + \phi(t) + \varphi(\omega(t))\right]}\right\} =$$

$$= U_m K(\omega(t)) \cos\left[\omega_0 t + \phi(t) + \varphi\left(\omega(t)\right)\right],$$

откуда видно, что выходной сигнал имеет переменную амплитуду $U_m K(\omega(t))$, меняющуюся по закону, зависящему от изменений мгновенной частоты входного сигнала (происходит *паразитная амплитудная модуляция*). Закон изменения начальной фазы выходного сигнала также отличается от начальной фазы входного сигнала на величину, зависящую от времени и определяемую изменениями мгновенной частоты входного сигнала $\omega(t)$ и фазочастотной характеристикой цепи $\varphi(\cdot)$. Таким образом, ЛИС-цепь вносит искажения и в закон угловой модуляции сигнала. Мгновенная частота выходного сигнала отличается от мгновенной частоты входного сигнала (5.39) и равна

$$\omega_{\text{вых}}(t) = \omega_0 + \frac{d\phi(t)}{dt} + \frac{d\varphi\left(\omega(t)\right)}{dt}.$$

Пример 5.1. На колебательный контур, настроенный на частоту несущего колебания, поступает УМ-сигнал с мгновенной частотой, определяемой выражением (5.40)

$$\omega(t) = \omega_0 + \omega_{\text{д}} \cos\Omega t.$$

Тогда мгновенная частота выходного сигнала

$$\omega_{\text{вых}}(t) = \omega_0 + \frac{d\phi(t)}{dt} + \frac{d\varphi\left(\omega(t)\right)}{dt} =$$

$$= \omega_0 + \omega_{\text{д}} \cos\Omega t + \frac{d}{dt}\left[\varphi\left(\omega_0 + \omega_{\text{д}} \cos\Omega t\right)\right],$$

где $\xi(t) = \dfrac{d}{dt}\left[\varphi\left(\omega_0 + \omega_{\text{д}} \cos\Omega t\right)\right]$ – периодическая функция времени, описывающая искажение закона изменения частоты УМ-ко-

лебания. Поскольку колебательный контур настроен на частоту несущего колебания ω_0, ФЧХ цепи представляет собой функцию, практически антисимметричную (нечетную) относительно ω_0, поэтому выражение в квадратных скобках представляется рядом Фурье по косинусоидальным составляющим с нечетными гармониками частоты Ω (обдумайте это!). Тогда [13]

$$\xi(t) = \xi_1 \sin\Omega t + \xi_3 \sin 3\Omega t + \xi_5 \sin 5\Omega t + \ldots$$

и мгновенная частота выходного сигнала

$$\omega_{\text{вых}}(t) = \omega_0 + \omega_\text{д} \cos\Omega t + \xi_1 \sin\Omega t + \xi_3 \sin 3\Omega t + \xi_5 \sin 5\Omega t + \ldots =$$

$$= \omega_0 + \sqrt{\omega_\text{д}^2 + \xi_1^2}\cos\left(\Omega t - \gamma\right) + \xi_3 \sin 3\Omega t + \xi_5 \sin 5\Omega t + \ldots ,$$

где $\gamma = \operatorname{arctg}\left(\xi_1 / \omega_\text{д}\right)$. Таким образом, при прохождении УМ-сигнала через настроенный колебательный контур имеют место некоторое запаздывание закона модуляции, определяемое фазовым сдвигом γ, увеличение девиации частоты (от $\omega_\text{д}$ до $\sqrt{\omega_\text{д}^2 + \xi_1^2}$), а также появление высших (3-й, 5-й и т.д.) гармоник, т.е. *нелинейное искажение* закона модуляции. ◄

5.5.3. ПОЛУЧЕНИЕ КОЛЕБАНИЙ С УГЛОВОЙ МОДУЛЯЦИЕЙ

Для угловой модуляции можно использовать изменение в соответствии с первичным сигналом параметров частотно-задающей цепи генератора несущего гармонического колебания. Чаще всего в этом качестве используют полупроводниковый диод, называемый варикапом или варактором, включенный в обратном направлении. При этом $p-n$-переход диода функционирует как управляемый конденсатор, емкость которого зависит от приложенного напряжения. Тогда, подавая на диод первичный сигнал, можно управлять мгновенной частотой колебаний, т.е. осуществлять частотную модуляцию. Если мгновенная частота колебаний зависит от первичного сигнала линейно[81], то получаемое колебание имеет вид

$$u_\text{ЧМ}(t) = U_m \cos\left\{\left[\omega_0 + kb(t)\right]t + \varphi\right\},$$

[81] Линейность модуляционной характеристики является желательным свойством, которое может выполняться лишь приближенно.

где k – параметр, называемый крутизной *модуляционной характеристики*.

Фазовая модуляция может быть выполнена аналогично, если варактор включить в контур, являющийся *нагрузкой* резонансного усилителя, на вход которого подается несущее колебание. В этом случае изменение напряжения на варакторе не может изменить частоту колебания, но изменяет резонансную частоту контура и, следовательно, приводит к его расстройке относительно частоты колебания (что, очевидно, равносильно прохождению через резонансную цепь с неизменной настройкой «гармонического колебания с медленно меняющейся частотой»). Поэтому в соответствии с изменениями первичного сигнала будут изменяться амплитуда и начальная фаза напряжения на выходе усилителя. Вид этих изменений можно приближенно рассчитать по методу мгновенной частоты. Нежелательную («паразитную») амплитудную модуляцию можно устранить при помощи усилителя-ограничителя. Качество фазовой модуляции будет тем выше, чем ближе ФЧХ контура к линейной в диапазоне изменения частоты настройки контура при воздействии первичного сигнала на варактор. Допущение о линейности ФЧХ справедливо при небольших индексах модуляции (20…30°, или около 0,5 рад).

Другой способ, реализуемый в *модуляторе Армстронга*[82], заключается в суммировании балансно-модулированного АМ-колебания и несущего колебания, повернутого по фазе на 90°, что соответствует рис. 5.22.

В самом деле, на выходе перемножителя (рис. 5.24) имеет место напряжение БМ-сигнала, который превратился бы в обычный АМ-сигнал, если бы на сумматор подавалось то же несущее колебание, что и на перемножитель. Поскольку косинусоида отстает от синусоиды на 90°, результат совпадает с колебанием, векторная диаграмма которого показана на рис. 5.22. Модуляционная характеристика близка к линейной при малых индексах модуляции. Приближенно можно считать, что

$$u_{\text{вых}}(t) = U_m \cos\left[\omega_0 t - b(t)\right],$$

что соответствует фазовой модуляции. Тогда

$$u_{\text{вых}}(t) = U_m \cos\left[\omega_0 t - b(t)\right] =$$
$$= U_m \cos b(t) \cos\omega_0 t + U_m \sin b(t) \sin\omega_0 t \ .$$

[82] Э. Армстронг – известный американский инженер, автор идеи *супергетеродинного* приемника [15].

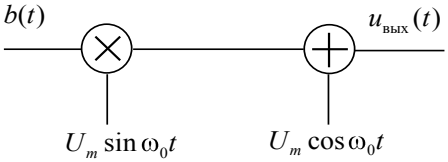


Рис. 5.24. Структура модулятора Армстронга

Учитывая, что при малом индексе модуляции $\cos b(t) \approx 1$ и $\sin b(t) \approx b(t)$, получаем

$$u_{\text{вых}}(t) = U_m \cos\omega_0 t + U_m b(t) \sin\omega_0 t,$$

что и реализуется схемой Армстронга.

Увеличить девиацию частоты УМ-колебания можно, подав его на умножитель частоты (см. разд. 5.3.1). При этом преобразовании часто́ты всех составляющих сигнала увеличиваются в определенное целое число раз. Поэтому во столько же раз увеличивается и ширина спектра. Если нужно, несущую частоту затем можно понизить путем переноса спектра вниз (см. разд. 5.3.4).

5.5.4. ДЕТЕКТИРОВАНИЕ УМ-КОЛЕБАНИЙ

Детектирование УМ-сигналов осуществляется несколькими способами.

Синхронное детектирование ФМ-колебаний

При малом индексе модуляции для демодуляции ФМ-сигналов можно использовать синхронный детектор, подобный применяемому для детектирования АМК, за исключением фазы опорного колебания (рис. 5.25). Если на вход перемножителя подать напряжение $u_{\text{ФМ}}(t) = U_m \sin\left[\omega_0 t + b(t)\right]$, а в качестве опорного колебания использовать напряжение $u_{\text{оп}}(t) = U \cos\omega_0 t$, то напряжение на выходе перемножителя будет равно

$$u_{\text{вых}}(t) = \frac{U_m U}{2} \sin\left[b(t)\right] + \frac{U_m U}{2} \sin\left[2\omega_0 t + b(t)\right].$$

При условии, что индекс модуляции не превышает 20…30°, можно принять $\sin\left[b(t)\right] \approx b(t)$; ВЧ составляющая может быть подавлена фильтром нижних частот

$$u_{\text{вых}}(t) \approx \frac{U_m U}{2} b(t).$$

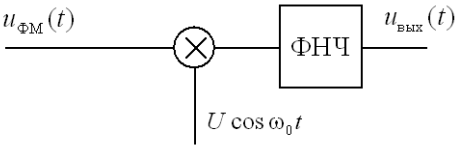


Рис. 5.25. Синхронное детектирование ФМ-колебаний

Диодное детектирование ФМ-сигналов

Для детектирования ФМ-сигналов может быть использован диодный детектор, на вход которого подается сумма ФМ-сигнала и опорного колебания (рис. 5.26, *а*).

Принцип детектирования поясняется векторной диаграммой (рис. 5.26, *б*), где опорное напряжение равно

$$u_{оп}(t) = U_{оп}\cos\omega_0 t\,,$$

а ФМ-сигнал $u_{ФМ}(t) = U_m \sin\left[\omega_0 t + b(t)\right] = U_m \cos\left[\omega_0 t - \pi/2 + b(t)\right]$; угол ϕ определяется первичным сигналом $b(t)$. При изменении угла амплитуда суммарного напряжения, приложенного ко входу диодного детектора, изменяется в соответствии с выражением

$$U_{вх} = \sqrt{U_{ФМ}^2 + U_{оп}^2 + 2U_{ФМ}U_{оп}\cos\phi}\,.$$

При $\phi \sim \pm 90^\circ$ характеристика детектора близка к линейной.

Для расширения линейного участка применяют балансную схему фазового детектора (рис. 5.27), где амплитуды напряжений на входах диодов плеч определяются выражениями

$$U_{вх_1} = \sqrt{U_{ФМ_1}^2 + U_{оп}^2 + 2U_{ФМ_1}U_{оп}\cos\phi}\,, \qquad (5.45)$$

$$U_{вх_2} = \sqrt{U_{ФМ_2}^2 + U_{оп}^2 - 2U_{ФМ_2}U_{оп}\cos\phi}\,, \qquad (5.46)$$

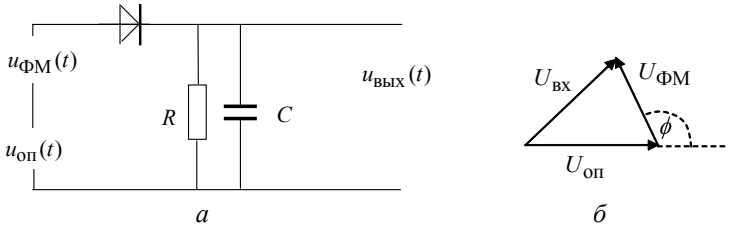


Рис. 5.26. Диодный детектор ФМ-колебания

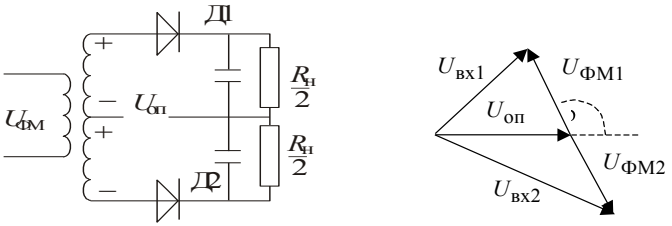


Рис. 5.27. Балансный фазовый детектор

где $U_{ФМ_1}$ и $U_{ФМ_2}$ – напряжения на секциях вторичной обмотки трансформатора. Выходное напряжение детектора пропорционально разности $U_{вх_1} - U_{вх_2}$. Зависимость выходного напряжения от угла показана сплошной линией на рис. 5.28 (амплитуды опорного и модулированного напряжений приняты равными друг другу). Штриховые линии соответствуют выражениям (5.45), (5.46). Видно, что детекторная характеристика практически линейна при изменении угла от 0 до 180°.

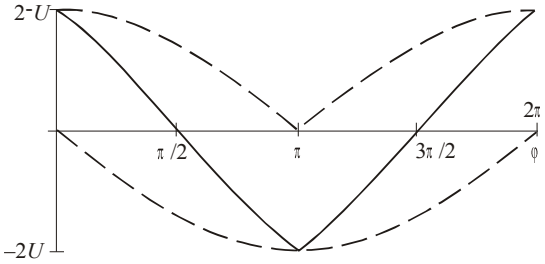


Рис. 5.28. Характеристика балансного фазового детектора

Детектирование ЧМ-сигналов

Детектирование ЧМ-сигналов можно выполнить при помощи фазового детектора, после чего выходной сигнал следует продифференцировать. (В самом деле, фазовый детектор вырабатывает напряжение, пропорциональное изменяющейся начальной фазе УМ-колебания. Но при частотной модуляции начальная фаза пропорциональна интегралу первичного сигнала, откуда и следует высказанное утверждение.)

Второй вариант заключается в преобразовании частотной модуляции в фазовую. Для этого ЧМ-сигнал подается на цепь с ли-

нейной ФЧХ. Роль такой цепи может выполнить резонансный усилитель, настроенный на среднюю частоту ЧМ-сигнала, если добротность колебательного контура не слишком высока, тогда изменения мгновенной частоты ЧМ-колебания происходят в пределах линейного участка ФЧХ. Выходной сигнал оказывается модулированным как по частоте, так и по фазе, поэтому для получения правильного результата демодуляции в качестве *опорного* колебания нужно использовать входной ЧМ-сигнал. Кроме того, при прохождении контура сигнал приобретает еще и паразитную амплитудную модуляцию, которую устраняют путем жесткого амплитудного ограничения сигнала (до фазового детектирования).

Еще один способ частотного детектирования состоит в преобразовании ЧМ-сигнала в АМ-сигнал, который затем детектируется обычным диодным детектором. Преобразование ЧМ-сигнала в АМ-сигнал производится путем подачи ЧМ-сигнала на резонансный усилитель с расстроенным контуром (рис. 5.29). Резонансная частота контура выбирается таким образом, чтобы изменения мгновенной частоты ЧМ-колебания происходили в пределах линейного участка на склоне резонансной кривой[83]. Перед таким преобразованием ЧМ-сигнал пропускают через усилитель-ограничитель с тем, чтобы избавиться от паразитной АМ, возникающей при прохождении сигнала через канал связи, в котором действуют помехи, замирания и другие вредные факторы, приводящие к изменениям амплитуды сигнала. Повышение качества преобразования ЧМ в АМ достигается противофазным (балансным) включением двух усилительных каскадов с резонансными нагрузками,

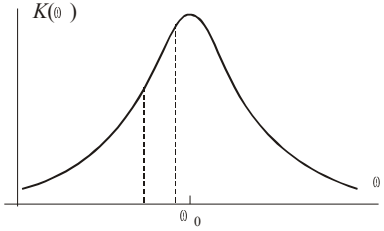


Рис. 5.29. АЧХ резонансного каскада. Штриховыми линиями выделен рабочий участок

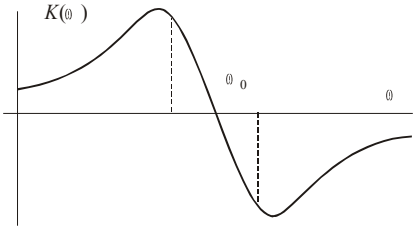


Рис. 5.30. Совместная дискриминаторная характеристика двух резонансных каскадов. Штриховыми линиями выделен рабочий участок

[83] Сигналы ЧМ-радиостанции, таким образом, можно принимать при помощи АМ-приемника, если его слегка расстроить относительно несущей частоты (правда, качество приема будет невысоким).

расстроенными симметрично относительно несущей частоты, тогда их общая характеристика имеет больший линейный участок (рис. 5.30).

5.6. ДИСКРЕТНАЯ МОДУЛЯЦИЯ

Особую роль в современных средствах связи играют методы дискретной (цифровой) модуляции[84], когда модулирующий (первичный) сигнал принимает в пределах временно́го интервала определенной длины постоянное значение, а при переходе к следующему такому же интервалу меняется скачкообразно. Таким образом, модулированный сигнал имеет вид последовательности элементарных сигналов, называемых *посылками* и отличающихся друг от друга некоторыми параметрами (амплитудой, частотой, начальной фазой). В соответствии с этим различают амплитудную, частотную и фазовую манипуляции, кроме того, находят применение комбинированные виды дискретной модуляции (например, применяется цифровая амплитудно-фазовая модуляция).

Большой практический интерес представляет выбор элементарного сигнала (посылки). Обозначим посылку $v(t, b)$, где b – значение цифрового модулирующего сигнала (канальный кодовый символ), так что сигнал цифровой модуляции имеет вид

$$b_{\text{ц}}(t) = \sum_{n=-\infty}^{\infty} v\left(t - nT_{\text{п}}, b_n\right),$$

где b_n – n-й символ в кодовой последовательности, $T_{\text{п}}$ – длительность посылки.

Часто полагают, что посылка имеет прямоугольную форму. Это объясняется тем, что в этом случае сигнал цифровой модуляции имеет наиболее простой вид. Однако спектральная плотность прямоугольного импульса нефинитна, а любой канал связи имеет ограниченную полосу пропускания. Поэтому при распространении такого сигнала неизбежно происходит его «размазывание» по временной оси, что приводит к *межсимвольной интерференции* – влиянию на значение сигнала в некоторый момент времени предшествующих посылок. Межсимвольная интерференция может приводить к ошибкам при принятии решения о принимаемом символе (подробнее см. разд. 9).

[84] Дискретная, или цифровая, модуляция называется также манипуляцией.

Другой крайний случай – посылка с прямоугольной спектральной плотностью, занимающей всю полосу частот данного канала. Предположим, что эта полоса сосредоточена в интервале $\left(-F_{\text{в}}, F_{\text{в}}\right)$, тогда посылка имеет вид

$$v(t) = 2F_{\text{в}} \frac{\sin\left(2\pi F_{\text{в}} t\right)}{2\pi F_{\text{в}} t}$$

и бесконечную длительность. Если решение о каждом переданном символе принимается по единственному отсчету наблюдаемого колебания и этот отсчет берется через интервал $T = 1/(2F_{\text{в}})$, то межсимвольная интерференция не влияет на правильность решений, так как функция $v(t)$ принимает нулевые значения при всех значениях $t = nT$ (см. разд. 2.11).

Часто в качестве $v(t)$ используют посылку гауссовской формы $v(t) = Ae^{-\alpha t^2}$, имеющую спектральную плотность также гауссовского вида $V(f) = Be^{-\beta f^2}$, где $B = A\sqrt{\pi/\alpha}$, $\beta = \pi^2/\alpha$. Такой сигнал не финитен ни по времени, ни по частоте, но он имеет минимальную эффективную «площадь» на плоскости время – частота (произведение эффективной длительности на эффективную ширину спектра).

5.6.1. ЦИФРОВАЯ (ДИСКРЕТНАЯ) АМПЛИТУДНАЯ МОДУЛЯЦИЯ (ЦАМ, ДАМ), ИЛИ АМПЛИТУДНАЯ МАНИПУЛЯЦИЯ

Модулированный сигнал при амплитудной манипуляции гармонической несущей имеет вид

$$u_{\text{ЦАМ}}(t) = \left(U_m + K_{\text{АМ}} \sum_{n=-\infty}^{\infty} b_n v\left(t - nT_{\text{п}}\right) \right) \cos(\omega_0 t + \phi_0), \quad (5.47)$$

где U_m – амплитуда несущего колебания, $K_{\text{АМ}}$ – коэффициент, управляющий глубиной амплитудной манипуляции, b_n – значение цифрового сигнала, отображающее n-й символ последовательности. Демодуляцию сигнала можно осуществить при помощи синхронного детектора либо нелинейных амплитудных детекторов, как для обычной амплитудной модуляции, с последующим принятием решений о переданных кодовых символах. Однако из-за на-

личия помех в канале этот способ не является наилучшим. Задача построения наилучшего (оптимального) демодулятора для ЦАМ-сигналов рассматривается в разд. 9.

Полагая в (5.47) $U_m = 0$, получим сигнал ЦАМ без несущей (ЦБАМ), который можно демодулировать синхронным детектором или – после восстановления несущей – нелинейным детектором.

При условии соблюдения когерентности (т.е. при известной и неизменной начальной фазе несущего колебания) можно передавать по одному каналу два ЦАМ-сигнала по квадратурной схеме (цифровая квадратурная модуляция ЦКАМ):

$$u_{\text{ЦКАМ}}(t) = b_{\text{ц1}}(t)\cos(\omega_0 t + \phi_0) + b_{\text{ц2}}(t)\sin(\omega_0 t + \phi_0), \quad (5.48)$$

при этом на приемной стороне разделение сигналов $b_{\text{ц1}}(t)$ и $b_{\text{ц2}}(t)$ осуществляется парой синхронных детекторов. Если за сигнал $b_{\text{ц1}}(t)$ принять сопряженный по Гильберту сигнал $\overline{b}_{\text{ц2}}(t)$, то выражение (5.48) даст *однополосную* ЦАМ.

5.6.2. ЦИФРОВАЯ (ДИСКРЕТНАЯ) ФАЗОВАЯ МОДУЛЯЦИЯ (ЦФМ, ДФМ), ИЛИ ФАЗОВАЯ МАНИПУЛЯЦИЯ

Фазовая манипуляция используется очень широко благодаря своим преимуществам перед другими видами цифровой модуляции [30]. Простейшим видом ЦФМ является двоичная фазовая манипуляция, когда модулированный сигнал имеет вид

$$u_{\text{ЦФМ}}(t) = \cos(\omega_0 t + \phi_0 + \theta(t)),$$

где $\theta(t)$ равно 0 или π радиан в зависимости от передаваемого символа.

Эквивалентной формой описания двоичного ЦФМ-сигнала является произведение

$$u_{\text{ЦФМ}}(t) = b_{\text{ц}}(t)\cos(\omega_0 t + \phi_0),$$

где $b_{\text{ц}}(t)$ принимает значения +1 или –1. Это означает, что двоичный ЦФМ-сигнал совпадает с результатом балансной амплитудной модуляции гармонического переносчика ступенчатым сигналом.

Демодуляция может быть выполнена синхронным детектором, для работы которого необходимо знать несущую частоту ω_0 и начальную фазу ϕ_0. Поскольку при балансной амплитудной модуля-

ции модулированный сигнал не содержит несущего колебания, на приемной стороне канала используется восстановление несущей частоты при помощи возведения сигнала в квадрат и последующего деления частоты полученного колебания на 2 [30]. Получаемое гармоническое колебание совпадает с несущим по частоте и может использоваться в когерентном детекторе в качестве опорного колебания. При этом, однако, имеет место неоднозначность его начальной фазы с точностью до 180º, что приводит к так называемой *обратной работе*, когда все двоичные символы при приеме заменяются на обратные. Для преодоления этого недостатка применяют периодическое зондирование канала специальным *пилот-сигналом*, по которому приемное устройство определяет действительную начальную фазу опорного колебания. Другим способом борьбы с этим явлением служит применение относительной фазовой манипуляции (ОФМ). При ОФМ символ 1 передается радиоимпульсом с той же начальной фазой, что и предыдущий, а при передаче символа 0 передается импульс с начальной фазой, отличающейся от предыдущего на 180º (или наоборот). При этом случайный «перескок» фазы опорного колебания приводит к ошибке при демодуляции только в одном символе и обратная работа не возникает.

5.6.3. ЦИФРОВАЯ (ДИСКРЕТНАЯ) ЧАСТОТНАЯ МОДУЛЯЦИЯ (ЦЧМ, ДЧМ), ИЛИ ЧАСТОТНАЯ МАНИПУЛЯЦИЯ

Частотная манипуляция может быть выполнена, например, путем поочередного подключения к входу канала связи выходов нескольких генераторов гармонических колебаний разных частот. При каждом переключении фаза канального сигнала в общем случае терпит разрыв. Прохождение такого сигнала через инерционные линейные устройства (например, фильтры) сопровождается переходными процессами, приводящими к возникновению паразитной амплитудной модуляции и ухудшающими *пик-фактор*[85]. Поэтому на практике получили распространение методы цифровой частотной модуляции с непрерывной фазой (ЧМНФ); при этом изменение частоты в соответствии с дискретным модулирующим сигналом производится не скачком, а по непрерывному (например,

[85] Пик-фактором называют отношение максимальной (пиковой) мощности сигнала к его средней мощности.

гауссовскому) закону путем изменения частоты генератора колебания. Демодуляция частотно-манипулированных сигналов может быть выполнена путем обычного частотного детектирования или на основе квадратурного приема [11].

5.7. ИМПУЛЬСНАЯ МОДУЛЯЦИЯ

Импульсной модуляцией называется модуляция переносчика, имеющего вид периодической последовательности импульсов одинаковой формы. Модулирующий сигнал при этом является аналоговым. Фактически при импульсной модуляции параметрами переносчика управляют *дискретные отсчеты первичного сигнала*, поэтому, для того чтобы была возможна передача информации без потерь, частоту следования импульсов переносчика следует выбирать исходя из ширины спектра модулирующего сигнала в соответствии с требованиями теоремы отсчетов.

Как и в случае модуляции гармонического переносчика, виды модуляции различаются в зависимости от изменяемых параметров. Если в соответствии с первичным сигналом изменяется амплитуда импульсов, модуляция называется амплитудно-импульсной (АИМ), если изменяется длительность (ширина) импульсов – широтно-импульсной (ШИМ, ДИМ), если изменяется временной сдвиг (относительно положения импульса в немодулированной последовательности) – времяимпульсной (ВИМ) или частотно-импульсной (ЧИМ). Два последних вида модуляции аналогичны фазовой и частотной модуляции гармонического переносчика в том смысле, что при изменении временно́го сдвига в соответствии с первичным сигналом частота следования импульсов меняется пропорционально его производной.

Рассмотрим более подробно сигнал амплитудно-импульсной модуляции. Выясним, как связаны спектральные плотности АИМ-сигнала $x_{\text{и}}(t)$ и исходного аналогового сигнала $x(t)$. Примем в качестве переносчика колебание $s(t) = \sum_{n=-\infty}^{\infty} d(t - nT)$, где $d(t)$ – короткий импульс известной формы. Преобразование Фурье для краткости будем обозначать в операторной форме $\mathrm{F}\{\cdot\}$.

Периодическое колебание $s(t)$ можно записать в виде ряда Фурье

$$s(t) = \sum_{k=-\infty}^{\infty} C_k e^{jk\Omega t},$$

где $\Omega = \dfrac{2\pi}{T}$, C_k – коэффициенты ряда, определяемые формой импульса $d(t)$. Спектральная плотность переносчика $S(\omega) = \mathrm{F}\{s(t)\}$, очевидно, равна сумме спектральных плотностей гармонических составляющих (комплексных экспонент) с теми же весовыми коэффициентами, т. е.

$$S(\omega) = \mathrm{F}\left\{ \sum_{k=-\infty}^{\infty} C_k e^{jk\Omega t} \right\} = 2\pi \sum_{k=-\infty}^{\infty} C_k \delta(\omega - k\Omega).$$

Поскольку АИМ-сигнал получается умножением $x_{\text{и}}(t) = x(t)s(t)$, его спектральная плотность равна свертке спектральных плотностей сомножителей

$$X_{\text{и}}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\sigma) S(\omega - \sigma) d\sigma =$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\sigma) 2\pi \sum_{k=-\infty}^{\infty} C_k \delta(\omega - \sigma - k\Omega) d\sigma = \sum_{k=-\infty}^{\infty} C_k X(\omega - k\Omega)$$

и представляет собой совокупность *копий* спектральной плотности первичного сигнала $X(\cdot)$, сдвинутых по частотной оси на величины $k\Omega = k\dfrac{2\pi}{T}$, $k = \overline{-\infty, \infty}$ и умноженных на весовые коэффициенты, определяемые *формой* импульса $d(t)$. (В частности, при $d(t) = \delta(t)$ все коэффициенты равны 1.)

Очевидно, если каждая копия $X(\cdot)$ занимает на частотной оси интервал, ширина которого меньше $\Omega = \dfrac{2\pi}{T}$, то копии *не перекрываются* и можно выделить путем фильтрации нижних частот единственную копию $X(\omega - 0 \cdot \Omega) = X(\omega)$, тем самым восстановив первичный сигнал из АИМ-сигнала (т.е. выполнить демодуляцию). Таким образом, демодуляция АИМ-сигнала выполняется ЛИС-цепью; это исключение из общего правила[86] возможно потому, что

[86] Напомним, что в общем случае операции, связанные с модуляцией и демодуляцией, выполняются с использованием нелинейных или параметрических цепей.

в спектре модулированного сигнала содержатся спектральные компоненты полезного сигнала и обогащение спектра не требуется.

При этом в интервале $-\frac{\pi}{T} \le \omega \le \frac{\pi}{T}$ спектральная плотность

$$X_{\text{и}}(\omega) = C_0 X(\omega).$$

Для восстановления первичного сигнала из АИМ-сигнала должен быть скомпенсирован весовой коэффициент C_0. Если $d(t)$ – прямоугольный импульс с амплитудой U и длительностью τ, симметрично расположенный относительно момента $t = 0$, то

$$C_k = \frac{U}{T}\int_{-\tau/2}^{\tau/2} e^{-jk\Omega t}dt = \frac{U\tau}{T}\frac{\sin(k\Omega\tau/2)}{k\Omega\tau/2}.$$

Коэффициент $C_0 = \frac{U\tau}{T}$, поэтому восстанавливающий ФНЧ должен иметь прямоугольную АЧХ вида

$$K(\omega) = \begin{cases} \dfrac{T}{U\tau}, & \dfrac{-\pi}{T} < \omega < \dfrac{\pi}{T}, \\ 0 \quad \text{в противном случае} \end{cases}$$

и импульсную характеристику

$$h(t) = F^{-1}\{K(\omega)\} = \frac{1}{U\tau}\sin\left(\frac{\pi}{T}t\right)\Big/\frac{\pi}{T}t.$$

Заметим, что полученная функция совпадает (с точностью до амплитудного постоянного множителя) с нулевой базисной функцией ряда Котельникова (все остальные базисные функции образуются ее сдвигами на величины, кратные шагу дискретизации). Таким образом, демодуляция АИМ-сигнала по смыслу близка к восстановлению аналогового сигнала по последовательности его отсчетов (сравните с разд. 2).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое угол отсечки? Как выбрать оптимальный угол отсечки?
2. Для чего используются двухтактные (балансные) схемы?
3. Почему нельзя осуществить модуляцию при помощи ЛИС-цепи?
4. В чем состоит принцип преобразования частоты?

5. Как выбирают параметры нагрузки для диодного детектора?

6. Каким должен быть спектр переносчика, чтобы можно было осуществить демодуляцию при помощи ЛИС-цепи?

7. При помощи каких схем можно реализовать угловую модуляцию?

8. Как выполнить демодуляцию УМ колебаний?

9. Какие частоты будут присутствовать в спектре тока, протекающего через параметрический элемент, если крутизна меняется по закону $s(t) = S_0 + S_1 \cos(\omega_1 t) + S_2 \cos(\omega_2 t)$, а напряжение – по закону $u(t) = U_0 + U_1 \cos(\omega_0 t)$?

УПРАЖНЕНИЯ

1. Найдите формулы, связывающие коэффициенты в выражениях (5.2) и (5.9) для $N = 2$.

2. Постройте колебательную характеристику нелинейного усилителя при кусочно-линейной аппроксимации ВАХ нелинейного элемента (примите $U_н = 1$ В, $U_0 = 1{,}5$ В, $S = 2$ мА/В).

3. Повторите расчеты для $U_н = 1$ В, $U_0 = 0{,}5$ В, $S = 2$ мА/В. Сравните результаты.

4. Тональное АМ-колебание характеризуется следующими параметрами. Амплитуда колебания верхней боковой частоты 10 В. Коэффициент модуляции 0,25. Определите среднюю мощность несущего колебания и отношение суммарной мощности боковых к мощности несущего. Постройте спектральную диаграмму.

5. Радиопередатчик с амплитудной модуляцией излучает мощность 10 Вт в отсутствие модулирующего колебания. Найдите пиковую (максимальную) мощность при тональной модуляции, если коэффициент модуляции равен 0,5; 1.

6. Нелинейный элемент, используемый в амплитудном модуляторе, имеет вольт-амперную характеристику $i = 0{,}01u^2$. К диоду приложено напряжение $u(t) = 0{,}5 + 0{,}2\cos\omega t + 0{,}1\cos\Omega t$. Определите коэффициент модуляции тока (все величины даны в системе СИ, ω, Ω – частоты несущего и модулирующего колебаний соответственно).

7. Запишите выражение для напряжения на параллельном колебательном контуре диодного модулятора, описанного в разд. 5.4.2, учитывающее зависимость сопротивление контура от частоты. Получите выражение, связывающее уменьшение коэффициента модуляции с частотой модулирующего гармонического колебания и с добротностью контура.

6. ЦЕПИ С ОБРАТНОЙ СВЯЗЬЮ

В некоторых случаях полезные изменения свойств цепей достигаются введением так называемой обратной связи (ОС). При этом к основной цепи (прямому каналу) подключается цепь обратной связи, через которую выходной сигнал прямого канала воздействует на его вход (рис. 6.1). Такая ОС называется *внешней*. Иногда обратная связь возникает за счет соответствующего соединения элементов схемы и в ней явно не выделяется цепь ОС; такая обратная связь называется *внутренней*. В некоторых случаях обратная связь возникает против воли создателя устройства и является нежелательной; такая ОС называется *паразитной*.

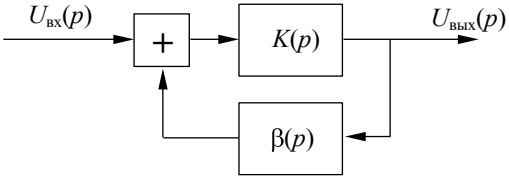


Рис. 6.1. Цепь с обратной связью

При анализе цепей с ОС принято рассматривать входные и выходные сигналы в операторной форме; соответственно цепи характеризуются операторными передаточными функциями.

Выразим изображение выходного сигнала через передаточную функцию прямого канала и изображение сигнала на его входе

$$U_{вых}(p) = K(p)\left[U_{вх}(p) + \beta(p)U_{вых}(p)\right].$$

Решая это уравнение относительно $U_{вых}(p)$, получаем

$$U_{вых}(p) = \frac{U_{вх}(p)K(p)}{1-\beta(p)K(p)}.$$

Передаточная функция цепи, охваченной обратной связью, находится как отношение изображений выходного и входного сигналов

$$K_{\text{ос}}(p) = \frac{K(p)}{1-\beta(p)K(p)}. \qquad (6.1)$$

От изображений перейдем к частотному описанию цепи; комплексная частотная характеристика цепи с ОС

$$K_{\text{ос}}(j\omega) = \frac{K(j\omega)}{1-\beta(j\omega)K(j\omega)}. \qquad (6.2)$$

Из полученного выражения видно, что, во-первых, КЧХ цепи с ОС отличается от КЧХ прямого канала, и, во-вторых, это отличие, определяемое знаменателем дроби, различно на разных частотах. Таким образом, обратная связь в общем случае является *частотно-зависимой*.

6.1. ВИДЫ ОБРАТНОЙ СВЯЗИ

Влияние ОС на характеристики цепи принципиально различно в зависимости от того, увеличивается (по модулю) коэффициент передачи цепи при введении ОС или уменьшается. Различают два частных случая:

1) если $|1-\beta(j\omega)K(j\omega)|>1$, обратная связь называется *отрицательной*;

2) если $|1-\beta(j\omega)K(j\omega)|<1$, обратная связь называется *положительной*.

Эти два частных случая не исчерпывают всех видов ОС. Очевидно, возможна ситуация, когда обратная связь является положительной для одних частотных составляющих и отрицательной для других.

Исторически первой в радиотехнике и связи стали использовать положительную ОС, так как она увеличивает коэффициент усиления усилителя и используется для самовозбуждения автогенераторов[88]. Отрицательная ОС была впервые применена Х. Блэком для расширения полосы пропускания усилителя[89].

[88] Примером более раннего использования отрицательной ОС *в технике вообще* можно считать центробежный регулятор Уатта.

[89] Интересные исторические подробности см. в [14].

6.1.1. ПОЛОЖИТЕЛЬНАЯ ОБРАТНАЯ СВЯЗЬ

Рассмотрим резонансный усилитель, в котором обратная связь создается катушкой обратной связи $L_{св}$, индуктивно связанной с колебательным контуром, служащим нагрузкой усилителя (рис. 6.2). КЧХ усилителя без ОС определяется в окрестности резонансной частоты ω_0 выражением [8]

$$K(j\omega) = \frac{-SR_{экв}}{1 + j\xi},$$

где $\xi = 2Q\dfrac{\omega - \omega_0}{\omega_0}$ – обобщенная расстройка, Q – добротность контура, $R_{экв}$ – эквивалентное сопротивление контура на резонансной частоте. Введем обозначение $\dfrac{2Q}{\omega_0} = \dfrac{2}{\Delta\omega_{0.7}} = \tau_к$ для *постоянной времени* контура, тогда

$$K(j\omega) = \frac{-SR_{экв}}{1+j(\omega - \omega_0)\tau_к}. \qquad (6.3)$$

Введем положительную обратную связь с коэффициентом β_0 (знак ОС определяется полярностью подключения катушки связи, а абсолютная величина β_0 – соотношением чисел витков катушек), тогда КЧХ усилителя согласно (6.2) будет равна

$$K_{ос}(j\omega) = \frac{-SR_{экв}/\left[1+j(\omega - \omega_0)\tau_к\right]}{1 - \beta_0 \dfrac{SR_{экв}}{1+j(\omega - \omega_0)\tau_к}} = \frac{-SR_{экв}}{1 - \beta_0 SR_{экв} + j(\omega - \omega_0)\tau_к} =$$

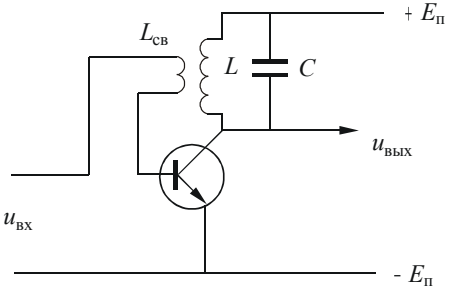


Рис. 6.2. Усилитель с ОС

$$= \frac{\dfrac{-SR_{\text{экв}}}{1-\beta_0 SR_{\text{экв}}}}{1+j(\omega-\omega_0)\dfrac{\tau_{\text{к}}}{1-\beta_0 SR_{\text{экв}}}}.$$

Сравнение полученного выражения с (6.3) свидетельствует о том, что введение ПОС эквивалентно увеличению $R_{\text{экв}}$ и $\tau_{\text{к}}$ в $\dfrac{1}{1-\beta_0 SR_{\text{экв}}}$ раз, что, в свою очередь, равносильно повышению добротности контура; степень повышения определяется тем, насколько величина $\beta_0 SR_{\text{экв}}$ близка к единице. На рис. 6.3 показано изменение АЧХ резонансного усилителя после введения положительной ОС. Для этого случая $\beta_0 SR_{\text{экв}} = 0.75$, поэтому коэффициент усиления на резонансной частоте увеличился в 4 раза, также в 4 раза уменьшилась полоса пропускания по уровню 0,707. Усилитель с ПОС называется *регенеративным* (иногда положительная ОС также называется регенеративной).

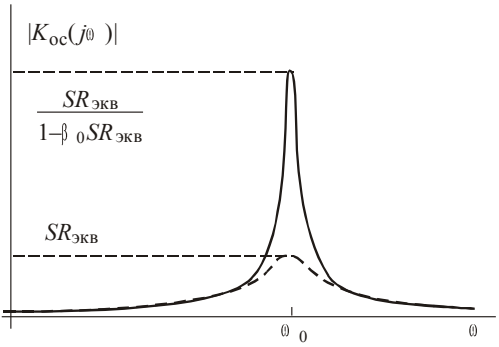


Рис. 6.3. Изменение АЧХ усилителя после введения ПОС

6.1.2. ОТРИЦАТЕЛЬНАЯ ОБРАТНАЯ СВЯЗЬ

Для простоты положим, что цепь обратной связи является частотно-независимой и коэффициент обратной связи равен $\beta(j\omega) = -\beta_0$. Тогда выражение (6.2) принимает вид

$$K_{\text{ОС}}(j\omega) = \frac{K(j\omega)}{1+\beta_0 K(j\omega)}. \qquad (6.4)$$

Одно из полезных свойств отрицательной ОС состоит в снижении чувствительности цепи к нестабильности характеристик прямого канала. В большинстве случаев прямой канал содержит активные (усилительные) элементы, подверженные старению, температурному дрейфу и другим вредным факторам. Кроме того, все активные элементы имеют значительный производственный разброс параметров, который приводит к изменениям характеристик цепи при замене активного элемента (например, при ремонте). Отрицательная ОС уменьшает чувствительность цепи к подобным факторам. В самом деле, если в выражении (6.4) положить

$$\beta_0 K(j\omega) \gg 1,$$

то, очевидно, КЧХ цепи с отрицательной обратной связью (ООС)

$$K_{\text{ОС}}(j\omega) \approx \frac{1}{\beta_0}$$

практически не зависит от прямого канала, а значит, и от его нестабильности.

Для более точного анализа рассмотрим полный дифференциал функции $K_{\text{ОС}}(j\omega)$ по отношению к коэффициентам передачи прямого и обратного каналов (зависимость от частоты для простоты выкладок опускается):

$$dK_{\text{ОС}} = \frac{\partial K_{\text{ОС}}}{\partial K} dK + \frac{\partial K_{\text{ОС}}}{\partial \beta_0} d\beta_0 =$$

$$= \frac{1 + \beta_0 K - \beta_0 K}{(1 + \beta_0 K)^2} dK + \frac{K^2}{(1 + \beta_0 K)^2} d\beta_0,$$

откуда получаем

$$\frac{dK_{\text{ОС}}}{K_{\text{ОС}}} = \frac{1}{1 + \beta_0 K} \frac{dK}{K} - \frac{\beta_0 K}{1 + \beta_0 K} \frac{d\beta_0}{\beta_0}.$$

Это выражение демонстрирует зависимость относительной нестабильности коэффициента усиления цепи с ООС от относительных

нестабильностей прямого и обратного каналов. Видно, что при $\beta_0 K \gg 1$

$$\frac{dK_{\text{ОС}}}{K_{\text{ОС}}} \approx -\frac{d\beta_0}{\beta_0},$$

т.е. относительная нестабильность коэффициента усиления цепи с глубокой ООС определяется только нестабильностью цепи ОС. Поскольку цепь ОС, как правило, не содержит активных элементов, ее стабильность на порядки превосходит стабильность прямого канала.

Пример 6.1. Пусть имеется усилитель мощности с коэффициентом усиления 10 и относительной нестабильностью $\frac{dK}{K} = 0,2$. Необходимо обеспечить относительную нестабильность в сто раз меньше [14].

Для этого следует включить каскадно с усилителем мощности предварительный усилитель с коэффициентом усиления 100 и охватить оба усилителя отрицательной ОС с коэффициентом $\beta = 0,099$. Тогда общий коэффициент усиления останется равным 10:

$$K_{\text{ОС}} = \frac{1000}{1 + 0,099 \cdot 1000} = 10.$$

При этом его относительная нестабильность

$$\frac{dK_{\text{ОС}}}{K_{\text{ОС}}} \approx \frac{1}{1 + \beta_0 K} \frac{dK}{K} = 0,002.$$

(Здесь принято, что нестабильность цепи ОС пренебрежимо мала.)◄

Ранее мы увидели, что положительная ОС сужает полосу пропускания цепи. Естественно поэтому предположить, что отрицательная ОС должна расширять полосу пропускания, или, что равнозначно, повышать равномерность АЧХ.

Рассмотрим резистивный усилитель, показанный на рис. 6.4.

Комплексная частотная характеристика усилителя определяется выражением

$$K(j\omega) = \frac{-SR_{\text{экв}}}{1 + j\omega\tau_{\text{экв}}}, \qquad (6.5)$$

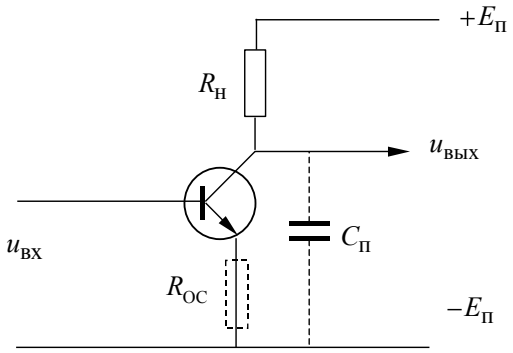


Рис. 6.4. Резистивный усилитель с отрицательной ОС

постоянная времени $\tau_{\text{экв}} = R_{\text{экв}} C_{\text{П}}$, где $C_{\text{П}}$ – паразитная емкость, а $R_{\text{экв}}$ образуется параллельным соединением сопротивления нагрузки $R_{\text{Н}}$ и внутреннего сопротивления транзистора R_i, т.е. $R_{\text{экв}} = R_{\text{Н}} R_i / (R_{\text{Н}} + R_i)$. Амплитудно-частотная характеристика усилителя характеризуется граничной частотой $\omega_{\text{гр}} = 1/\tau_{\text{экв}}$.

Отрицательная ОС может быть введена путем включения сопротивления обратной связи в цепь эмиттера (на схеме показано штриховой линией). Полярность напряжения переменного тока, падающего на сопротивлении ОС, совпадает с полярностью входного напряжения, поэтому напряжение между базой и эмиттером уменьшается за счет введения ОС, т.е. ОС действительно является отрицательной. Коэффициент ОС равен $R_{\text{ос}} / R_{\text{Н}}$.

При введении отрицательной ОС

$$K_{\text{ОС}}(j\omega) = \frac{\dfrac{-SR_{\text{экв}}}{1 + j\omega\tau_{\text{экв}}}}{1 + \beta_0 \dfrac{SR_{\text{экв}}}{1 + j\omega\tau_{\text{экв}}}} = \frac{-SR_{\text{экв}}}{1 + \beta_0 SR_{\text{экв}} + j\omega\tau_{\text{экв}}} =$$

$$= \frac{\dfrac{-SR_{\text{экв}}}{1 + \beta_0 SR_{\text{экв}}}}{1 + j\omega \dfrac{\tau_{\text{экв}}}{1 + \beta_0 SR_{\text{экв}}}}.$$

Из полученного выражения следует, что введение отрицательной ОС приводит к уменьшению коэффициента усиления на нулевой частоте, но при этом во столько же раз расширяется полоса пропускания, так как увеличивается граничная частота, которая теперь равна

$$\omega_{\text{гр.ОС}} = \frac{1}{\dfrac{\tau_{\text{экв}}}{1+\beta_0 SR_{\text{экв}}}} = \left(1+\beta_0 SR_{\text{экв}}\right)\omega_{\text{гр}}.$$

6.2. УСТОЙЧИВОСТЬ ЦЕПЕЙ С ОБРАТНОЙ СВЯЗЬЮ

Особенность цепей с обратной связью состоит в том, что при определенных условиях ОС может вызвать *неустойчивость*. Говоря нестрого, цепь называется устойчивой, если малые изменения входного сигнала приводят к малым изменениям выходного сигнала[90]. Для линейных цепей это означает, что ограниченному воздействию соответствует ограниченная реакция[91].

Из теории цепей известно, что для устойчивости ЛИС-цепи необходимо и достаточно, чтобы полюсы передаточной функции находились слева от мнимой оси p-плоскости (для пассивных цепей это выполняется всегда). Поскольку полюсы – это такие значения переменной p, при которых передаточная функция обращается в бесконечность, из выражения (6.1) видно, что полюсы цепи с ОС не совпадают с полюсами прямого канала. Таким образом, если прямой канал устойчив, то введение ОС может сместить полюсы в правую p-полуплоскость и нарушить устойчивость (привести к самовозбуждению).

Если выражения для передаточных функций известны и являются рациональными (т.е. числитель и знаменатель представляют собой полиномы), то для выяснения вопроса об устойчивости цепи с ОС можно воспользоваться критериями устойчивости, среди которых наиболее известными являются алгебраический *критерий Рауса – Гурвица* и графоаналитический *критерий Михайлова* [10]. Однако на практике иногда приходится исследовать устойчивость

[90] Более подробно об устойчивости см., например, [14].

[91] Иногда так определенную устойчивость называют ОВОВ-устойчивостью (от слов «ограниченный вход – ограниченный выход»).

цепей на основе их характеристик, полученных экспериментально и представленных в виде таблиц или графиков. В этом случае может быть использован графоаналитический *критерий Найквиста*. Для его применения достаточно знать передаточные функции (или КЧХ) прямого и обратного каналов[92], представленных в любом виде.

Для выяснения устойчивости цепи необходимо построить так называемый *годограф* Найквиста, т.е. графическое изображение на комплексной плоскости комплексной частотной характеристики *разомкнутой цепи* (прямого и обратного каналов, соединенных каскадно) $\beta(j\omega)K(j\omega)$, при этом частота рассматривается как параметр, изменяющийся от $-\infty$ до ∞. Для такого построения, очевидно, достаточно знать АЧХ и ФЧХ разомкнутой цепи: при каждом значении частоты на комплексную плоскость наносится точка, отстоящая от нулевой точки плоскости на расстояние, равное значению АЧХ при данной частоте, а от вещественной оси по углу – на величину ФЧХ при данной частоте.

Согласно критерию Найквиста, *цепь является устойчивой, если построенный годограф не охватывает точку* $1 + j \cdot 0$.

Физический смысл критерия Найквиста прост: если при прохождении по разомкнутой цепи гармоническое колебание какой-нибудь частоты приобретает фазовый сдвиг, кратный 2π, и его амплитуда не уменьшается, то такой сигнал в замкнутой цепи будет поддерживать сам себя (а если АЧХ разомкнутой цепи на этой частоте больше единицы, то в замкнутой цепи произойдет *самовозбуждение,* т. е. возникновение колебаний[93] и их неограниченный рост). В *автогенераторах* колебаний неустойчивость создается *преднамеренно*, а условия неустойчивости называют условиями *баланса фаз*

$$\arg\{\beta(j\omega)K(j\omega)\} = \pm 2\pi k$$

и *баланса амплитуд*

$$|\beta(j\omega)K(j\omega)| > 1.$$

Пример 6.2. Рассмотрим усилитель, схема которого показана на рис. 6.4 (без сопротивления, показанного штриховой линией). Его КЧХ описывается выражением (6.5). Введем 100 %-ю обрат-

---

[92] Очевидно, предполагается *внешняя* ОС.

[93] Роль начального слабого колебания, с которого начинается процесс самовозбуждения, в таких случаях играет тепловой шум, который имеет широкий спектр и содержит, таким образом, гармоники *всех* частот.

ную связь, соединив коллектор транзистора с базой, при этом $\beta(\omega) \equiv 1$, и годограф разомкнутой цепи строится по КЧХ усилителя. Рассматривая по отдельности АЧХ и ФЧХ, можно записать

$$K(j\omega) = \frac{-SR_{\text{экв}}}{1 + j\omega\tau_{\text{экв}}} = \frac{SR_{\text{экв}}}{\sqrt{1+\omega^2\tau_{\text{экв}}^2}} \exp(j\pi - j\,\text{arctg}\,\omega\tau_{\text{экв}}).$$

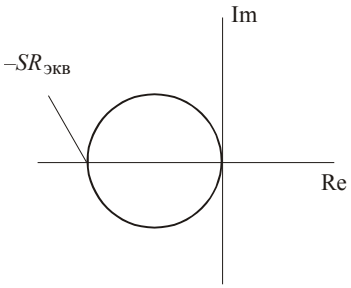


Рис. 6.5. Годограф Найквиста

Годограф Найквиста для этой КЧХ показан на рис. 6.5. Очевидно, цепь устойчива при любых значениях S и $R_{\text{экв}}$.

Устойчивость усилителя со 100 %-й отрицательной обратной связью естественна, так как сигнал на коллекторе противофазен сигналу на базе, и при прохождении по цепи ОС колебания ослабляются. ◄

Пример 6.3. Рассмотрим вопрос об устойчивости двух таких же усилителей, соединенных каскадно и охваченных 100 %-й обратной связью. КЧХ разомкнутой цепи в этом случае имеет вид

$$K(j\omega) = \frac{S^2R_{\text{экв}}^2}{1+\omega^2\tau_{\text{экв}}^2} \exp\left(-j \cdot 2\,\text{arctg}\,\omega\tau_{\text{экв}}\right).$$

Годограф Найквиста для этого случая показан на рис. 6.6. Очевидно, цепь неустойчива, если $S^2R_{\text{экв}}^2 > 1$. Это естественно, так как каждый из каскадов инвертирует сигнал, значит, обратная связь является положительной, и если значение АЧХ разомкнутой цепи (*петлевое усиление*) оказывается для некоторых частот больше 1, то гармонические составляющие соответствующих частот при прохождении цепи ОС усиливают себя – усилитель превращается в генератор незатухающих колебаний (*автоколебаний*). ◄

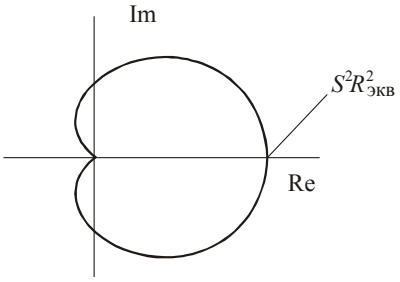


Рис. 6.6. Годограф Найквиста для двухкаскадного усилителя

6.3. АВТОГЕНЕРАТОРЫ КОЛЕБАНИЙ

Генерирование колебаний – одна из важнейших задач, касающихся построения систем передачи информации. Устройства, предназначенные для этой цели, называются *автогенераторами,* или генераторами с самовозбуждением[94]. Назначение автогенератора (рис. 6.7) состоит в преобразовании мощности источника питания ИП в мощность незатухающих периодических колебаний. Различают автогенераторы гармонических и негармонических (прямоугольных, пилообразных и т.д.) колебаний.

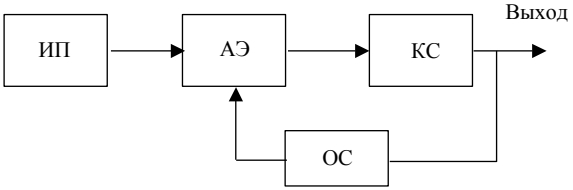


Рис. 6.7. Структура автогенератора

Необходимыми элементами автогенератора являются активный элемент АЭ (транзистор, электронная лампа и т.п.), колебательная система КС (контур, резонатор и т.п.) и положительная обратная связь (внутренняя или внешняя).

Рассмотрим *условия самовозбуждения* автогенератора гармонических колебаний, в котором в качестве АЭ используется полевой транзистор (рис. 6.8). Входное сопротивление полевого транзистора велико, и это позволяет пренебречь влиянием АЭ на характеристики колебательной системы, которая в данном случае представляет собой параллельный колебательный контур. Обратная связь в этом генераторе является внешней и обеспечивается катушкой связи $L_{\text{св}}$.

Обозначим ток в цепи истока – стока транзистора i, ток, протекающий по колебательному контуру $i_{\text{к}}$, напряжение на катушке индуктивности u_L, напряжение на конденсаторе u_C, напряжение на резисторе u_R, коэффициент взаимоиндукции M. Учитывая известные соотношения

$$i_{\text{к}} = C\frac{du_C}{dt}, \quad u_L = L\frac{di_{\text{к}}}{dt}, \quad u_R = Ri_{\text{к}},$$

[94] Так называемые *генераторы с внешним возбуждением* представляют собой фактически *усилители* колебаний, поступающих с автогенератора.

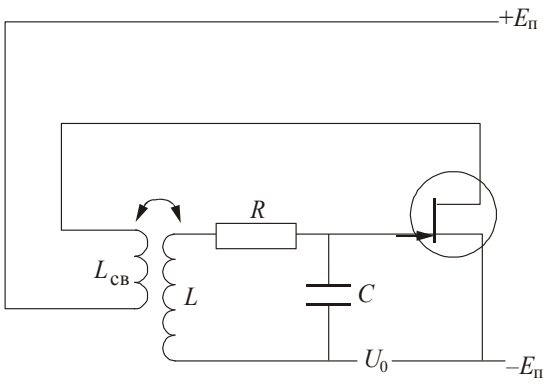


Рис. 6.8. Автогенератор на полевом транзисторе

а также $u_L + u_R + u_C = 0$ для одиночного колебательного контура, запишем уравнение, учитывающее обратную связь:

$$LC\frac{d^2u_C}{dt^2} + RC\frac{du_C}{dt} + u_C = \pm M\frac{di}{dt},$$

где $\frac{di}{dt} = S\frac{du_C}{dt}$ (поскольку рассматриваются *условия самовозбуждения* автогенератора, оправданно предположение о малом сигнале, и транзистор считается линейным устройством с крутизной ВАХ, равной S). Знак при M определяется порядком подключения выводов катушки связи.

Уравнение можно переписать в виде

$$\frac{d^2u_C}{dt^2} + \left(\frac{R}{L} \mp \frac{MS}{LC}\right)\frac{du_C}{dt} + \frac{1}{LC}u_C = 0.$$

Обозначим $2\alpha = \frac{R}{L} \mp \frac{MS}{LC}$, $\omega_0^2 = \frac{1}{LC}$, тогда общее решение данного уравнения имеет вид

$$u_C(t) = C_1 e^{-\alpha t + j\omega_к t} + C_2 e^{-\alpha t - j\omega_к t}, \qquad (6.6)$$

где константы C_1 и C_2 определяются начальными условиями, частота колебаний $\omega_к^2 = \sqrt{\omega_0^2 - \alpha^2}$.

Самовозбуждение соответствует возрастанию амплитуды колебаний, поэтому из $\alpha < 0$ получаем *условие самовозбуждения*

$$\frac{R}{L} - \frac{MS}{LC} < 0, \tag{6.7}$$

откуда следует, что обратная связь должна быть положительной и обеспечивать уменьшение сопротивления потерь за счет вносимого *отрицательного* сопротивления

$$R_{\text{вн}} = -\frac{MS}{C}.$$

Для этого коэффициент взаимоиндукции должен удовлетворять выражению

$$M > \frac{RC}{S}. \tag{6.8}$$

При выполнении условия (6.8) слабые колебания, описываемые выражением (6.6), будут со временем возрастать[95] и их амплитуда будет стремиться к бесконечности. Ясно, что в реальных устройствах это невозможно, поэтому любой реальный автогенератор обязан быть *нелинейным,* тогда в выражении (6.8) вместо крутизны S линейного активного элемента фигурирует *средняя к*рутизна по первой гармонике, т.е отношение амплитуды первой гармоники тока к амплитуде гармонического входного напряжения. Таким образом, можно автогенератор рассматривать как нелинейный усилитель, охваченный положительной обратной связью. Такой усилитель описывается колебательной характеристикой (см. разд. 5.2) $I_1 = f(U_m)$, где I_1 – амплитуда первой гармоники тока стока, а U_m – амплитуда гармонического напряжения на затворе транзистора. В зависимости от выбора рабочего режима транзистора возможны два вида колебательной характеристики (рис. 6.9).

Перепишем условие возникновения колебаний (6.8) в виде

$$S > \frac{RC}{M}. \tag{6.9}$$

Обеспечить выполнение этого условия можно изменением степени связи катушек индуктивности M (или изменением сопротивления потерь контура, подключая параллельно контуру переменное

[95] Напомним, что роль начальных колебаний малой амплитуды в автогенераторах играют флюктуационные тепловые шумы.

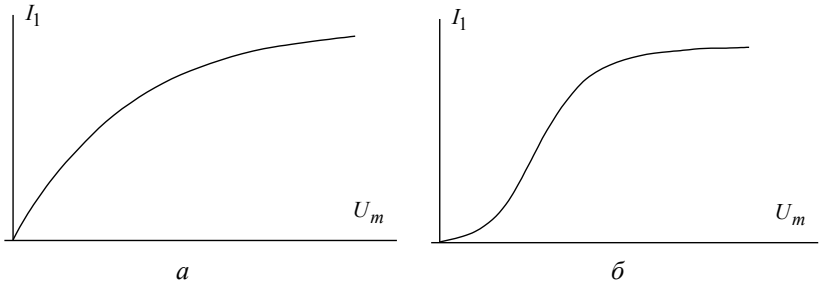


Рис. 6.9. Виды колебательной характеристики

шунтирующее сопротивление). На графике зависимости средней крутизны от амплитуды U_m напряжения на затворе (рис. 6.10, *а*) величина RC/M отображена горизонтальной линией. Если линия занимает положение *1*, условие самовозбуждения не выполнено и автоколебания отсутствуют. По мере усиления связи прямая опускается и в положении *2* условие (6.9) выполняется для бесконечно малых значений амплитуды U_m. Поскольку в реальных устройствах всегда есть тепловой шум, возникают автоколебания бесконечно малой амплитуды; если связь усиливается (прямая опускается), то происходит *монотонное* увеличение амплитуды. При некотором положении *3* прямой автоколебания имеют амплитуду стационарного режима U_{m0}. Изменяя степень связи (перемещая прямую RC/M вверх или вниз), можно плавно регулировать амплитуду стационарного режима от нуля до максимума, определяемого практическими целями. При этом рабочая точка (точка пересечения прямой и графика средней крутизны) перемещается по графику в направлениях, указанных стрелками. Такой режим работы автогенератора называется *мягким*.

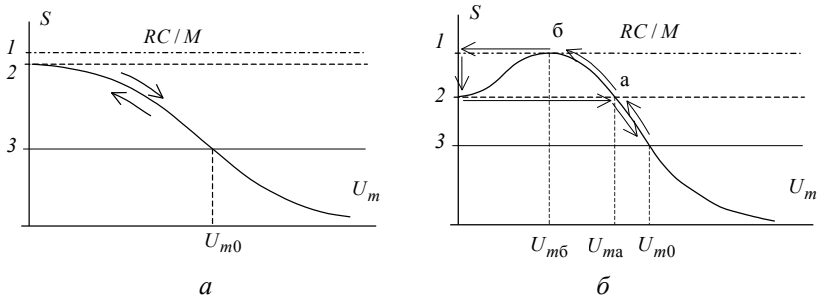


Рис. 6.10. Мягкий и жесткий режимы самовозбуждения

Заметим, что поддержание постоянной амплитуды колебаний в стационарном режиме можно рассматривать как проявление *отрицательной обратной связи*. Действительно, увеличение амплитуды колебаний приводит к снижению крутизны, что способствует уменьшению амплитуды колебаний, и наоборот. Поэтому установившийся режим колебаний амплитуды U_{m0} можно считать состоянием *устойчивого равновесия*, так как при спонтанном увеличении амплитуды снижение крутизны приведет к ослаблению колебаний, а при уменьшении амплитуды, наоборот, вследствие увеличения крутизны амплитуда колебаний увеличится.

Иная картина возникновения автоколебаний имеет место в случае, показанном на рис. 6.10, *б*. Здесь увеличение степени связи приводит к тому, что в положении *1* прямой RC/M условия самовозбуждения выполняются для колебаний конечной (и довольно большой) амплитуды $U_{m\text{б}}$, но, поскольку колебания такой амплитуды в генераторе отсутствуют, возбуждение не происходит. То же самое относится ко всем положениям прямой между положениями *1* и *2*.

Ситуация меняется, когда прямая находится в положении *2*. При этом условия самовозбуждения выполняются для бесконечно малых колебаний, но они также выполняются для колебаний всех амплитуд, меньших, чем значение $U_{m\text{а}}$, определяемое пересечением прямой и графика средней крутизны. Поэтому происходит лавинообразное нарастание амплитуды генерируемых колебаний от 0 до $U_{m\text{а}}$ и рабочая точка занимает положение а. Дальнейшее увеличение связи приводит к плавному увеличению амплитуды до желаемого значения U_{m0}.

Уменьшение амплитуды колебаний при уменьшении степени связи происходит плавно до точки б, где условие (6.9) перестает выполняться, при этом колебания лавинообразно затухают (происходит *срыв* колебаний). Таким образом, рабочая точка при изменениях связи движется в направлениях, указанных на рис. 6.10, *б* стрелками, описывая траекторию гистерезисного типа. Колебания амплитуды меньше чем $U_{m\text{б}}$ невозможны. Такой режим самовозбуждения называется *жестким*.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какой вид обратной связи стабилизирует коэффициент усиления?
2. Какой вид обратной связи увеличивает коэффициент усиления?
3. Можно ли сузить полосу пропускания резонансного усилителя при помощи ОС? Какая это ОС?
4. Какая ОС применяется в автогенераторах?
5. Что такое мягкий и жесткий режимы?
6. Что такое вносимое сопротивление и какова его природа?
7. Что такое устойчивость и как выяснить, устойчива ли цепь?

УПРАЖНЕНИЯ

1. Постройте качественно график зависимости средней крутизны по первой гармонике от амплитуды гармонического напряжения для нелинейного усилителя при кусочно-линейной аппроксимации ВАХ нелинейного элемента (примите $U_{\text{н}} = 1$ В, $U_0 = 1{,}5$ В, $S = 2$ мА/В, при напряжении 3 В наступает насыщение ВАХ). Какой режим возбуждения возможен, если этот усилитель охватить положительной ОС?
2. Повторите задание упражнения 1 для $U_{\text{н}} = 1$ В, $U_0 = 0.5$ В, $S = 2$ мА/В. Сравните результаты.
3. Постройте (качественно) графики зависимости амплитуды генерируемых колебаний от коэффициента взаимоиндукции M для мягкого и жесткого режимов.

7. КАНАЛЫ СВЯЗИ

Каналом связи называется совокупность устройств и линий связи, которые сигнал проходит последовательно между *любыми* двумя точками системы связи. Понятие канала является, таким образом, очень широким: канал в простейшем случае может состоять из пары проводов, в то же время при передаче сообщения с одного континента на другой канал включает многочисленные формирующие, коммутирующие, преобразующие, усиливающие, фильтрующие устройства и различные среды распространения колебаний (волноводы, кабели, свободное пространство, кварцевое стекло оптоволоконных линий и т.д.).

При передаче по каналу сигнал претерпевает изменения, как правило, весьма значительные, и в этом смысле канал, как и цепь, можно математически описать некоторым отображением множества входных сигналов на множество выходных сигналов (см. разд. 2.7). Во многих случаях канал связи можно представить *каскадным* соединением достаточно простых звеньев – устройств, а также участков среды распространения. Каждое такое звено может быть линейным или нелинейным, стационарным или нестационарным (параметрическим), инерционным или безынерционным. В такой ситуации можно анализировать прохождение сигнала через звенья последовательно, рассматривая выходной сигнал одного звена как входной для последующего и при этом использовать для анализа методы, рассмотренные в разд. 2 и 5. В редких случаях канал связи *в целом* можно считать линейным стационарным; чаще в канале имеют место нелинейные преобразования и/или временная зависимость характеристик. Иногда удобно рассматривать канал как каскадное соединение нескольких каналов, а в других случаях (например, при многолучевом распространении колебаний) – как более сложное, включающее участки, соединенные как каскадно, так и параллельно. В общем случае каналы связи заслуживают отдельного рассмотрения, которому и посвящен этот раздел.

Напомним, что каналы связи могут быть классифицированы по различным признакам, в частности, по назначению (телеграфные, фототелеграфные, телефонные и т.д.), по виду используемой среды (проводные и радиоканалы), по характеру связи входных и выходных сигналов (линейные и нелинейные, стационарные и нестационарные, детерминированные и случайные); по типу входных и выходных сигналов (аналоговые, цифровые, аналогово-цифровые и цифро-аналоговые), по количеству независимых переменных в описании сигналов – временные и пространственно-временные (в этом случае сигнал описывается функцией не только времени, но и пространственных координат, т.е. представляет собой поле; тогда и канал связи требует пространственно-временного описания). В дальнейшем, если тип канала явно не указан, подразумевается временно́й аналоговый канал.

Неотъемлемым свойством любого реального канала связи является наличие в нем источников помех. Помехи могут быть классифицированы по следующим признакам[96].

По происхождению помехи делятся на *естественные* и *преднамеренные* (*искусственные*). Источниками естественных помех являются природные процессы – грозовые явления, космические излучения, шум океана в подводной акустической связи, тепловые шумы, имеющие место во всех без исключения устройствах и средах, а также излучения техногенного характера (наводки от промышленных установок, рентгеновских аппаратов, автомобильного и железнодорожного транспорта и т.п.). Искусственные помехи создаются сознательно, чтобы помешать или полностью воспрепятствовать противнику (военному, политическому, экономическому и т.д.) передавать и получать информацию.

По месту возникновения помехи делятся на внутренние и внешние. К *внутренним* помехам относятся тепловые шумы устройств, входящих в канал, дробовые шумы электронных и полупроводниковых приборов, обусловленные дискретностью носителей зарядов, фликкер-шум полупроводниковых приборов, а также сигналы, попадающие в канал по внутренним цепям вследствие плохого экранирования или развязки между каскадами. *Внешние* помехи порождаются различными электромагнитными процессами в атмосфере, космическом пространстве (космические шумы), а

[96] Эта классификация, как и всякая другая, не является полной. Так, в радиолокации, кроме описанных здесь типов помех, рассматривают *активные* и *пассивные* помехи, и т.д.

также посторонними радиостанциями (*станционные* помехи), промышленными установками, транспортом и т.п.

По характеру проявления помехи могут быть *шумовыми* (*флюктуационными*), а также *сосредоточенными* (по времени или по спектру). Шумовая помеха представляет собой случайный процесс, который во многих каналах можно считать стационарным в широком смысле эргодическим процессом. Сосредоточенная по времени помеха, называемая также импульсной, имеет более сложное описание, которое включает случайный поток событий, управляющий моментами появления импульсов, а также набор случайных параметров, определяющих форму импульсов. Помеха, сосредоточенная по спектру, обычно описывается как узкополосный случайный процесс.

По способу взаимодействия с сигналом помехи подразделяются на *аддитивные* (от английского *add* – складывать), *мультипликативные* (от английского *multiply* – умножать) и смешанные (этим термином обозначаются сложные функциональные взаимодействия, не сводимые к сложению или умножению).

Для полного описания канала связи следовало бы задать множество допустимых входных сигналов и для каждого допустимого сигнала указать соответствующий выходной сигнал. Наличие помех и случайный характер любого сигнала делают эту задачу вероятностной, таким образом, исчерпывающее описание канала связи должно представлять собой условное распределение вероятностей *для всевозможных выходных* сигналов при *любом заданном допустимом входном* сигнале. Получение и использование такого описания в большинстве случаев представляет собой слишком сложную задачу, поэтому на практике используются упрощенные модели, позволяющие при умеренных вычислительных затратах получать результаты приемлемой точности.

Ниже кратко рассматриваются математические модели некоторых каналов связи, наиболее широко используемые в настоящее время.

7.1. КАНАЛ С АДДИТИВНЫМ ШУМОМ

Это простейший канал, единственное воздействие которого на сигнал $x(t)$ состоит в сложении его с шумом $\xi(t)$, в результате чего получается выходное колебание $z(t)$ (рис. 7.1). Чаще всего полагают, что шум имеет нормальное распределение, тогда канал

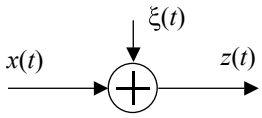


Рис. 7.1. Модель канала с аддитивным шумом

называется каналом с аддитивным гауссовским шумом (АГШ). Иногда для учета затухания сигналов в канале модель дополняют масштабным звеном с частотно-независимым коэффициентом передачи γ, тогда выходное колебание имеет вид

$$z(t) = \gamma x(t) + \xi(t). \qquad (7.1)$$

Некоторое дальнейшее усложнение модели производится путем учета задержки сигнала, вносимой каналом, когда

$$z(t) = \gamma x(t - \tau) + \xi(t). \qquad (7.2)$$

В некоторых случаях предполагается, что коэффициент передачи зависит от времени детерминированным образом, тогда $z(t) = \gamma(t)x(t - \tau) + \xi(t)$. Несмотря на простоту модели канала с аддитивным шумом, она часто используется для описания проводных каналов, а также радиоканалов при связи в пределах прямой видимости [10].

7.2. ЛИНЕЙНЫЙ СТАЦИОНАРНЫЙ (ФИЛЬТРОВОЙ) КАНАЛ

Эта модель учитывает частотно-избирательные свойства устройств и физических сред, входящих в канал. Строго говоря, все реальные устройства и среды обладают временной инерционностью, а следовательно, и частотной избирательностью. Во многих случаях этими свойствами пренебречь нельзя, и они учитываются в модели фильтрового канала. Так, в проводной телефонной связи используются фильтры, которые предназначены для частотного разделения сигналов, передаваемых по общей линии[97], и частотные свойства фильтров должны быть учтены в модели. Линейный стационарный канал, как и ЛИС-цепь, полностью описывается импульсной характеристикой $h(t)$ во временно́й области или комплексной частотной характеристикой $H(f)$ – в частотной (рис. 7.2). Поэтому для анализа таких каналов пригодны методы анализа ЛИС-цепей, рассмотренные ранее (временной метод, основанный на интеграле Дюамеля, спектральный и операторный методы, а также приближенные методы). Полагая шум пренебрежимо малым,

[97] Подробнее о частотном разделении каналов см. разд. 11

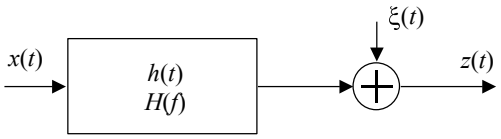


Рис. 7.2. Модель фильтрового канала

получаем модель *идеального канала без помех*, которую иногда используют при рассмотрении каналов малой протяженности с закрытым распространением (волновод, кабель, световод) [10]. Следует отметить, что идеальный канал вносит искажения сигнала вследствие инерционности, которые, например, при цифровой передаче могут приводить к *межсимвольной интерференции* – наложению друг на друга соседних посылок, если длительность посылки меньше, чем длительность импульсной характеристики (время памяти канала).

7.3. ЛИНЕЙНЫЙ НЕСТАЦИОНАРНЫЙ КАНАЛ

Во многих случаях канал при выполнении свойства линейности (принципа суперпозиции) нельзя считать стационарным (инвариантным к сдвигу по времени). Таковы, например, каналы подводной связи, ионосферные каналы, радиоканалы в системах подвижной связи и т.п. Тогда используется модель линейного нестационарного канала с аддитивным шумом (рис. 7.3), где выходное колебание определяется выражением

$$z(t) = \int_{-\infty}^{\infty} h(t, \tau) x(t - \tau) d\tau + \xi(t) ; \qquad (7.3)$$

здесь $h(t, \tau)$ – весовая функция (ядро) линейного оператора, описывающего канал, которая имеет физический смысл отклика канала в момент времени t на сигнал в виде δ -функции, действующий на

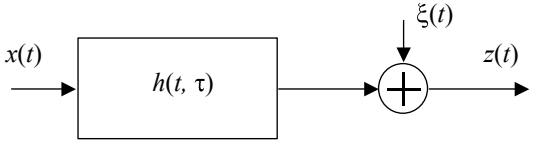


Рис. 7.3. Модель линейного нестационарного канала

вход канала в момент времени $(t-\tau)$. Частотные свойства канала можно описать комплексной частотной характеристикой

$$H(f,t)=\int_{-\infty}^{\infty} h(t,\tau)e^{-j\cdot 2\pi f\tau}\,d\tau, \qquad (7.4)$$

зависящей от времени.

7.4. СЛУЧАЙНЫЙ ЛИНЕЙНЫЙ КАНАЛ

Многие каналы, которые можно считать с достаточной для практики точностью линейными, изменяют во времени свои свойства, причем эти изменения нельзя предсказать заранее. В такой ситуации естественно считать изменения свойств канала случайными. При этом канал по-прежнему описывается выражениями (7.3) и (7.4), где, однако, ядро $h(t,\tau)$ и КЧХ $H(f,t)$ являются случайными функциями. Приведем несколько примеров случайного линейного канала.

7.4.1. КАНАЛ СО СЛУЧАЙНЫМИ ЗАТУХАНИЕМ И ЗАДЕРЖКОЙ

Линейный случайный канал можно описать в простейших случаях выражениями (7.1) и (7.2), в которых коэффициент γ и задержка τ рассматриваются как случайные величины или медленные случайные процессы $\gamma(t)$ и $\tau(t)$. Имеется в виду, что интервалы корреляции процессов $\gamma(t)$ и $\tau(t)$ значительно превосходят интервал корреляции входного сигнала. Причиной таких медленных флюктуаций может быть изменение физических условий распространения в линии (например, изменение со временем температуры, влажности и других характеристик при ионосферной и тропосферной связи, изменение расстояния при связи с подвижными объектами и т.п.). Если распределение шума нормальное, то канал называют гауссовским каналом с неопределенной амплитудой и фазой [10].

7.4.2. КАНАЛ С МНОГОЛУЧЕВЫМ РАСПРОСТРАНЕНИЕМ

Такая модель предполагает, что сигнал распространяется по нескольким траекториям (лучам), причем каждый луч представляет собой рассмотренный только что канал со случайным затуханием и случайной задержкой, а сигналы на выходах таких *парциальных* каналов складываются. Рассмотрим канал с многолучевым распространением по N путям при гармоническом комплексном воздействии $e^{j\omega t}$. Парциальные каналы будем характеризовать комплексными коэффициентами передачи $\dot{\gamma}_i = \gamma_i e^{-j\omega\tau_i}$, $i = \overline{1,N}$. Тогда выходной сигнал равен

$$y(t) = \mathrm{Re}\left\{\sum_{i=1}^{N}\gamma_i e^{j\omega(t-\tau_i)}\right\} = \mathrm{Re}\left\{\sum_{i=1}^{N}\dot{\gamma}_i e^{j\omega t}\right\} = \mathrm{Re}\left\{e^{j\omega t}\sum_{i=1}^{N}\dot{\gamma}_i\right\}.$$

Таким образом, многолучевой канал описывается комплексной частотной характеристикой $H(\omega) = \sum_{i=1}^{N}\gamma_i e^{-j\omega\tau_i} = \sum_{i=1}^{N}\dot{\gamma}_i$. Если затухание и задержка в парциальных каналах флюктуируют, то КЧХ представляет собой случайную функцию частоты, меняющуюся со временем. Поэтому даже если входной сигнал является детерминированным, а уровень помех в канале пренебрежимо мал, выходное колебание имеет случайный характер. В частности, при гармоническом входном сигнале амплитуда и начальная фаза выходного колебания определяются случайными величинами – значениями АЧХ и ФЧХ (модуля и аргумента КЧХ) на частоте входного сигнала.

Предположим, что количество путей N велико, все величины γ_i независимы и имеют один порядок (т.е. среди них нет преобладающих), а дисперсии независимых задержек τ_i настолько велики, что вносимые фазовые сдвиги имеют распределение, практически равномерное в интервале $(0, 2\pi)$. Тогда в силу центральной предельной теоремы вещественная и мнимая части КЧХ имеют распределение, близкое к нормальному, и одинаковые дисперсии σ^2.

Распределение вероятности модуля γ КЧХ является в этом случае рэлеевским (см. разд. 3.6) и имеет плотность

$$W(\gamma) = \frac{\gamma}{\sigma^2} e^{-\frac{\gamma^2}{2\sigma^2}}.$$

Если условия распространения таковы, что кроме многочисленных флюктуирующих лучей, имеющих одинаковый порядок затухания, имеется также регулярный канал с малым затуханием, то распределение γ оказывается обобщенным рэлеевским с плотностью

$$W(\gamma) = \frac{\gamma}{\sigma^2} e^{-\frac{\gamma^2+\Gamma^2}{2\sigma^2}} I_0\left(\frac{\gamma\Gamma}{\sigma^2}\right),$$

где Γ^2/σ^2 – отношение средних мощностей регулярной и флюктуирующих составляющих, $I_0(\cdot)$ – модифицированная функция Бесселя нулевого порядка.

7.5. НЕЛИНЕЙНЫЙ КАНАЛ

Любой реально действующий канал может рассматриваться как линейный лишь приближенно и только при определенных условиях (например, при не очень больших уровнях сигнала). Учитывая то, что все реальные каналы обладают также инерционностью, следовало бы рассмотреть самую общую модель – нелинейный инерционный канал. В принципе такой подход возможен, но он настолько трудоемок, что обычно идут по другому пути: представляют канал каскадным соединением линейных инерционных и нелинейных безынерционных звеньев. Анализ прохождения сигнала через такие звенья в отдельности сравнительно прост, а получаемые результаты имеют достаточную для практики точность.

Нелинейное безынерционное звено, как было показано в разд. 5, при воздействии на него гармонического колебания обогащает спектр сигнала кратными гармониками, а при бигармоническом воздействии – составляющими с кратными и комбинационными частотами. В каналах связи сигналы всегда имеют широкий непрерывный (сплошной) спектр, поэтому нелинейность в канале приводит к появлению продуктов взаимодействия различных гармоник сигнала, взаимодействия между собой различных частотных составляющих шума, перекрестного взаимодействия гармонических составляющих

сигнала и шума, причем спектр этих новых составляющих также является сплошным и занимает практически ту же полосу частот, что и полезный сигнал, поэтому подавить их путем фильтрации невозможно. При построении систем связи стремятся сделать канал по возможности близким к линейному. Требования к линейности канала несколько ослабляются при использовании временно́го уплотнения (см. разд. 11).

7.6. ДИСКРЕТНО-НЕПРЕРЫВНЫЕ КАНАЛЫ

Дискретно-непрерывный канал характеризуется дискретным входным сигналом и непрерывным выходным, что соответствует совокупности технических средств от выхода кодера до входа демодулятора[98] (рис. 1.3). Для полного описания дискретно-непрерывного канала необходимо задать алфавит входных символов (кодовых символов), априорные вероятности их появления, а также *условные* многомерные плотности вероятности для реализаций случайного процесса на входе демодулятора для каждого символа, который может быть передан по каналу. Полагая, что непрерывный канал (от выхода модулятора до входа демодулятора) имеет полосу пропускания $F_к$, а длительность элементарного сигнала (посылки) равна $T_с$, на основании теоремы отсчетов можно полагать, что колебание $z(t)$ на входе демодулятора может быть заменено его отсчетами, взятыми через интервал дискретизации $T_d = 1/(2F_к)$. Таких отсчетов на протяжении сигнала оказывается $M = 2F_к T_с$, таким образом, упомянутые условные плотности должны быть M -мерными[99]. Заметим, что число M имеет смысл приближенной размерности пространства сигналов, которые могут поступить на вход демодулятора.

Если условные плотности не зависят от времени, дискретно-непрерывный канал является *стационарным*. Если условные плотности не зависят от символов, передававшихся ранее, канал называется каналом *без памяти*. Реальные каналы обычно обладают памятью и нестационарны, тем не менее модель стационарного канала без памяти часто применяется ввиду ее сравнительной простоты.

---

[98] Заметим, что внутри этого канала содержится непрерывный канал.

[99] Если не накладывать ограничения на полосу канала, то приходится вместо условной многомерной плотности рассматривать *функционал плотности вероятности.*

7.7. ДИСКРЕТНЫЕ КАНАЛЫ

Дискретный канал имеет дискретный вход и дискретный выход, что соответствует каналу от выхода кодера до выхода демодулятора (входа декодера) (рис. 1.3). Для описания дискретного канала необходимо задать алфавит входных символов $\alpha_k, k=\overline{1,K}$, априорные вероятности их появления $p(\alpha_k), k=\overline{1,K}$, алфавит выходных символов (который, вообще говоря, не обязан совпадать с входным алфавитом[100]) $\beta_l, l=\overline{1,L}$, а также набор всех переходных (условных) вероятностей появления каждого выходного символа при условии передачи любого входного $p(\beta_l|\alpha_k),\ k=\overline{1,K},\ l=\overline{1,L}$. Входной алфавит и набор априорных вероятностей определяются источником дискретных сообщений и кодером, выходной алфавит – устройством (алгоритмом работы) декодера, а переходные вероятности – характеристиками непрерывного канала (в частности, уровнем помех) и устройством (алгоритмом работы) демодулятора. Очевидно, помехоустойчивость системы тем выше, чем ближе к единице условные вероятности правильного приема символов. Задача оптимального синтеза демодулятора играет важнейшую роль в построении систем связи и подробно рассматривается в разд. 9.

Если переходные вероятности не зависят от времени, дискретный канал является *стационарным*. Если переходные вероятности не зависят от символов, передававшихся ранее, канал называется каналом *без памяти*.

Если входной и выходной алфавиты совпадают $\alpha_k=\beta_k$, $k=\overline{1,K}$, вероятность ошибочного приема любого символа $p_{\text{ош}}$, а в случае ошибки может быть с равной вероятностью принят любой другой символ, т.е.

$$p(\alpha_l|\alpha_k)=\begin{cases} p_{\text{ош}}/(K-1), & \text{если } l\neq k,\\ 1-p_{\text{ош}}, & \text{если } l=k,\end{cases}$$

канал называют *симметричным*. Если, кроме того, вероятность ошибки не зависит от времени, имеет место *стационарный* симметричный канал. Наиболее проста модель стационарного симмет-

[100] Например, в *канале со стиранием* выходной алфавит содержит, кроме символов входного алфавита, специальный символ стирания, который появляется на выходе дискретного канала, когда демодулятор не может с уверенностью принять решение в пользу одного из входных символов.

ричного канала *без памяти*, в котором ошибки при приеме различных символов являются статистически независимыми. Для такого канала вероятность получения r ошибок при передаче n символов подчиняется биномиальному закону [23]

$$P_n(r) = C_n^r p_{\text{ош}}^r \left(1 - p_{\text{ош}}\right)^{n-r}.$$

Из этого выражения можно найти такие характеристики, как вероятность правильного приема блока из n символов $P_n(0) = (1 - p_{\text{ош}})^n$, вероятность приема блока, содержащего хотя бы одну ошибку $P_n(r \geq 1) = 1 - P_n(0) = 1 - (1 - p_{\text{ош}})^n$, вероятность появления в блоке m и более ошибок и т.д.

Память реального дискретного канала проявляется в том, что вероятность ошибки приема символа зависит от того, какие символы передавались ранее. Эта зависимость может возникнуть, например, вследствие межсимвольной интерференции в непрерывном фильтровом канале. Простейшей моделью дискретного канала с памятью является *марковская* модель, согласно которой дискретный канал может находиться в двух состояниях, каждому из которых соответствует определенная вероятность ошибки; состояние канала при приеме очередного символа определяется предыдущим символом. Более сложной является марковская модель порядка N, в которой состояние канала определяется N последними принятыми символами (последовательность состояний канала представляет собой N -связную *цепь Маркова*[101]).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое канал связи? Как описать канал?
2. Существуют ли линейные стационарные каналы?
3. Что такое многолучевость?
4. Как описываются дискретные каналы?
5. Что такое дискретно-непрерывный канал? Как он описывается?
6. К каким последствиям приводит нелинейность канала?

УПРАЖНЕНИЕ

Выведите формулу определения вероятности появления в блоке из n символов m и более ошибок для стационарного симметричного дискретного канала без памяти.

[101] Андрей Андреевич Марков (1856 – 1922) – выдающийся русский математик, известен своими достижениями в области теории вероятностей и др.

8. ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

8.1. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ

Информация относится к предельно широким понятиям, которым трудно или невозможно дать строгое определение, поэтому приходится прибегать к интуиции, объясняя термин «информация» через синонимичные понятия – «данные», «сведения» и т.п. Однако для решения инженерных задач требуется *количественное* определение информации. В теории и технике связи в настоящее время используется определение количества информации, предложенное К. Ше́нноном[102].

Для введения этого определения необходимо абстрагироваться от физического воплощения источника и семантического (смыслового) содержания сообщений. Дискретный источник сообщений тогда полностью определяется набором символов (*алфавитом*) $\mathrm{A} = \{\alpha_1, ..., \alpha_K\}$ (K – объем алфавита) и распределением вероятностей $P(\mathbf{a})$, заданным на множестве всех возможных *последовательностей* символов $\mathbf{a} = (a_1, ..., a_n)$, $a_i \in \mathrm{A}$ произвольной длины. В простейшем случае *источника без памяти* символы в последовательности являются независимыми, и распределение $P(\mathbf{a})$ полностью определяется набором *априорных* вероятностей отдельных символов $\{p(\alpha_k),\ k = 1, ..., K\}$. В более сложных моделях *источников с памятью* условная вероятность появления в последовательности определенного символа зависит от того, какие символы ему предшествуют. Например, в тексте телеграммы на русском языке после буквы «щ» можно ожидать букв «а», «у», «е», но не «ю»,

[102] Клод Элвуд Шеннон (1916 – 2001) – выдающийся американский математик и инженер, один из основоположников теории информации.

«я», «й» и т.п. Далее, как правило, рассматриваются источники без памяти.

В процессе передачи *информационная* последовательность символов, представляющая собой сообщение, может быть заменена другой, *кодовой* последовательностью, состоящей из символов кодового алфавита. Целью кодирования может быть более полное использование канала связи (экономное кодирование) или повышение достоверности передачи (помехоустойчивое кодирование). Естественно, кодовые последовательности характеризуются другими распределениями вероятностей, нежели информационные последовательности.

Канал связи (дискретный) формально описывается входным и выходным алфавитами $\mathbf{X} = \{x_1, ..., x_L\}$ и $\mathbf{Y} = \{y_1, ..., y_M\}$ разных в общем случае объемов L и M и условным распределением вероятностей $P(\mathbf{y} \mid \mathbf{x})$, заданным для всех возможных последовательностей $\mathbf{y}$ и $\mathbf{x}$ произвольной длины. Условное распределение $P(\mathbf{y} \mid \mathbf{x})$ описывает вероятностный механизм действия помех в канале. В простейшем случае *канала без памяти* распределение $P(\mathbf{y} \mid \mathbf{x})$ полностью определяется набором условных вероятностей для всех пар отдельных символов $P(y_j \mid x_i)$, $x_i \in \mathbf{X}$, $y_j \in \mathbf{Y}$.

Информация, согласно современным представлениям, – это *свойство* сообщения снимать (или уменьшать) неопределенность относительно исхода некоторого случайного опыта (например, относительно переданного символа). Действительно, во всех реальных случаях получатель сообщения что-то знает о некотором объекте или событии до опыта («a priori»), но ему известно не все, иначе не было бы необходимости передавать сообщение. Например, футбольный болельщик знает, с кем сегодня играла его любимая команда, но не знает, кто победил. Таким образом, до опыта (до получения сообщения) налицо некоторая неопределенность. После приема сообщения неопределенность исчезает (или, по крайней мере, уменьшается) вследствие получения информации. *Количество* получаемой информации, очевидно, должно быть связано со степенью снимаемой неопределенности. Так, принимая сообщение о событии, которое *достоверно* известно, информации мы не получаем.

Количественная мера информации должна удовлетворять следующим интуитивно очевидным требованиям:

- если исход опыта единствен (достоверное событие), то количество информации в сообщении о нем должно быть равно нулю;

- количество получаемой информации тем больше, чем более неожиданным является исход;
- общее количество информации в нескольких сообщениях об исходах опытов, независимых в вероятностном смысле, должно равняться сумме количеств информации в отдельных сообщениях (*аддитивность* информации).

Мера неожиданности сообщения $\mathbf{a}$ в виде $1/P(\mathbf{a})$ удовлетворяет второму требованию, однако она не равна нулю для достоверного события и не обладает свойством аддитивности: неожиданность двух независимых сообщений $\mathbf{a}_1$ и $\mathbf{a}_2$ равна, очевидно, $1/[P(\mathbf{a}_1)P(\mathbf{a}_2)]$. Чтобы обеспечить выполнение всех требований, необходимо определить *частное* (*индивидуальное*) количество информации в сообщении выражением

$$\log_m \frac{1}{P(\mathbf{a})} = -\log_m P(\mathbf{a}).$$

Основание логарифма может быть произвольным и определяет лишь масштаб (единицу измерения). Общепринятым является основание 2, при этом единица называется *битом*[103]. Учитывая это, в дальнейшем всюду будем использовать двоичный логарифм без явного указния его основания.

Поскольку событие, состоящее в выдаче сообщения $\mathbf{a}$, случайно и происходит с вероятностью $P(\mathbf{a})$, количество информации, связанное с этим сообщением, также является случайной величиной. Введем величину

$$I(\alpha_i) = -\log p(\alpha_i),$$

называемую *собственной информацией* символа α_i.

Информационная производительность дискретного источника характеризуется *средним количеством информации на символ*, которое определяется как математическое ожидание этой случайной величины. Рассматривая для простоты источник без памяти, запишем среднее количество информации, приходящееся на один символ и называемое *энтропией* дискретного источника A, в виде

$$H(\mathrm{A}) = -\sum_{k=1}^{K} p(\alpha_k)\log p(\alpha_k). \qquad (8.1)$$

[103] В литературе упоминаются единицы *нат* и *хартли*, соответствующие основаниям логарифма *e* и 10.

Пример 8.1. Предположим, что передается сообщение о карте, вытащенной наугад из идеально перетасованной колоды в 32 карты (вероятность вытащить любую карту равна при этих условиях 1/32). Очевидно, это сообщение несет количество информации, равное 5 битам. Если это сообщение разбить на два так, что вначале сообщается масть карты, а затем ее достоинство, то это количество информации будет передано частями – сначала 2 бита, затем еще 3. (Убедитесь, что это действительно так!) ◄

8.2. ЭНТРОПИЯ И ИНФОРМАЦИЯ

Рассмотрим *основные свойства энтропии*.

1. Энтропия любого источника A неотрицательна $H(\mathrm{A}) \geq 0$. Это следует из того, что вероятность любого события неотрицательна и не превосходит единицы. Энтропия источника равна нулю в том случае, если один из символов имеет вероятность 1, а остальные – 0. Неопределенность, возникающая вследствие того, что $\log p \to -\infty$ при $p \to 0$, может быть раскрыта с применением правила Лопиталя:

$$\lim_{p\to 0}(-p\log p) = \lim_{p\to 0}\frac{\log 1/p}{1/p} = \lim_{q\to\infty}\frac{\log q}{q} = \lim_{q\to\infty}\frac{1/q\log e}{1} = 0.$$

2. При заданном объеме алфавита K энтропия максимальна, если все символы равновероятны $p(\alpha_k) = p_k = 1/K$.

Доказательство состоит в нахождении условия максимума энтропии при ограничении $\sum_{k=1}^{K} p_k = 1$. Задачу поиска экстремума функции при наличии ограничения можно свести к обычному нахождению экстремума (без ограничения) другой функции, состоящей из двух слагаемых[104]. Первое слагаемое представляет собой в нашем случае энтропию, которую необходимо максимизировать, а второе слагаемое равно нулю, когда выполняется ограничивающее условие. Составим целевую функцию в виде

$$\Phi(p_1,...,p_K) = -\sum_{k=1}^{K} p_k \log p_k + \lambda\left(\sum_{k=1}^{K} p_k - 1\right),$$

[104] Этот метод называется методом *неопределенных множителей Лагранжа.*

где λ – неопределенный *множитель Лагранжа*, и запишем условие достижения ее экстремума

$$\frac{\partial}{\partial p_i}\left[-\sum_{k=1}^{K} p_k \log p_k + \lambda\left(\sum_{k=1}^{K} p_k - 1\right)\right] = 0\,,\ i = \overline{1,K}\,.$$

Решая уравнения относительно p_i, получаем

$$-p_i \frac{1}{\ln 2}\frac{1}{p_i} - \frac{\ln p_i}{\ln 2} + \lambda = 0 \Rightarrow -\frac{1}{\ln 2}\left(1 + \ln p_i\right) + \lambda = 0\,,$$

откуда $p_i = \exp(\lambda \ln 2 - 1)$ независимо от i, а это и означает равновероятность символов. Максимальное значение энтропии равно $H_{\max} = \log K$. В частности, при $K = 2$ энтропия максимальна при $p_1 = p_2 = 1/2$ и равна 1 биту. Таким образом, *1 бит – это количество информации, доставляемое одним из двух равновероятных символов, вырабатываемых источником без памяти*.

Два источника A и B, рассматриваемых в совокупности, характеризуются *совместной энтропией*

$$H(\mathrm{A},\mathrm{B}) = -\sum_i \sum_j p(\alpha_i, \beta_j) \log p(\alpha_i, \beta_j)\,,$$

где $p(\alpha_i, \beta_j)$ – совместная вероятность символов; суммирование проводится по всем возможным значениям индексов. Совместная энтропия характеризуется свойством коммутативности $H(\mathrm{A},\mathrm{B}) = H(\mathrm{B},\mathrm{A})$, что прямо следует из равенства $p(\alpha_i, \beta_j) = p(\beta_j, \alpha_i)$.

Используя выражение для совместной вероятности, перепишем совместную энтропию в виде

$$H(\mathrm{A},\mathrm{B}) = -\sum_i \sum_j p(\alpha_i)\, p(\beta_j \mid \alpha_i) \log\left[\, p(\alpha_i)\, p(\beta_j \mid \alpha_i)\right] =$$

$$= -\sum_i \sum_j p(\alpha_i)\, p(\beta_j \mid \alpha_i) \log p(\alpha_i) - \sum_i \sum_j p(\alpha_i)\, p(\beta_j \mid \alpha_i) \log p(\beta_j \mid \alpha_i)\,.$$

Заметим, что $\sum_j p(\alpha_i)\, p(\beta_j \mid \alpha_i) = \sum_j p(\alpha_i, \beta_j) = p(\alpha_i)$, тогда первое слагаемое принимает вид $-\sum_i p(\alpha_i) \log p(\alpha_i) = H(\mathrm{A})$, а второе слагаемое представляет собой *условную* энтропию

$$-\sum_i \sum_j p(\alpha_i, \beta_j) \log p(\beta_j \mid \alpha_i) = H(\mathrm{B} \mid \mathrm{A})\,.$$

Таким образом, совместная энтропия

$$H(\mathrm{A},\mathrm{B}) = H(\mathrm{A}) + H(\mathrm{B}\,|\,\mathrm{A}) = H(\mathrm{B}) + H(\mathrm{A}\,|\,\mathrm{B}). \qquad (8.2)$$

Если источники статистически независимы, то

$$H(\mathrm{A},\mathrm{B}) = H(\mathrm{A}) + H(\mathrm{B}),$$

что согласуется с интуитивным представлением об аддитивности информации от независимых источников.

Рассмотрим более подробно понятие условной энтропии. Предположим, что имеется дискретный канал связи, на входе которого задан алфавит A, а на выходе алфавит B; канал описывается условным распределением $P(\mathrm{B}\,|\,\mathrm{A})$. Можно считать, что на входе действует источник с алфавитом A и энтропией $H(\mathrm{A})$, а на выходе – источник с алфавитом B и энтропией $H(\mathrm{B})$, причем эти источники статистически связаны.

Условное распределение $P(\mathrm{B}\,|\,\mathrm{A})$ описывает вероятностную связь входных и выходных символов. Чем сильнее эта связь, тем более уверенно можно судить о *входных* символах на основании наблюдения *выходных*, тем лучше канал передает информацию. Количество информации в символе β_j относительно символа α_i определяется выражением

$$I(\alpha_i;\beta_j) = \log\frac{p(\alpha_i\,|\,\beta_j)}{p(\alpha_i)}. \qquad (8.3)$$

В самом деле, если символы независимы, то $p(\alpha_i\,|\,\beta_j) = p(\alpha_i)$ и $I(\alpha_i;\beta_j) = 0$ (символ β_j не несет информации о символе α_i). И, наоборот, при жесткой (детерминированной) связи между символами α_i и β_j, очевидно, $p(\alpha_i\,|\,\beta_j) = 1$, поэтому $I(\alpha_i;\beta_j) = \log\dfrac{1}{p(\alpha_i)} = I(\alpha_i)$, т. е. количество информации в символе β_j относительно символа α_i равно собственному количеству информации в символе α_i (или, что эквивалентно, в символе β_j).

Используя известные формулы для совместных и условных вероятностей, легко видеть, что

$$I(\alpha_i;\beta_j) = \log\frac{p(\alpha_i\,|\,\beta_j)p(\beta_j)}{p(\alpha_i)p(\beta_j)} = \log\frac{p(\alpha_i,\beta_j)}{p(\alpha_i)p(\beta_j)} =$$

$$= \log \frac{p(\beta_j \mid \alpha_i) p(\alpha_i)}{p(\alpha_i) p(\beta_j)} = \log \frac{p(\beta_j \mid \alpha_i)}{p(\beta_j)} = I(\beta_j; \alpha_i) .$$

Количество информации в символе β_j относительно символа α_i равно количеству информации в символе α_i относительно символа β_j. Поэтому величина $I(\alpha_i; \beta_j) = I(\beta_j; \alpha_i)$ называется *взаимной информацией* указанных символов.

Очевидно, в силу вероятностной связи входных и выходных символов наблюдение выходной последовательности символов не снимает полностью неопределенности относительно переданного сообщения. Иными словами, представляет интерес вопрос: какова *энтропия входного алфавита при условии* наблюдения выходных символов? Очевидно, что *чем меньше эта условная энтропия, тем лучше* канал передает информацию. Частное количество информации во входном символе определяется, как и раньше, но с заменой безусловных вероятностей *условными*, усреднение же производится по всем возможным сочетаниям входного и выходного символов (по совместному распределению вероятностей):

$$H(\mathrm{A} \mid \mathrm{B}) = -\sum_{i=1}^{L} \sum_{j=1}^{M} p(\alpha_i, \beta_j) \log p(\alpha_i \mid \beta_j) . \qquad (8.4)$$

Пример 8.2. Предположим, что на входе двоичного канала действует источник с равновероятными символами 0 и 1, а искажения символов при передаче происходят с некоторыми вероятностями $p_0 = p(y = 1 \mid x = 0)$ и $p_1 = p(y = 0 \mid x = 1)$.

Найдем количество информации в выходном символе относительно входного. Безусловные вероятности выходных символов

$$p(y=0) = p(y=0|x=0)p(x=0) + p(y=0|x=1)p(x=1) = \frac{1 - p_0 + p_1}{2} ;$$

$$p(y=1) = p(y=1|x=0)p(x=0) + p(y=1|x=1)p(x=1) = \frac{1 - p_1 + p_0}{2} .$$

Взаимная информация переданного символа $x = 0$ и наблюдаемого символа $y = 0$ равна

$$I(0;0) = \log \frac{p(y=0 \mid x=0)}{p(y=0)} = \log \frac{2(1-p_0)}{1 - p_0 + p_1} ,$$

аналогично взаимная информация переданного символа $x=1$ и наблюдаемого символа $y=0$ равна

$$I(1;0)=\log\frac{p(y=0\,|\,x=1)}{p(y=0)}=\log\frac{2p_1}{1-p_0+p_1}.$$

Так же находятся два оставшихся количества информации:

$$I(0;1)=\log\frac{p(y=1\,|\,x=0)}{p(y=1)}=\log\frac{2p_0}{1-p_1+p_0},$$

$$I(1;1)=\log\frac{p(y=1\,|\,x=1)}{p(y=1)}=\log\frac{2(1-p_1)}{1-p_1+p_0}.$$

Особый интерес представляют некоторые частные случаи.

Первый случай соответствует каналу без помех и характеризуется вероятностями $p_0=p_1=0$. Тогда, очевидно, $I(0;0)=I(1;1)=1$. Поскольку энтропия источника равна 1 биту и взаимная информация входных и выходных символов равна также 1 биту при их совпадении, такой канал обеспечивает передачу информации без потерь.

Второй частный случай имеет место при $p_0=p_1=0.5$. Тогда $I(0;0)=I(1;1)=I(0;1)=I(1;0)=0$ и канал не передает информации (такая ситуация называется «обрывом канала»). ◄

Рассмотрим *основные свойства условной энтропии.*

1. Если источники сообщений A и B являются независимыми, то условная энтропия равна безусловной:

$$H(\mathrm{A}\,|\,\mathrm{B})=H(\mathrm{A}),\quad H(\mathrm{B}\,|\,\mathrm{A})=H(\mathrm{B}).$$

Действительно, если источники независимы, то $p(\alpha_i\,|\,\beta_j)=p(\alpha_i)$ при всех i, j. Тогда выражение (8.4) можно переписать в виде

$$H(\mathrm{A}\,|\,\mathrm{B})=-\sum_{i=1}^{L}\sum_{j=1}^{M}p(\alpha_i,\beta_j)\log p(\alpha_i)=-\sum_{i=1}^{L}\log p(\alpha_i)\sum_{j=1}^{M}p(\alpha_i,\beta_j).$$

Но $\sum\limits_{j=1}^{M}p(\alpha_i,\beta_j)=p(\alpha_i)$, откуда немедленно следует

$H(\mathrm{A}\,|\,\mathrm{B})=-\sum\limits_{i=1}^{L}p(\alpha_i)\log p(\alpha_i)=H(\mathrm{A})$, что и требовалось доказать.

2. Если символы источников A и B жестко связаны, то условная энтропия равна нулю. В самом деле, при жесткой связи в выражении (8.4) некоторые условные вероятности равны 1, а остальные 0. Но как было показано выше, в этом случае сумма равна нулю.

Для условий примера 8.2 жесткая (детерминированная) связь входных и выходных символов соответствует вероятностям ошибок $p_0 = p_1 = 0$ (или $p_0 = p_1 = 1$).

3. Условная энтропия входного алфавита относительно выходного характеризует передаваемую по каналу информацию следующим образом. Если энтропия входного источника в отсутствие передачи равна $H(\mathrm{A})$, а после приема выходного символа она становится равной $H(\mathrm{A}\,|\,\mathrm{B})$, то, очевидно, *среднее* количество передаваемой информации на символ равно разности

$$I(\mathrm{A},\mathrm{B}) = H(\mathrm{A}) - H(\mathrm{A}\,|\,\mathrm{B}).$$

Величина $I(\mathrm{A},\mathrm{B})$ представляет собой *взаимную информацию* входа и выхода.

Если потери информации отсутствуют (канал без помех), то условная энтропия источника после передачи равна 0, количество передаваемой информации равно $H(\mathrm{A})$. Величина $H(\mathrm{A}\,|\,\mathrm{B})$, таким образом, характеризует потери информации в канале и называется *ненадежностью* [10].

Заметим, что из выражения (8.2) для совместной энтропии следует

$$H(\mathrm{A}) + H(\mathrm{B}\,|\,\mathrm{A}) = H(\mathrm{B}) + H(\mathrm{A}\,|\,\mathrm{B}),$$

поэтому

$$I(\mathrm{A},\mathrm{B}) = H(\mathrm{A}) - H(\mathrm{A}\,|\,\mathrm{B}) = H(\mathrm{B}) - H(\mathrm{B}\,|\,\mathrm{A}) = I(\mathrm{B},\,\mathrm{A}). \quad (8.5)$$

При очень высоком уровне помех условные энтропии равны безусловным ($H(\mathrm{A}\,|\,\mathrm{B}) = H(\mathrm{A})$, $H(\mathrm{B}\,|\,\mathrm{A}) = H(\mathrm{B})$) и количество информации, передаваемой по каналу, становится равным нулю.

4. Из выражения для совместной энтропии $H(\mathrm{B}\,|\,\mathrm{A}) = H(\mathrm{A},\mathrm{B}) - H(\mathrm{A})$ и $H(\mathrm{A}\,|\,\mathrm{B}) = H(\mathrm{A},\mathrm{B}) - H(\mathrm{B})$. Подставляя эти выражения в (8.5), получаем среднее количество передаваемой информации на символ

$$I(\mathrm{A},\,\mathrm{B}) = I(\mathrm{B},\,\mathrm{A}) = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{A},\,\mathrm{B}). \quad (8.6)$$

Приведем выражение (8.6) к более удобному виду, для чего подставим в него формулы для вычисления безусловной и совместной энтропии.

$$I(\mathrm{A},\mathrm{B}) = -\sum_{i=1}^{L}\sum_{j=1}^{M} p(\alpha_i,\beta_j)\log p(\alpha_i) - \sum_{i=1}^{L}\sum_{j=1}^{M} p(\alpha_i,\beta_j)\log p(\beta_j) +$$

$$+\sum_{i=1}^{L}\sum_{j=1}^{M} p(\alpha_i,\beta_j)\log p(\alpha_i,\beta_j) = \sum_{i=1}^{L}\sum_{j=1}^{M} p(\alpha_i,\beta_j)\log\frac{p(\alpha_i,\beta_j)}{p(\alpha_i)p(\beta_j)}. \quad (8.7)$$

8.3. ПРОПУСКНАЯ СПОСОБНОСТЬ ДИСКРЕТНОГО КАНАЛА

Если источник вырабатывает символы со скоростью $v_{\text{п}} = 1/T_{\text{п}}$, где $T_{\text{п}}$ – время передачи одного символа, то *производительность* источника определяется как $H' = Hv_{\text{п}} = H/T_{\text{п}}$ и имеет размерность бит/с. Поскольку количество информации на один символ составляет при передаче по каналу величину $I(\mathrm{A},\mathrm{B})$, определяемую выражением (8.7), *скорость передачи* информации по каналу

$$I'(\mathrm{A},\mathrm{B}) = \frac{I(\mathrm{A},\mathrm{B})}{T_{\text{п}}} \quad \text{бит/с.}$$

Рассмотрим выражение (8.5), которое характеризует количество информации на символ, передаваемое по дискретному каналу связи, на входе которого действует источник с алфавитом A, а на выходе образуются символы из алфавита B. Заметим, что энтропия $H(\mathrm{A})$ определяется только источником входных символов, в то время как $H(\mathrm{B})$, $H(\mathrm{B}|\mathrm{A})$ и $H(\mathrm{A}|\mathrm{B})$ зависят также от свойств канала. Таким образом, скорость передачи информации по каналу зависит и от свойств источника, и от свойств канала. Для того чтобы охарактеризовать *только* канал, находят максимум скорости передачи информации по данному каналу при всевозможных источниках (имеется в виду, что при одном и том же алфавите перебираются всевозможные распределения вероятностей его символов). Максимальная скорость передачи информации, которая может быть достигнута *для данного канала*, называется его *пропускной способностью*

$$C = \max_{P(A)} I'(\mathrm{A},\mathrm{B}) = \frac{1}{T_{\text{п}}}\max_{P(A)} I(\mathrm{A},\mathrm{B}) \quad \text{бит/с.}$$

Заметим, что нахождение пропускной способности реального канала связи представляет собой сложную задачу. В простейшем случае бинарного канала без помех (см. пример 8.2) пропускная способность численно равна *скорости модуляции* $v_{\text{п}} = 1/T_{\text{п}}$.

Очевидно, скорость передачи информации по определению не может быть больше пропускной способности канала. Можно рассматривать также пропускную способность канала на символ [10]

$$C_{\text{симв}} = \max_{P(\text{A})} I(\text{A}, \text{B}).$$

Пример 8.3. Найдем пропускную способность стационарного симметричного канала без памяти. Как было указано в разд. 7, для такого канала входной и выходной алфавиты совпадают $\alpha_k = \beta_k$, $k = \overline{1, K}$, а вероятность ошибки одинакова для всех символов и при этом выполняется условие

$$p\left(\alpha_l \middle| \alpha_k\right) = \begin{cases} p_{\text{ош}}/(K-1), & \text{если} \quad l \neq k, \\ 1 - p_{\text{ош}}, & \text{если} \quad l = k, \end{cases}$$

Найдем согласно (8.4) условную энтропию

$$H(\text{B} \mid \text{A}) = -\sum_{i=1}^{K} \sum_{j=1}^{K} p(\alpha_i, \beta_j) \log p(\beta_j \mid \alpha_i) =$$

$$= -\sum_{i=1}^{K} p(\alpha_i) \sum_{j=1}^{K} p(\beta_j \mid \alpha_i) \log p(\beta_j \mid \alpha_i) =$$

$$= -\sum_{i=1}^{K} p(\alpha_i) \left[\sum_{\substack{j=1 \\ j \neq i}}^{K} \frac{p_{\text{ош}}}{K-1} \log \frac{p_{\text{ош}}}{K-1} + (1 - p_{\text{ош}}) \log (1 - p_{\text{ош}}) \right] =$$

$$= -(1 - p_{\text{ош}}) \log (1 - p_{\text{ош}}) - p_{\text{ош}} \log \frac{p_{\text{ош}}}{K-1}. \qquad (8.8)$$

В последнем преобразовании учтено, что выражение в квадратных скобках не зависит от i, поэтому сумма вероятностей $p(\alpha_i)$, равная 1, как сомножитель исчезает, а суммирование в квадратных скобках по $j \neq i$ эквивалентно умножению на $(K-1)$. Очевидно, выражение (8.8) не зависит от распределения вероятностей передаваемых символов, поэтому выражение (8.5)

$$I(\text{A}, \text{B}) = H(\text{B}) - H(\text{B} \mid \text{A})$$

достигает максимума, когда максимальна энтропия $H(\mathrm{B})$, что означает равновероятность символов выходного алфавита, а это, в свою очередь, имеет место, когда равновероятны символы входного алфавита (что очевидно в силу симметрии канала). Таким образом, пропускная способность стационарного симметричного канала без памяти (на символ) равна

$$C_{\text{симв}} = \log K + (1 - p_{\text{ош}}) \log(1 - p_{\text{ош}}) + p_{\text{ош}} \log \frac{p_{\text{ош}}}{K - 1}. \blacktriangleleft$$

8.4. КОДИРОВАНИЕ ИСТОЧНИКА

Реальные источники редко обладают максимальной энтропией, поэтому их принято характеризовать так называемой *избыточностью*, определяемой выражением

$$\kappa = \frac{H_{\max} - H}{H_{\max}}.$$

Для независимых источников (источников без памяти) избыточность равна нулю (а энтропия максимальна) при равновероятности символов. Для источников с памятью избыточность тем больше, чем выше степень статистической зависимости символов в сообщении, при этом неопределенность относительно очередного символа в сообщении уменьшается, соответственно уменьшается и количество информации, переносимое этим символом. Например, в естественном английском языке после буквы *q* всегда следует буква *u*, поэтому при передаче такого текста буква *u*, следующая за буквой *q*, информации не несет. (В реальном английском тексте могут встречаться аббревиатуры, например, «QWERTY», а также иноязычные, например французские слова, для которых указанная закономерность не выполняется.)

Объем алфавита источника и количество различных символов, передаваемых по каналу (*канальных* символов), могут не совпадать. В таких случаях один символ источника представляется (*кодируется*) последовательностью из нескольких *кодовых* символов (*кодовым словом*, или кодовой комбинацией). Если для всех символов источника длина кодовых слов одинакова, код называют *равномерным*, в противном случае – *неравномерным*. Примером равномерного кода является код *Бодо́*, смысл которого состоит в представлении каждой из букв алфавита двоичным числом фиксированной разрядности (например, для алфавита из 32 символов,

включающего 26 латинских букв и знаки препинания, достаточно пятиразрядного кода Бодо). При передаче сообщений неравномерным кодом говорят о *средней* длине кодового слова (усреднение длин кодовых слов производится по соответствующему распределению вероятностей).

Шеннону принадлежит следующая теорема (доказательство см., например, в [10]), называемая основной теоремой о кодировании в отсутствие шумов.

ТЕОРЕМА. *Среднюю длину кодовых слов для передачи символов источника* A *при помощи кода с основанием* m *можно как угодно приблизить к величине* $H(\mathrm{A})/\log m$.

Смысл теоремы состоит в том, что она определяет *нижнюю границу* длины кодовых слов и устанавливает принципиальную возможность достичь этой границы, однако она не указывает способов достижения.

Пример 8.4. Если источник имеет объем алфавита 32, то при равновероятных символах его энтропия равна 5 битам. Тогда для двоичного кода наименьшая средняя длина составляет 5, следовательно, пятизначный код Бодо является оптимальным кодом. Однако при *неравных* вероятностях символов энтропия источника меньше чем 5 бит (избыточность источника отлична от нуля), следовательно, можно найти код со средней длиной кодового слова меньше пяти и таким образом *повысить скорость* передачи информации. Текст на русском языке, например, имеет энтропию около 2,5 бит, поэтому путем соответствующего кодирования можно увеличить скорость передачи информации вдвое против пятиразрядного равномерного кода Бодо (чтобы использовать код Бодо для передачи русского текста, можно отождествить буквы «е» и «ё», а также «ь» и «ъ»). ◄

Практическое значение теоремы Шеннона заключается в возможности повышать эффективность систем передачи информации (систем связи) путем применения экономного кодирования (*кодирования источника*).

Очевидно, что экономный код должен быть в общем случае неравномерным. Общее правило кодирования источника (без памяти) состоит в том, что *более вероятным символам источника ставятся в соответствие менее длинные кодовые слова* (последовательности канальных символов).

Пример 8.5. Известный код *Морзе* служит примером неравномерного кода. Кодовые слова состоят из трех различных символов: точки • (передается короткой посылкой), тире — (передается относительно длинной посылкой) и пробела (паузы). Наиболее частой

букве в русском тексте – букве «е» – соответствует самое короткое кодовое слово, состоящее из одной точки, относительно редкая буква «ш» передается кодовым словом из четырех тире, разделенных пробелами, и т.д. ◄

Кодирование источника по методу Шеннона – Фано

Принцип построения кода Шеннона – Фано состоит в упорядочении всех символов алфавита (назовем их для краткости «буквами») по убыванию вероятностей. Затем все буквы делятся на две (неравные в общем случае) группы так, что сумма вероятностей букв для обеих групп одинакова или примерно одинакова, и в качестве первого символа кодового слова каждой букве первой группы присваивается кодовый символ 0, а каждой букве второй группы – символ 1 (или наоборот). Далее первая и вторая группы делятся на подгруппы в соответствии с принципом равной вероятности, и эта процедура продолжается до тех пор, пока алфавит источника не будет исчерпан. Пример построения кода Шеннона – Фано приведен в табл. 8.1.

Таблица 8.1

Построение кода Шеннона – Фано

Символ и его вероятность		Комбинация кодовых символов						Длина комбинации
α_i	$p(\alpha_i)$	1-й	2-й	3-й	4-й	5-й	6-й	μ_i
α_1	1/4	0	0					2
α_2	1/4	0	1					2
α_3	1/4	1	0					2
α_4	1/8	1	1	0				3
α_5	1/16	1	1	1	0			4
α_6	1/32	1	1	1	1	0		5
α_7	1/64	1	1	1	1	1	0	6
α_8	1/64	1	1	1	1	1	1	6

На первом шаге процедуры все символы алфавита источника делятся на две группы, причем в первую группу входят символы α_1 и α_2, которым соответствует суммарная вероятность 1/2, а во вторую – все остальные символы. Символам α_1 и α_2 приписывается в качестве первого кодового символа символ 0, а всем осталь-

ным символам источника – кодовый символ 1. На втором шаге первая и вторая группы рассматриваются по отдельности, при этом в первой группе содержатся всего два символа, которые получают в качестве второго кодового символа 0 и 1 соответственно. Таким образом, символу источника α_1 ставится в соответствие кодовое слово 00, а символу α_2 – слово 01. Вторая группа символов источника, включающая символы α_3, α_4, α_5 и α_6, делится на две части в соответствии с их вероятностями, при этом символ α_3, которому соответствует вероятность 1/4, получает в качестве второго символа кодового слова символ 0, а остальные символы источника – символ 1. Далее процесс продолжается до тех пор, пока не останется группа из двух символов – в данном примере это символы α_7 и α_8, – которым присваиваются кодовые символы 0 и 1.

Необходимо обратить внимание на следующее свойство полученного кода: *ни одна кодовая комбинация не является началом какой-либо другой кодовой комбинации* (так называемое *префиксное правило*). Такие коды называются *неперекрываемыми* (неприводимыми). Декодирование неприводимого кода может быть осуществлено на основе *дерева декодирования*[105] (рис. 8.1), отвечающего некоторому *конечному автомату*, который переходит из начального состояния в другие состояния в соответствии с очередным символом кодовой последовательности.

Перед декодированием конечный автомат устанавливается в начальное состояние НС, а дальнейшие переходы зависят только

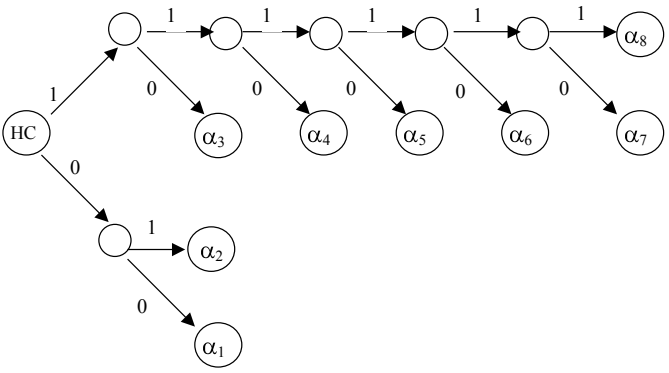


Рис. 8.1. Дерево декодирования для кода Шеннона– Фано (табл. 8.1)

[105] Деревом называется граф, не содержащий циклов (замкнутых путей).

от поступающих символов кода, при этом все концевые состояния («листья» дерева) соответствуют декодированным символам алфавита источника; по достижении листа автомат переходит вновь в начальное состояние. Поскольку с поступлением последнего кодового символа декодирование кодового слова всегда заканчивается, префиксные коды называют также мгновенными [16]. Существуют однозначно декодируемые коды, не обладающие префиксным свойством и не являющиеся мгновенными, однако их декодирование требует больших объемов памяти декодера и приводит к большим задержкам результата.

Средняя длина кодовой комбинации для построенного кода

$$\mu = \sum_{i=1}^{8} p(\alpha_i)\mu_i =$$

$$= 0,75 \cdot 2 + 0,125 \cdot 3 + 0,0625 \cdot 4 + 0,03125 \cdot 5 + 0,03125 \cdot 6 = 2,469 .$$

Согласно теореме Шеннона при оптимальном кодировании можно достичь средней длины

$$\mu_{\min} = H(\mathrm{A}) / \log 2 = -\sum_{i=1}^{8} p(\alpha_i) \log p(\alpha_i) = 2.469 .$$

Таким образом, построенный код является оптимальным. Так получилось вследствие того, что на каждом шаге процедуры построения кода удавалось разделить символы на группы с равными вероятностями. Заметим, что восемь различных символов источника можно представить восемью комбинациями равномерного двоичного кода (Бодо), при этом длина каждой кодовой комбинации равняется, очевидно, трем. Уменьшение средней длины кодовой комбинации (и, следовательно, увеличение скорости передачи информации) составляет в данном примере около 22 %. Если при делении символов на группы их суммарные вероятности оказываются неравными, выигрыш может быть не столь значительным.

Определим вероятность появления определенного символа в кодовой комбинации (пусть это будет символ 1). Очевидно, ее можно найти следующим образом: а) подсчитать количества единиц во всех кодовых словах; б) умножить эти количества на вероятности соответствующих кодовых слов; в) просуммировать полученные величины; г) отнести результат к средней длине кодового слова. Таким образом,

$$p(1) = \frac{0,25 + 0,25 + 2 \cdot 0,125 + 3 \cdot 0,0625 + 4 \cdot 0,03125 + (5+6) \cdot 0,015625}{2,469} = 0,5 .$$

Итак, при оптимальном кодировании источника кодовые символы равновероятны; такое кодирование является *безызбыточным.* Источник вместе с кодером можно рассматривать как новый источник с алфавитом, состоящим из кодовых символов; энтропия и избыточность этого источника – это энтропия и избыточность кода. Оптимальный код имеет максимальную энтропию и нулевую избыточность.

Кодирование источника по методу Хаффмена.

Другим широко известным методом кодирования источника является метод Хаффмена[106]. Процедура кодирования состоит из следующих шагов.

1. Все символы алфавита записываются в порядке убывания вероятностей.

2. Два нижних символа соединяются скобкой, из них верхнему приписывается символ 0, нижнему 1 (или наоборот).

3. Вычисляется сумма вероятностей, соответствующих этим символам алфавита.

4. Все символы алфавита снова записываются в порядке убывания вероятностей, при этом только что рассмотренные символы «склеиваются», т.е. учитываются как единый символ с суммарной вероятностью.

5. Повторяются шаги 2, 3 и 4 до тех пор, пока не останется ни одного символа алфавита, не охваченного скобкой.

Скобки в совокупности образуют дерево. Код Хаффмена для некоторого символа алфавита находится путем последовательной записи нулей и единиц, встречающихся на пути от корня дерева (корню соответствует суммарная вероятность 1) до листа, соответствующего данному символу. Полученное дерево, очевидно, является деревом декодирования.

Для примера, показанного на рис. 8.2, получаются следующие кодовые комбинации: $\alpha_1 \to 11$; $\alpha_2 \to 01$; $\alpha_3 \to 101$: $\alpha_4 \to 100$; $\alpha_5 \to 001$; $\alpha_6 \to 0000$; $\alpha_7 \to 00011$; $\alpha_8 \to 00010$.

Энтропия алфавита $H(\mathrm{A}) = 2{,}628$. Средняя длина кодового слова

$$\mu = \sum_{i=1}^{n} p_i \mu_i = 0{,}3 \cdot 2 + 0{,}2 \cdot 2 + 0{,}15 \cdot 3 + 0{,}15 \cdot 3 + 0{,}1 \cdot 3 +$$

$$+ 0{,}04 \cdot 4 + 0{,}03 \cdot 5 + 0{,}03 \cdot 5 = 2{,}66.$$

[106] Доказана оптимальность кода Хаффмена в смысле наименьшей средней длины кодовых слов [16].

Следовательно, код не оптимален, но очевидно, что он довольно близок к оптимальному. Для сравнения: равномерный код для этого случая имеет среднюю длину кодового слова 3 (совпадающую для равномерного кода с длиной каждого кодового слова).

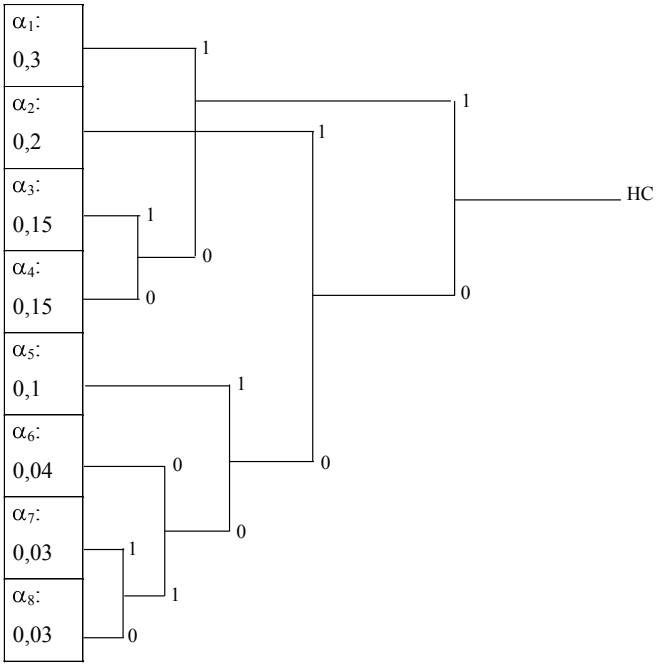


Рис. 8.2. Дерево декодирования для кода Хаффмена

Вероятность символа 1 в последовательности кодовых комбинаций находится как среднее количество единиц, отнесенное к средней длине кодового слова

$$p(1) = \frac{2 \cdot 0,3 + 0,2 + 2 \cdot 0,15 + 0,15 + 0,1 + 2 \cdot 0,03 + 0,03}{2,66} = 0,541 .$$

Неоптимальность кода проявляется в неравенстве вероятностей кодовых символов (0,541 для 1 и 0,459 для 0). Избыточность кода, как легко видеть, равна 0.005.

ЗАМЕЧАНИЯ

1. *Из рассмотренных примеров не должно составиться ложное впечатление, будто код Шеннона – Фано оптимален, а код Хаффмена – нет. Оптимальность кода Шеннона – Фано в рассмотренном примере объясняется специально подобранными вероятностями символов, так, что на каждом шаге вероятности делятся ровно пополам.*

2. *В приведенных примерах предполагалось, что символы в сообщении являются независимыми* (*т. е. вероятность появления в сообщении любых двух символов рядом равна произведению вероятностей появления каждого из символов в отдельности*). *В реальных сообщениях на естественных языках символы не являются независимыми; в таких случаях следует кодировать не отдельные символы* (*буквы*), *а группы букв или слова. Это уменьшает зависимость и повышает эффективность кода.*

3. *Кодирование групп символов вместо отдельных символов также повышает эффективность кодирования в случае независимого источника с сильно различающимися вероятностями символов, как это видно из следующего примера.*

Пример 8.6. Рассмотрим источник, вырабатывающий два независимых символа с вероятностями 0,1 и 0,9. В этом тривиальном случае методы кодирования Хаффмена и Шеннона – Фано приводят к одинаковому коду: символы алфавита кодируются символами 0 и 1. Пусть для определенности $\alpha_1 \to 0$; $\alpha_2 \to 1$. Энтропия источника равна $H(\mathrm{A}) = 0{,}469$; средняя длина кодового слова равна 1, избыточность источника и избыточность кода одинаковы и равны

$$\kappa = \frac{H_{\max}(\mathrm{A}) - H(\mathrm{A})}{H_{\max}(\mathrm{A})} = \frac{1 - 0{,}469}{1} = 0{,}531.$$

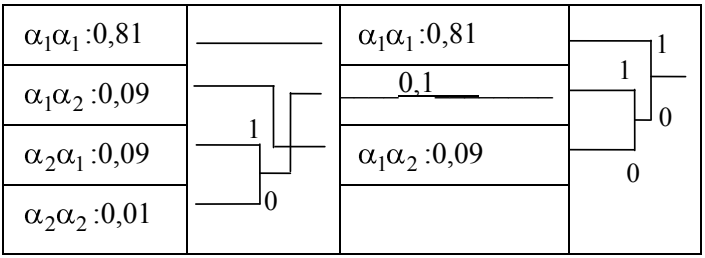
Применим кодирование группами по 2 символа алфавита. Составим всевозможные пары и запишем их в табл. 8.2 с соответствующими вероятностями (найденными как произведения вероятностей отдельных символов, поскольку символы независимы).

Согласно построенному дереву код Хаффмена для указанных групп содержит следующие кодовые комбинации:

$$\alpha_1\alpha_1 \to 1;\ \alpha_1\alpha_2 \to 00;\ \alpha_2\alpha_1 \to 011;\ \alpha_2\alpha_2 \to 010.$$

Таблица 8.2

Кодирование группами



Средняя длина кодовой комбинации, приведенная к одному символу алфавита (для этого взвешенная сумма делится на 2), равна

$$\mu = \frac{0,81 + 0,09 \cdot 2 + 0,09 \cdot 3 + 0,01 \cdot 3}{2} = 0,645 .$$

Вероятность символа 1 в последовательности кодовых комбинаций находится как среднее количество единиц, отнесенное к средней длине кодового слова:

$$p(1) = \frac{0,81 + 0,09 \cdot 2 + 0,01}{0,645 \cdot 2} = 0,775 .$$

Энтропия кода находится как энтропия случайной величины, принимающей два значения (0 и 1) с вероятностями 0,225 и 0,775:

$$H_{\text{к}} = -0,225 \log 0,225 - 0,775 \log 0,775 = 0,769 .$$

Избыточность кода

$$\text{к} = \frac{H_{\text{к max}} - H_{\text{к}}}{H_{\text{к max}}} = \frac{1 - 0,769}{1} = 0,231 .$$

Сравнение с кодированием одиночных символов показывает, что кодирование групп является более эффективным: уменьшаются избыточность кода и средняя длина кодового слова, вероятности символов 0 и 1 сближаются.

Еще более эффективные коды для данного источника можно получить, объединяя символы алфавита в группы по три, четыре и т. д. В пределе, согласно теореме Шеннона, средняя длина кодовой комбинации, приведенная к одному символу алфавита, должна стремиться к значению 0,469, избыточность кода – к нулю, а вероятности кодовых символов 0 и 1 – к значению 0,5. ◄

8.5. ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ

Кодирование источника, называемое также *статистическим* или *экономным* кодированием[107], преследует цель *повышения эффективности* передачи информации, под которым понимается максимально быстрая передача. Экономное кодирование можно рассматривать как замену исходного источника другим источником с меньшей (в пределе нулевой) избыточностью. Если в канале действуют помехи, то при приеме сигналов возникают ошибки, приводящие к неправильному декодированию сообщений. В таких случаях выдвигается на передний план задача *повышения верности* передачи. Одним из путей ее решения является *помехоустойчивое* (*канальное*) кодирование. Помехоустойчивыми, или корректирующими, кодами называются коды, обеспечивающие автоматическое обнаружение и/или исправление ошибок в кодовых комбинациях. Такая возможность достигается целенаправленным *введением избыточности* в передаваемые сообщения. Наиболее простой способ повышения помехоустойчивости путем введения избыточности состоит в многократной передаче каждого символа, например, вместо слова *связь* можно передавать слово *сссввваяяззззььь,* тогда одиночные ошибки могут быть исправлены путем «голосования» среди символов каждой тройки. На практике применяются более сложные и более эффективные методы кодирования.

Теоретическим обоснованием применения канального кодирования служит следующая *основная теорема кодирования* Шеннона для каналов с помехами (шумами) [10].

ТЕОРЕМА. *Если производительность источника* $H'(\mathrm{A})$ *меньше пропускной способности канала* C*, то существует по крайней мере одна процедура кодирования/декодирования, при которой вероятность ошибочного декодирования и ненадежность* $H(\mathrm{A}\,|\,\mathrm{B})$ *могут быть сколь угодно малы. Если* $H'(\mathrm{A}) > C$*, то такой процедуры не существует.*

Содержание теоремы кажется парадоксальным: интуиция говорит о том, что для того чтобы вероятность ошибки стремилась к нулю, также должна стремиться к нулю скорость передачи (это ясно для случая многократной повторной передачи, описанной выше). Тем не менее теорема верна, но, к сожалению, она не указывает практических путей построения соответствующих кодов. Известно лишь, что по мере приближения скорости передачи к

[107] Широко употребляется также термин *сжатие*.

пропускной способности канала длины кодовых комбинаций и сложность кодера и декодера возрастают; также возрастает время декодирования.

В настоящее время известно множество кодов, которые с большим или меньшим успехом применяются для канального кодирования. Эти коды подразделяются на классы в соответствии с различными признаками.

Если информационная последовательность символов источника (возможно, после экономного кодирования) разбивается на сегменты (блоки), кодируемые независимо друг от друга, то код называется блочным (блоковым), если же информационная последовательность кодируется без разбиения, то код называют непрерывным[108]. Блочные коды, как правило, являются равномерными.

Если в кодовом слове можно выделить *информационные* символы, служащие для передачи сообщения, и *проверочные* (контрольные) символы, предназначенные только для обнаружения и исправления ошибок, такой код называют *разделимым*; если такое разбиение осуществить нельзя, код является *неразделимым*. Примерами неразделимых кодов являются так называемые *коды с постоянным весом,* в частности, код «3 из 7» (стандартный телеграфный код № 3 [10]), а также коды на основе матриц Адамара (коды *Рида–Мюллера*).

Разделимые коды, в свою очередь, подразделяются на *линейные* и *нелинейные*.

В качестве примера рассмотрим один класс помехоустойчивых кодов – линейные блочные коды.

Блочный равномерный код состоит из кодовых слов (комбинаций) одинаковой длины n. Элементы кодовых слов выбираются из некоторого алфавита (канальных) символов объемом q. Если $q = 2$, код называется двоичным. Далее для простоты считается, что $q = 2$. Поскольку все кодовые слова имеют одинаковую длину, удобно считать их векторами, принадлежащими линейному пространству размерности n. Для линейных кодов справедливо утверждение: линейная комбинация кодовых слов является кодовым словом.

Всего можно образовать 2^n n-мерных векторов с двоичными компонентами (кодовых комбинаций или слов). Из них только $M = 2^k$, $k < n$ комбинаций являются *разрешенными* и составляют

[108] Широкое распространение получили непрерывные коды, принадлежащие подклассу *сверточных* кодов.

код, который называется (n, k)-*кодом* (отношение $k/n = R$ называется *скоростью кода*). Остальные комбинации в кодере образоваться не могут (являются *запрещенными*), но могут получиться из разрешенных под воздействием помех в канале. Поэтому если в приемнике имеет место запрещенная комбинация, то это означает, что при передаче по каналу произошла ошибка. Разрешенные комбинации, как векторы линейного пространства, должны отстоять друг от друга достаточно далеко. Чем больше расстояние между разрешенными комбинациями, тем меньше вероятность преобразования их друг в друга под действием помех, тем выше способность кода к *обнаружению* ошибок. Более того, при приеме запрещенной комбинации можно не только обнаруживать, но и исправлять ошибки. Для этого декодер должен принимать решение о переданной комбинации на основе расстояния между принятой запрещенной комбинацией и ближайшей разрешенной. Таким образом, чем дальше друг от друга разрешенные комбинации, тем выше *корректирующая* способность кода. Алгоритм работы декодера формально сводится к разбиению всего пространства на области A_i, $i = 1,...,M$, каждая из которых содержит одну разрешенную комбинацию x_i. Если принятая комбинация принадлежит области A_k, то декодер принимает решение о том, что передавалась разрешенная комбинация x_k.

Для кодирования и декодирования линейных блоковых кодов применяются действия, описываемые операциями над векторами в линейном пространстве над конечным полем целых чисел [2]. Сложение и умножение в конечном поле понимаются как сложение и умножение по модулю q. Простейшее из таких полей, называемых *полями Галуа́* – поле по модулю 2, обозначаемое $GF(2)$. Сложение и умножение в этом поле описываются следующими таблицами сложения и умножения (табл. 8.3, 8.4).

Заметим, что вычитание по модулю 2 совпадает со сложением по модулю 2 (это легко увидеть из таблицы сложения).

Мерой различия между векторами линейного пространства, как известно, может служить некоторая функция (функционал), называемая метрикой, или расстоянием [2]. В теории кодирования часто

Таблица 8.3

Таблица сложения в поле $GF(2)$

+	0	1
0	0	1
1	1	0

Таблица 8.4

Таблица умножения в поле $GF(2)$

×	0	1
0	0	0
1	0	1

используется *метрика Хэмминга,* определяемая для двух двоичных кодовых векторов x и y выражением

$$d(x, y) = \sum_{i=1}^{n} (x_i - y_i) \bmod 2 .$$

Легко видеть, что расстояние по Хэммингу между двумя двоичными векторами равно количеству несовпадающих элементов (например, для векторов 00011100 и 11000110 расстояние равно 5).

В n-мерном пространстве двоичных векторов можно определить скалярное произведение выражением $(x, y) = \sum_{i=1}^{n} x_i y_i$, где сумма понимается как сумма по модулю 2. Если для некоторой пары векторов скалярное произведение равно 0, то векторы являются *ортогональными*.

Таким образом, множество всех двоичных кодовых слов длины n можно рассматривать как n-мерное линейное пространство над конечным полем скаляров $GF(2)$. Хотя это пространство содержит лишь конечное множество векторов, а именно 2^n, оно удовлетворяет всем аксиомам векторного пространства [2].

Линейные коды являются разделимыми, поэтому из n символов только k являются информационными, а остальные $(n-k)$ – проверочными. Тогда, очевидно, в n-мерном пространстве S_n всех комбинаций можно выделить k-мерное подпространство S_k разрешенных комбинаций. Таким образом, пространство S_n можно представить *прямой суммой* k-мерного подпространства S_k и $(n-k)$-мерного подпространства S_{n-k}, так что любой вектор из S_{n-k} ортогонален любому вектору, принадлежащему S_k:

$$S_n = S_k \oplus S_{n-k}, \quad S_k \perp S_{n-k},$$

где $\oplus$ – символ прямой суммы, а знак $\perp$ обозначает ортогональность подпространств.

Предположим, что блок из k информационных двоичных символов кодируется словом из n канальных двоичных символов. Обозначим информационный k-мерный вектор[109] через $\mathbf{X} = (x_1, ..., x_k)$, кодовый n-мерный вектор через $\mathbf{C} = (c_1, ..., c_n)$. Ко-

[109] В кодировании принято записывать векторы, как векторы-строки.

дирование описывается линейным преобразованием (оператором), отображающим векторы, принадлежащие подпространству S_k, в векторы из S_n

$$\mathbf{C} = \mathbf{XG}, \tag{8.9}$$

где $\mathbf{G}$ – матрица кодирования (порождающая матрица кода) вида

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & \cdot & \cdot & \cdot & g_{1n} \\ g_{21} & g_{22} & \cdot & \cdot & \cdot & g_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{k1} & g_{k2} & \cdot & \cdot & \cdot & g_{kn} \end{pmatrix}.$$

Уравнение (8.9) можно рассматривать и как систему из n линейных уравнений вида

$$c_j = x_1 g_{1j} + x_2 g_{2j} + \ldots + x_k g_{kj}, \quad j = 1,\ldots,n,$$

где сложение понимается по модулю 2.

Нетрудно видеть, что любое кодовое слово – это не что иное, как линейная комбинация строк матрицы $\mathbf{G}$ с весовыми коэффициентами, равными информационным символам. Отсюда следует, что, хотя разрешенные кодовые слова принадлежат всему пространству S_n, они также принадлежат k-мерному подпространству S_k, натянутому на векторы – строки матрицы $\mathbf{G}$, какова бы ни была эта матрица (если, конечно, у нее n столбцов и k строк, которые, очевидно, должны быть линейно независимыми). Путем линейных операций над строками и перестановки столбцов любую такую матрицу можно привести к *систематическому* виду:

$$\mathbf{G} = \left(\mathbf{I}_k \;\vdots\; \mathbf{P}\right) = \begin{pmatrix} 1 & 0 & 0 & . & . & 0 & p_{11} & p_{12} & \cdot & p_{1(n-k)} \\ 0 & 1 & 0 & . & . & 0 & p_{21} & p_{22} & \cdot & p_{2(n-k)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & . & . & 1 & p_{k1} & p_{k2} & \cdot & p_{k(n-k)} \end{pmatrix}, \tag{8.10}$$

где $\mathbf{I}_k$ – единичная матрица размера $k \times k$, а $\mathbf{P}$ – матрица размера $k \times (n-k)$. Воздействие такого преобразования на информационный вектор приводит к формированию кодового вектора, k первых символов которого повторяют символы информационного вектора, а остальные $(n-k)$ символов формируются из информационных матрицей $\mathbf{P}$ и являются проверочными (паритетными). В этом случае код называют *систематическим*. Любой линейный код можно преобразованием матрицы привести к систематическому коду, эквивалентному в смысле помехоустойчивости, которая определяется расстояниями между кодовыми словами, инвариантными к таким преобразованиям. Все порождающие матрицы эквивалентных кодов представляют собой наборы векторов-строк, являющиеся различными базисами одного и того же подпространства.

Пример 8.7. Систематический (7, 4)-код порождается матрицей

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Кодовые слова имеют структуру $\mathbf{C} = (x_1, x_2, x_3, x_4, c_5, c_6, c_7)$, где

$$c_5 = x_1 + x_2 + x_3,$$

$$c_6 = x_2 + x_3 + x_4,$$

$$c_7 = x_1 + x_2 + x_4$$

(подразумевается сложение по модулю 2).

Реализовать такое кодирование можно при помощи устройства, структурная схема которого показана на рис. 8.2.

Устройство включает два сдвиговых регистра объемом 4 и 3 разряда, а также три сумматора по модулю 2. Информационная последовательность поступает на вход первого регистра и записывается в его разрядах. На выходах сумматоров по модулю 2 формируются проверочные символы, которые запоминаются в разрядах второго сдвигового регистра. Последним шагом формирования кода является считывание вначале четырех информационных символов, а затем – трех проверочных, при этом на выходе устройства получается семиразрядное кодовое слово. ◄

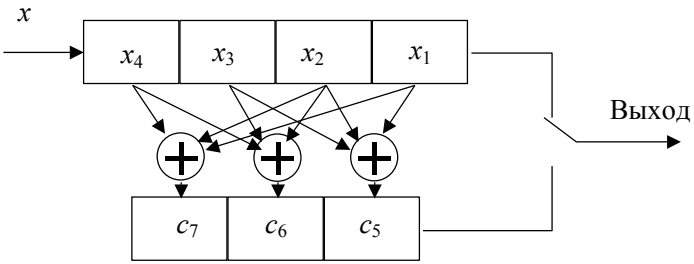


Рис. 8.2. Структура кодера для систематического (7, 4)-кода

Применение любого кода предполагает достаточно простую реализацию не только кодирования, но и декодирования. Декодирование систематического линейного блочного кода могло бы заключаться в простом отбрасывании проверочных символов, но это не обеспечивало бы обнаружения и исправления ошибок.

Вернемся к структуре пространства S_n. Подпространство S_k представляет собой множество всех разрешенных кодовых комбинаций – линейную оболочку совокупности векторов-строк порождающей матрицы $\mathbf{G}$. Другими словами, подпространство S_k и есть (n, k)-код. Тогда подпространство S_{n-k}, ортогональное к нему, также можно считать некоторым $(n, n-k)$-кодом, *дуальным* к данному. Порождающая матрица $\mathbf{H}$ дуального кода содержит $(n-k)$ линейно независимых строк длины n.

Любое кодовое слово (n, k)-кода ортогонально любому кодовому слову $(n, n-k)$-кода, следовательно,

$$\mathbf{G}\mathbf{H}^{\mathbf{T}} = \mathbf{0},$$

где $\mathbf{0}$ – матрица размера $k \times (n-k)$, состоящая из нулей, $(\cdot)^{\mathbf{T}}$ – символ транспонирования. С учетом (8.10) можно записать

$$\mathbf{H} = \left(-\mathbf{P}^{\mathbf{T}} \ \vdots \ \mathbf{I}_{n-k}\right), \tag{8.11}$$

причем для двоичного кода минус можно опустить, так как сложение и вычитание по модулю 2 совпадают.

Матрица $\mathbf{H}$ является порождающей матрицей дуального кода; в то же время она может использоваться для обнаружения ошибок. В самом деле, если принятая кодовая комбинация $\mathbf{Y}$ является раз-

решенной, то она ортогональна к подпространству[110] S_{n-k}, или, что то же самое, ко всем строкам матрицы $\mathbf{H}$, поэтому $\mathbf{Y}\mathbf{H}^{\mathbf{T}} = \vec{0}$, где $\vec{0}$ – нулевой вектор размерности $(n-k)$. Таким образом, умножая слева вектор-строку, соответствующую принятой комбинации, на транспонированную матрицу $\mathbf{H}^{\mathbf{T}}$, получаем вектор (называемый *синдромом*), который равен нулевому вектору в том и только в том случае, если комбинация является разрешенной. В противном случае комбинация является запрещенной, следовательно, при передаче произошла ошибка. По значению синдрома можно определить, какой именно разряд кодового слова содержит ошибку.

Коды Хэмминга

Одним из наиболее известных классов помехоустойчивых линейных блочных кодов являются коды Хэмминга. Коды Хэмминга представляют собой (n, k) -коды, удовлетворяющие условию

$$(n, k) = (2^m - 1, 2^m - 1 - m)$$

при некотором целом m.

В частности, рассмотренный (7, 4)-код является кодом Хэмминга.

Особое свойство кодов Хэмминга заключается в строении проверочной матрицы. Для любого линейного кода проверочная матрица содержит $(n-k)$ строк и n столбцов; для кода Хэмминга $n = 2^m - 1$ и проверочная матрица содержит в качестве столбцов все возможные комбинации нулей и единиц, исключая нулевой вектор.

Для (7, 4)-кода, рассмотренного в примере 8.7, проверочная матрица в соответствии с выражением (8.11), очевидно, имеет вид

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Если передается кодовая комбинация $\mathbf{C}$ и в канале происходит ее искажение, то принятую комбинацию $\mathbf{Y}$ можно представить в виде $\mathbf{Y} = \mathbf{C} + \mathbf{e}$, где $\mathbf{e}$ – вектор ошибки, содержащий единичные

[110] Это подпространство также называют нуль-пространством [15].

компоненты в тех позициях, в которых произошли ошибки, т. е. нули были заменены единицами или наоборот (напомним, что суммирование всюду понимается по модулю 2).

Умножим принятую комбинацию на транспонированную проверочную матрицу

$$\mathbf{Y}\mathbf{H}^{\mathbf{T}} = \mathbf{C}\mathbf{H}^{\mathbf{T}} + \mathbf{e}\mathbf{H}^{\mathbf{T}} = \vec{0} + \mathbf{e}\mathbf{H}^{\mathbf{T}} = \boldsymbol{\sigma},$$

здесь вектор $\boldsymbol{\sigma}$ представляет собой синдром, который равен нулевому вектору в том и только в том случае, если вектор ошибки ортогонален всем строкам проверочной матрицы, т. е. подпространству S_{n-k}. Это означает, что не могут быть обнаружены ошибки, составляющие вектор, который сам является разрешенной комбинацией кода.

Чтобы убедиться в корректирующих свойствах кода Хэмминга, рассмотрим пример обнаружения ошибки в кодовой комбинации.

Пример 8.8. Предположим, что передавалась разрешенная кодовая комбинация 0100111 (напомним, что разрешенными комбинациями являются все линейные комбинации строк порождающей матрицы кода). Предположим также, что при передаче произошла ошибка, скажем, во втором символе, так что принята комбинация 0000111.

Умножая вектор-строку, соответствующую принятой комбинации, слева на транспонированную проверочную матрицу $\mathbf{H}^{\mathbf{T}}$, получим синдром

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix},$$

который совпадает со второй строкой матрицы $\mathbf{H}^{\mathbf{T}}$. Это указывает на то, что ошибочным является второй символ принятой комбинации. ◄

То обстоятельство, что синдром позволяет определить номер «испорченного» символа, фактически означает возможность исправления ошибок. В самом деле, если точно известно, что во вто-

ром символе имела место ошибка, декодер может ее исправить, прибавив (по модулю 2) к ошибочному символу единицу. Поэтому код Хэмминга принадлежит к кодам, *исправляющим ошибки,* или корректирующим.

Границы корректирующей способности кода Хэмминга иллюстрируются следующим примером.

Пример 8.9. Предположим, что при передаче разрешенной кодовой комбинации 0100111 произошли две ошибки, скажем, в третьем и пятом символах, так что принята комбинация 0110011. Найдем синдром:

$$\begin{pmatrix}0 & 1 & 1 & 0 & 0 & 1 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 1\\ 1 & 1 & 1\\ 1 & 1 & 0\\ 0 & 1 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}=\begin{pmatrix}0 & 1 & 0\end{pmatrix}.$$

Синдром указывает на 6-й символ, как на ошибочный. Таким образом, в случае двукратной ошибки факт ошибки обнаруживается (синдром оказывается ненулевым), но исправить ее нельзя, так как синдром оказывается таким же, как в случае однократной ошибки в другом символе. Итак, код Хэмминга (7,4) обнаруживает одно- и двукратные ошибки и исправляет однократные. ◄

Помехоустойчивость рассмотренного кода Хэмминга просто объясняется с геометрической точки зрения. Легко убедиться, что расстояние между любыми двумя разрешенными комбинациями этого кода не менее 3. Поэтому при приеме запрещенной комбинации она заменяется той разрешенной комбинацией, расстояние до которой равно 1. Двукратная ошибка отдаляет принимаемую комбинацию на расстояние, равное 2, что и приводит к ошибочному «исправлению» ошибки. При этом «исправляется» один символ, поэтому «исправленная» комбинация отстоит от принятой на расстояние 1.

Коды, обнаруживающие ошибки, но не исправляющие их, могут использоваться в системах с решающей обратной связью (*системах с переспросом* [10]). В таких системах при обнаружении ошибки во время декодирования по каналу обратной связи передается сигнал переспроса, и тогда передающее устройство повторяет передачу забракованной комбинации.

В заключение отметим, что при решении вопроса о целесообразности помехоустойчивого кодирования и выборе помехоустойчивого кода следует руководствоваться критерием скорости передачи информации при заданной достоверности. Дело в том, что введение избыточных символов приводит к увеличению времени передачи кодовой комбинации или к укорочению элементарных посылок, что ведет к повышению вероятности ошибочного приема символа. Поэтому применение помехоустойчивого кодирования или некоторого конкретного кода может оказаться нецелесообразным.

8.6. ИНФОРМАТИВНОСТЬ НЕПРЕРЫВНЫХ ИСТОЧНИКОВ СООБЩЕНИЙ

Наряду с дискретными источниками сообщений часто встречаются непрерывные источники, которые вырабатывают сообщения, обычно описываемые функциями, принимающими значения из непрерывного множества. Ярким примером непрерывного сообщения является речевое сообщение, описываемое вещественной функцией непрерывного времени. Значение непрерывного сообщения в некоторый отдельный момент времени представляет собой непрерывную случайную величину x, описываемую функцией распределения

$$F(x) = \mathbf{P}\{\xi \le x\},$$

где ξ – реализация случайной величины x, или плотностью распределения

$$w(x) = dF(x)/dx.$$

Очевидно, введенное ранее понятие энтропии неприменимо к непрерывному источнику, так как неопределенность относительно любого конкретного значения непрерывной случайной величины равна бесконечности.

Действительно, разобьем область определения непрерывной случайной величины $(-\infty,\infty)$ на отрезки одинаковой длины Δx и пронумеруем их при помощи индекса $i = \overline{-\infty,\infty}$. Поставим в соответствие каждому отрезку значение x_i', равное его середине, и вероятность $P(x_i')$, равную вероятности попадания в данный интервал исходной непрерывной случайной величины x. Таким образом получается дискретная случайная величина, которая тем точнее описывает непрерывную случайную величину, чем меньше интервал Δx.

Для этой дискретной случайной величины можно записать энтропию

$$H(X') = -\sum_{i=-\infty}^{\infty} P(x'_i)\log P(x'_i)\,.$$

Подставив вместо вероятности $P(x'_i)$ ее приближенное значение $w(x'_i)\Delta x$, получим в пределе при $\Delta x \to 0$

$$H(X) = \lim_{\Delta x\to 0} H(X') = \lim_{\Delta x\to 0}\left\{-\sum_{i=-\infty}^{\infty} w(x'_i)\Delta x \log\left[w(x'_i)\Delta x\right]\right\} =$$

$$= \lim_{\Delta x\to 0}\left\{-\sum_{i=-\infty}^{\infty} w(x'_i)\log\left[w(x'_i)\right]\Delta x\right\} + \lim_{\Delta x\to 0}\left\{-\sum_{i=-\infty}^{\infty} w(x'_i)\log\left[\Delta x\right]\Delta x\right\} =$$

$$= -\int_{-\infty}^{\infty} w(x)\log w(x)dx - \lim_{\Delta x\to 0}\log\left[\Delta x\right] = \infty\,.$$

Из полученного выражения следует, что энтропия непрерывного распределения равна бесконечности за счет второго слагаемого, которое одинаково для всех непрерывных распределений, заданных на интервале $(-\infty,\infty)$. «Индивидуальность» распределения определяется первым слагаемым, которое и используют в качестве меры информативности непрерывного источника и называют *относительной*, или *дифференциальной* энтропией

$$h(X) = -\int_{-\infty}^{\infty} w(x)\log w(x)dx\,.$$

Дифференциальная энтропия, в отличие от энтропии дискретного источника, самостоятельного смысла не имеет и служит для сравнения информативности различных непрерывных источников между собой [10].

Пример 8.10. Дифференциальная энтропия источника, описываемого гауссовской плотностью вероятности $w(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$, равна

$$h(X) = -\int_{-\infty}^{\infty} w(x)\log w(x)dx = \int_{-\infty}^{\infty} w(x)\log\frac{1}{w(x)}dx =$$

$$= \int_{-\infty}^{\infty} w(x)\left[\log\sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}(x-m)^2\right]dx =$$

$$= \log\sqrt{2\pi\sigma^2}\int_{-\infty}^{\infty} w(x)dx + \frac{\log e}{2\sigma^2}\int_{-\infty}^{\infty}(x-m)^2 w(x)dx =$$

$$= \log\sqrt{2\pi\sigma^2} + \frac{\log e}{2} = \log\sqrt{2\pi e\sigma^2}\,.$$

Отметим, что дифференциальная энтропия нормального распределения не зависит от математического ожидания и она тем больше, чем больше дисперсия. Это вполне соответствует пониманию дифференциальной энтропии как меры неопределенности, которая, очевидно, возрастает с ростом дисперсии случайной величины. ◀

Определим взаимную информацию двух непрерывных случайных величин x и y. Разобьем области их значений на интервалы Δx и Δy, перейдем к дискретным случайным величинам x' и y', после чего воспользуемся формулой (8.7) и выполним предельный переход:

$$I(X,Y) = \lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\left\{\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right\} =$$

$$= \lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\left\{\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} w(x_i,y_i)\Delta x\Delta y\log\frac{w(x_i,y_j)\Delta x\Delta y}{w(x_i)\Delta x w(y_j)\Delta y}\right\} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\log\frac{w(x,y)}{w(x)w(y)}dxdy\,.$$

Полученное выражение можно переписать следующим образом:

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\log\frac{w(y)w(x\,|\,y)}{w(x)w(y)}dxdy =$$

$$= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\log w(x)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\log w(x\,|\,y)dxdy =$$

$$= h(X) - h(X\,|\,Y)\,,$$

где $h(X|Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\log w(x|y)dxdy$ – условная дифференциальная энтропия.

Эпсилон-энтропия. Энтропия источника непрерывных сообщений, как было показано, равна бесконечности. Это означает по существу то, что для передачи непрерывного сообщения с *абсолютной* (бесконечной) точностью необходимо передать бесконечное количество информации. В то же время ясно, что на практике это и не требуется, так как любой получатель сообщений обладает ограниченной разрешающей способностью: достаточно воспроизвести сообщение с конечной точностью, характеризуемой некоторым малым числом ε. При этом количество передаваемой информации оказывается конечным и зависит от параметра ε. В качестве критерия можно использовать, например, средний квадрат разности между принятым $\tilde{x}(t)$ и переданным $x(t)$ сообщениями

$$\overline{\varepsilon^2(t)} = \overline{\left[\tilde{x}(t) - x(t)\right]^2}.$$

При этом сообщения считаются эквивалентными, если $\overline{\varepsilon^2(t)}$ не превосходит заданного уровня ε_0^2.

Средняя взаимная информация между $\tilde{x}(t)$ и $x(t)$

$$I(X,\tilde{X}) = h(X) - h(X|\tilde{X})$$

зависит не только от свойств сообщения $x(t)$, которыми определяется дифференциальная энтропия, но и от ε. Величина

$$H_\varepsilon(X) = \min I(X,\tilde{X}) = h(X) - \max h(X|\tilde{X}), \qquad (8.12)$$

где минимум выбирается по всем возможным условным распределениям $w(x|\tilde{x})$, для которых $\varepsilon < \varepsilon_0$, называется *эпсилон-энтропией*. Это *минимальное* количество информации в сообщении $\tilde{x}(t)$ о сообщении $x(t)$ при условии, что они эквиваленты в смысле критерия ε.

Положим

$$X(t) = \tilde{X}(t) - \varepsilon(t),$$

тогда условная дифференциальная энтропия $h(X|\tilde{X})$ при известном $\tilde{x}(t)$ полностью определяется дифференциальной энтропией

$h(\varepsilon)$ отсчета шума воспроизведения $\varepsilon(t)$. При этом $\max h\left(X \mid \tilde{X}\right) = \max h(\varepsilon)$. Поскольку мощность шума воспроизведения *ограничена* значением ε_0^2, очевидно, что дифференциальная энтропия $h(\varepsilon)$ максимальна, когда отсчет $\varepsilon(t)$ – гауссовская случайная величина с нулевым средним, и это максимальное значение равно

$$\max h(\varepsilon) = \log\sqrt{2\pi e\varepsilon_0^2}\,. \tag{8.13}$$

Подставляя это выражение в (8.12), получим

$$H_\varepsilon(X) = h(X) - \log\sqrt{2\pi e\varepsilon_0^2}\,.$$

Эпсилон-энтропия максимальна для гауссовского источника с дисперсией σ_x^2:

$$H_\varepsilon(X) = \log\sqrt{2\pi e\sigma_0^2} - \log\sqrt{2\pi e\varepsilon_0^2} = \frac{1}{2}\log\frac{\sigma_x^2}{\varepsilon_0^2}\,. \tag{8.14}$$

8.7. ПРОПУСКНАЯ СПОСОБНОСТЬ НЕПРЕРЫВНОГО КАНАЛА С АДДИТИВНЫМ БЕЛЫМ ГАУССОВСКИМ ШУМОМ

Пусть колебание $z(t)$ на выходе непрерывного канала представляет собой сумму сигнала $x(t)$ и шума $\xi(t)$:

$$z(t) = x(t) + \xi(t)\,, \tag{8.15}$$

причем сигнал и шум статистически независимы. Предположим, что канал имеет полосу частот, ограниченную частотой $F_к$, и действует в течение временного интервала T. Тогда согласно теореме отсчетов каждый процесс, входящий в выражение (8.15), может быть представлен совокупностью $M = 2F_кT$ отсчетов. Совокупность отсчетов сигнала, которую можно представить вектором $\mathbf{x} = x_1, x_2, ..., x_M$, имеет совместную плотность распределения вероятностей $w(\mathbf{x}) = w(x_1, x_2, ..., x_M)$; статистические свойства шумового вектора $\boldsymbol{\xi} = \xi_1, \xi_2, ..., \xi_M$ описываются совместной ПРВ $w(\boldsymbol{\xi}) = w\left(\xi_1, \xi_2, ..., \xi_M\right)$.

Пропускную способность непрерывного канала определим как

$$C = \lim_{T\to\infty} \frac{1}{T} \max_{w(\mathbf{x})} I(X, Z), \tag{8.16}$$

где $I(X, Z)$ – количество информации о реализации сигнала $x(t)$ длительности T, содержащееся (в среднем) в реализации процесса $z(t)$ той же длительности (максимум ищется среди всевозможных распределений $w(\mathbf{x})$).

Средняя взаимная информация сигнала и наблюдаемого процесса равна

$$I(X, Z) = h(Z) - h(Z \mid X). \tag{8.17}$$

С учетом (8.15) условная плотность распределения вероятности $w(\mathbf{z} \mid \mathbf{x}) = w(\boldsymbol{\xi})$, а условная дифференциальная энтропия

$$h(Z \mid X) = -\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} w(\mathbf{z}, \mathbf{x}) \log w(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} = h(\Xi), \tag{8.18}$$

где Ξ – обозначение векторной случайной величины, составленной из шумовых отсчетов. Таким образом, с учетом (8.16), (8.17) и (8.18) пропускная способность непрерывного канала с аддитивным шумом

$$C = \lim_{T\to\infty} \frac{1}{T} \max_{w(\mathbf{x})} \left[h(Z) - h(\Xi) \right]. \tag{8.19}$$

Пример 8.11. Рассмотрим пропускную способность непрерывного канала с аддитивным квазибелым гауссовским шумом, имеющим одностороннюю спектральную плотность мощности N_0 в полосе частот от 0 до $F_к$. Отсчеты шума статистически независимы, и дифференциальная энтропия

$$h(\Xi) = 2F_к T \log \sqrt{2\pi e \sigma_\xi^2} = F_к T \log\left(2\pi e \sigma_\xi^2\right) = F_к T \log\left(2\pi e P_ш\right), \tag{8.20}$$

где $P_ш = \sigma_\xi^2 = N_0 F_к$ – мощность (дисперсия) шума.

Дифференциальная энтропия $h(Z_i)$ случайной величины Z_i максимальна, если Z_i – гауссовская случайная величина с нулевым средним, а это означает, что X_i – тоже гауссовская случайная величина с нулевым средним. Дифференциальная энтропия совокуп-

ности отсчетов максимальна, если отсчеты статистически независимы (это выполняется, так как отсчеты квазибелого шума некоррелированные гауссовские, а значит, независимые), тогда

$$h(Z) = F_{\text{к}} T \log\left[2\pi e(P_{\text{с}} + P_{\text{ш}})\right]. \qquad (8.21)$$

Подставляя (8.21) и (8.20) в (8.19), получаем формулу Шеннона для пропускной способности непрерывного канала с АБГШ [10]:

$$C = F_{\text{к}} \log\left(1 + \frac{P_{\text{с}}}{P_{\text{ш}}}\right) = F_{\text{к}} \log\left(1 + \frac{P_{\text{с}}}{F_{\text{к}} N_0}\right). \qquad (8.22)$$

При расширении полосы пропускания пропускная способность непрерывного канала с АБГШ стремится к пределу

$$C_{\infty} = \frac{P_{\text{с}}}{N_0} \log e \approx 1{,}443 \frac{P_{\text{с}}}{N_0}.$$

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем состоит цель экономного кодирования?
2. Что такое избыточность дискретного источника?
3. Какой источник обладает минимальной избыточностью и чему равна его энтропия?
4. Может ли равномерный код быть оптимальным (безызбыточным)?
5. В результате применения процедуры экономного кодирования получился троичный код с вероятностями символов 0.5, 0.2 и 0.3 соответственно. Можно ли считать такой код оптимальным?
6. Можно ли применять коды, для которых префиксное правило не выполняется?
7. Может ли помехоустойчивый код быть безызбыточным?
8. На чем основано корректирующее свойство помехоустойчивых кодов?
9. Что такое кодовое расстояние?
10. В какой связи находятся порождающая и проверочная матрицы линейного кода?
11. Какой геометрический смысл имеют строки порождающей матрицы?
12. Какова размерность линейного пространства, натянутого на строки порождающей матрицы? Каково количество разрешенных комбинаций кода? Каково количество запрещенных комбинаций?
13. Что такое синдром?

УПРАЖНЕНИЯ

1. Рассчитайте для частных случаев, указанных в конце примера 8.2, а также для $p_0 = p_1 = 0.25$ условную энтропию согласно (8.4).

2. Даны источники с алфавитами, содержащими по три символа и с распределениями вероятностей $\begin{matrix}\alpha_1 & \alpha_2 & \alpha_3 \\ 1/3 & 1/3 & 1/3\end{matrix}$ и $\begin{matrix}\beta_1 & \beta_2 & \beta_3 \\ 1/2 & 1/4 & 1/4\end{matrix}$. Найдите энтропии источников.

3. Имеются два дискретных источника с матрицами $\begin{pmatrix}X \\ P\end{pmatrix} = \begin{pmatrix}x_1 & x_2 \\ p_1 & p_2\end{pmatrix}$, $\begin{pmatrix}Y \\ Q\end{pmatrix} = \begin{pmatrix}y_1 & y_2 & y_3 \\ q_1 & q_2 & q_3\end{pmatrix}$ (верхняя строка матрицы содержит символы, нижняя – их вероятности). Определите, какой источник обладает большей неопределенностью в случае, если: а) $p_1 = p_2$, $q_1 = q_2 = q_3$; б) $p_1 = q_1$, $p_2 = q_2 + q_3$.

4. По каналу связи передается один из двух символов x_1 или x_2 с одинаковыми вероятностями. На выходе они преобразуются в символы y_1 и y_2, причем из-за помех в среднем два символа из ста принимаются неверно. Определите среднее количество информации на один символ, передаваемое по такому каналу. Сравните с аналогичной величиной при отсутствии помех.

5. Марковский источник сообщений вырабатывает символы x_1, x_2 и x_3 с вероятностями 0,4, 0,5 и 0,1 соответственно. Вероятности появления пар заданы таблицей

$x_i x_j$	x_1x_1	x_1x_2	x_1x_3	x_2x_1	x_2x_2	x_2x_3	x_3x_1	x_3x_2	x_3x_3
$P(x_i, x_j)$	0,1	0,2	0,1	0,2	0,3	0	0,1	0	0

Определите энтропию источника и сравните ее с энтропией источника без памяти с такими же вероятностями символов.

Указание. Энтропия марковского источника первого порядка при объеме алфавита K находится по формуле

$$H(X) = -\sum_{i=1}^{K}\sum_{j=1}^{K} P(x_j, x_i)\log P(x_i \mid x_j).$$

6. Две двоичные случайные величины X и Y имеют совместные вероятности $P(x = y = 0) = P(x = 0, y = 1) = P(x = y = 1) = 1/3$. Найдите $H(X)$, $H(Y)$, $H(X \mid Y)$, $H(Y \mid X)$ и $H(X, Y)$.

7. Сообщения x_1, x_2, x_3 и x_4 появляются на выходе источника с вероятностями 1/2, 1/4, 1/8 и 1/8. Постройте двоичный код Шеннона – Фано и определите вероятности символов 0 и 1, а также среднюю длину кодового слова.

8. Источник вырабатывает два независимых символа α_1 и α_2 с вероятностями 0,9 и 0,1 соответственно. Постройте коды Хаффмена для отдельных символов и групп по два символа. Найдите и сравните для двух полученных кодов:

− среднюю длину кодового слова,

− избыточность кода,

− вероятность появления символа 0 (1) в кодовой последовательности,

− скорость передачи информации (длительность посылки примите равной 1 мкс).

9. Для условий упражнения 8 постройте код Хаффмена для групп по три символа. Найдите среднюю длину кодового слова, избыточность кода, вероятность появления символа 0 (1) в кодовой последовательности и скорость передачи информации. Сравните с аналогичными показателями для случая кодирования отдельных символов и групп по два символа.

10. Найдите две разрешенные кодовые комбинации кода Хэмминга (7, 4), не совпадающие со строками порождающей матрицы, и убедитесь в том, что расстояние между ними не менее трех.

11. Измените в одной из комбинаций два символа, найдите синдром и «исправьте» в принятой комбинации символ, на который он укажет. Найдите расстояние между «исправленной» и принятой комбинациями.

12. Код Хэмминга (7, 4) обнаруживает одно- и двукратные ошибки и исправляет однократные. Считая, что ошибки при приеме двоичных посылок независимы и происходят с вероятностью p, найдите вероятность появления двукратной ошибки в пределах 7-разрядной кодовой комбинации. Найдите вероятность появления не более чем одной ошибки.

9. ОСНОВЫ ТЕОРИИ ПОМЕХОУСТОЙЧИВОСТИ ПЕРЕДАЧИ ДИСКРЕТНЫХ СООБЩЕНИЙ

9.1. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ

В процессе передачи сообщений в системах связи выполняются различные преобразования, основные из которых показаны на упрощенной структурной схеме дискретной системы связи (рис. 9.1).

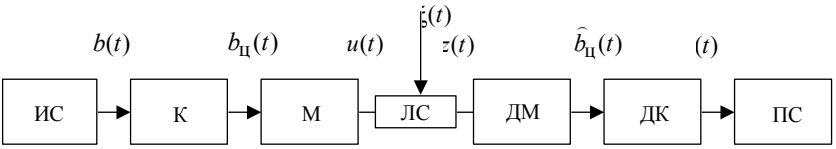


Рис. 9.1. Упрощенная структурная схема дискретной системы связи

Источник сигнала ИС включает в себя источник сообщений и преобразователь сообщения $a(t)$ в первичный сигнал $b(t)$. Первичный сигнал подвергается кодированию (экономному и/или помехоустойчивому) в кодере К, после чего сигнал $b_{\text{ц}}(t)$, называемый цифровым, поступает в модулятор М (передатчик), вырабатывающий сигнал $u(t)$, приспособленный по своим характеристикам для передачи по линии связи ЛС. В линии связи происходит искажение сигнала и его взаимодействие с помехой $\xi(t)$ (в простейшем случае аддитивное), в результате чего на вход демодулятора ДМ (приемника) поступает наблюдаемое колебание $z(t)$. Демодулятор выполняет функцию, обратную модуляции, поэтому на его выходе

должен быть выработан в идеальном случае сигнал $b_{\text{ц}}(t)$. Однако вследствие воздействия помех результат демодуляции $\widehat{b_{\text{ц}}}(t)$ отличается в общем случае от сигнала $b_{\text{ц}}(t)$, поэтому результат декодирования $\widehat{b}(t)$ также не совпадает с первичным сигналом $b(t)$.

В двоичной системе связи с *амплитудной телеграфией* (АТ) канальный сигнал, соответствующий передаваемому символу 1, представляет собой радиоимпульс с прямоугольной огибающей, а символу 0 соответствует отсутствие сигнала (*пауза*)[111]. При *частотной* (*фазовой*) телеграфии различные символы передаются сигналами одинаковой формы с несущей частотой (начальной фазой), меняющейся скачком от посылки к посылке. Для простоты здесь полагается, что система является изохронной, т. е. моменты начала и окончания элементарных посылок точно известны.

Для облегчения восприятия в дальнейшем рассматривается идеализированный канал связи без памяти, в котором отсутствуют искажения сигнала, тогда наблюдаемое колебание

$$z(t) = \sum_{k=-\infty}^{\infty} b_{\text{ц}}(t)s(t - k\tau) + \xi(t), \qquad (9.1)$$

где $s(t)$ – посылка длительности τ, $\xi(t)$ – помеха. Полагая, что отсутствует перекрытие посылок по времени (называемое *межсимвольной интерференцией*), можно считать, что в каждый момент времени $z(t) = s(t, b_i) + \xi(t)$, где b_i – одно из возможных значений цифрового сигнала[112].

Задача демодулятора состоит в том, чтобы по наблюдаемому колебанию $z(t)$ принять решение $\hat{b}_{\text{ц}}(t)$ о переданном сигнале $b_{\text{ц}}(t)$, такое, чтобы обеспечить максимальную *верность*. Правило (алгоритм) принятия решения – это закон преобразования $z(t)$ в $\hat{b}_{\text{ц}}(t)$. Поскольку помеха является случайной, задача построения оптимального (наилучшего) демодулятора представляет собой *статистическую* задачу и решается на основе методов теории вероятностей и математической статистики (теории статистических решений).

---

[111] Такой способ модуляции называют амплитудной телеграфией с пассивной паузой.

[112] Отметим, что выражение (9.1) представляет *частный* случай модуляции.

Перед принятием решения с целью повышения его качества (верности) часто наблюдаемое колебание подвергают дополнительной *обработке*. Если обработка линейная, то ее результат $y(t)$ может быть записан в форме

$$y(t) = \int_0^T z(\theta)\phi(t,\theta)d\theta = \int_0^T s(\theta, b_i)\phi(t,\theta)d\theta + \int_0^T \xi(\theta)\phi(t,\theta)d\theta,$$

где для простоты принято, что колебание наблюдается на интервале времени от 0 до T, $\phi(t,\theta)$ – ядро линейного оператора, описывающего устройство обработки (2.30). Видно, что результат обработки представляет собой сумму сигнальной и шумовой составляющих.

В простейшем случае $\phi(t,\theta) = \delta(\theta - t_0)$, тогда сигнальная составляющая равна величине

$$\int_0^T s(\theta, b_i)\phi(t,\theta)d\theta = \int_0^T s(\theta, b_i)\delta(\theta - t_0)d\theta = s(t_0, b_i),$$

т. е. отсчету канального сигнала (посылки) в момент времени t_0 (рис. 9.2).

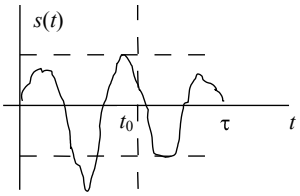


Рис. 9.2. Взятие отсчета в момент времени t_0

Очевидно, такой способ «обработки» плохо использует посылку: фактически правильность решения зависит не от энергии, а только от одного мгновенного значения сигнала. При этом очень важно, чтобы отсчет был взят точно в тот момент, когда значение сигнала достигает максимума. Улучшить эффективность решения можно путем «накопления» нескольких (K) отсчетов, взятых в i-е моменты времени, $i = 1,...,K$; при этом $\phi(t,\theta) = \sum_{i=1}^{K} \delta(\theta - t_{i0})$. Учесть различную *значимость* отсчетов для принятия решения можно, введя весовые коэффициенты при δ-функциях, тогда $\phi(t,\theta) = \sum_{i=1}^{K} h_i \delta(\theta - t_{i0})$. Увеличивая K, в пределе получаем непрерывное ядро оператора обработки $\phi(t,\theta) = h(t,\theta)$ – весовую функцию *линейного фильтра* (см. разд. 2.7). Вообще говоря, оптимальная обработка может быть нелинейной.

Материалом для принятия решения в демодуляторе служит в рассматриваемом случае реализация колебания $z(t)$ на интервале длительности T. Если бы помеха отсутствовала, то эта реализация совпадала бы с элементарным сигналом (посылкой), который можно считать точкой в гильбертовом пространстве сигналов, определенных на заданном временном интервале. Все возможные в данной системе связи посылки изображаются различными точками, и демодулятор должен вырабатывать свои решения в зависимости от того, какой именно точке соответствует принятая реализация $z(t)$. Реализация помехи, взаимодействуя с посылкой, смещает точку, изображающую принятую реализацию, причем смещение случайно вследствие случайного характера помехи. Если смещения будут значительными, демодулятор может ошибаться. Ошибка является случайным событием, поэтому качество решения можно характеризовать вероятностью ошибки.

Задача синтеза оптимального демодулятора (приемника) ставится следующим образом: нужно найти оптимальный алгоритм обработки и оптимальное правило решения, обеспечивающие максимальную вероятность безошибочного (правильного) решения. Максимум этой вероятности В.А. Котельников назвал *потенциальной помехоустойчивостью*, а приемник, реализующий этот максимум, – идеальным приемником.

Алгоритм работы приемника состоит в разбиении гильбертова пространства реализаций входного колебания на области, так что решение принимается в соответствии с тем, какой области принадлежит принятая реализация. Количество областей равно количеству различных кодовых символов данной системы связи. Ошибка возникает в том случае, если в результате воздействия помехи реализация попадает в «чужую» область. Оптимальный приемник разбивает пространство реализаций наилучшим образом, так что средняя вероятность ошибки минимальна среди всех возможных разбиений.

Каждая область соответствует предположению (*гипотезе*) о том, что передан был один из возможных сигналов. Поэтому каждая *простая* гипотеза есть предположение о том, что наблюдаемое колебание представляет собой реализацию случайного процесса, описываемого определенной многомерной плотностью распределения вероятностей[113] или функционалом плотности распределения.

---

[113] Часто гипотезе соответствует не одно распределение, а класс распределений, тогда гипотеза называется *сложной*.

Пример 9.1. Предположим, что результатом обработки в двоичной системе связи с амплитудной телеграфией является значение y, соответствующее окончанию интервала наблюдения. Если в колебании $z(t)$ присутствует только шум, имеющий гауссово распределение с нулевым математическим ожиданием, то плотность распределения величины y имеет вид

$$w_0(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}}; \qquad (9.2)$$

если кроме шума на вход приемника поступает сигнал, то результат обработки имеет ненулевое (для определенности – положительное) среднее a, и плотность распределения величины y имеет вид

$$w_1(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-a)^2}{2\sigma^2}}. \qquad (9.3)$$

Гипотезы, соответствующие выражениям (9.2) и (9.3), являются простыми. Если среднеквадратическое отклонение σ неизвестно, гипотезы являются *сложными*. ◄

Рассмотрим систему связи, в которой используются K различных символов. Тогда демодулятор должен различать K различных гипотез. При этом возможны ошибки: может быть принято решение D_j в пользу j-й гипотезы, в то время как справедливой является i-я гипотеза. Такая ситуация характеризуется *условной вероятностью ошибки* $p_{ij} = \mathbf{P}\{D_j \mid H_i\}$. Различные ошибки могут наносить разный вред, поэтому вводится численная характеристика Π_{ij}, называемая *риском*, или потерей. Иногда потери объединяют в квадратную $K \times K$-матрицу $\{\Pi_{ij}\}$, называемую *матрицей потерь*, при этом ее главная диагональ обычно содержит нули, что соответствует нулевым потерям при правильных решениях.

Символы, которым соответствуют разные гипотезы, могут иметь разные вероятности появления в сообщении. Поэтому каждая (i-я) гипотеза характеризуется некоторой вероятностью p_i осуществления, которая называется *априорной* вероятностью. Итак, суммируя, можно ввести усредненную характеристику (критерий) качества принятия решения, называемую *средним риском*

$$R = \sum_{i=1}^{K} \sum_{j=1}^{K} p_i p_{ij} \Pi_{ij}.$$

Средний риск представляет собой математическое ожидание потерь, связанных с принятием решения.

Если априорные вероятности гипотез точно известны, а потери назначены обоснованно, то приемник, обеспечивающий *наименьший средний риск,* будет наиболее выгодным. Критерий минимума среднего риска называют также критерием Байеса[114].

Иногда потери, связанные с различными ошибками, принимают равными друг другу $\Pi_{ij} = \Pi$, $\Pi_{ii} = 0$, $i = 1,...,K$, тогда оптимальный байесовский приемник обеспечивает минимальную *среднюю вероятность* ошибки (критерий *идеального наблюдателя*)

$$p_{\text{ош}} = \sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} p_i p_{ij}$$

и называется идеальным приемником Котельникова.

Если также принять равными априорные вероятности гипотез $p_i = 1/K$, $i = 1,...,K$, то критерий Байеса сводится к критерию *минимума суммарной условной вероятности ошибки*

$$p_{\text{ош усл}} = \sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} p_{ij} \,. \tag{9.4}$$

Проблема синтеза оптимального демодулятора состоит в нахождении границ областей, разбивающих пространство наблюдений наилучшим образом в соответствии с выбранным критерием качества. Ниже эта задача рассматривается для простейшего случая двух *простых гипотез*, что соответствует АТ-системе связи с пассивной паузой.

9.2. БИНАРНАЯ ЗАДАЧА ПРОВЕРКИ ПРОСТЫХ ГИПОТЕЗ

Наиболее просто задача построения оптимального демодулятора (приемника) решается для случая амплитудной телеграфии с пассивной паузой, что соответствует принятию решения о том, что передавался символ 0 (сигнала нет) или символ 1 (сигнал есть). Таким образом, решается задача *обнаружения* сигнала в наблюдае-

[114] Томас Байес (1702 – 1761) – английский математик, один из основоположников теории вероятностей и математической статистики.

мом колебании. Далее предполагается, что помеха в канале представляет собой гауссовский шум с нулевым средним и известной дисперсией, который взаимодействует с сигналом аддитивно (суммируется). Результатом обработки наблюдаемого колебания является случайная величина y, которая может иметь различное распределение в зависимости от того, есть ли сигнал в наблюдаемом колебании, а именно: распределение при гипотезе H_0 – «сигнала нет» – является гауссовским с нулевым средним, а распределение при гипотезе H_1 – «сигнал есть» – отличается сдвигом на величину a, зависящую от способа обработки (например, если обработка сводится к взятию отсчета в момент, когда несущее колебание достигает максимума, величина a представляет собой его амплитуду). Значение a предполагается известным. Таким образом, проверяемые гипотезы описываются двумя *условными* плотностями распределения вероятностей $w(y \mid H_0)$ и $w(y \mid H_1)$, изображенными на рис. 9.3.

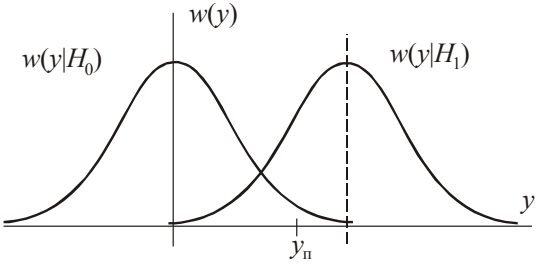


Рис. 9.3. Условные плотности распределения вероятностей величины y при простых гипотезах

В данной постановке демодулятор (приемник) может принимать решение, основываясь только на наблюдаемом значении y: очевидно, чем больше наблюдаемое значение, тем с большей уверенностью можно утверждать, что сигнал в принятом колебании есть. Приемник в таком случае должен сравнить y с некоторым фиксированным значением (порогом) $y_{\text{п}}$ и если y больше порога, принять решение о наличии сигнала, в противном случае – о его отсутствии, что можно кратко записать в следующей символической форме:

$$y > y_{\text{п}} \to \text{"1"},$$
$$y \le y_{\text{п}} \to \text{"0"}.$$

Каким бы ни был порог $y_{\text{п}}$, очевидно, есть некоторая ненулевая вероятность p_{01} принять решение о наличии сигнала при его фактическом отсутствии. Эта вероятность называется условной вероятностью *ошибки первого рода* («ложной тревоги») и определяется выражением

$$p_{01} = \int_{y_{\text{п}}}^{\infty} w(y \mid H_0) dy .$$

Аналогично, существует ненулевая вероятность принять решение об отсутствии сигнала, в то время как на самом деле он есть (условная вероятность *ошибки второго рода*, или пропуска сигнала)

$$p_{10} = \int_{-\infty}^{y_{\text{п}}} w(y \mid H_1) dy .$$

Анализ рис. 9.3 показывает, что сумма указанных условных вероятностей минимальна, если порог $y_{\text{п}}$ находится как абсцисса точки пересечения условных плотностей $w(y \mid H_0)$ и $w(y \mid H_1)$. Очевидно, при таком выборе порога приемник является оптимальным по критерию *минимума суммарной условной вероятности ошибки* (9.4) и принятие решения основывается на сравнении значений функций $w(y \mid H_0)$ и $w(y \mid H_1)$ при наблюдаемом значении y:

$$w(y \mid H_0) < w(y \mid H_1) \rightarrow \text{"1"};$$

$$w(y \mid H_0) \geq w(y \mid H_1) \rightarrow \text{"0"}.$$

Это правило принятия решения можно переписать также в форме

$$\frac{w(y \mid H_1)}{w(y \mid H_0)} > 1 \rightarrow \text{"1"}; \quad \frac{w(y \mid H_1)}{w(y \mid H_0)} \leq 1 \rightarrow \text{"0"}. \qquad (9.5)$$

Решение, таким образом, принимается в пользу той гипотезы, которая представляется более *правдоподобной* при данном значении y, поэтому отношение $\dfrac{w(y \mid H_1)}{w(y \mid H_0)}$ называется *отношением правдоподобия* и обозначается $\Lambda(y)$. Правило (9.5) называют правилом *максимального правдоподобия*. Заметим, что критерий (9.4) часто называют *критерием максимума правдоподобия*.

Критерий идеального наблюдателя предполагает учет *априорных* вероятностей гипотез, и оптимальный в смысле этого критерия

приемник обеспечивает минимум средней вероятности ошибки, т. е. наименьшую *сумму безусловных вероятностей ошибок* первого и второго рода. Иначе говоря, сравнению подлежат функции $w(y \mid H_0)$ и $w(y \mid H_1)$, умноженные на соответствующие априорные вероятности. Правило принятия решения в таком приемнике можно записать в форме

$$\frac{p_1 w(y \mid H_1)}{p_0 w(y \mid H_0)} > 1 \to \text{"1"}; \quad \frac{p_1 w(y \mid H_1)}{p_0 w(y \mid H_0)} \le 1 \to \text{"0"}.$$

Используя понятие отношения правдоподобия, можно записать правило в виде

$$\Lambda(y) > \frac{p_0}{p_1} \to \text{"1"}, \quad \Lambda(y) \le \frac{p_0}{p_1} \to \text{"0"},$$

при этом отношение правдоподобия сравнивается с пороговым значением, зависящим от априорных вероятностей.

Наконец, в случае байесовского критерия решение принимается по правилу

$$\frac{\Pi_{10} p_1 w(y \mid H_1)}{\Pi_{01} p_0 w(y \mid H_0)} > 1 \to \text{"1"}; \quad \frac{\Pi_{10} p_1 w(y \mid H_1)}{\Pi_{01} p_0 w(y \mid H_0)} \le 1 \to \text{"0"},$$

или

$$\Lambda(y) > \frac{p_0 \Pi_{01}}{p_1 \Pi_{10}} \to \text{"1"}, \quad \Lambda(y) \le \frac{p_0 \Pi_{01}}{p_1 \Pi_{10}} \to \text{"0"}.$$

Итак, во всех случаях оптимальный приемник (демодулятор, или решающее устройство) «устроен одинаково»: для наблюдаемого значения y, зависящего от принятой реализации $z(t)$, вычисляется значение отношения правдоподобия, которое сравнивается с порогом; порог равен $\frac{p_0 \Pi_{01}}{p_1 \Pi_{10}}$ для приемника, оптимального в смысле критерия минимума среднего риска, p_0 / p_1 для идеального приемника Котельникова и 1 для приемника максимального правдоподобия.

В заключение отметим, что иногда удобнее вычислять не отношение правдоподобия, а его логарифм. В силу монотонности логарифмической функции это не влияет на условные вероятности ошибок, если порог также прологарифмировать.

9.3. ПРИЕМ ПОЛНОСТЬЮ ИЗВЕСТНОГО СИГНАЛА (КОГЕРЕНТНЫЙ ПРИЕМ)

Рассмотрим принятие решения в системе связи при следующих условиях: синхронизация является точной и форма сигнала на интервале наблюдения точно известна, неизвестен лишь сам факт наличия либо отсутствия сигнала в наблюдаемом колебании. (Эта ситуация наиболее близка к реальности в кабельных линиях связи, где условия распространения сигналов известны и практически неизменны.)

Будем считать, что на интервале наблюдения независимо от сигнала присутствует гауссовский шум с нулевым средним и спектральной плотностью мощности $N_0/2$, постоянной в некоторой полосе частот $-F < f < F$ («квазибелый» шум). Полагая, что длительность интервала наблюдения равна T, возьмем n отсчетов наблюдаемого колебания с шагом $\Delta t = \dfrac{1}{2F} = \dfrac{T}{n}$, при этом отсчеты шума являются некоррелированными вследствие того, что корреляционная функция квазибелого шума (вида "$\sin x/x$") пересекает ось абсцисс при значениях времени, кратных Δt. Поэтому совместная плотность распределения вероятностей взятых отсчетов (*выборочных* значений) равна в отсутствие сигнала

$$w(z_1,...,z_n \mid H_0) = \frac{1}{\left(\sqrt{2\pi}\cdot\sigma\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n} z_k^2},$$

где $\sigma^2 = N_0 F = N_0/(2\Delta t)$. Напомним, что для гауссовских случайных величин некоррелированность влечет независимость.

Если сигнал присутствует и принимает в моменты взятия отсчетов значения $s_k = s(t_k)$, то совместная плотность распределения вероятностей выборочных значений

$$w(z_1,...,z_n \mid H_1) = \frac{1}{\left(\sqrt{2\pi}\cdot\sigma\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n} (z_k - s_k)^2}.$$

Отношение правдоподобия

$$\Lambda = \frac{w(z_1,...,z_n \mid H_1)}{w(z_1,...,z_n \mid H_0)} = e^{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{n}(z_k - s_k)^2 - \sum_{k=1}^{n} z_k^2\right]}.$$

Подставляя в это выражение $2\sigma^2 = N_0 / \Delta t$, получим

$$\Lambda = e^{-\frac{1}{N_0}\left[\sum_{k=1}^{n}(z_k - s_k)^2 \Delta t - \sum_{k=1}^{n} z_k^2 \Delta t\right]}. \qquad (9.6)$$

Устремляя Δt к нулю ($n \to \infty$), запишем логарифм отношения правдоподобия

$$\ln \Lambda = -\frac{1}{N_0}\int_0^T \left[z(t) - s(t)\right]^2 dt + \frac{1}{N_0}\int_0^T z^2(t)dt =$$

$$= \frac{2}{N_0}\int_0^T z(t)s(t)dt - \frac{1}{N_0}\int_0^T s^2(t)dt\ . \qquad (9.7)$$

Поскольку логарифм является монотонной функцией, правило обнаружения сигнала известной формы на фоне гауссовского квазибелого шума, оптимальное в смысле критерия максимума правдоподобия, основано на сравнении с нулевым порогом величины

$$\int_0^T y(t)s(t)dt - \frac{E}{2}, \qquad (9.8)$$

где $E = \int_0^T s^2(t)dt$ – энергия сигнала. Первое слагаемое в выражении (9.8) называется *корреляционным интегралом*, так как совпадает по форме с выражением взаимно корреляционной функции сигнала и наблюдаемого процесса при нулевом сдвиге. Так как энергия сигнала известна, то при обнаружении можно сравнивать значение корреляционного интеграла (случайное в силу случайности реализации $z(t)$) с порогом, равным $E/2$.

Правило *различения* M сигналов известной формы на фоне гауссовского квазибелого шума, оптимальное в смысле критерия максимума правдоподобия, основано на сравнении *между собой* величин $\int_0^T z(t)s_i(t)dt - E_i/2$, $i = 1,...,M$. Решение принимается в

пользу того сигнала, для которого эта величина максимальна. Структура оптимального приемника для различения M сигналов показана на рис. 9.4.

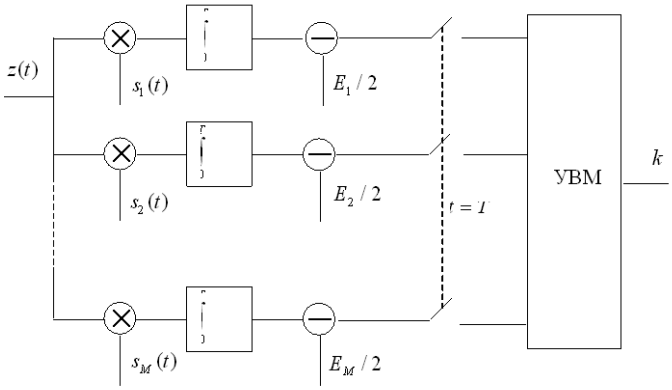


Рис. 9.4. Структура приемника максимального правдоподобия

Устройство выбора максимума УВМ выдает на выход номер k канала, в котором величина (9.8) максимальна. Приемник упрощается, когда энергии всех сигналов равны.

Пример 9.2. В проводных системах связи с амплитудной телеграфией могут применяться посылки в форме прямоугольного видеоимпульса. Предположим, что сигнал, соответствующий символу «1», представляет собой прямоугольный видеоимпульс с амплитудой a и длительностью T. Тогда корреляционный интеграл имеет вид

$$\int_0^T z(t)s(t)dt = a\int_0^T z(t)dt\,,$$

а порог равен $E/2 = a^2T/2$, тогда решающее правило имеет вид

$$\int_0^T z(t)dt > aT/2 \to \text{"1"}, \qquad \int_0^T z(t)dt \le aT/2 \to \text{"0"}.$$

Структурная схема приемника показана на рис. 9.5.

Постоянная времени интегрирующей цепи должна быть много больше длительности посылки T. В этом случае начальный участок экспоненты $a(1-e^{-t/(RC)})$, отображающей заряд емкости, можно аппроксимировать прямой линией с тангенсом угла наклона

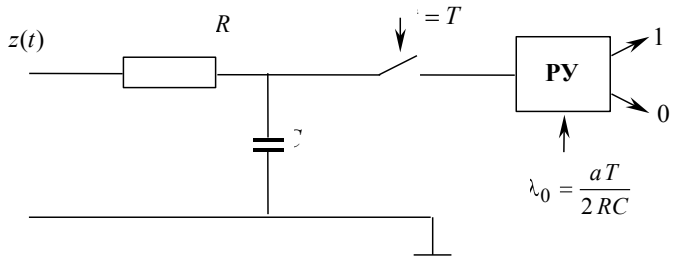


Рис. 9.5. Структурная схема приемника прямоугольного ви деоимпульса на фоне гауссовского шума

$a/(RC)$, равным производной экспоненты в нуле. Тогда за время T напряжение на входе решающего устройства, обусловленное сигналом, составит $aT/(RC)$, а значение порога должно быть равно $aT/(2RC)$. ◄

Пример 9.3. Предположим, что в двоичной системе связи с амплитудной телеграфией сигнал, соответствующий символу «1», представляет собой прямоугольный *радиоимпульс* с амплитудой a и длительностью T. Тогда $s(t) = a\cos(\omega_0 t + \varphi)$, корреляционный интеграл имеет вид

$$\int_0^T z(t)s(t)dt = a\int_0^T z(t)\cos(\omega_0 t + \varphi)dt,$$

а порог равен $E/2 = a^2T/4$. Сокращая на a и применяя реальный интегратор в виде RC-цепи, получаем структуру приемника, показанную на рис. 9.6. ◄

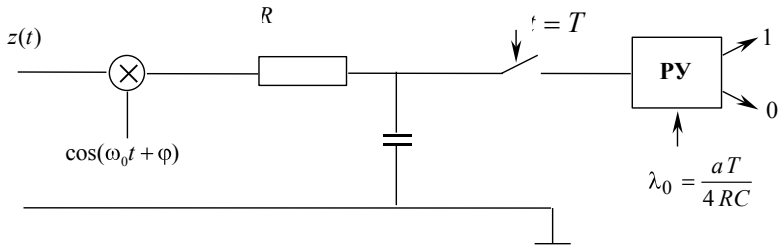


Рис. 9.6. Структурная схема приемника прямоугольного радиоимпульса на фоне гауссовского шума

Пример 9.4. В двоичной системе связи с фазовой телеграфией сигналы $s_1(t)$ и $s_2(t)$, соответствующие символам «1» и «0», являются противоположными

$$s_1(t) = a\cos(\omega_0 t + \varphi);$$

$$s_2(t) = a\cos(\omega_0 t + \varphi + \pi) = -s_1(t).$$

Принятие решения основано на сравнении величин $\int_0^T z(t)s_1(t)dt - E_1/2$ и $\int_0^T z(t)s_2(t)dt - E_2/2$. С учетом равенства энергий правило принятия решения упрощается и принимает вид

$$\int_0^T z(t)\cos(\omega_0 t + \varphi)dt > 0 \to \text{"1"}, \quad \int_0^T z(t)\cos(\omega_0 t + \varphi)dt \le 0 \to \text{"0"}. \blacktriangleleft$$

9.4. СОГЛАСОВАННАЯ ФИЛЬТРАЦИЯ

В случае приема сигнала известной формы, как было показано, устройство принятия решения (демодулятор) должно вычислять значение корреляционного интеграла, которое и сравнивается с порогом, выбираемым в соответствии с принятым критерием эффективности. Устройство, вычисляющее корреляционный интеграл, называется *коррелятором* (рис. 9.7).

Коррелятор является *нестационарным* (параметрическим) устройством и включает генератор опорного колебания, совпадающего по форме с ожидаемым сигналом на интервале наблюдения, и интегратор, на выходе которого в момент окончания интервала наблюдения формируется значение, сравниваемое с порогом. В некоторых случаях удобнее использовать линейную стационарную (инвариантную к сдвигу) цепь, которая вычисляет значение корреляционного интеграла и называется *согласованным фильтром*. Этот фильтр, как и любая линейная инвариантная к сдвигу цепь, исчерпывающим образом описывается импульсной характеристикой $h_{\text{сф}}(t)$, при этом выходной сигнал определяется сверткой (интегралом Дюамеля), которая для

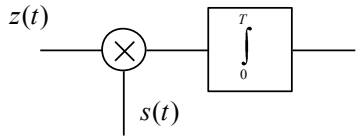


Рис. 9.7. Структура коррелятора

момента t_0 сравнения с порогом равна $\int_{-\infty}^{\infty} z(t)h_{\text{сф}}(t_0 - t)dt$, а с учетом финитности посылки $\int_{0}^{T} z(t)h_{\text{сф}}(t_0 - t)dt$.

Учитывая, что в момент t_0 на выходе согласованного фильтра должно быть выработано значение корреляционного интеграла, приходим к выводу, что должно выполняться равенство

$$\int_{0}^{T} z(t)h_{\text{сф}}(t_0 - t)dt = \int_{0}^{T} z(t)s(t)dt,$$

откуда $h_{\text{сф}}(t_0 - t) = s(t)$, следовательно, $h_{\text{сф}}(t) = s(t_0 - t)$. Импульсная характеристика согласованного фильтра, таким образом, совпадает по форме с ожидаемым сигналом, обращенным во времени и задержанным на время t_0. Для выполнения требования *каузальности* (причинности, физической реализуемости) фильтра, очевидно, необходимо, чтобы t_0 было не меньше, чем T (рис. 9.8).

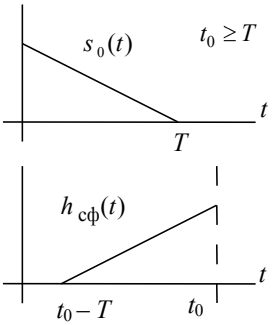


Рис. 9.8. Сигнал и импульсная характеристика согласованного фильтра

Найдем комплексную частотную характеристику согласованного фильтра:

$$H_{\text{сф}}(\omega) = \int_{-\infty}^{\infty} h_{\text{сф}}(t)e^{-j\omega t}dt = \int_{-\infty}^{\infty} s(t_0 - t)e^{-j\omega t}dt =$$

$$= \int_{-\infty}^{\infty} s(\tau)e^{-j\omega(t_0-\tau)}d\tau = e^{-j\omega t_0}\int_{-\infty}^{\infty} s(\tau)e^{j\omega\tau}d\tau =$$

$$= e^{-j\omega t_0}\left(\int_{-\infty}^{\infty} s(\tau)e^{-j\omega\tau}d\tau\right)^{*} = e^{-j\omega t_0}S^{*}(\omega).$$

Таким образом, КЧХ согласованного фильтра является комплексно-сопряженной функцией по отношению к спектральной плотности ожидаемого сигнала, умноженной на фазовый множитель, соответствующий задержке на t_0, необходимой для обеспечения каузальности.

Для понимания физического смысла согласованной фильтрации целесообразно рассмотреть отдельно составляющие КЧХ: амплитудно-частотную и фазочастотную характеристики.

Амплитудно-частотная характеристика совпадает по форме с модулем спектральной плотности сигнала. Это означает, что согласованный фильтр имеет больший коэффициент передачи для более интенсивных частотных компонент сигнала («подчеркивает» сильные гармоники и подавляет слабые).

Фазочастотная характеристика состоит из двух сомножителей, а именно: аргумента функции $S^*(\omega)$, обратного фазовому спектру (спектральной плотности фазы) сигнала, и фазового множителя $e^{-j\omega t_0}$. Первый сомножитель обеспечивает суммирование всех частотных компонент сигнала «в фазе», благодаря чему в момент времени t_0, обусловленный множителем $e^{-j\omega t_0}$, имеет место максимальное значение отклика, *численно* равное энергии сигнала[115]

$$\int_0^T s(t)h_{\text{сф}}(t_0 - t)dt = \int_0^T s(t)s(t)dt = E_s .$$

Для произвольного момента времени $t \in (0, T)$ отклик согласованного фильтра на «свой» сигнал

$$\int_0^T s(\tau)h_{\text{сф}}(t - \tau)d\tau = \int_0^T s(\tau)s(t_0 - t + \tau)d\tau = B_s(t_0 - t),$$

где $B_s(t)$ – автокорреляционная функция сигнала, которая, как известно, достигает максимума, равного энергии сигнала, при нулевом значении аргумента.

Согласованный фильтр для сигнала произвольной формы может быть реализован (приближенно) на основе линии задержки с отводами (рис. 9.9).

При подаче на вход 1 линии задержки с отводами ЛЗО короткого импульса (в идеале – δ-функции) на вход ФНЧ поступают (с интервалом Δt, обусловленным конструкцией линии задержки) такие же импульсы с амплитудами, определяемыми коэффициентами усиления a_0, a_1, a_2, … a_{n-1}. Тогда на выходе ФНЧ формируется сигнал, равный взвешенной сумме функций, получаемых сдвигами импульсной характеристики ФНЧ. В частности, если

[115] Очень важно понимать, что здесь имеется в виду значение *напряжения* на выходе фильтра.

ФНЧ является идеальным с П-образной КЧХ и частотой среза $F_в$, то его импульсная характеристика имеет вид

$$\frac{\sin(2\pi F_в t)}{2\pi F_в t},$$

а отклик устройства на короткий импульс, поданный на вход 1, представляет собой конечную сумму ряда Котельникова

$$\hat{s}(t) = \sum_{k=0}^{n-1} a_k \frac{\sin[2\pi F_в (t - k\Delta t)]}{2\pi F_в (t - k\Delta t)},$$

аппроксимирующую сигнал $s(t)$ требуемого вида. Нетрудно видеть, что если короткий импульс подать на вход 2, то отклик будет зеркальной копией сигнала $s[(n-1)\Delta t - t]$. Коэффициенты a_0, a_1, $a_2, \ldots, a_{n-1}$ представляют собой отсчеты сигнала $s(t)$ с шагом, определяемым верхней частотой $F_в$ спектра сигнала.

Следует иметь в виду, что такой способ реализации согласованного фильтра является хотя и универсальным, но заведомо приближенным, так как любой сигнал конечной длительности имеет нефинитную спектральную плотность, а идеальный ФНЧ нереализуем (см. разд. 2.11). Тем не менее этот способ применяется на практике; например, для согласованной фильтрации сигналов с линейной частотной модуляцией используют в качестве линий задержки с отводами интегральные устройства на поверхностных акустических волнах (ПАВ).

Очевидно, форма сигнала на выходе согласованного фильтра отличается от формы сигнала на его входе. Это естественно, так

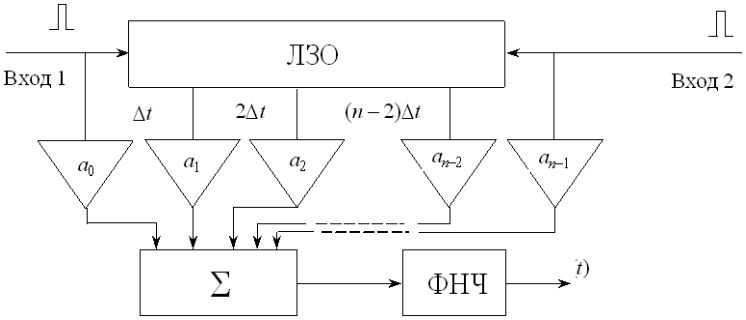


Рис. 9.9. Согласованный фильтр на основе линии задержки с отводами

как назначение согласованного фильтра состоит в вычислении корреляционного интеграла для наиболее надежного *принятия решения* о наличии или отсутствии сигнала на входе приемника. Иными словами, согласованный фильтр должен обеспечивать максимальное отношение сигнал/шум в момент времени t_0. Убедимся, что это действительно так при условии, что входной шум является белым стационарным процессом с нулевым средним.

Пусть на вход фильтра с импульсной характеристикой $h(t)$ воздействует процесс $z(t) = s(t) + \xi(t)$, где $\xi(t)$ – белый стационарный шум с нулевым средним, тогда сигнальная составляющая выходного процесса

$$u_{\text{с}}(t_0) = \int_0^{t_0} s(\tau) h(t_0 - \tau) d\tau,$$

а шумовая составляющая

$$u_{\text{ш}}(t_0) = \int_0^{t_0} \xi(\tau) h(t_0 - \tau) d\tau.$$

Так как $\overline{\xi(t)} = 0$, то $\overline{u_{\text{ш}}(t_0)} = \int_0^{t_0} \overline{\xi(\tau)} h(t_0 - \tau) d\tau = 0$, поэтому дисперсия шумовой составляющей выходного процесса равна среднему квадрату, а поскольку $\xi(t)$ – белый шум,

$$\sigma_{\text{ш}}^2 = \overline{u_{\text{ш}}^2(t_0)} =$$

$$= \int_0^{t_0} \int_0^{t_0} \overline{\xi(\tau_1)\xi(\tau_2)} h(t_0 - \tau_1) h(t_0 - \tau_2) d\tau_1 d\tau_2 =$$

$$= \int_0^{t_0} \int_0^{t_0} \frac{N_0}{2} \delta(\tau_1 - \tau_2) h(t_0 - \tau_1) h(t_0 - \tau_2) d\tau_1 d\tau_2 =$$

$$= \frac{N_0}{2} \int_0^{t_0} h^2(t_0 - \tau) d\tau = \frac{N_0}{2} E_h,$$

где E_h – энергия импульсной характеристики.

Отношение сигнал/шум по мощности в момент отсчета составляет

$$q = \frac{2u_c^2(t_0)}{N_0 E_h} = \frac{2\left|\int_0^{t_0} s(\tau)h(t_0-\tau)d\tau\right|^2}{N_0 E_h}.$$

Заметим, что согласно неравенству Шварца

$$\left|\int_0^{t_0} s(\tau)h(t_0-\tau)d\tau\right|^2 \le \int_0^{t_0} s^2(\tau)d\tau \int_0^{t_0} h^2(t_0-\tau)d\tau$$

и равенство достигается лишь тогда, когда $h(t) = As(t_0 - t)$ при произвольном коэффициенте A. Таким образом, в момент t_0 среди всех ЛИС-цепей именно согласованный фильтр обеспечивает максимальное отношение сигнал/шум на выходе. Умножение импульсной характеристики на коэффициент A не влияет на отношение сигнал/шум (почему?).

9.5. ПОТЕНЦИАЛЬНАЯ ПОМЕХОУСТОЙЧИВОСТЬ КОГЕРЕНТНОГО ПРИЕМА

Напомним, что по определению В.А. Котельникова потенциальной помехоустойчивостью называется максимум вероятности правильного решения, достижимый при заданных условиях приема сигналов на фоне помех (шумов). Определим потенциальную помехоустойчивость приема двух сигналов $s_1(t)$ и $s_0(t)$ известной формы на фоне белого гауссовского шума при равных априорных вероятностях сигналов.

Алгоритм принятия решения в приемнике, реализующем критерий максимума правдоподобия, кратко запишем в виде

$$\int_0^T z(t)s_1(t)dt - E_1/2 \underset{0}{\overset{1}{\gtrless}} \int_0^T z(t)s_0(t)dt - E_0/2.$$

Это выражение можно привести к виду

$$\int_0^T z(t)[s_1(t) - s_0(t)]dt \underset{0}{\overset{1}{\gtrless}} (E_1 - E_0)/2.$$

Ошибки при приеме состоят в том, что при передаче первого сигнала принимается решение о приеме второго и наоборот. Поскольку гауссово распределение симметрично и априорные вероятности равны, легко видеть, что суммарная (средняя) вероятность ошибки равна любой из условных вероятностей ошибок (убедитесь в этом!).

Найдем условную вероятность ошибки, т. е. вероятность события, заключающегося в принятии решения о наличии сигнала $s_0(t)$ при условии, что в наблюдаемом колебании присутствует сигнал $s_1(t)$. Это событие соответствует выполнению неравенства

$$\int_0^T [s_1(t)+\xi(t)][s_1(t)-s_0(t)]dt \lessdot (E_1-E_0)/2,$$

которое можно переписать в виде

$$\int_0^T s_1^2(t)dt+\int_0^T \xi(t)[s_1(t)-s_0(t)]dt-\int_0^T s_1(t)s_0(t)dt \lessdot$$

$$\lessdot \frac{1}{2}\int_0^T s_1^2(t)dt-\frac{1}{2}\int_0^T s_0^2(t)dt.$$

Проведя очевидные преобразования, получим

$$\int_0^T \xi(t)[s_1(t)-s_0(t)]dt \lessdot -\frac{1}{2}\int_0^T [s_1(t)-s_0(t)]^2 dt. \qquad (9.9)$$

Левая часть неравенства представляет собой случайную величину (так как это интеграл по времени от случайного процесса $\xi(t)$ с весом, равным разности сигналов $s_\Delta(t)=[s_1(t)-s_0(t)]$), имеющую нормальное распределение (поскольку процесс $\xi(t)$ гауссов) с нулевым средним (очевидно); обозначим ее ν и найдем ее средний квадрат, равный дисперсии:

$$D_\nu=\overline{\nu^2}=\int_0^T\int_0^T \overline{\xi(t_1)\xi(t_2)}s_\Delta(t_1)s_\Delta(t_2)dt_1dt_2=$$

$$=\frac{N_0}{2}\int_0^T\int_0^T \delta(t_1-t_2)s_\Delta(t_1)s_\Delta(t_2)dt_1dt_2=\frac{N_0}{2}\int_0^T s_\Delta^2(t)dt=\frac{N_0E_\Delta}{2}. \qquad (9.10)$$

Вероятность выполнения неравенства (9.9) – это, очевидно, вероятность того, что нормальная случайная величина с нулевым

средним и дисперсией $N_0E_\Delta/2$ принимает значение меньше, чем $-E_\Delta/2$. Эта вероятность равна

$$p_{10}=\int\limits_{-\infty}^{-E_\Delta/2}\frac{1}{\sqrt{2\pi D_\nu}}e^{-\frac{\nu^2}{2D_\nu}}d\nu=\int\limits_{\alpha}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt\,,$$

где $t=-\nu/\sqrt{D_\nu}$ – центрированная нормальная случайная величина с единичной дисперсией, а $\alpha=E_\Delta/\left(2\sqrt{D_\nu}\right)$ – положительное число. Очевидно, p_{10} зависит только от $\alpha=E_\Delta/\left(2\sqrt{D_\nu}\right)=\sqrt{\frac{E_\Delta}{2N_0}}$, поэтому можно ввести функцию

$$Q(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{x}^{\infty}e^{-\frac{t^2}{2}}dt=1-\Phi(x)\,,$$

где $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$ – интеграл вероятности, и записать $p_{10}=Q\left(\sqrt{\frac{E_\Delta}{2N_0}}\right)$. (Напомним, что в силу симметрии гауссовского распределения $p_{10}=p_{01}$.)

Таким образом, условная вероятность ошибки, равная средней вероятности ошибки при когерентном приеме сигналов на фоне белого шума, определяется энергией разностного сигнала $s_\Delta(t)$ и спектральной плотностью мощности шума N_0.

Рассмотрим потенциальную помехоустойчивость двоичного когерентного приемника максимального правдоподобия для различных способов модуляции, считая, что энергия сигнала E фиксирована.

1. *Амплитудная телеграфия с пассивной паузой*

В этом случае $s_0(t)=0$ и энергия разностного сигнала равна E (норма равна $\sqrt{E}$), рис. 9.10, *а*. Следовательно, потенциальная помехоустойчивость определяется средней вероятностью ошибки

$$p_{\text{ош}}^{\text{АТ-ПП}}=Q\left(\sqrt{\frac{E}{2N_0}}\right).$$

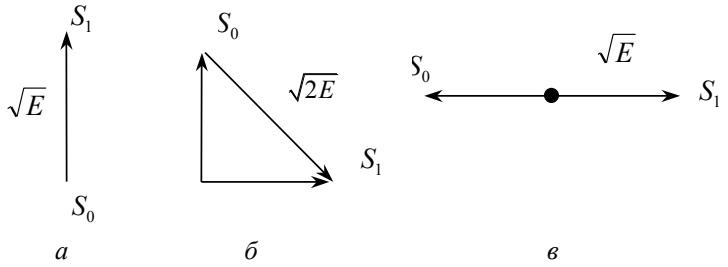


Рис. 9.10. К помехоустойчивости приема двух сигналов

2. *Частотная телеграфия с ортогональными сигналами*

Два сигнала представляют собой радиоимпульсы одинаковой формы с различными несущими частотами, так что сигналы взаимно ортогональны (рис. 9.10, *б*). Энергия разностного сигнала равна $2E$, а средняя вероятность ошибки

$$p_{\text{ош}}^{\text{ЧТ}} = Q\left(\sqrt{\frac{E}{N_0}}\right).$$

Повышение потенциальной помехоустойчивости при переходе от АТ-ПП к частотной телеграфии представляется естественным, так как во втором случае вдвое возрастает средняя мощность передатчика. Однако средняя вероятность ошибки может быть дополнительно понижена без увеличения мощности передатчика, если перейти к взаимно обратным сигналам.

3. *Фазовая телеграфия с манипуляцией фазы на 180°*

В случае фазовой телеграфии с взаимно обратными сигналами (рис. 9.10, *в*) энергия разностного сигнала составляет $4E$, средняя вероятность ошибки равна

$$p_{\text{ош}}^{\text{ФТ}} = Q\left(\sqrt{\frac{2E}{N_0}}\right)$$

и дальнейшее повышение потенциальной помехоустойчивости за счет выбора сигналов при заданной энергии, очевидно, невозможно.

Заметим, что если используются *три* сигнала одинаковой энергии, то для достижения максимальной помехоустойчивости они должны иметь взаимный фазовый сдвиг 120°, т. е. соответствующие сигналам точки должны располагаться на окружности радиуса

$\sqrt{E}$ в вершинах равностороннего треугольника (рис. 9.11). Если сигналов *четыре*, то оптимальным является их размещение в вершинах правильного тетраэдра, вписанного в сферу радиуса $\sqrt{E}$. В общем случае оптимальный выбор системы из n сигналов соответствует их расположению в вершинах правильного $(n-1)$-мерного *симплекса*, вписанного в $(n-1)$-мерную сферу[116].

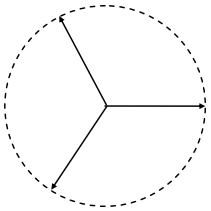


Рис. 9.11. К помехоустойчивости приема трех сигналов

9.6. НЕКОГЕРЕНТНЫЙ ПРИЕМ

На практике иногда не удается обеспечить условия для когерентного приема сигналов, так как один или несколько параметров принимаемого сигнала оказываются неизвестными. Такая ситуация типична, например, для систем спутниковой связи, радиосвязи с подвижными объектами, и т.п., поскольку расстояние между передатчиком и приемником изменяется случайным образом. Это приводит, в частности, к тому, что меняется начальная фаза несущего колебания. Если изменение происходит настолько медленно, что соседние посылки имеют практически одинаковую начальную фазу, то ее можно *оценить* и оценку использовать вместо точного значения при организации приема. Такой прием называют *квазикогерентным*. Если же начальная фаза изменяется (флюктуирует) быстро или устройство оценивания оказывается слишком сложным, тогда рассматривается задача приема сигнала со случайной начальной фазой, или *некогерентного* приема.

Перепишем выражение (9.7) для логарифма отношения правдоподобия при приеме сигнала $s(t)$:

$$\ln\Lambda = \frac{2}{N_0}\int_0^T z(t)s(t)dt - \frac{1}{N_0}\int_0^T s^2(t)dt\,. \qquad (9.11)$$

Сигнал при некогерентном приеме известен с точностью до начальной фазы, поэтому обозначим его $s(t,\phi)$ и запишем

$$s(t,\phi) = \mathrm{Re}\left\{\dot{s}(t)e^{-j\phi}\right\}.$$

[116] Отрезок, треугольник и тетраэдр являются одномерным, двумерным и трехмерным симплексами.

В этом выражении неизвестная начальная фаза сигнала представлена комплексным фазовым множителем $e^{-j\phi}$ при аналитическом комплексном сигнале $\dot{s}(t)$, который определяется выражением

$$\dot{s}(t) = s(t) + j\hat{s}(t),$$

где вещественная и мнимая части связаны парой преобразований Гильберта

$$\hat{s}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau,$$

$$s(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\hat{s}(\tau)}{\tau-t}d\tau.$$

Тогда, очевидно,

$$s(t,\phi) = \mathrm{Re}\left\{\left[s(t) + j\hat{s}(t)\right]\left[\cos\phi - j\sin\phi\right]\right\} = s(t)\cos\phi + \hat{s}(t)\sin\phi.$$

Корреляционный интеграл в выражении (9.11) в таком случае приобретает вид

$$\int_0^T z(t)s(t,\phi)dt = \int_0^T z(t)s(t)\cos\phi\,dt + \int_0^T z(t)\hat{s}(t)\sin\phi\,dt =$$

$$= \mathrm{Re}\left\{\int_0^T z(t)\left[s(t) + j\hat{s}(t)\right]\left[\cos\phi - j\sin\phi\right]dt\right\} =$$

$$= \mathrm{Re}\left\{\int_0^T z(t)\dot{s}(t)e^{-j\phi}dt\right\} = \mathrm{Re}\left\{e^{-j\phi}\int_0^T z(t)\dot{s}(t)dt\right\} =$$

$$= \mathrm{Re}\left\{e^{-j\phi}\dot{V}\right\} = \mathrm{Re}\left\{e^{-j\phi}Ve^{-j\psi}\right\}. \qquad (9.12)$$

В полученном выражении фигурирует комплексная величина $\dot{V}$, имеющая смысл корреляционного интеграла для аналитического сигнала $\dot{s}(t)$:

$$\dot{V} = \int_0^T z(t)\dot{s}(t)dt = \int_0^T z(t)s(t)dt + j\int_0^T z(t)\hat{s}(t)dt,$$

где, очевидно,

$$V = \sqrt{\left(\int_0^T z(t)s(t)dt\right)^2 + \left(\int_0^T z(t)\hat{s}(t)dt\right)^2};$$

$$\psi = -\operatorname{arctg}\frac{\int_0^T z(t)\hat{s}(t)dt}{\int_0^T z(t)s(t)dt}.$$

Корреляционный интеграл согласно (9.12) можно переписать в виде

$$\int_0^T z(t)s(t,\phi)dt = V\cos(\psi+\phi),$$

тогда логарифм отношения правдоподобия

$$\ln\Lambda = \frac{2}{N_0}V\cos(\psi+\phi) - \frac{1}{N_0}E,$$

а само отношение правдоподобия

$$\Lambda = e^{\frac{2}{N_0}V\cos(\psi+\phi)}e^{-\frac{1}{N_0}E}.$$

Считая, что начальная фаза сигнала является случайной величиной, имеющей равномерное в интервале $(0, 2\pi)$ распределение, выполним усреднение отношения правдоподобия по ансамблю:

$$\bar{\Lambda} = e^{-\frac{E}{N_0}}\frac{1}{2\pi}\int_0^{2\pi} e^{\frac{2V}{N_0}\cos(\psi+\phi)}d\phi.$$

Учтем известное соотношение

$$\frac{1}{2\pi}\int_0^{2\pi} e^{a\cos(\psi+\phi)}d\phi = I_0(a),$$

где $I_0(a)$ – модифицированная функция Бесселя нулевого порядка, тогда

$$\bar{\Lambda} = e^{-\frac{E}{N_0}}I_0\left(\frac{2V}{N_0}\right).$$

Правило некогерентного приема сигнала со случайной равновероятной начальной фазой на фоне гауссовского шума должно быть основано на сравнении величины $\bar{\Lambda}$ с некоторым порогом, а

правило различения двух сигналов – на сравнении двух отношений правдоподобия между собой. Предположим, что рассматривается прием двух сигналов $s_1(t)$ и $s_0(t)$. Сравнение усредненных отношений правдоподобия можно заменить сравнением их логарифмов

$$\ln I_0\left(\frac{2V_1}{N_0}\right)-\frac{E_1}{N_0}\underset{0}{\overset{1}{\gtrless}}\ln I_0\left(\frac{2V_0}{N_0}\right)-\frac{E_0}{N_0},$$

или сравнением с порогом разности логарифмов

$$\ln I_0\left(\frac{2V_1}{N_0}\right)-\ln I_0\left(\frac{2V_0}{N_0}\right)\underset{0}{\overset{1}{\gtrless}}\frac{E_1-E_0}{N_0}.$$

Алгоритм сильно упрощается, если энергии сигналов равны, в этом случае в силу монотонности функции I_0 можно сравнивать между собой величины V_1 и V_0:

$$V_1\underset{0}{\overset{1}{\gtrless}}V_0.$$

Структурная схема корреляционного приемника, реализующего это правило, показана на рис. 9.12. Для каждого из сигналов реализуется корреляционный прием раздельно по двум квадратурным составляющим, после чего квадраты огибающих поступают на решающее устройство РУ, выполняющее их сравнение.

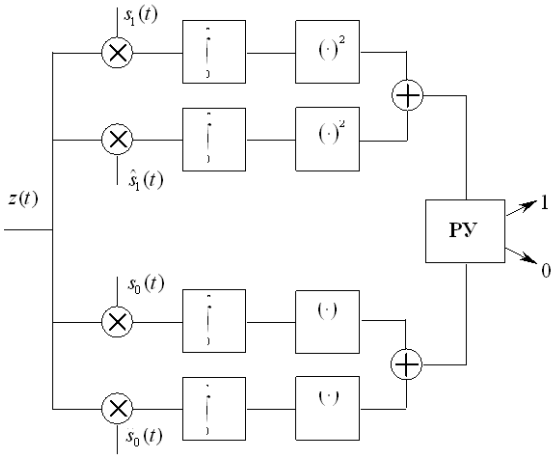


Рис. 9.12. Структура некогерентного приемника двух сигналов с равными энергиями

То же правило можно реализовать с использованием согласованных фильтров по схеме рис. 9.13. Здесь вычисление величин V_1 и V_0 производится устройством, называемым детектором огибающей ДО.

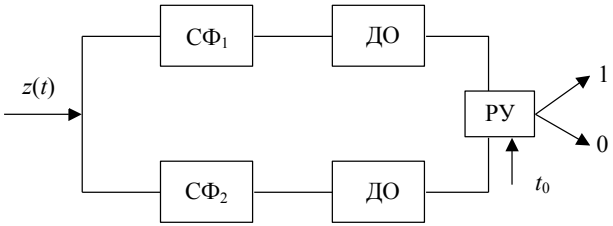


Рис. 9.13. Структура некогерентного приемника двух сигналов с использованием согласованных фильтров

9.7. ПОТЕНЦИАЛЬНАЯ ПОМЕХОУСТОЙЧИВОСТЬ НЕКОГЕРЕНТНОГО ПРИЕМА

Определим потенциальную помехоустойчивость некогерентного приема на примере системы с пассивной паузой при равных априорных вероятностях посылок

$$s_1(t) = A\cos(\omega t + \phi), \quad s_0(t) = 0; \quad p_1 = p_0 = 0{,}5.$$

Средняя вероятность ошибки равна

$$p_{\text{ош}} = 0{,}5p_{01} + 0{,}5p_{10} =$$

$$= 0{,}5\int_0^{V_\text{п}} w_1\left(V \mid H_1\right)dV + 0{,}5\int_{V_\text{п}}^{\infty} w_0\left(V \mid H_0\right)dV.$$

Здесь $w_1\left(V \mid H_1\right)$ и $w_0\left(V \mid H_0\right)$ – условные плотности распределения вероятности огибающей корреляционного интеграла при условии гипотез о передаче сигналов $s_1(t)$ и $s_0(t)$ соответственно, $V_\text{п}$ – порог (рис. 9.14).

При гипотезе H_0 значение огибающей обусловлено только шумом, тогда квадратурные составляющие являются независимыми нормальными случайными величинами с нулевыми средними и дисперсиями $N_0E/2$ [см. разд. 3.6, а также выражение (9.10)].

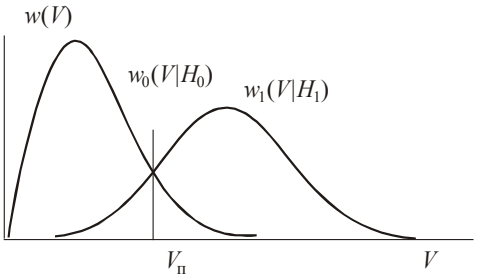


Рис. 9.14. Выбор порога при некогерентном приеме

Условная плотность распределения вероятностей огибающей имеет рэлеевский вид

$$w_0(V \mid H_0) = \frac{2V}{EN_0} e^{-\frac{V^2}{EN_0}}.$$

Если наблюдаемое колебание содержит сигнал $s_1(t)$, то огибающая имеет обобщенное рэлеевское распределение (распределение Рэлея – Райса)

$$w_1(V \mid H_1) = \frac{2V}{EN_0} e^{-\frac{V^2+E^2}{EN_0}} I_0\left(\frac{2V}{N_0}\right).$$

Средняя вероятность ошибки равна

$$p_{\text{ош}} = \frac{1}{2}\int_0^{V_\text{п}} \frac{2V}{EN_0} e^{-\frac{V^2+E^2}{EN_0}} I_0\left(\frac{2V}{N_0}\right) dV + \frac{1}{2}\int_{V_\text{п}}^{\infty} \frac{2V}{EN_0} e^{-\frac{V^2}{EN_0}} dV. \qquad (9.13)$$

Второй интеграл берется по частям, при этом

$$p_{\text{ош}} = \frac{1}{2}\int_0^{V_\text{п}} \frac{2V}{EN_0} e^{-\frac{V^2+E^2}{EN_0}} I_0\left(\frac{2V}{N_0}\right) dV + \frac{1}{2} e^{-\frac{V_\text{п}^2}{EN_0}}.$$

Оптимальное значение порога, при котором достигается потенциальная помехоустойчивость некогерентного приема, является решением уравнения $dp_{\text{ош}} / dV_\text{п} = 0$.

Взяв производную и приравняв ее нулю, получим

$$\frac{2V_{\text{п}}}{EN_0}e^{-\frac{V_{\text{п}}^2+E^2}{EN_0}}I_0\left(\frac{2V_{\text{п}}}{N_0}\right)-\frac{2V_{\text{п}}}{EN_0}e^{-\frac{V_{\text{п}}^2}{EN_0}}=0$$

или

$$I_0\left(\frac{2V_{\text{п}}}{N_0}\right)=e^{\frac{E}{N_0}}.$$

Точно решить полученное уравнение не удается. Прологарифмируем обе части выражения:

$$\ln I_0\left(\frac{2V_{\text{п}}}{N_0}\right)=\frac{E}{N_0}.$$

Известно, что

$$\ln I_0(x)\approx\begin{cases}x, x\gg 1,\\ x^2/4, x\ll 1.\end{cases}$$

Поэтому оптимальный порог определяется приближенными выражениями

$$V_{\text{п opt}}=\begin{cases}E/2 & \text{при больших отношениях сигнал/шум,}\\ \sqrt{EN_0} & \text{при малых отношениях сигнал/шум.}\end{cases}$$

Подставляя в (9.13) порог $E/2$, получим среднюю вероятность ошибки при больших отношениях сигнал/шум (ОСШ):

$$p_{\text{ош}}=\frac{1}{2}\int_0^{E/2}\frac{2V}{EN_0}e^{-\frac{V^2+E^2}{EN_0}}I_0\left(\frac{2V}{N_0}\right)dV+\frac{1}{2}e^{-\frac{E}{4N_0}}.$$

При больших ОСШ ($E/N_0\geq 10$) первым слагаемым можно пренебречь, тогда

$$p_{\text{ош}}=\frac{1}{2}e^{-\frac{E}{4N_0}}.$$

Аналогично можно проанализировать помехоустойчивость приема двух ортогональных частотно-манипулированных сигналов; для этого случая средняя вероятность ошибки

$$p_{\text{ош}}=\frac{1}{2}e^{-\frac{E}{2N_0}}.$$

Сигналы с фазовой манипуляцией при случайной начальной фазе каждой посылки, очевидно, применять при некогерентном приеме нельзя. Однако при медленных изменениях фазы можно использовать *относительную фазовую манипуляцию*, при которой начальная фаза следующей посылки совпадает с начальной фазой предыдущей посылки при передаче символа «0» и отличается от нее на 180° – при передаче символа «1». При этом средняя вероятность ошибки [10]

$$p_{\text{ош}} = \frac{1}{2} e^{-\frac{E}{N_0}}.$$

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое потенциальная помехоустойчивость?
2. Как формулируется задача синтеза оптимального демодулятора?
3. В чем состоит сущность критерия Байеса?
4. Что такое средний риск?
5. Что такое гипотеза?
6. Что такое отношение правдоподобия?
7. Запишите правила проверки гипотез на основе логарифма отношения правдоподобия для критериев Байеса, идеального наблюдателя и максимального правдоподобия.
8. Чем отличается когерентный прием от некогерентного?
9. Что такое согласованный фильтр?
10. Какую форму имеет сигнал на выходе согласованного фильтра, когда на его вход воздействует «свой» сигнал? «чужой» сигнал? шум?
11. Что удобнее применять на практике – коррелятор или согласованный фильтр?
12. Можно ли реализовать согласованный фильтр для сигнала произвольной формы с любой заданной точностью?
13. Как следует выбирать совокупность сигналов одинаковой энергии для обеспечения максимальной помехоустойчивости?

УПРАЖНЕНИЯ

1. Для когерентного приема сигнала в системе амплитудной телеграфии с пассивной паузой методом однократного отсчета выбран порог, равный 2 В. Известно, что порог оптимален с точки зрения критерия максимального правдоподобия, в то же время ап-

риорные вероятности символов равны $p_1 = 0.6$ и $p_0 = 0.4$. Насколько изменится средняя вероятность ошибки при выборе порога по критерию идеального наблюдателя, если дисперсия шума равна 9 $В^2$?

2. В системе амплитудной телеграфии с пассивной паузой используется согласованный фильтр. Определите изменение средней вероятности ошибки по сравнению с методом однократного отсчета, если амплитуда радиоимпульса равна $a = 0{,}5$ В, длительность $\tau = 1$ мкс, среднеквадратическое отклонение шума 0.3 В.

3. Определите, насколько изменится средняя вероятность ошибки при переходе от когерентного приема к некогерентному, если в системе амплитудной телеграфии с пассивной паузой амплитуда радиоимпульса равна $a = 2$ В, длительность $\tau = 10$ мкс, среднеквадратическое отклонение шума 0,5 В.

10. ОСНОВЫ ТЕОРИИ ПОМЕХОУСТОЙЧИВОСТИ ПЕРЕДАЧИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

10.1. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ

Непрерывные сообщения (например, речь, музыка и т.п.) могут передаваться по каналу связи непосредственно (например, по местной проводной радиосети, по телефонному кабелю) или при помощи модуляции. В первом случае сигнал $s(t)$, передаваемый по каналу, может совпадать с сообщением (первичным сигналом) $b(t)$ или быть связан с ним простой пропорциональной зависимостью, во втором – передаваемый сигнал $s\left[t, b(t)\right]$ является функцией сообщения, в общем случае нелинейной (рис. 10.1).

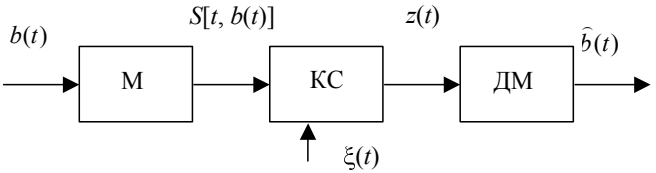


Рис. 10.1. Структура системы передачи непрерывных сообщений

Колебание на входе демодулятора

$$z(t) = s[t, b(t)] + \xi(t)$$

представляет собой в простейшем случае сумму передаваемого сигнала и шума $\xi(t)$. Задача демодулятора состоит в нахождении

такого *выходного* первичного сигнала $\hat{b}(t)$, который был бы *близок* к передаваемому сообщению $b(t)$. Для строгой постановки задачи необходимо указать количественную меру (критерий) близости двух указанных сигналов, тогда демодулятор должен найти *оценку* первичного сигнала (сообщения), наилучшую в смысле выбранного критерия близости. В качестве критерия часто используют средний квадрат ошибки

$$\overline{\varepsilon^2} = \overline{\left[\hat{b}(t) - b(t)\right]^2}, \qquad (10.1)$$

где черта означает статистическое усреднение по ансамблю. В системах телеметрии используется критерий максимальной ошибки

$$\delta = \left|\hat{b}(t) - b(t)\right|_{\max},$$

в радиовещании – увеличение выходного отношения сигнал/шум по сравнению с входным, критерий разборчивости речевых сообщений[117] и т.п.

10.2. ОПТИМАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ СИГНАЛА

Оценивание сигнала, как функции времени, – достаточно сложная задача. Во многих случаях ее можно свести к более простой задаче оценивания одного или нескольких параметров сигнала.

Простейшей задачей, связанной с оцениванием параметров сигнала, является оценка параметра, постоянного или настолько медленно меняющегося во времени, что на интервале наблюдения его можно считать постоянным. Такие задачи встречаются в системах телеуправления и телеметрии, когда сообщение представляет собой *значение* управляющего сигнала или *результат измерения* некоторой физической величины. Рассмотрим задачу оценивания единственного скалярного параметра λ, который до опыта рассматривается как случайная величина, имеющая *априорное* распределение с плотностью $w(\lambda)$. В конкретном опыте реализация

[117] Разборчивость речи крайне трудно оценить количественно; обычно применяют качественную оценку, определяемую группой слушателей-экспертов (метод экспертных оценок).

этой случайной величины представляет собой значение, постоянное на интервале $(0, T)$ наблюдения колебания

$$z(t) = s(t, \lambda) + \xi(t).$$

Правило оценивания[118] – это алгоритм обработки наблюдаемого колебания, результатом выполнения которого является значение $\tilde{\lambda}$ оценки параметра λ. Для оценивания одного и того же параметра может существовать множество алгоритмов, вырабатывающих различные оценки. Для сравнения алгоритмов оценивания между собой и выбора наилучшего используют следующие показатели.

1. Несмещенность

Оценка называется *несмещенной*, если выполняется условие $\overline{\tilde{\lambda} - \lambda} = 0$, означающее, что при любом значении параметра условное математическое ожидание оценки равно этому значению $\overline{\tilde{\lambda}} = \lambda$. Другими словами, несмещенность означает отсутствие *систематической* ошибки оценивания. В противном случае оценка называется смещенной. Следует отметить, что смещенные оценки также находят применение, если смещение достаточно мало или стремится к нулю при увеличении времени наблюдения или мощности сигнала.

2. Состоятельность

Оценка называется *состоятельной*, если при неограниченном возрастании времени наблюдения оценка *сходится по вероятности* к значению параметра:

$$\lim_{T \to \infty} \mathbf{P}\left\{\left|\tilde{\lambda} - \lambda\right| \geq \Delta\right\} = 0 \quad \text{при любом} \quad \Delta > 0,$$

(здесь $\mathbf{P}\{A\}$ обозначает вероятность события A).

Смещенная оценка может быть состоятельной, если ее смещение стремится к нулю при $T \to \infty$. Для состоятельной оценки, очевидно, *дисперсия* ошибки стремится к нулю $\lim_{T \to \infty} \overline{\left|\tilde{\lambda} - \lambda\right|^2} = 0$.

3. Эффективность

Несмещенная оценка называется *эффективной*, если среди всех оценок, полученных при заданном времени наблюдения *всевозможными алгоритмами* оценивания, она обеспечивает наименьшую дисперсию ошибки

$$\overline{\left|\tilde{\lambda} - \lambda\right|^2} = \min.$$

[118] Часто в литературе *правило оценивания* также называют *оценкой*.

Эффективность представляет собой очень сильное свойство, и во многих случаях эффективную оценку не удается найти или доказано, что она не существует[119].

Классический подход к оцениванию параметров сигналов основывается на формуле Байеса для *апостериорной* плотности распределения вероятностей оцениваемого параметра [18]

$$w(\lambda \mid z) = \frac{w(\lambda) w(z \mid \lambda)}{w(z)}, \qquad (10.2)$$

где $w(\lambda)$ – априорная ПРВ параметра λ; $w(z \mid \lambda)$ – условная ПРВ наблюдаемого процесса при заданном значении λ, рассматриваемая как функция от λ при данном z (*функция правдоподобия*); $w(z)$ – при фиксированной реализации z постоянная величина. Выражение (10.2) показывает, что, зная априорную плотность $w(\lambda)$ и наблюдая реализацию процесса z, можно получить уточненное представление о значении параметра λ. На рис. 10.2 показаны примеры априорной и апостериорной ПРВ параметра λ (истинное значение параметра обозначено λ_0).

Влияние функции правдоподобия на апостериорное распределение выражается в его *обострении* по сравнению с априорным распределением, что естественно, так как, наблюдая реализацию z, мы получаем дополнительную информацию о параметре, что уменьшает исходную *неопределенность*, заключенную в априорной ПРВ.

Апостериорное распределение содержит *всю информацию о параметре, которую можно получить из наблюдаемой реализации*

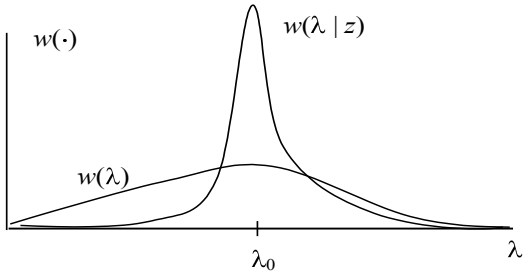


Рис. 10.2. Априорная и апостериорная ПРВ оцениваемого параметра

[119] Это означает, что не существует правила оценивания, которое являлось бы наилучшим *среди всех возможных* правил.

и априорных данных. Поэтому правило оценивания должно использовать апостериорную ПРВ, а способ ее использования зависит от выбранного критерия качества оценки.

Ошибки оценивания параметра в общем случае приводят к различным последствиям, поэтому естественным способом их учета является введение *функции потерь* (штрафной функции) $L\left(\tilde{\lambda}-\lambda\right)$, зависящей от разности оценки и истинного значения параметра. Усредняя функцию потерь по апостериорному распределению параметра, получаем количественную характеристику, называемую *апостериорным (условным) риском*

$$r(\tilde{\lambda}, z)=\int\limits_{(\lambda)} L\left(\tilde{\lambda}-\lambda\right) w(\lambda \mid z) d\lambda, \qquad (10.3)$$

описывающим потери, связанные с получением оценки $\tilde{\lambda}$ при наблюдении реализации z. Усреднение апостериорного риска (10.3) по всевозможным реализациям приводит к *среднему риску*

$$R(\tilde{\lambda})=\int\limits_{(z)} w(z)\left[\int\limits_{(\lambda)} L\left(\tilde{\lambda}-\lambda\right) w(\lambda \mid z) d\lambda\right] dz .$$

Правило оценивания, которому соответствует наименьший средний риск, называется байесовским, а соответствующая оценка – байесовской, или оценкой по критерию *минимума среднего риска*. Правило, оптимальное в смысле минимума среднего риска, находится из условия минимизации условного риска (10.3).

Часто используют квадратичную функцию потерь

$$L\left(\tilde{\lambda}-\lambda\right)=\left(\tilde{\lambda}-\lambda\right)^2,$$

тогда

$$r(\tilde{\lambda}, z)=\int\limits_{(\lambda)}\left(\tilde{\lambda}-\lambda\right)^2 w(\lambda \mid z) d\lambda=\overline{\left(\tilde{\lambda}-\lambda\right)^2}, \qquad (10.4)$$

т. е. апостериорный риск равен среднему квадрату ошибки (а если оценка несмещенная, то дисперсии ошибки). Байесовская оценка в этом случае становится оценкой *минимума среднеквадратической ошибки*. Для нахождения правила раскроем скобки в выражении (10.4):

$$\overline{\left(\tilde{\lambda}-\lambda\right)^2}=\tilde{\lambda}^2-2\tilde{\lambda}\int\limits_{(\lambda)} \lambda w(\lambda \mid z) d\lambda+\int\limits_{(\lambda)} \lambda^2 w(\lambda \mid z) d\lambda .$$

Дифференцируя полученное выражение по λ и приравнивая результат нулю, получаем правило

$$\tilde{\lambda} = \int\limits_{(\lambda)} \lambda w(\lambda \mid z) d\lambda .$$

Таким образом, оценка, оптимальная в смысле минимума среднеквадратической ошибки, равна *апостериорному среднему* значению параметра.

Кроме квадратичной, на практике часто используется *простая* функция потерь

$$L\left(\tilde{\lambda} - \lambda\right) = \text{const} - \delta\left(\tilde{\lambda} - \lambda\right). \qquad (10.5)$$

Подставляя (10.5) в (10.4), получаем

$$\int\limits_{(\lambda)} \left[\text{const} - \delta\left(\tilde{\lambda} - \lambda\right)\right] w(\lambda \mid z) d\lambda = \text{const} - w(\lambda \mid z)\big|_{\lambda=\tilde{\lambda}} .$$

Очевидно, это выражение достигает минимума, если в качестве оценки $\tilde{\lambda}$ использовать значение параметра, доставляющее *максимум апостериорной ПРВ* $w(\lambda \mid z)$. Такая оценка называется МАВ-оценкой (оценкой максимума апостериорной вероятности).

Во многих задачах априорная ПРВ параметра неизвестна, тогда принимают ее равной константе и максимизируют функцию правдоподобия $w(z \mid \lambda)$. Получаемые таким образом оценки называются оценками *максимального правдоподобия*, или МП-оценками.

Пример 10.1. Пусть наблюдается колебание

$$z(t) = \gamma s(t) + \xi(t) ,$$

где $s(t)$ – точно известный сигнал; γ – амплитудный множитель, подлежащий оцениванию; $\xi(t)$ – гауссовский шум с нулевым средним и спектральной плотностью мощности $N_0/2$, постоянной в полосе частот $-F < f < F$ («квазибелый» шум). Найдем правило оценивания параметра γ, оптимальное по критерию максимального правдоподобия.

Как в разд. 9.3, возьмем n отсчетов наблюдаемого колебания на интервале наблюдения T с шагом $\Delta t = \dfrac{1}{2F} = \dfrac{T}{n}$, при этом отсче-

ты шума являются некоррелированными. Совместная плотность распределения вероятности взятых отсчетов поэтому равна

$$w(z_1,...,z_n \mid \gamma) = \frac{1}{\left(\sqrt{2\pi}\cdot\sigma\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(z_k-\gamma s_k)^2},$$

где $\sigma^2 = N_0 F = N_0/(2\Delta t)$. Устремляя Δt к нулю ($n \to \infty$), запишем функцию правдоподобия

$$w(z \mid \gamma) = C \exp\left\{-\frac{1}{N_0}\int_0^T \left[z(t)-\gamma s(t)\right]^2 dt\right\},$$

где C – константа, несущественная для задачи оценивания.

Для нахождения правила оценивания следует продифференцировать функцию правдоподобия или, что проще, ее логарифм и приравнять результат нулю. Получаемое при этом *уравнение правдоподобия*

$$\frac{d[\ln w(z \mid \gamma)]}{d\gamma} = 0$$

для данного случая имеет вид

$$\int_0^T \left[z(t)-\gamma s(t)\right] s(t)dt = 0,$$

откуда

$$\gamma \int_0^T s^2(t)dt = \int_0^T z(t)s(t)dt.$$

Решением этого уравнения относительно γ является оценка $\tilde{\gamma}$, определяемая выражением

$$\tilde{\gamma} = \frac{1}{E}\int_0^T z(t)s(t)dt, \qquad (10.6)$$

где $E = \int_0^T s^2(t)dt$ – энергия сигнала, известная по условию задачи.

Качество полученной МП-оценки можно оценить, подставив в (10.6) выражение для $z(t)$:

$$\tilde{\gamma} = \frac{\int_0^T [\gamma s(t)+\xi(t)]s(t)dt}{E} = \frac{\gamma E}{E} + \frac{\int_0^T \xi(t)s(t)dt}{E} = \gamma + \frac{1}{E}\int_0^T \xi(t)s(t)dt. \qquad (10.7)$$

Второе слагаемое представляет собой ошибку оценивания, причем дисперсия интеграла равна $N_0E/2$ [см. разд. 9.5, выражение (9.10)], поэтому дисперсия ошибки равна $N_0/(2E)$. Таким образом, оценка тем точнее, чем больше энергия сигнала (для гармонического сигнала $s(t)$ увеличение энергии эквивалентно увеличению длительности интервала наблюдения) и чем меньше спектральная плотность мощности. Из выражения (10.7) видно, что оценка *несмещенная*, так как $\xi(t)$ имеет нулевое математическое ожидание. Учитывая несмещенность и стремление дисперсии к нулю при увеличении интервала наблюдения, можно заключить, что оценка является *состоятельной*. Кроме того, можно показать, что оценка также эффективна. ◄

Полученный алгоритм оценивания может быть реализован в виде структурной схемы, показанной на рис. 10.3.

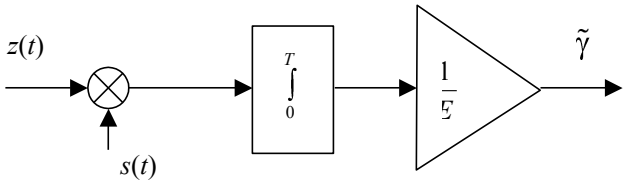


Рис. 10.3. Структура устройства оценивания амплитуды сигнала

Полученное правило оценивания амплитуды сигнала можно использовать и при медленном изменении этого параметра; вместо интегратора можно применить фильтр нижних частот (ФНЧ), и при гармоническом сигнале схема рис. 10.3 превращается в схему синхронного детектора амплитудно-модулированных колебаний (рис. 10.4). Масштабирующее звено (усилитель) для задачи детектирования не обязательно, поэтому оно показано пунктиром.

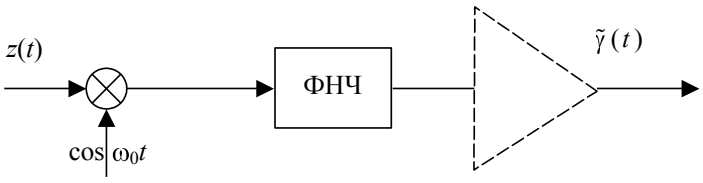


Рис. 10.4. Синхронный детектор АМ-колебаний

10.3. ОПТИМАЛЬНАЯ ФИЛЬТРАЦИЯ СЛУЧАЙНОГО СИГНАЛА

Более сложной и общей, чем задача оценивания постоянного параметра, является задача оценивания *изменяющегося сообщения* (первичного сигнала) на основе наблюдаемой реализации. Такое оценивание принято называть *фильтрацией*. Сообщение рассматривается как реализация случайного процесса, множество всевозможных сообщений – как ансамбль реализаций с некоторым вероятностным распределением. Сообщение (первичный сигнал) модулирует несущее колебание, поэтому сигнал на выходе канала связи также случаен. Таким образом, ставится задача по наблюдаемому случайному колебанию оценить другое случайное колебание (первичный сигнал, или закон модуляции), связанное с наблюдаемым в общем случае нелинейным образом (задача *нелинейной фильтрации, или демодуляции*). Эта задача может быть весьма сложной.

В этом подразделе рассматривается наиболее простой случай оптимальной *линейной* фильтрации. При этом с самого начала предполагается, что фильтр представляет собой ЛИС-цепь, и задача состоит в подборе такого ЛИС-фильтра, который при подаче на вход наблюдаемой реализации обеспечивает выходной сигнал, наилучшим образом соответствующий выбранному критерию.

На практике линейная фильтрация может применяться, например, для повышения отношения сигнал/шум на входе демодулятора Д (рис. 10.5).

Предположим, что модулированный сигнал с выхода модулятора М, представляющий собой стационарный случайный процесс $s(t)$ со спектральной плотностью мощности $G_s(\omega)$, суммируется в канале связи КС со стационарным шумом $\xi(t)$, имеющим спектральную плотность мощности $G_\xi(\omega)$, причем оба процесса имеют нулевые средние. Задача состоит в том, чтобы найти характеристики линейной стационарной цепи (оптимального фильтра ОФ), такой, чтобы процесс $\tilde{s}(t)$ на ее выходе был наиболее близок к процессу

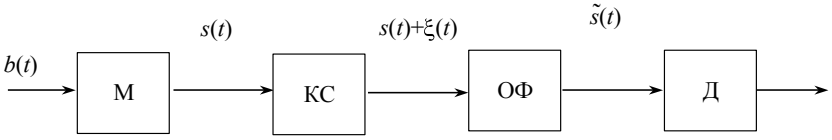


Рис. 10.5. К задаче оптимальной линейной фильтрации

$s(t)$ при условии, что на вход воздействует смесь $z(t) = s(t) + \xi(t)$. Примем за критерий близости дисперсию разности

$$\varepsilon(t) = s(t) - \tilde{s}(t), \tag{10.8}$$

которая представляет собой ошибку фильтрации.

Поскольку фильтр линейный стационарный, его отклик на смесь $z(t)$ представляется сверткой

$$\tilde{s}(t) = \int\limits_{(\tau)} z(\tau)h(t-\tau)d\tau = \int\limits_{(\tau)} s(\tau)h(t-\tau)d\tau + \int\limits_{(\tau)} \xi(\tau)h(t-\tau)d\tau .$$

Здесь (τ) обозначает множество всех допустимых значений переменной τ. Обозначим импульсную характеристику оптимального фильтра, которую предстоит найти, $h_0(t)$.

Поскольку и сигнал, и шум имеют нулевые средние, а фильтр линеен, ошибка (10.8) также имеет нулевое математическое ожидание, а ее средний квадрат совпадает с дисперсией.

Средний квадрат ошибки для оптимального фильтра

$$e = \overline{\varepsilon^2(t)} = \overline{[s(t) - \tilde{s}(t)]^2}$$

представляет собой *минимальное* значение, достижимое при фильтрации любым линейным устройством. Для *произвольного* линейного фильтра импульсную характеристику можно представить в виде $h_0(t) + \alpha h_\delta(t)$, где α и $h_\delta(t)$ – некоторые, пока неопределенные, константа и функция. Тогда средний квадрат ошибки для произвольного фильтра

$$e_\alpha = \overline{[s(t) - \int\limits_{(\tau)} \{h_0(\tau) + \alpha h_\delta(\tau)\} z(t-\tau)d\tau]^2} . \tag{10.9}$$

Поскольку при $\alpha = 0$ произвольный фильтр превращается в оптимальный, достигается минимум среднего квадрата ошибки (10.9), тогда можно записать уравнение

$$\left.\frac{\partial e_\alpha}{\partial \alpha}\right|_{\alpha=0} = 0 ,$$

решением которого и будет искомая импульсная характеристика оптимального фильтра. Дифференцируя (10.9) по α и приравнивая результат нулю (при $\alpha = 0$), получаем уравнение

$$\overline{\left[s(t) - \int\limits_{(\tau)} h_0(\tau)z(t-\tau)d\tau\right]\left[\int\limits_{(\theta)} h_\delta(\theta)z(t-\theta)d\theta\right]} = 0 . \tag{10.10}$$

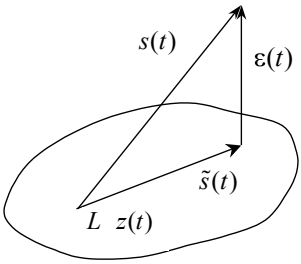


Рис. 10.6. Принцип ортогональности

Выражение в первых квадратных скобках представляет собой ошибку оптимального фильтра (10.8), во вторых квадратных скобках заключен отклик на наблюдаемый процесс линейного фильтра с *произвольной* импульсной характеристикой $h_\delta(t)$. В силу того, что $h_\delta(t)$ – произвольная функция, выражение во вторых квадратных скобках формулы (10.10) есть отклик произвольной ЛИС-цепи на наблюдаемый процесс. Все такие отклики при всех мыслимых $h_\delta(t)$ являются линейными комбинациями всех мгновенных отсчетов процесса $z(t)$. Другими словами, все такие отклики в совокупности составляют линейное пространство $L\{z(t)\}$, натянутое на все мгновенные отсчеты процесса $z(t)$, как на базис (отсчеты представляют собой случайные величины, которые можно рассматривать как векторы; при нулевых средних ортогональность в этом пространстве равнозначна некоррелированности случайных величин). Оцениваемый сигнал $s(t)$ в общем случае этому пространству не принадлежит, рис. 10.6.

Выражение (10.10) составляет математическую запись *принципа ортогональности* [20], смысл которого состоит в том, что для оптимального фильтра ошибка фильтрации должна быть некоррелированна с наблюдаемым процессом во все моменты времени.

Раскрывая скобки в (10.10), получаем

$$\int\limits_{(\theta)} h_\delta(\theta)\overline{s(t)z(t-\theta)}d\theta - \int\limits_{(\theta)} h_\delta(\theta)\int\limits_{(\tau)} h_o(\tau)\overline{z(t-\tau)z(t-\theta)}d\tau d\theta = 0,$$

$$\int\limits_{(\theta)} h_\delta(\theta)\left\{\overline{s(t)z(t-\theta)} - \int\limits_{(\tau)} h_0(\tau)\overline{z(t-\tau)z(t-\theta)}d\tau\right\}d\theta = 0,$$

откуда снова в силу произвольности $h_\delta(t)$ следует, что выражение в фигурных скобках должно быть равно нулю при всех θ. Учитывая, что результатами усреднения являются взаимно корреляционная $R_{sz}(\cdot)$ и автокорреляционная $R_z(\cdot)$ функции, запишем *уравнение Винера – Хопфа*

$$R_{sz}(t) = \int\limits_{(\tau)} h_0(\tau)R_z(t-\tau)d\tau \qquad (10.11)$$

относительно импульсной характеристики оптимального фильтра $h_0(t)$.

Решение уравнения Винера – Хопфа легко находится для случая, когда все процессы рассматриваются на бесконечной временной оси и являются стационарными (в широком смысле). Тогда к левой и правой частям уравнения (10.11) можно применить преобразование Фурье, в результате чего получается уравнение

$$G_{sz}(\omega) = G_z(\omega)H(\omega),$$

где $G_z(\omega)$ – спектральная плотность мощности наблюдаемого процесса; $G_{sz}(\omega)$ – взаимная СПМ полезного сигнала и наблюдаемого процесса; $H(\omega)$ – комплексная частотная характеристика оптимального фильтра, которая находится как

$$H(\omega) = G_{sz}(\omega) / G_z(\omega).$$

Если сигнал и шум некоррелированны, то их взаимная СПМ $G_{s\xi}(\omega)$ равна нулю, тогда $G_{sz}(\omega) = G_s(\omega) + G_{s\xi}(\omega) = G_s(\omega)$, $G_z(\omega) = G_s(\omega) + G_\xi(\omega)$ и

$$H(\omega) = \frac{G_s(\omega)}{G_s(\omega) + G_\xi(\omega)}.$$

Полученный фильтр известен как *винеровский фильтр* (*фильтр Колмогорова – Винера*)[120]. Видно, что коэффициент передачи фильтра меньше на тех частотах, где больше СПМ шума, и в этом состоит сходство фильтра Колмогорова – Винера с согласованным фильтром. В заключение отметим, что в случае, когда полезный сигнал и шум являются *совместно гауссовскими* процессами, винеровский фильтр является оптимальным среди *всех* фильтров (а не только среди линейных). Следует также иметь в виду, что полученный фильтр некаузален (физически нереализуем). Условие каузальности усложняет нахождение характеристик винеровского фильтра и увеличивает дисперсию ошибки фильтрации [20].

---

[120] Задача построения линейного фильтра, оптимального по критерию минимума дисперсии ошибки, была решена А.Н. Колмогоровым в 1939 г. для стационарных случайных процессов с дискретным временем; в 1942 г. американский математик Н. Винер решил эту же задачу в непрерывном времени.

10.4. ЦИФРОВАЯ ПЕРЕДАЧА НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Непрерывные сообщения, представленные сигналами с непрерывным временем (континуальными, или аналоговыми), можно передавать по *цифровым* каналам связи. Теоретическим основанием для этого служит теорема отсчетов (Котельникова), которая утверждает, что континуальный сигнал с финитным (ограниченным по ширине) спектром можно без потери информации представить последовательностью его отсчетов, взятых с достаточно малым шагом, определяемым верхней частотой спектра сигнала. Отсчеты сигнала, т.е. числа, можно закодировать и передать при помощи последовательности кодовых символов по цифровому каналу связи.

Цифровые сигналы имеют перед аналоговыми ряд общеизвестных преимуществ. Одно из них заключается в *большей помехоустойчивости* цифровых сигналов. В самом деле, континуальный сигнал, искаженный сколь угодно малым шумом, уже невозможно восстановить точно. Причина этого заключается в том, что комбинация (например, сумма) аналогового сигнала и аналогового шума ничем принципиально не отличается от исходного аналогового сигнала. Поскольку ни форма сигнала, ни форма помехи заранее не известны, разделить их в общем случае практически невозможно[121]. Таким образом, поражение аналогового сигнала шумом необратимо. Цифровой сигнал, который по определению может принимать значения только из дискретного множества, может быть искажен шумом только в том случае, если шум имеет достаточно большую интенсивность, чтобы перевести сигнал с одного допустимого уровня на другой.

Второе преимущество цифровых сигналов заключается в *возможности помехоустойчивого кодирования* отсчетов, что позволяет повысить помехоустойчивость передачи цифровых сигналов. Третье преимущество состоит в возможности использования для обработки цифровых сигналов универсальных цифровых вычислителей (процессоров), позволяющих реализовать *практически любые* алгоритмы. Учитывая широкое распространение средств цифровой обработки информации, можно утверждать, что значение цифровой передачи будет в обозримом будущем возрастать.

---

[121] Разделение, в частности, возможно, если сигнал и помеха строго ортогональны.

10.5. ИМПУЛЬСНО-КОДОВАЯ МОДУЛЯЦИЯ

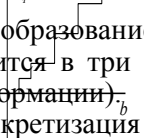


Рис. 10.7. Неравномерное квантование

Преобразование аналогового сигнала в цифровой производится в три этапа и сопровождается искажениями (потерями информации).

1. Дискретизация состоит в замене аналогового сигнала дискретным, т. е. последовательностью его значений, измеренных с бесконечной точностью в моменты времени, следующие строго периодически. В реальных устройствах дискретизации (устройствах выборки – хранения, УВХ) происходит случайное смещение моментов взятия отсчетов (*джиттер*)[122], а также искажение сигнала за счет конечного времени запоминания его уровня (подробности см., например, в [19]).

2. Квантование заключается в замене точного значения отсчета его приближенным значением, имеющим конечную разрядность. Эта операция является неизбежной, так как любое реальное устройство цифровой обработки сигналов имеет *конечную разрядность*. Квантование выполняется путем округления бесконечной дроби, представляющей вещественное число, или ее усечения до заданного числа разрядов. Размер шага квантования определяется номером младшего разряда цифрового устройства. Если сигнал является случайным процессом достаточно большой интенсивности, то квантование эквивалентно сложению сигнала с *шумом квантования*.

3. Кодирование состоит в представлении полученной квантованной величины в виде некоторой кодовой комбинации. Чаще всего используется двоичный код, соответствующий обычному представлению полученного значения в двоичной системе счисления. Полученный двоичный код можно непосредственно передавать по двоичному каналу связи; описанное преобразование аналогового сигнала в цифровой в теории связи называют *импульсно-кодовой модуляцией* (ИКМ, иначе КИМ – кодоимпульсная модуляция). На практике операции квантования и кодирования предстают в неразрывной связи и осуществляются в одном устройстве, называемом *аналого-цифровым преобразователем* (АЦП).

Обратное преобразование цифрового сигнала в аналоговый производится в устройстве, называемом цифроаналоговым преоб-

---

[122] Заметим, что в литературе по связи джиттером также называют пиковое дрожание переходов данных [30] – смещение моментов пересечения сигналом на выходе канала заданного порогового уровня вследствие нарушений синхронизации.

разователем (ЦАП) и выполняющем декодирование (преобразование кода в квантованный уровень напряжения) и сглаживание полученного ступенчатого сигнала при помощи фильтра нижних частот.

Аналоговый сигнал, *восстановленный* из цифрового после передачи его по цифровому каналу, отличается от *передаваемого* аналогового сигнала, во-первых, вследствие квантования (шум квантования) и, во-вторых, вследствие ошибок, которые случайным образом искажают отдельные символы кодовых комбинаций при передаче по каналу, подверженному действию случайных помех (*шум ложных импульсов*).

Шум квантования. Обозначим буквой Δ наименьшее значение сигнала, представимое двоичным кодом заданной разрядности. Если квантование производится путем округления исходного значения до ближайшего возможного при заданной разрядности, то можно считать, что к исходному значению прибавляется случайная величина, имеющая равномерное распределение в интервале от $-\Delta/2$ до $\Delta/2$. Дисперсия этой случайной величины равна, как нетрудно видеть, $\Delta^2/12$. Если сигнал близок по своим характеристикам к белому шуму, то и шум квантования будет близок к белому шуму, некоррелированному с сигналом. Качество квантования обычно оценивают отношением сигнал/шум квантования, которое увеличивается примерно на 6 децибел при увеличении разрядности цифрового кода на 1. Следует иметь в виду, что увеличение разрядности влечет не только повышение требований к быстродействию устройств цифровой обработки сигналов, но и расширение полосы частот, требуемой для передачи сигнала (так как при прочих равных условиях увеличение разрядности двоичного числа, передаваемого с помощью ИКМ за тот же промежуток времени, требует уменьшения длительности посылки, т.е. расширения ее спектра).

На практике часто применяют *неравномерное* квантование, при котором шаг квантования зависит от уровня квантуемого сигнала: чем больше (по модулю) значение сигнала, тем больше шаг квантования (рис. 10.7). Таким образом, более слабые участки сигнала квантуются более подробно. Основанием для неравномерного квантования служит стремление поддерживать на постоянном (или почти постоянном) уровне относительную погрешность квантования. Практически такое неравномерное квантование можно осуществить, например, при помощи схемы *компандирования* (рис. 10.8). Компандер представляет собой каскадное соединение трех узлов: *компрессора*, равномерного квантователя и *экспандера*. Ампли-

тудные характеристики компрессора и экспандера являются взаимно обратными, поэтому результирующая характеристика указанных трех узлов имеет вид, показанный на рис. 10.7.

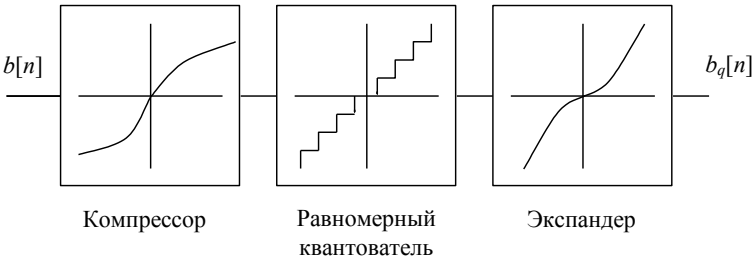


Рис.10.8. Принцип компандирования

Одним из желательных свойств неравномерного квантования является постоянство относительной погрешности, вызванной квантованием. Очевидно, для этого требуется, чтобы характеристика компрессора была логарифмической (а экспандера – соответственно экспоненциальной). Однако логарифмическая характеристика при малых значениях сигнала стремится к $-\infty$, что нежелательно и нереализуемо. Поэтому на практике используют составную характеристику, которая совпадает с логарифмической при больших значениях сигнала и является линейной при малых. Одна из таких аппроксимаций, применяемая в США (μ-закон [31]), имеет вид

$$y = y_{\max} \frac{\ln\left[1+\mu\left(|x|/x_{\max}\right)\right]}{\ln(1+\mu)} \operatorname{sgn} x\,, \qquad (10.12)$$

где μ – положительная константа, x и y – напряжения на входе и выходе компрессора, $x_{\max}$ и $y_{\max}$ – их максимальные (амплитудные) значения, а функция $\operatorname{sgn}(\cdot)$ определяется выражением

$$\operatorname{sgn} x = \begin{cases} 1, & x \ge 0 \\ -1, & x < 0 \end{cases}.$$

Другая аппроксимация (A-закон), применяемая в Европе, имеет вид

$$y = \begin{cases} y_{\max} \dfrac{A\left(|x|/x_{\max}\right)}{1+\ln A} \operatorname{sgn} x, & 0 \le \dfrac{|x|}{x_{\max}} \le \dfrac{1}{A}, \\ y_{\max} \dfrac{\ln\left[A\left(|x|/x_{\max}\right)\right]}{1+\ln A} \operatorname{sgn} x, & \dfrac{1}{A} \le \dfrac{|x|}{x_{\max}} \le 1, \end{cases}$$

где A – положительная константа, а остальные обозначения такие же, как в формуле (10.12). При стандартных значениях констант [31], которые равны соответственно $A = 87.56$ и $\mu = 255$, обе функции практически совпадают.

Представление аналоговых величин двоичными кодами характеризуется не только минимально возможной (равной шагу квантования), но и максимально возможной величиной, представимой кодом данной разрядности. При превышении аналоговым сигналом этой предельной величины двоичный код перестает зависеть от сигнала (происходит его ограничение вследствие насыщения квантователя).

Шум ложных импульсов. Шум ложных импульсов – это ошибка, возникающая в приемнике при декодировании кодовых комбинаций, искаженных в канале действием помех. Очевидно, что влияние указанных ошибок на восстанавливаемый сигнал зависит от места, занимаемого «испорченным» символом в кодовой комбинации. Если считать искажения различных символов независимыми случайными событиями, происходящими с вероятностью p, то вероятность ошибки кратности k в кодовой комбинации длины n можно рассчитать по формуле биномиального распределения

$$P(k) = C_n^k p^k (1-p)^{n-k}.$$

Если вероятность p мала, то вероятность того, что в комбинации произойдет хотя бы одна ошибка, равна

$$1-(1-p)^n \approx np \text{ при } np \ll 1.$$

При правильном построении системы связи вероятность p, определяемая отношением сигнал/шум в канале, мала, и шумом ложных импульсов можно пренебречь в сравнении с шумом квантования.

10.6. КОДИРОВАНИЕ С ПРЕДСКАЗАНИЕМ

Если передаваемый сигнал близок по своим статистическим характеристикам к белому шуму, т. е. имеет в ограниченном частотном диапазоне примерно постоянную спектральную плотность мощности, то дискретизация его в соответствии с требованиями теоремы Котельникова обеспечивает некоррелированность отсчетов. На практике часто приходится передавать сигналы с неравномерным спектром, а также производить дискретизацию с бóльшей частотой, что приводит к заметной корреляции между отсчетами. Таким образом, передаваемый дискретный сигнал обладает *избыточностью*, что приводит к неэффективному использованию канала. Один из способов повысить эффективность – передача и прием с использованием метода, называемого кодированием с предсказанием. Основная идея такого кодирования заключается в том, что если между передаваемыми отсчетами имеется статистическая связь, то ее можно использовать для предсказания следующего отсчета на основании известных предыдущих отсчетов. Очевидно, предсказанное значение никакой новой информации не содержит. Предсказание не может быть точным, поэтому разность истинного $b[n]$ и предсказанного $b_{\text{пр}}[n]$ значений сигнала в момент дискретного времени n представляет собой ошибку предсказания $\varepsilon[n] = b[n] - b_{\text{пр}}[n]$, которая и содержит всю информацию о текущем отсчете и передается по каналу. В приемнике на основе предыдущих отсчетов предсказывается текущий отсчет $\tilde{b}_{\text{пр}}[n]$ и к нему прибавляется принятое значение ошибки предсказания $\tilde{\varepsilon}[n]$ (рис. 10.9). Если бы в канале не действовали помехи, выходной сигнал совпадал бы со входным. На самом деле в результате действия помех имеют место ошибки.

Чем сильнее корреляция между отсчетами, тем точнее предсказание и тем меньше мощность (дисперсия) ошибки предсказания. Поэтому в таких случаях для передачи данных по каналу требуется меньшее количество кодовых символов (этим и достигается повышение эффективности, т. е. снижаются требования к пропускной

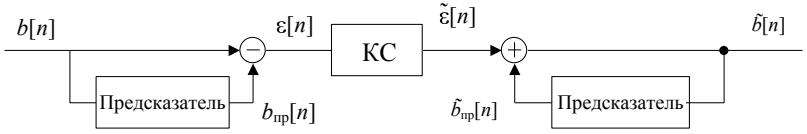


Рис. 10.9. Структура системы связи с предсказанием

способности канала). Во многих случаях алгоритм работы предсказателя может быть линейным, при этом очередное значение сигнала формируется как линейная комбинация некоторого числа предшествующих отсчетов. В частности, кодирование речевых сигналов на основе линейного предсказания применяется в современных системах мобильной связи (стандарты GSM, D-AMPS).

Способ передачи сигналов путем кодирования *ошибки* предсказания получил название *дифференциальной импульсно-кодовой модуляции* (ДИКМ). В таких системах применяют неравномерное квантование, это дает дополнительное преимущество, так как более вероятны малые значения ошибок. Выигрыш ДИКМ по сравнению с ИКМ тем больше, чем выше корреляция между соседними отсчетами сигнала.

Предельным случаем ДИКМ можно считать *дельта-модуляцию*, при которой число уровней квантования равно двум и передаваемый сигнал ошибки $d[n]$ содержит лишь информацию о полярности (знаке) ошибки. Если кодируемый сигнал $b(t)$ на n-м тактовом интервале возрастает (ошибка $\varepsilon[n]$ больше нуля), то передается сигнал +1, иначе –1 (рис. 10.10). Дельта-модуляция применима в тех случаях, когда шаг дискретизации сигнала много меньше, чем интервал корреляции.

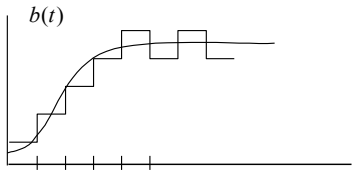


Рис. 10.10. Преобразования сигнала при дельта-модуляции

Преимуществом дельта-модуляции является простота кодера и декодера (рис. 10.11). Для восстановления сигнала $\tilde{b}(t)$ достаточно «проинтегрировать» сигнал $\tilde{d}[n]$ («интегрирование» состоит в суммировании с накоплением последовательности $\tilde{d}[n]$ нулей и единиц и преобразовании полученной последовательности чисел в ступенчатую функцию, которая затем сглаживается фильтром нижних частот). Однако дельта-модуляции свойственны специфические искажения, связанные с отставанием изменения ступенчатой

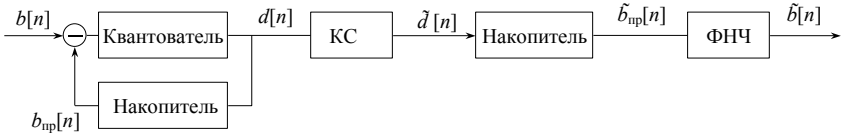


Рис. 10.11. Структура системы дельта-модуляции

аппроксимации сигнала от истинной сигнальной функции (*перегрузка по наклону*) и с колебаниями на участках, где сигнал изменяется слабо (*шум дробления*), см. рис. 10.10. Способ борьбы с этими явлениями состоит в *адаптации* шага квантования к виду сигнала: если несколько соседних значений ошибки $d[n]$ совпадают, это считается признаком монотонного изменения сигнала и шаг квантования увеличивается, если же на некотором интервале времени отсчеты сигнала $d[n]$ принимают поочередно значения +1 и –1, то это говорит о слабом изменении сигнала и шаг квантования уменьшается.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Перечислите свойства согласованного фильтра и фильтра Колмогорова–Винера.
2. Назовите общие черты и различия в постановке задач синтеза этих фильтров.
3. Назовите преимущества цифровых сигналов перед аналоговыми.
4. Что такое шум квантования? Что нужно для его уменьшения?
5. Что такое шум ложных импульсов?
6. В каких случаях целесообразно применение кодирования с предсказанием? Дельта-модуляции?
7. Что такое компандирование и для чего его применяют?

УПРАЖНЕНИЯ

1. Покажите, что дисперсия шума квантования равна $\Delta^2/12$, если его распределение равномерно в интервале от $-\Delta/2$ до $\Delta/2$.
2. Дан 10-разрядный аналого-цифровой преобразователь с динамическим диапазоном ±5 В. Определите:
а) величину шага квантования;
б) среднеквадратическое отклонение шума квантования;
в) отношение сигнал/шум на выходе АЦП, если на вход подается синусоида амплитудой 5 В;
г) отношение сигнал/шум на выходе АЦП, если на вход подается синусоида амплитудой 0,05 В.
3. Дан 8-разрядный аналого-цифровой преобразователь с динамическим диапазоном ±10 В. Определите:
а) величину шага квантования;
б) среднеквадратическое отклонение шума квантования;
в) отношение сигнал/шум на выходе АЦП, если на вход подается гауссовский случайный сигнал со среднеквадратическим отклонением 0,5 В;
г) вероятность попадания сигнала в область насыщения, если его СКО равно 3 В.

11. ПРИНЦИПЫ МНОГОКАНАЛЬНОЙ СВЯЗИ И РАСПРЕДЕЛЕНИЯ ИНФОРМАЦИИ

Во многих случаях необходимость передачи информации от источника к приемнику существует не постоянно, а лишь периодически, или возникает от случая к случаю. В то же время наиболее дорогостоящими частями многих систем и сетей передачи информации являются линии связи – кабельные, волноводные, световодные, радиорелейные и т.п. Поэтому естественно возникает задача совместного использования этого оборудования многими пользователями (абонентами), т.е. многоканальной связи или уплотнения. Тем самым повышается эффективность использования ресурсов линии.

Многоканальная связь возможна, очевидно, лишь тогда, когда пропускная способность совместно используемого оборудования больше суммарной информационной производительности всех источников. При этом ресурсы линии связи должны быть некоторым образом распределены между пользователями. Способы этого распределения (разделения каналов) и особенности построения многоканальных систем будут рассмотрены в этом разделе.

11.1. СТРУКТУРА МНОГОКАНАЛЬНОЙ СИСТЕМЫ СВЯЗИ

Упрощенная структурная схема многоканальной системы связи показана на рис. 11.1. Сообщения a_1, a_2, ... a_N вырабатываются источниками сообщений $ИС_1$, $ИС_2$,...,$ИС_N$ и поступают на индивидуальные модуляторы (передатчики), где преобразуются в канальные сигналы $u_1(t), u_2(t),...,u_N(t)$. Устройство объединения Σ

образует из них групповой сигнал $u_{\Sigma}(t)$, который в групповом модуляторе (передатчике) М преобразуется в линейный сигнал $u_{л}(t)$, поступающий в линию связи ЛС. Линейный сигнал в линии связи подвергается искажениям и воздействию помех, в результате на выходе линии имеет место наблюдаемое колебание $z(t)$, которое групповым приемником П преобразуется в групповой сигнал. Индивидуальные (канальные) приемники $П_1$, $П_2$,..., $П_N$ выделяют из группового сигнала соответствующие канальные сигналы, которые затем преобразуются в сообщения $\hat{a}_1, \hat{a}_2, ..., \hat{a}_N$, предназначенные для получателей $ПС_1$, $ПС_2$,...,$ПС_N$.

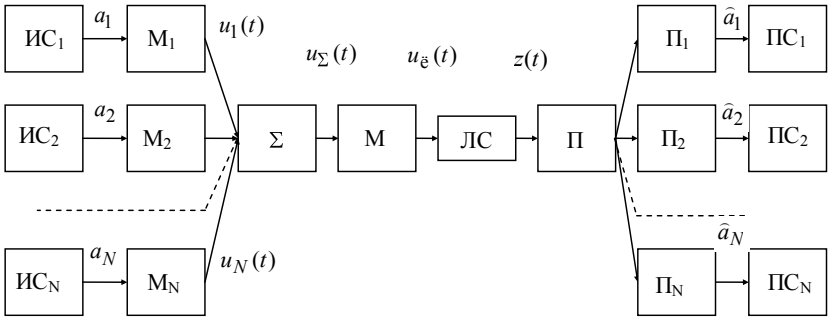


Рис. 11.1. Упрощенная структура многоканальной системы связи

Канальные передатчики вместе с устройством объединения образуют аппаратуру уплотнения каналов; групповой модулятор, линия связи и групповой приемник составляют групповой канал, канальные приемники образуют устройство разделения.

Очевидно, для того чтобы канальные сигналы можно было выделить из группового сигнала, они должны отличаться какими-либо признаками. Наиболее распространены линейные методы разделения, когда устройством объединения является суммирующий усилитель, а селекторы канальных сигналов, входящие в канальные приемники, представляют собой линейные устройства с постоянными или переменными параметрами.

При линейном разделении групповой сигнал равен сумме канальных сигналов

$$u_{\Sigma}(t) = \sum_{i=1}^{N} u_i(t) ,$$

а j-й селектор канального сигнала, описываемый линейным оператором $\mathbb{L}_j\{\cdot\}$, должен выделять из группового сигнала j-й канальный сигнал, т.е. должно выполняться условие

$$\mathbb{L}_j\{u_\Sigma(t)\} = \mathbb{L}_j\left\{\sum_{i=1}^{N} u_i(t)\right\} = \sum_{i=1}^{N}\mathbb{L}_j\{u_i(t)\} = \begin{cases} u_i(t), & i = j \\ 0, & i \neq j \end{cases},$$

которое означает линейную независимость канальных сигналов.

Представляя канальные сигналы векторами в пространстве сигналов, легко видеть, что они в силу линейной независимости образуют базис N-мерного подпространства. Групповой сигнал, как линейная комбинация базисных векторов, всегда принадлежит этому подпространству, а оператор $\mathbb{L}_j\{\cdot\}$ должен проецировать групповой сигнал на вектор, ортогональный всем базисным векторам (т.е. всем канальным сигналам), кроме j-го. Другими словами, оператор $\mathbb{L}_j\{\cdot\}$ находит скалярное произведение группового сигнала на j-й вектор базиса, *взаимного* по отношению к базису, образованному канальными сигналами (см. разд. 2).

Условие линейной разделимости канальных сигналов принято записывать как условие неравенства нулю *определителя Грама*

$$\begin{vmatrix} (u_1,u_1) & (u_1,u_2) & (u_1,u_3) & \dots & (u_1,u_N) \\ (u_2,u_1) & (u_2,u_2) & (u_2,u_3) & \dots & (u_2,u_N) \\ (u_3,u_1) & (u_3,u_2) & (u_3,u_3) & \dots & (u_3,u_N) \\ \dots & \dots & \dots & \dots & \dots \\ (u_N,u_1) & (u_N,u_2) & (u_N,u_3) & \dots & (u_N,u_N) \end{vmatrix} \neq 0 .$$

Очевидно, что в отсутствие помех любая линейно независимая совокупность сигналов одинаково пригодна для многоканальной связи. Однако в реальных каналах связи помехи есть всегда, поэтому наилучшими помехоустойчивыми свойствами обладают ортогональные системы сигналов; тогда проекции, выделяемые селекторами канальных сигналов, совпадают с канальными сигналами, а реализация самих селекторов оказывается наиболее простой.

Систему ортогональных сигналов можно выбрать многими способами. Наиболее очевидными вариантами выбора являются временной и частотный способы разделения, когда ортогональность обеспечивается тем, что сигналы не перекрываются во временной или частотной области.

11.2. ЧАСТОТНОЕ РАЗДЕЛЕНИЕ КАНАЛОВ

Канальные сигналы при частотном разделении каналов (ЧРК) занимают неперекрывающиеся полосы частот, поэтому их разделение обеспечивается полосовыми фильтрами.

Сообщения с выходов источников (рис. 11.1) поступают на канальные модуляторы, где происходит модуляция гармонических колебаний с различными частотами, называемых *поднесущими*. Частоты поднесущих колебаний должны различаться настолько, чтобы спектры модулированных сигналов не накладывались друг на друга во избежание взаимных помех. После модуляции информационные сигналы занимают ограниченные полосы частот, которые могут отличаться по ширине от спектров исходных колебаний (например, при частотной или фазовой модуляции), или совпадать с ними (при ОБП-модуляции). Важно, чтобы полосы частот, занимаемые различными сигналами, не только не перекрывались, но и отстояли друг от друга на ширину некоторого *защитного* интервала, что облегчает их последующее разделение при помощи реальных фильтров, имеющих конечную крутизну АЧХ в переходной полосе.

Индивидуальные канальные (модулированные) сигналы суммируются и поступают на групповой передатчик, где происходит модуляция несущего колебания групповым сигналом, после чего модулированный *линейный* сигнал передается в линию связи. Групповой приемник производит демодуляцию линейного сигнала, после чего каждый канальный приемник выделяет при помощи полосового фильтра «свой» канальный сигнал, демодулирует его и выделяет сообщение.

Как видно, частотное разделение каналов основано на распределении одного из ресурсов – полосы пропускания группового канала – между различными индивидуальными каналами. Частотному разделению каналов свойственны следующие недостатки.

Во-первых, из-за неидеальности полосовых фильтров необходимы защитные интервалы, которые суммарно составляют около 20 % полосы пропускания группового канала связи. Например, в многоканальных телефонных системах для передачи речевых сигналов установлена полоса частот 3100 Гц (считается, что при передаче речи для обеспечения разборчивости с сохранением индивидуальных голосовых признаков достаточен диапазон от 300 до 3400 Гц), а ширина защитного интервала составляет 900 Гц; таким образом, при объединении N телефонных каналов общая ширина полосы частот группового канала составляет $4N$ кГц.

Во-вторых, предъявляются очень жесткие требования к линейности канала (нелинейность приводит к появлению кратных и комбинационных составляющих, а поскольку спектры канальных сигналов имеют ширину значительно больше защитного интервала, эти составляющие попадают в «чужие» каналы и разделить их путем фильтрации или каким-либо другим способом невозможно).

11.3. ВРЕМЕННО́Е РАЗДЕЛЕНИЕ КАНАЛОВ

Временно́е уплотнение (*временно́е разделение каналов*, ВРК) основано на распределении временно́го ресурса группового канала между различными индивидуальными каналами. Все время действия канала разбивается на короткие (тактовые) интервалы, и канал предоставляется различным абонентам поочередно в периодическом порядке. Таким образом, каждый пользователь, передающий информацию, получает канал в свое пользование многократно на короткое время. Очевидно, таким образом можно передавать лишь отсчеты сигнала, взятые с шагом, равным периоду следования тактовых интервалов. Таким образом, в основе ВРК лежит использование теоремы отсчетов, и передавать можно лишь сигналы с финитным спектром.

Для формирования канальных сигналов используются различные виды импульсной модуляции (АИМ, ВИМ, ШИМ). Групповой сигнал может передаваться непосредственно по линии или модулировать гармоническую несущую.

На рис. 11.2 представлена упрощенная структурная схема многоканальной системы связи с временны́м разделением каналов. Источники первичных сигналов ИС соединены с коммутатором передатчика $К_{пер}$. Коммутатор поочередно подключает источники к импульсному модулятору ИМ, который модулирует сигнал-переносчик – периодическую последовательность импульсов. В результате этого получается групповой сигнал, который поступает в канал связи КС, который может включать модулятор гармонической несущей, линию связи и общий демодулятор. После общей демодуляции групповой сигнал разделяется коммутатором приемника $К_{пр}$ на канальные сигналы $s_1(t),...,s_N(t)$, которые после демодуляции в импульсных демодуляторах ИД поступают к получателям сигналов ПС. Следует отметить, что в системах связи с ВРК во избежание межканальных помех необходима синхронизация приемной и передающей станций, поэтому в линейный сигнал, кроме канальных сигналов, добавляется периодическая последовательность

синхроимпульсов, которые должны достаточно сильно отличаться от канальных импульсов, чтобы их можно было легко выделить.

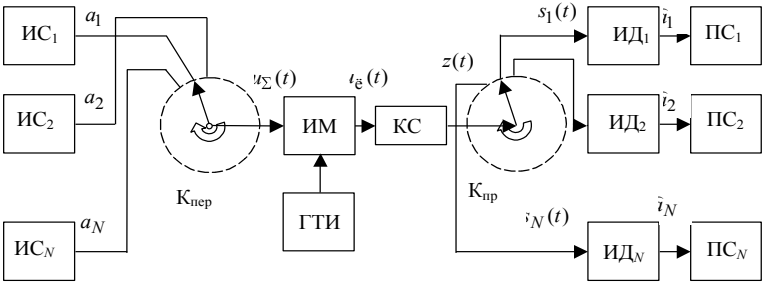


Рис.11.2. Структура системы связи с ВРК

При распространении импульсной последовательности (группового сигнала) по групповому каналу появляются переходные искажения вследствие инерционности (ограниченной полосы частот) любой физически осуществимой системы связи. При этом происходит искажение формы импульсов за счет переходных процессов, в результате фронты импульсов «затягиваются» и соседние импульсы, принадлежащие разным каналам, начинают перекрываться. Второй причиной взаимных помех является несовершенство системы синхронизации.

Для снижения уровня взаимных помех приходится вводить защитные временны́е интервалы между соседними импульсами, что приводит или к уменьшению числа каналов при сохранении длительности канальных импульсов, или к укорочению канальных импульсов, что ведет к расширению требуемой полосы частот группового канала. Например, при передаче речевых сигналов, исходя из ширины спектра речевого сигнала 3100 Гц, минимальная частота дискретизации должна быть равна 6200 Гц. Однако в действительности частоту дискретизации принимают равной 8 кГц, что требует для передачи канальной импульсной последовательности полосы частот около 4 кГц. В результате общая полоса частот группового канала практически совпадает с полосой, требуемой при ЧРК. Если учесть необходимость дополнительной передачи синхроимпульсов, то сравнение получается не в пользу ВРК.

Вместе с тем системы с ВРК имеют неоспоримое преимущество, состоящее в их нечувствительности к нелинейности группового канала. Кроме того, роль временно́го разделения возрастает в связи с широчайшим распространением цифровых систем связи.

11.4. РАЗДЕЛЕНИЕ КАНАЛОВ ПО ФОРМЕ СИГНАЛОВ

Как было показано, временной и частотный способы разделения каналов представляют собой лишь частные случаи линейного разделения, которое основано на вычислении скалярных произведений группового сигнала и некоторых опорных сигналов (векторов в сигнальном пространстве), причем это разделение наиболее эффективно, когда все такие векторы являются взаимно ортогональными. В случае временно́го разделения ортогональность обеспечивается тем, что все канальные сигналы отличны от нуля на различных носителях – временны́х интервалах, при частотном разделении не пересекаются носители сигналов на частотной оси. И в том, и в другом случае это гарантирует равенство нулю соответствующих скалярных произведений (см. разд. 2). Однако скалярное произведение сигналов может быть равно нулю и в том случае, когда оба сигнала занимают один и тот же временно́й и/или частотный интервал, отличаясь друг от друга по форме. Этот факт и лежит в основе рассматриваемого метода разделения каналов.

Рассмотрим N -канальную систему связи, в которой передаваемые сообщения (сигналы) $x_i(t),\ i=\overline{1,N}$, дискретизируются с шагом T_d. Выберем в качестве канальных сигналов N функций $\psi_i(t),\ i=\overline{1,N}$, взаимно ортогональных на интервале T_d. Будем на каждом интервале длительности T_d передавать по групповому каналу сумму канальных сигналов, умноженных на отсчеты $x_i(kT_d)$ соответствующих индивидуальных сигналов

$$u_\Sigma(t)=\sum_{i=1}^{N}x_i(kT_d)\psi_i(t-kT_d).$$

Выделение j-го сигнала на k-м временно́м интервале основано на свойстве ортогональности:

$$\mathbb{L}_j\left\{u_\Sigma(t)\right\}=\int_{kT_d}^{(k+1)T_d}\left(\sum_{i=1}^{N}x_i(kT_d)\psi_i(t-kT_d)\right)\psi_j(t-kT_d)dt=x_j(kT_d)E_j,$$

где E_j – энергия j-го сигнала. Таким образом, селекторы канальных сигналов в случае разделения по форме представляют собой корреляторы (или согласованные фильтры). И в том, и в другом случае необходима синхронизация системы.

В качестве канальных сигналов можно использовать любые системы функций, ортогональных на конечном временно́м интервале (например, системы ортогональных полиномов Лежандра, Чебышева и др.). В настоящее время для этой цели нередко используются кусочно-постоянные функции Радемахера, Уолша и др. [10].

11.5. АСИНХРОННЫЕ АДРЕСНЫЕ СИСТЕМЫ СВЯЗИ

Общим недостатком рассмотренных методов разделения каналов является необходимость синхронизации. В некоторых случаях этот недостаток не является критическим и сравнительно легко преодолевается путем передачи по групповому каналу дополнительного синхронизирующего сигнала. Но иногда обеспечить синхронизацию с достаточной точностью слишком трудно (или она обходится слишком дорого).

В таких случаях используют асинхронные системы связи, в которых сигналы различных каналов передаются одновременно в одной среде в общей полосе частот без синхронизации. Эти системы называют системами *со свободным доступом* или системами *с незакрепленными каналами*. Отличительным признаком каждого канального сигнала, позволяющим приемнику выделить «свой» сигнал из группового, является его форма. Как и при разделении по форме, на канальные сигналы накладывается требование взаимной ортогональности, но теперь ортогональность понимается в усиленном смысле: сигналы должны быть ортогональны друг другу при *любых временны́х сдвигах*. Строго говоря, это невозможно, однако можно получить системы сигналов, удовлетворяющие этому требованию приближенно. Такие сигналы по своим корреляционным свойствам напоминают реализации белого шума, поэтому их называют шумоподобными сигналами (ШПС)[1]. За каждым пользователем закрепляется ШПС определенной формы, представляющий своеобразный «адрес» абонента, по которому канальный сигнал может быть выделен из наблюдаемой смеси. Поэтому системы связи, основанные на свободном доступе к каналу и не требующие синхронизации, называются *асинхронными адресными системами*.

[1] Широко используются ШПС на основе m-последовательностей, кодов Голда, Касами и др. [30].

Асинхронные адресные системы оказываются очень эффективными, когда они объединяют большое количество малоактивных абонентов. Тогда количество одновременно работающих передатчиков может быть сравнительно небольшим. По мере увеличения числа активных абонентов растет уровень взаимных переходных помех («шумов неортогональности» [10]), и с некоторого момента качество связи падает. Расчет вероятного числа активных абонентов и соответствующих ресурсов системы свободного доступа основывается на статистических данных.

11.6. КОМБИНАЦИОННОЕ РАЗДЕЛЕНИЕ КАНАЛОВ

Принцип комбинационного уплотнения, используемый для многоканальной передачи дискретных сообщений, состоит в следующем.

Предположим, что необходимо передавать по одному групповому тракту сообщения N источников, каждое из которых состоит из символов некоторого кода по основанию m. Очевидно, в произвольный момент времени символы N источников образуют одно из m^N возможных сочетаний. Можно поставить в соответствие каждому из этих сочетаний один символ m^N-значного кода, так что приняв один такой символ, на приемной стороне, можно восстановить все N символов, относящихся к отдельным каналам.

Например, при $m = N = 2$ канальные символы могут образовать четыре возможных сочетания: 00, 01, 10, 11, которым можно поставить в соответствие четыре символа 0, 1, 2, 3 четырехзначного кода группового канала. Ясно, что эти символы можно считать просто номерами сочетаний канальных символов и поэтому возможность их однозначного восстановления по принятому номеру не представляет трудности. Номера (символы кода группового канала) могут передаваться, например, посредством кодоимпульсной модуляции.

Устройство объединения каналов при комбинационном разделении представляет собой комбинационную[2] схему, преобразующую комбинацию входных m-значных символов в один выходной символ m^N-значного кода, а устройство разделения – комбинационное устройство, выполняющее обратное преобразование.

[2] В цифровой технике комбинационными называют логические устройства без памяти.

11.7. МНОГОПОЗИЦИОННЫЕ СИГНАЛЫ

При кодовом уплотнении представляет интерес вопрос передачи символов m^N-значного кода по групповому каналу. Как было сказано ранее, они могут передаваться, например, посредством кодоимпульсной модуляции. Другой способ основан на использовании многопозиционных сигналов.

Многопозиционные сигналы образуются путем манипуляции различными параметрами гармонического переносчика. Например, четырехпозиционный сигнал, известный как КАФМ-4, получается манипуляцией амплитуд квадратурных составляющих значениями +1 и –1 (рис. 11.3, *а*) (буквами *s* и *q* обозначены синфазная и квадратурная составляющие). Очевидно, такой же вид имеет сигнал, полученный четырехуровневой манипуляцией фазы при постоянной амплитуде (ФМ-4). На рис. 11.3, *б* показана амплитудно-фазовая диаграмма сигнала КАФМ-16, получаемого квадратурной амплитудно-фазовой манипуляцией с уровнями +1, –1, 0.5 и –0.5.

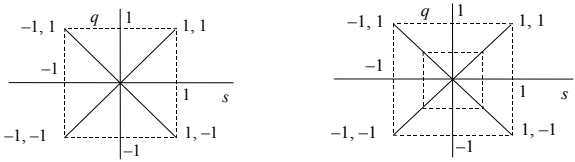


Рис. 11.3. Многопозиционные сигналы КАФМ-4 и КАФМ-16

Рассмотренные сигналы задаются в декартовой системе координат и формируются путем сложения квадратурных компонент после их амплитудной манипуляции (дискретной модуляции). Используя последовательно манипуляцию фазы и манипуляцию амплитуды, формируют многопозиционные сигналы, заданные в *полярных* координатах (рис. 11.4).

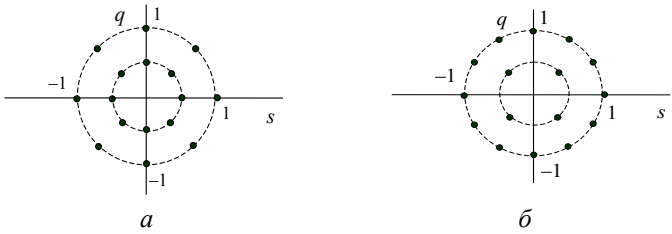


Рис.11.4. Многопозиционные сигналы, формируемые в полярных координатах

11.8. КОММУТАЦИЯ В СЕТЯХ СВЯЗИ

Для обмена информацией между многими пользователями (абонентами) создаются *сети связи*, в которых производится распределение информации в соответствии с заданными адресами. Сети связи подразделяются на *некоммутируемые*, в которых связь абонентов осуществляется по принципу «каждый с каждым» по закрепленным каналам, и *коммутируемые*, в которых связь осуществляется по временно выделяемым каналам. Выделение каналов парам абонентов производится *узлами коммутации*. Таким образом, сеть связи состоит из оконечных (абонентских) устройств, каналов связи и узлов коммутации. Сети могут иметь различную структуру: линейную, радиальную, кольцевую, радиально-узловую и т.д. Построение и оптимизация сетей связи осуществляются на основе теории графов и теории массового обслуживания.

Наиболее широко известными узлами коммутации можно считать автоматические телефонные станции (АТС). Основной составной частью АТС как узла коммутации является коммутационное *поле*. Коммутационное поле может быть пространственным, характерным для аналоговых систем, или пространственно-временным. В первом случае коммутационное поле электрически соединяет отдельные линии на все время соединения. Во втором случае линии соединяются на короткие временны́е интервалы в соответствии с методом временно́го уплотнения. В цифровых сетях связи применяется цифровая коммутация канальных интервалов, которая осуществляется записью принятых сегментов сообщений в память и считыванием их в определенном порядке [10].

Для эффективного использования имеющихся каналов связи и узлов коммутации форма представления информации должна быть стандартизована. Существующий стандарт имеет иерархическую (многоуровневую) структуру, основанную на модели *взаимодействия открытых систем*.

Эталонная модель содержит семь *уровней*. Каждый уровень позволяет рассматривать некоторый аспект функционирования системы или сети связи, абстрагируясь от содержания остальных уровней (рис. 11.5).

Физический уровень модели непосредственно взаимодействует с физической средой распространения сигналов и обеспечивает передачу сигналов между двумя узлами. Сетевой уровень обеспечивает установление адреса и маршрута для передачи пакета данных от узла передачи до узла назначения. Транспортный уровень соответствует передаче данных между абонентами сети и характеризуется

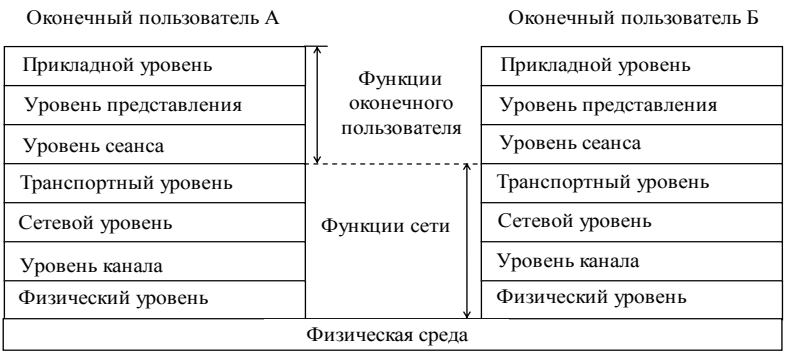


Рис. 11.5. Эталонная модель

максимальным временем установления соединения, пропускной способностью, временем задержки при передаче сообщений и т.п. Четыре нижних уровня реализуют функции сети, оставшиеся три уровня ориентированы на услуги, предоставляемые оконечным пользователям. Уровень сеанса обеспечивает организацию диалога, очередность передачи данных, приоритеты и т.д. Уровень представления определяет коды, форматы данных, способы сжатия и т.п. Прикладной уровень служит для реализации услуг, предоставляемых сетью пользователям (электронная почта, телетекст, факс, электронные переводы, пакетная передача речи и т.п.) [10].

Уровни модели взаимодействуют со смежными (верхним и нижним соседними) уровнями; кроме того, возможно взаимодействие различных пользователей на одинаковых уровнях. Правила взаимодействия одного уровня называются *протоколом*. Взаимодействие на некотором уровне обеспечивается предоставлением ему услуг смежным нижележащим уровнем.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Объясните, почему в системах с ВРК нет жестких требований к нелинейности группового канала.
2. В чем заключаются причины межканальных помех в системах с частотным и временным разделением каналов?
3. Почему необходимы защитные интервалы при ВРК и ЧРК? Чем они различаются, и в чем состоит их сходство?
4. Для чего нужна синхронизация в системах с разделением сигналов по форме при реализации селекторов: 1) в форме корреляторов? 2) в форме согласованных фильтров?

5. Какими свойствами должны обладать сигналы, применяемые в асинхронных адресных системах? Можно ли в них использовать сигналы Баркера (см. разд. 2)?

6. Какие сигналы называются многопозиционными?

7. Для чего нужны узлы коммутации?

8. Что такое протокол?

УПРАЖНЕНИЯ

1. Рассмотрите в качестве канальных сигналов $u_1(t) = \cos(\omega_0 t)$ и $u_2(t) = \cos(\omega_0 t + \phi)$, где ω_0 – частота, ϕ – начальная фаза. Выясните, можно ли использовать такие сигналы в двухканальной системе связи (воспользуйтесь определителем Грама). Если да, то какое значение ϕ является наилучшим с точки зрения помехоустойчивости?

2. Проверьте непосредственными вычислениями, можно ли построить систему связи с фазовым разделением каналов при числе каналов более двух.

3. Предложите амплитудно-фазовую диаграмму многопозиционного сигнала, пригодного для передачи 3-значного кода; 8-значного кода; 12-значного кода.

12. ОСНОВЫ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ

Под обработкой сигналов в широком смысле можно понимать совокупность преобразований, направленную на наиболее эффективную передачу, хранение и извлечение информации. В последние десятилетия все более широко применяется цифровая обработка сигналов (ЦОС), которой свойственны следующие преимущества перед аналоговой обработкой:

– принципиальная возможность реализации практически *любых* алгоритмов обработки (в аналоговой технике могут быть реализованы далеко не всякие алгоритмы); развитие элементной базы обеспечивает реализуемость все более широкого класса алгоритмов обработки в *реальном масштабе времени*;

– потенциально сколь угодно высокая *точность* реализации алгоритмов, определяемая разрядностью цифровых устройств;

– принципиальная возможность *безошибочного* воспроизведения сигналов при передаче и хранении на основе помехоустойчивого кодирования, которое применимо только к цифровым сигналам.

Реализация перечисленных преимуществ на практике возможна лишь на основе глубокого понимания теории дискретных сигналов и цепей, которая во многом сходна с аналоговой теорией, но в то же время имеет и существенные особенности [5, 19].

12.1. ОСНОВНЫЕ ПОНЯТИЯ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ. ДИСКРЕТНЫЕ И ЦИФРОВЫЕ СИГНАЛЫ

Центральным в теории ЦОС является понятие дискретного сигнала. Математической моделью дискретного сигнала служит решетчатая функция, или *последовательность* $x[n]$, где n – аргумент, принимающий значения из дискретного множества, а функ-

ция $x[\cdot]$ может принимать значения из непрерывного множества вещественных или комплексных чисел (впрочем, в обработке пространственно-временных сигналов функция $x[\cdot]$ принимает векторные значения). Как следует из теоремы отсчетов, аналоговый сигнал $x_{\text{а}}(t)$ с финитным спектром, сосредоточенным в полосе частот $\left(-F_{\text{в}}, F_{\text{в}}\right)$, может быть без потерь информации заменен дискретной последовательностью $x[n] = x_{\text{а}}(nT_d)$ своих значений, взятых с шагом $T_d < 1/(2F_{\text{в}})$; эта последовательность и представляет собой дискретный сигнал. В дальнейшем, если не сказано иное, предполагается, что аргумент n последовательности $x[n]$ принимает целые значения от $-\infty$ до $+\infty$ (обозначается $n = \overline{-\infty, +\infty}$). Напомним, что дискретный аргумент принято заключать в квадратные скобки.

Под *цифровым* сигналом понимают дискретный сигнал, *квантованный* по уровню. Другими словами, цифровой сигнал – это последовательность, принимающая значения из дискретного (к тому же, как правило, конечного) множества. Это связано с тем, что цифровые устройства всегда имеют ограниченную разрядность и отсчеты сигналов, подлежащих цифровой обработке, неизбежно округляются (квантуются). Для изучения большинства вопросов цифровой обработки сигналов удобнее считать, что сигнал принимает значения из непрерывного множества, поэтому всюду, где возможно, используется модель *дискретного* сигнала. Моделью цифрового сигнала пользуются обычно лишь в тех случаях, когда рассматриваются специфические эффекты, связанные с квантованием сигнала, округлением промежуточных результатов, ограничением разрядной сетки цифрового устройства и т.п.

При доказательстве теоремы отсчетов было установлено, что при умножении аналогового сигнала на периодическую последовательность δ -функций происходит периодическое продолжение спектральной плотности сигнала по частотной оси с периодом, равным частоте дискретизации.

Для последовательности (дискретного сигнала) $x[n]$, $n = \overline{-\infty, \infty}$ можно определить *преобразование Фурье*

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]\, e^{-j\omega n}, \qquad (12.1)$$

где ω – вещественный параметр[3]. Легко видеть, что изменение ω на величину $\pm k \cdot 2\pi$ при любом целом k никак не влияет на результат преобразования. Таким образом, величину ω можно понимать, как угол, а $e^{j\omega}$ – как точку на комплексной плоскости, находящуюся на окружности единичного радиуса (рис. 12.1). Поэтому выражение (12.1) определяет на единичной окружности функцию вещественной переменной ω, которая имеет смысл круговой частоты.

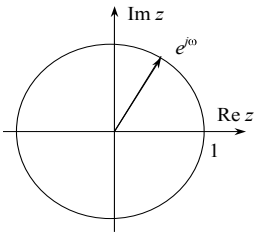


Рис. 12.1. Угловая интерпретация частоты

Вспомним, что при восстановлении аналогового сигнала моделью дискретного сигнала служит идеализированный АИМ-сигнал, состоящий из δ -функций, умноженных на отсчеты сигнала (2.58)

$$v(t) = \sum_{n=-\infty}^{\infty} x_{\text{a}}(nT_d)\delta(t - nT_d)\,.$$

Найдем преобразование Фурье этого *аналогового* сигнала, обозначив круговую частоту в его спектральном описании буквой Ω:

$$V(\Omega) = \int_{-\infty}^{\infty} v(t)e^{-j\Omega t}dt = \sum_{n=-\infty}^{\infty} x_{\text{a}}(nT_d)\int_{-\infty}^{\infty} \delta(t - nT_d)e^{-j\Omega t}dt =$$

$$= \sum_{n=-\infty}^{\infty} x_{\text{a}}(nT_d)e^{-j\Omega nT_d}\,. \qquad (12.2)$$

Сравнивая выражения (12.1) и (12.2), легко видеть, что при условиях

$$x[n] = x_{\text{a}}(nT_d)\,,$$

$$\omega = \Omega T_d\,,\ -\pi \le \omega \le \pi \qquad (12.3)$$

их левые части совпадают. Это означает, что выражение (12.1) определяет спектральную плотность дискретного сигнала, совпадающую

---

[3] Обозначение спектральной плотности в виде $X(e^{j\omega})$ общепринято в литературе по ЦОС и обусловлено связью преобразования Фурье с z-преобразованием (см. разд. 12.2).

по форме со спектральной плотностью идеального АИМ-сигнала (который при воздействии на идеальный ФНЧ с П-образной характеристикой позволяет точно восстановить исходный аналоговый сигнал)[4].

Из условия (12.3) следует, что необходимо выполнение неравенства $-\frac{\pi}{T_d} \le \Omega \le \frac{\pi}{T_d}$, или, что то же самое, $-\frac{\Omega_d}{2} \le \Omega \le \frac{\Omega_d}{2}$, где $\Omega_d = 2\pi F_d = \frac{2\pi}{T_d}$ – круговая частота дискретизации. Иными словами, мы снова получили условие выбора частоты дискретизации, как минимум, вдвое выше верхней частоты спектра аналогового сигнала (ср. разд. 2).

Таким образом, формальное совпадение левых частей (12.1) и (12.2) позволяет оперировать спектральной плотностью $X(e^{j\omega})$ последовательности $x[n]$ вместо спектральной плотности аналогового сигнала, но лишь при условии ее финитности и правильного выбора шага дискретизации, когда копии спектральной плотности периодически повторяются вдоль оси частот без перекрытия. При соблюдении этого условия любые действия над дискретным сигналом эквивалентны соответствующим действиям над аналоговым сигналом и обработка сигнала может производиться в цифровой форме.

Рассмотрим выражение (12.1) как разложение 2π -периодической функции аргумента ω в комплексный ряд Фурье по базисным функциям $e^{jn\omega}$, $n = \overline{-\infty, \infty}$. Тогда, очевидно, отсчеты $x[n]$ суть не что иное, как коэффициенты этого ряда и могут быть найдены по общей формуле для вычисления коэффициентов комплексного ряда Фурье:

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega, \qquad n = \overline{-\infty, \infty}. \tag{12.4}$$

Это выражение представляет собой обратное преобразование Фурье для последовательности (дискретного сигнала).

---

[4] Более детально взаимосвязь аналогового и дискретного сигналов рассматривается, например, в [19].

12.2. СТАЦИОНАРНЫЕ ЛИНЕЙНЫЕ ДИСКРЕТНЫЕ ЦЕПИ

Преобразования дискретных сигналов в процессе их обработки могут выполняться специализированными цифровыми устройствами или универсальными вычислителями (процессорами) под управлением программ; в любом случае удобно считать, что преобразование выполняется некоторой дискретной цепью. Таким образом, дискретной цепи соответствует отображение множества входных (дискретных) сигналов на множество выходных сигналов. Задать отображение – значит задать эти множества и каждому входному сигналу поставить в соответствие единственный выходной. Как и для аналоговых цепей, для упрощения этой задачи на отображение (цепь) накладываются определенные ограничения.

Прежде всего, положим, что множества входных и выходных сигналов совпадают (рассматривается задача фильтрации), тогда понятие отображения сужается до оператора. Будем также считать, что оператор цепи $\mathbb{L}\{\cdot\}$ линеен, т.е. удовлетворяет принципу суперпозиции

$$\mathbb{L}\{\alpha_1 x_1 + \alpha_2 x_2\} = \alpha_1 \mathbb{L}\{x_1\} + \alpha_2 \mathbb{L}\{x_2\},$$

где α_1, α_2 – скалярные коэффициенты (вещественные или комплексные), $x_1 = x_1[n]$, $x_2 = x_2[n]$ – дискретные сигналы. Произвольный дискретный сигнал (последовательность) $x[n]$ можно представить в виде обобщенного ряда Фурье относительно базиса, состоящего из сдвинутых δ-последовательностей (см. разд. 2)

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n-k],$$

где отсчеты этого сигнала $x[k]$ рассматриваются как постоянные коэффициенты при базисных функциях $\delta[n-k]$, $k = \overline{-\infty, \infty}$. Тогда результат воздействия линейного оператора (линейной цепи) на этот сигнал равен

$$y[n] = \mathbb{L}\left\{\sum_{k=-\infty}^{\infty} x[k]\delta[n-k]\right\} = \sum_{k=-\infty}^{\infty} x[k]\mathbb{L}\{\delta[n-k]\} = \sum_{k=-\infty}^{\infty} x[k]h[n,k],$$

где $h[n,k]$ представляет собой отклик цепи в момент времени n на δ-последовательность, имеющую единичное значение в момент

времени k . Если кроме линейности потребовать, чтобы весовая последовательность $h[n,k]$ зависела только от разности аргументов, $h[n,k]=h[n-k]$, то цепь станет инвариантной к сдвигу (стационарной), а формула нахождения выходного сигнала примет форму *дискретной свертки*

$$y[n]=\sum_{k=-\infty}^{\infty}x[k]h[n-k]=\sum_{k=-\infty}^{\infty}h[k]x[n-k]\,. \qquad (12.5)$$

Последовательность $h[n]$ называется импульсной характеристикой линейной инвариантной к сдвигу (ЛИС) цепи и является ее исчерпывающей характеристикой, так как позволяет найти сигнал на выходе данной ЛИС-цепи для произвольного входного сигнала.

Здесь уместно отметить одно важное свойство дискретных цепей, отличающее их от аналоговых. Дискретная свертка представляет не только метод анализа ЛИС-цепи, подобно интегралу Дюамеля для аналоговых цепей, но также *алгоритм* работы вычислительного устройства. Таким образом, задача анализа дискретных ЛИС-цепей оказывается тесно связанной с задачей синтеза (подробнее см., например [19]).

Рассмотрим ЛИС-цепь при воздействии на ее вход комплексной экспоненциальной последовательности $x[n]=e^{j\omega n}$ при $n=\overline{-\infty,\infty}$, тогда выходной сигнал в соответствии с (12.5)

$$y[n]=\sum_{k=-\infty}^{\infty}h[k]e^{j\omega n}e^{-j\omega k}=e^{j\omega n}\sum_{k=-\infty}^{\infty}h[k]e^{-j\omega k}=e^{j\omega n}H(e^{j\omega})\,,$$

где $H(e^{j\omega})=\sum_{n=-\infty}^{\infty}h[n]e^{-j\omega n}$ – *комплексная частотная характеристика* ЛИС-цепи.

Рассматривая выражение (12.4) как представление произвольного дискретного сигнала $x[n]$ суперпозицией несчетного множества комплексных экспоненциальных последовательностей $e^{j\omega n}$ ($\omega\in[-\pi,\pi]$), умноженных на весовые коэффициенты $\frac{1}{2\pi}X(e^{j\omega})$, легко видеть, что выходная последовательность получается домножением каждой из них на значение КЧХ:

$$y[n]=\frac{1}{2\pi}\int_{-\pi}^{\pi}H(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega\,,\ n=\overline{-\infty,\infty}\,. \qquad (12.6)$$

Сравнивая выражения (12.6) и (12.4), видим, что спектральная плотность выходного сигнала равна $Y(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$. Полученное выражение составляет основу *спектрального* метода анализа ЛИС-цепей.

Предположим, что импульсная характеристика некоторой цепи $h[n]$ имеет конечную длину N, т.е. $h[n] \neq 0$, $n = \overline{0, N-1}$. Тогда свертка (12.5) принимает вид конечной суммы

$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k] = \sum_{k=0}^{N-1} b_k x[n-k],$$

и может быть записана в виде *разностного уравнения*

$$y[n] = b_0 x[n] + b_1 x[n-1] + b_2 x[n-2] + \ldots + b_{N-1} x[n-N+1]. \quad (12.7)$$

Вычисление каждого значения выходного сигнала требует учета текущего и $N-1$ предшествующих отсчетов входного сигнала и может быть выполнено цепью, структурная схема которой показана на рис. 12.2. Такие цепи называются *трансверсальными*, или цепями с конечной импульсной характеристикой (*КИХ-цепями*).

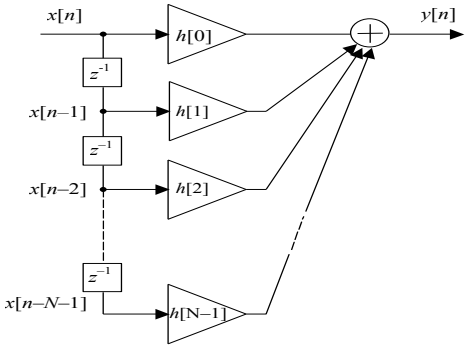


Рис. 12.2. Структура цепи с конечной импульсной характеристикой

Комплексная частотная характеристика КИХ-цепи имеет вид полинома порядка $N-1$ относительно $e^{-j\omega}$:

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h[n]e^{-j\omega n} = b_0 + b_1 e^{-j\omega} + b_2 e^{-j\cdot 2\omega} + \ldots + b_{N-1} e^{-j(N-1)\omega}.$$

Таким образом, КИХ-цепь *умножает* спектральную плотность входной последовательности на полином. Другой важный для

практики класс дискретных ЛИС-цепей составляют цепи, которые не умножают, а *делят* спектральную плотность входной последовательности на полином некоторого порядка $M-1$ относительно $e^{-j\omega}$. Обозначим этот полином $A(e^{j\omega}) = \alpha_0 + \alpha_1 e^{-j\omega} + \alpha_2 e^{-j\cdot 2\omega} + ...$ $... + \alpha_{M-1} e^{-j(M-1)\omega}$, тогда спектральные плотности входной и выходной последовательностей связаны выражением $Y(e^{j\omega}) = X(e^{j\omega}) / A(e^{j\omega})$, следовательно, $X(e^{j\omega}) = Y(e^{j\omega}) A(e^{j\omega})$, откуда по аналогии с (12.7) можно записать

$$x[n] = \alpha_0 y[n] + \alpha_1 y[n-1] + \alpha_2 y[n-2] + ... + \alpha_{M-1} y[n-M+1].$$

Решая это уравнение относительно выходного сигнала, получаем

$$y[n] = \frac{1}{\alpha_0} x[n] - \frac{\alpha_1}{\alpha_0} y[n-1] - \frac{\alpha_2}{\alpha_0} y[n-2] - ... - \frac{\alpha_{M-1}}{\alpha_0} y[n-M+1],$$

откуда, вводя обозначения $b = 1/\alpha_0$, $a_i = -\alpha_i / \alpha_0$, находим окончательно разностное уравнение *рекурсивной* цепи[5]

$$y[n] = bx[n] + a_1 y[n-1] + a_2 y[n-2] + ... + a_{M-1} y[n-M+1],$$

структура которой показана на рис. 12.3.

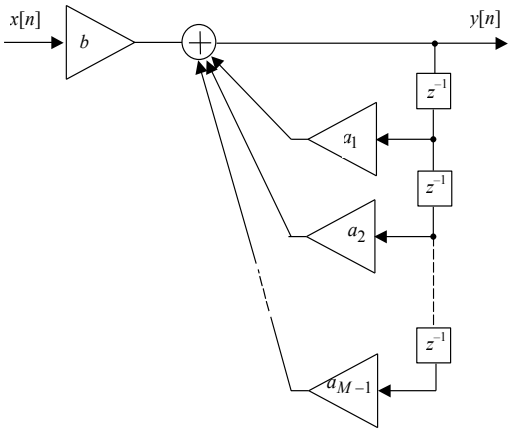


Рис. 12.3. Структура рекурсивной цепи

Обычно к ЛИС-цепям предъявляется требование *устойчивости*. Напомним, что линейная цепь называется устойчивой, если

[5] Отметим, что трансверсальные цепи иногда называют *нерекурсивными.*

отклик на воздействие, ограниченное по модулю, также ограничен. Для устойчивости ЛИС-цепи необходимо и достаточно, чтобы ее импульсная характеристика была *абсолютно суммируемой*, т.е. выполнялось условие [5, 19]

$$\sum_{n=-\infty}^{\infty} |h[n]| < \infty . \qquad (12.8)$$

Очевидно, для импульсных характеристик конечной длины это условие выполняется всегда, поэтому КИХ-цепи всегда устойчивы. Рекурсивные цепи могут быть неустойчивыми из-за наличия обратных связей. Анализ устойчивости ЛИС-цепей основан на использовании z-преобразования, которое формально может быть получено из преобразования Фурье (12.1) заменой величины $e^{j\omega}$ на комплексное переменное z:

$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n} . \qquad (12.9)$$

z-Преобразование может сходиться для одних значений комплексного переменного z и расходиться для других. Множество точек комплексной z-плоскости, в которых z-преобразование сходится, называется областью сходимости. Для абсолютно суммируемой импульсной характеристики область сходимости ее z-преобразования содержит единичную окружность. Если цепь является каузальной (физически реализуемой), то она устойчива в том и только в том случае, если все полюсы ее передаточной функции

$$H(z) = \sum_{n=0}^{\infty} h[n] z^{-n}$$

по модулю меньше единицы, т.е. находятся внутри единичной окружности.

Самый широкий класс ЛИС-цепей конечного порядка[6] образуют цепи, структура которых может быть сведена к каскадному соединению трансверсальной и рекурсивной частей, что соответствует разностному уравнению вида

$$y[n] = b_0 x[n] + b_1 x[n-1] + \ldots + b_{N-1} x[n-N+1] +$$

$$+ a_1 y[n-1] + a_2 y[n-2] + \ldots + a_{M-1} y[n-M+1] =$$

[6] ЛИС-цепи *бесконечного* порядка, очевидно, нереализуемы и представляют ограниченный интерес.

$$= \sum_{k=0}^{N-1} b_k x[n-k] + \sum_{r=1}^{M-1} a_r y[n-r], \qquad (12.10)$$

откуда следует выражение для КЧХ дробно-рационального вида

$$H(e^{j\omega}) = \frac{\sum_{k=0}^{N-1} b_k e^{-j\omega k}}{1 - \sum_{r=1}^{M-1} a_r e^{-j\omega r}}. \qquad (12.11)$$

В общем случае ЛИС-цепь конечного порядка с КЧХ вида (12.11) имеет бесконечно длинную импульсную характеристику (БИХ), но если полином-числитель делится на знаменатель без остатка, то результатом деления оказывается полином и импульсная характеристика имеет конечную длину (таковы, например, КИХ-фильтры на основе частотной выборки, см. ниже).

12.3. ДИСКРЕТНОЕ ПРЕОБРАЗОВАНИЕ ФУРЬЕ

Выражение (12.1) позволяет в принципе найти спектральную плотность последовательности бесконечной длины, но для этого должно быть известно ее описание в виде замкнутого выражения (формулы). В практике цифровой обработки сигналов чаще требуется вычислять спектральную плотность сигнала, заданного своими отсчетами (естественно, отсчетов может быть лишь конечное количество N). Тогда спектральная плотность согласно (12.1) имеет форму полинома порядка $N-1$:

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x[n]\, e^{-j\omega n}.$$

Известно, что для однозначного задания полинома порядка m относительно комплексного переменного достаточно задать его значения в $(m+1)$ точках комплексной плоскости, тогда полином может быть восстановлен при помощи интерполяционной формулы Лагранжа (см., например, [19]). Следовательно, для однозначного определения спектральной плотности сигнала длины N достаточно найти значения преобразования Фурье в N различных точках, которые можно выбрать произвольно. С точки зрения простоты

вычислений наилучший выбор состоит в размещении N точек равномерно на единичной окружности, так что $\omega_k = \frac{2\pi}{N}k$, $k = \overline{1, N-1}$ (рис. 12.4). Значения спектральной плотности в этих точках образуют последовательность длины N

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-i\frac{2\pi}{N}kn},\ k = \overline{1, N-1}. \qquad (12.12)$$

Это выражение носит название *дискретного преобразования Фурье* (ДПФ). Обратное ДПФ определяется выражением

$$x[n] = \frac{1}{N}\sum_{n=0}^{N-1} X[k]\, e^{i\frac{2\pi}{N}kn},\ n = \overline{1, N-1}. \qquad (12.13)$$

Нетрудно видеть, что в выражениях прямого и обратного ДПФ формально можно для переменных k и n задать бесконечные пределы, при этом левые части (12.12) и (12.13) определяют N -периодические последовательности. Таким образом, ДПФ связывает как последовательности конечной длины, так и периодические последовательности. Эта двойственность должна учитываться при применении ДПФ (см., например, п. 12.4.1).

Дискретное преобразование Фурье представляет собой не только инструмент анализа, но и алгоритм ЦОС. На его основе может быть реализована фильтрация сигналов в частотной области следующим образом: для входного сигнала вычисляется ДПФ, полученные спектральные отсчеты умножаются на КЧХ фильтра, а результат умножения подвергается обратному ДПФ. Этот метод фильтрации может быть более экономичным, чем вычисление свертки входного сигнала с импульсной характеристикой фильтра, благодаря существованию очень эффективных (быстрых) алгоритмов, которые получили название *быстрого преобразования Фурье* (БПФ).

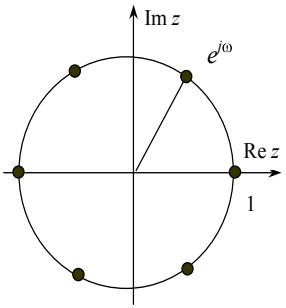


Рис. 12.4. Расположение точек вычисления ДПФ на 1-окружности для $N = 6$

12.4. ЦИФРОВЫЕ ФИЛЬТРЫ

Основное назначение дискретных ЛИС-цепей заключается в фильтрации дискретных сигналов, т.е. в избирательном воздействии на амплитуды и начальные фазы гармонических составляющих различных частот. Это фактически означает, что любая ЛИС-цепь представляет собой фильтр. Однако интерес представляет построение фильтров с *заданными* наперед частотно-избирательными и фазовыми свойствами. Построить (синтезировать) фильтр – значит найти его разностное уравнение (т.е. алгоритм вычисления выходного сигнала по известному входному) и/или структурную схему. Таким образом, под синтезом цифрового фильтра (ЦФ) обычно понимается построение дискретной ЛИС-цепи с КЧХ заданной формы. При решении задачи синтеза обычно не делают различия между дискретными и цифровым цепями, хотя, строго говоря, дискретная ЛИС-цепь становится цифровой в результате квантования коэффициентов ее разностного уравнения[7].

Ранее было показано, что ЛИС-цепь конечного порядка имеет в общем случае КЧХ дробно-рационального вида (12.11), поэтому, очевидно, задача синтеза ЦФ сводится к *аппроксимации* желаемой КЧХ функцией дробно-рационального вида, так как зная эту функцию, легко составить структурную схему цепи или записать разностное уравнение вида (12.10). Указанная аппроксимация сравнительно легко выполняется для КИХ-цепей, когда дробно-рациональная функция вырождается в полином, и представляет собой непростую задачу для общего случая. Поэтому методы синтеза ЦФ с конечными и бесконечными импульсными характеристиками совершенно различны.

12.4.1. МЕТОДЫ СИНТЕЗА КИХ-ФИЛЬТРОВ

Фильтры с конечной импульсной характеристикой имеют перед БИХ-фильтрами ряд преимуществ. Во-первых, КИХ-фильтры всегда устойчивы. Во-вторых, только КИХ-фильтр может иметь строго линейную фазочастотную характеристику [5, 19] (фильтр с линейной ФЧХ не искажает формы сигнала, если его спектр лежит в полосе частот, где амплитудно-частотная характеристика постоянна; при этом сигнал лишь задерживается на время, пропорциональное крутизне ФЧХ). Наконец, для КИХ-фильтров наиболее

[7] При этом квантуются также отсчеты сигнала, в результате чего цепь перестает быть *линейной*.

просто решается задача аппроксимации КЧХ желаемого вида реализуемой функцией (тригонометрическим полиномом). Однако КИХ-фильтры имеют существенный недостаток по сравнению с БИХ-фильтрами: для обеспечения сравнимых частотно-избирательных свойств, в частности крутизны АЧХ в переходной полосе частот, требуется КИХ-фильтр в десятки раз более высокого порядка, чем БИХ-фильтр. На практике в зависимости от конкретных обстоятельств применяются фильтры обоих типов. Ниже вкратце рассматриваются методы синтеза КИХ-фильтров.

Метод взвешивания (*метод функций окна*)

КЧХ трансверсального дискретного фильтра представляет собой тригонометрический полином, т.е. функцию вида

$$H(e^{j\omega}) = \sum_{n=-M}^{M} b_n e^{jn\omega} . \qquad (12.14)$$

Здесь не предполагается каузальность фильтра; если каузальность необходима, ее легко можно обеспечить умножением (12.14) на фазовый множитель $e^{-jM\omega}$. Если желаемая КЧХ имеет вид $H_{ж}(e^{j\omega})$, то синтез КИХ-фильтра состоит в нахождении тригонометрического полинома, близкого в каком-то смысле к $H_{ж}(e^{j\omega})$. Обычно в качестве критерия близости выбирается среднеквадратическая ошибка аппроксимации

$$\varepsilon = \int_{-\pi}^{\pi} \left| H(e^{j\omega}) - H_{ж}(e^{j\omega}) \right|^2 d\omega ,$$

тогда наилучшая аппроксимация обеспечивается, если коэффициентами полинома (12.14) являются коэффициенты разложения желаемой КЧХ в ряд Фурье

$$b_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_{ж}(e^{j\omega}) e^{j\omega n} d\omega , \; n = \overline{-M, M} . \qquad (12.15)$$

Эти коэффициенты представляют собой отсчеты импульсной характеристики КИХ-фильтра, в общем случае некаузального. После соответствующей задержки получается импульсная характеристика каузального фильтра $h[n] = b_{n-M}$, $n = \overline{0, N-1}$, где $N = 2M+1$. Поскольку всякая ЛИС-цепь однозначно определяется своей импульсной характеристикой, на этом синтез КИХ-фильтра можно было бы считать законченным. Однако если желаемая КЧХ раз-

рывна (например, как часто бывает на практике, требуется АЧХ прямоугольной формы), получаемая КЧХ, как сумма усеченного ряда Фурье (12.14), содержит гиббсовские осцилляции. Поэтому применяют дополнительное умножение импульсной характеристики на весовую последовательность («окно») подходящей формы. Причина явления Гиббса, как отмечалось в разд. 2, заключается в слишком медленном убывании коэффициентов фурье-разложения разрывной функции, поэтому все применяемые окна убывают от середины к краям [5, 19]. Для достижения приемлемых избирательных свойств длина импульсной характеристики, определяющая объем вычислений, на практике составляет обычно несколько сотен.

Кроме метода взвешивания, иногда применяют другой способ борьбы с гиббсовскими осцилляциями. На этапе формулирования требований к фильтру вводят переходную полосу, в которой задают закон *непрерывного* изменения АЧХ (например, линейный закон) [5]. Тогда ряд Фурье сходится равномерно и явление Гиббса отсутствует. Это не означает, что исчезает неравномерность АЧХ, просто осцилляции теперь убывают по амплитуде с увеличением порядка фильтра.

Следует также упомянуть машинные методы синтеза КИХ-фильтров на основе численной оптимизации. При этом подбором коэффициентов КИХ-фильтра минимизируется взвешенная среднеквадратическая ошибка

$$\varepsilon = \int_{-\pi}^{\pi} q(\omega)\left|H(e^{j\omega}) - H_{\text{ж}}(e^{j\omega})\right|^2 d\omega,$$

где $q(\omega)$ – весовая функция, позволяющая управлять относительной значимостью ошибок на разных участках частотной оси, или максимальная взвешенная погрешность

$$\varepsilon' = \max\left\{q(\omega)\left|H(e^{j\omega}) - H_{\text{ж}}(e^{j\omega})\right|\right\}.$$

Эти методы позволяют получить меньшие погрешности аппроксимации по сравнению с описанным выше методом оконного взвешивания, но их анализ значительно сложнее [5].

Метод частотной выборки

Метод синтеза фильтров с конечной импульсной характеристикой, получивший название метода частотной выборки, основан на задании значений желаемой КЧХ в точках, расположенных равно-

мерно на 1-окружности и соответствующих точкам частотной оси (отсюда название метода) и аппроксимации КЧХ интерполяционным полиномом Лагранжа [5]. Этот метод приводит к построению структуры, содержащей трансверсальную и рекурсивную части, которой, тем не менее, соответствует конечная импульсная характеристика (см. разд. 12.2). Благодаря наличию рекурсии такие фильтры при реализации требуют меньшего числа операций по сравнению с рассмотренными выше КИХ-фильтрами и оказываются предпочтительными.

Метод быстрой свертки

Фильтрация сигналов может быть выполнена в частотной области путем вычисления спектральной плотности входного сигнала, умножения ее на КЧХ фильтра и выполнения обратного преобразования Фурье (на практике для входного сигнала, который представляет собой реализацию случайного процесса, можно вычислить только *дискретное* преобразование Фурье). Этот на первый взгляд сложный способ нахождения выходного сигнала оказывается на практике более эффективным в вычислительном отношении, чем прямое вычисление свертки, благодаря существованию алгоритмов быстрого преобразования Фурье. Метод КИХ-фильтрации на основе БПФ получил название метода быстрой свертки.

При его реализации необходимо учитывать следующие два обстоятельства. Первое состоит в том, что дискретное преобразование Фурье обладает двойственностью – оно соответствует как последовательностям конечной длины, так и периодическим последовательностям. По этой причине перемножение коэффициентов ДПФ двух последовательностей (входного сигнала и импульсной характеристики) соответствует не обычной (апериодической), а так называемой циклической (круговой) свертке. Убедимся, что это действительно так.

Пусть $\tilde{x}[n]$ и $\tilde{h}[n]$ – периодические последовательности с периодом N. Их циклическая свертка определяется выражением

$$\tilde{y}[n] = \sum_{m=0}^{N-1} \tilde{x}[m]\tilde{h}[n-m] .$$

ДПФ результирующей последовательности

$$\tilde{Y}[k] = \sum_{n=0}^{N-1} \left(\sum_{m=0}^{N-1} \tilde{x}[m]\tilde{h}[n-m] \right) e^{-j\frac{2\pi}{N}nk} =$$

$$= \sum_{m=0}^{N-1} \tilde{x}[m] \left(\sum_{n=0}^{N-1} \tilde{h}[n-m] e^{-j\frac{2\pi}{N}(n-m)k} \right) e^{-j\frac{2\pi}{N}mk} = \tilde{H}[k]\tilde{X}[k], \quad (12.16)$$

$$\tilde{H}[k]\tilde{X}[k] = H[k]X[k],$$

где $\tilde{H}[k]$ и $\tilde{X}[k]$ находятся как ДПФ N-периодических последовательностей $\tilde{x}[n]$ и $\tilde{h}[n]$, а $H[k]$ и $X[k]$ – как ДПФ их конечных фрагментов длины N согласно (12.12).

При выводе (12.16) учтен тот факт, что сумма в круглых скобках во второй строке равна $\tilde{H}[k]$ независимо от m в силу периодичности суммируемых членов: при различных m суммируются одни и те же N слагаемых в разном порядке. Таким образом, видно, что циклическая свертка (или свертка периодических последовательностей) соответствует поточечному произведению ДПФ-спектров последовательностей. При фильтрации же должна выполняться обычная апериодическая свертка, определяемая выражением (12.5).

Преодолеть эту трудность можно следующим образом. Преобразование Фурье последовательности длины N дает полином (относительно $e^{-j\omega}$) степени $N-1$. Полагая, что $x[n]$ и $h[n]$ – последовательности длины N, видим, что произведение их фурье-образов есть полином степени $2(N-1)$. Но при поточечном перемножении ДПФ-спектров, имеющих по N отсчетов, получается всего N результирующих отсчетов, что соответствует полиному всего лишь $(N-1)$-й степени. Единственный способ получения правильного результата умножения двух полиномов состоит в том, чтобы точек вычисления ДПФ «хватило» для представления результата. Иначе говоря, если используется N-точечное ДПФ, то степень результирующего полинома должна быть не выше $(N-1)$. Это, в свою очередь, означает, что сумма длин последовательностей $x[n]$ и $h[n]$ (обозначим их M и L) должна удовлетворять очевидному соотношению $M-1+L-1 \le N-1$ (или $M+L-1 \le N$). Для того чтобы можно было для последовательности $x[n]$ длины $M < N$ получить N отсчетов ДПФ, следует перед вычислением ДПФ последовательность $x[n]$ дополнить $N-M$ нулевыми отсчетами (и то же проделать с другой последовательностью). Тогда ре-

зультат их циклической свертки, полученный применением БПФ, совпадает с результатом апериодической свертки.

Второе обстоятельство, которое должно учитываться при реализации КИХ-фильтрации методом быстрой свертки, относится к фильтрации последовательностей большой (в частности, бесконечной) длины. Если длина входной последовательности велика (сотни тысяч отсчетов и более), что типично для обработки сигналов, применяемых в радиотехнике и связи, то необходимое основание БПФ (число вычисляемых отсчетов) оказывается слишком большим, что влечет высокие требования к объему оперативной памяти вычислителя БПФ, а также приводит к большой задержке результирующего сигнала (результат может быть получен не ранее, чем поступит последний отсчет входной последовательности, плюс время, необходимое для вычисления прямого БПФ, умножения и обратного БПФ). Для того чтобы снизить требования к памяти и уменьшить задержку, применяют так называемое секционирование свертки [5].

Пусть $h[n]$ – импульсная характеристика фильтра, имеющая длину M, а $x[n]$ – сигнальная последовательность бесконечной длины. Представим $x[n]$ в виде

$$x[n]=\sum_{k=-\infty}^{\infty} x_k[n], \text{ где } x_k[n]=\begin{cases} x[n], & kL \le n < (k+1)L \\ 0, & \text{в противном случае.} \end{cases}$$

Нетрудно видеть, что таким образом входная последовательность разбивается (секционируется) на совокупность неперекрывающихся примыкающих друг к другу сегментов длиной L отсчетов каждый.

В силу билинейности свертки (линейности по каждому из операндов)

$$x[n]\otimes h[n]=\left(\sum_{k=-\infty}^{\infty} x_k[n]\right)\otimes h[n]=\sum_{k=-\infty}^{\infty}\left(x_k[n]\otimes h[n]\right)=\sum_{k=-\infty}^{\infty} y_k[n].$$

Как видно из полученного выражения, свертка последовательности бесконечной длины с конечной импульсной характеристикой может быть точно заменена бесконечной суммой сверток сегментов фиксированной длины с этой же ИХ. Каждая частичная свертка требует для своего вычисления $(L+M-1)$-точечного БПФ. Поскольку L выбирается произвольно, секционирование позволяет реализовать КИХ-фильтрацию длинных (потенциально – беско-

нечно длинных) последовательностей; при этом результат $y_k[n]$ фильтрации сегмента $x_k[n]$ появляется с задержкой, определяемой временем прямого и обратного БПФ и умножения, но эта задержка теперь отсчитывается от момента окончания сегмента, а не всей последовательности.

Заметим, что результат каждой частичной свертки имеет длину $(L+M-1)$, а следуют секции с относительным сдвигом L. Таким образом, результаты частичных сверток $y_k[n]$ суммируются с перекрытием носителей $kL \le n < (k+1)L+M-1$. Этот метод секционирования свертки известен как метод перекрытия с суммированием (overlap-add method) [5].

Альтернативный метод перекрытия с накоплением (overlap-save method) заключается в том, что перекрываются сегменты входной последовательности. Последовательность $x[n]$ разбивается на секции $x_k[n]$ длиной $(L+M-1)$, причем последние $(M-1)$ отсчетов каждой секции перекрываются с таким же количеством первых отсчетов следующей секции (M – по-прежнему длина импульсной характеристики, L выбирается произвольно). Каждая секция подвергается БПФ с основанием $(L+M-1)$. Импульсная характеристика длиной M дополняется нулями до основания БПФ. Результат фильтрации (циклической свертки) содержит всего $(L+M-1)$ отсчетов, из которых последние $(M-1)$ отсчетов – «ошибочные», не совпадающие с результатом апериодической свертки, а остальные L отсчетов являются «правильными». Суммируются результаты циклических частичных сверток после отбрасывания «ошибочных» отсчетов, т. е. секции выходной последовательности длиной L следуют друг за другом без перекрытия.

12.4.2. СИНТЕЗ БИХ-ФИЛЬТРОВ НА ОСНОВЕ АНАЛОГО-ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Сравнение БИХ-фильтров с КИХ-фильтрами показывает, что для получения примерно одинаковых частотно-избирательных свойств (имеется в виду крутизна спада АЧХ) КИХ-фильтр должен иметь в 10…20 раз более высокий порядок. Это вполне объяснимо, так как известно, что быстро изменяющиеся сигналы имеют широкий спектр, а благодаря двойственности (дуализму) времени и частоты отсюда следует, что круто изменяющейся функции частоты (АЧХ) должна соответствовать функция времени (импульсная характеристика) большой длительности. Но для КИХ-фильтра поря-

док – это количество отсчетов импульсной характеристики минус 1, в то время как БИХ-фильтр даже первого порядка имеет импульсную характеристику бесконечной длительности. Таким образом, в тех случаях, когда вид фазочастотной характеристики не играет определяющей роли для практического применения разрабатываемого фильтра, следует использовать БИХ-фильтр, так как при этом получается существенный выигрыш в быстродействии (или в аппаратурных затратах на реализацию) фильтра. То обстоятельство, что не всякий БИХ-фильтр оказывается устойчивым, не представляет такой заметной опасности, как может показаться на первый взгляд. Во-первых, устойчивость может быть проверена (и обеспечена) на этапе проектирования цифрового фильтра; во-вторых, характеристики цифровых фильтров не подвержены дрейфу с течением времени и при изменении внешних условий, следовательно, устойчивый фильтр останется устойчивым в течение всего времени работы (конечно, нужно учитывать возможность выхода аппаратуры из строя в результате *катастрофического отказа*; при этом может возникнуть неустойчивость). Следует также отметить, что в цифровых БИХ-фильтрах могут возникать незатухающие паразитные колебания (так называемые предельные циклы) вследствие своеобразного нарушения устойчивости при округлении дробных чисел; подробнее см. [5].

Наиболее широко применяются методы синтеза цифровых БИХ-фильтров, основанные на так называемой *аналого-цифровой трансформации*, т.е. на преобразовании аналогового фильтра с требуемыми характеристиками в цифровой (дискретный) фильтр. Это объясняется, во-первых, трудностью решения задачи прямой аппроксимации желаемых характеристик дробно-рациональными передаточными функциями и, во-вторых, наличием развитой теории синтеза аналоговых фильтров и простотой преобразования аналоговых фильтров-*прототипов* в дискретные фильтры.

В качестве фильтров-прототипов наиболее часто применяются аналоговые фильтры Баттерворта, Чебышёва, Золотарева – Кауэра (эллиптические) и Бесселя. Фильтры Баттерворта имеют при заданном порядке максимально гладкую АЧХ. Фильтры Чебышёва имеют АЧХ, пульсирующую в полосе пропускания (фильтры I рода) или в полосе заграждения (фильтры II рода). АЧХ эллиптического фильтра пульсирует как в полосе пропускания, так и в полосе заграждения и имеет поэтому максимальную крутизну спада. Все перечисленные фильтры характеризуются заметной нелинейностью фазочастотной характеристики. Фильтр Бесселя имеет ФЧХ, близкую к линейной в полосе пропускания.

Аналоговые фильтры принято описывать передаточными функциями, которые связаны с импульсными характеристиками преобразованием Лапласа [8]. Преобразование Лапласа связывает аналоговый сигнал $x(t)$ с его образом (*изображением*) в виде функции $X(p)$ комплексного переменного p. Мнимая ось комплексной p-плоскости представляет собой ось частот в описании аналогового сигнала. Аналогичную роль в описании дискретных сигналов играет единичная окружность z-плоскости.

Аналого-цифровая трансформация состоит в установлении связи комплексных переменных p и z. Выразив p в виде функции $p = f(z)$ и подставив в выражение передаточной функции $H_{\text{a}}(p)$ аналогового фильтра-прототипа, мы получили бы функцию комплексного переменного z, имеющую смысл передаточной функции $H(z)$ дискретного фильтра. Трудность состоит в том, что, во-первых, мнимая ось p-плоскости имеет бесконечную, а единичная окружность z-плоскости – конечную длину 2π. Во-вторых, реализуемы только ЛИС-цепи конечного порядка, поэтому подстановка $p = f(z)$ в дробно-рациональную функцию $H_{\text{a}}(p)$ должна давать также дробно-рациональную функцию.

Поскольку на единичной окружности $z = e^{j\omega}$, а при дискретизации должно обеспечиваться равенство $\omega = \Omega T_d$, связь комплексных переменных p и z, обусловленная дискретизацией аналоговых сигналов, описывается выражениями

$$z = e^{pT_d} \tag{12.17}$$

и

$$p = \frac{1}{T_d} \ln z, \tag{12.18}$$

которые не являются дробно-рациональными. Различные способы преодоления этой трудности лежат в основе двух рассмотренных ниже методов аналого-цифровой трансформации.

Метод инвариантности импульсной характеристики. Передаточная функция произвольного аналогового фильтра (с сосредоточенными параметрами) имеет вид дробно-рациональной функции комплексного переменного p. Такая функция может быть представлена в виде суммы дробей

$$H_{\text{a}}(p) = \sum_{k=1}^{N} \frac{A_k}{(p - p_k)},$$

где $p_k, k=\overline{1,N}$ – полюсы передаточной функции, а коэффициенты A_k определяются из условия равенства числителя передаточной функции $H_a(p)$ и числителя правой части после приведения ее к общему знаменателю. (Здесь мы рассматриваем лишь практически важный случай, когда степень числителя $H_{\text{a}}(p)$ меньше степени знаменателя, а все корни знаменателя некратные.)

Ввиду линейности преобразования Лапласа импульсная характеристика такого фильтра имеет вид суммы экспоненциальных функций непрерывного времени $h_{\text{a}}(t)=\sum_{k=1}^{N} A_k e^{p_k t}\cdot\sigma(t)$, где $\sigma(t)$ – функция Хевисайда, определяемая выражением (2.1).

Очевидно, для того чтобы импульсная характеристика затухала со временем (т. е. фильтр был устойчивым), необходимо и достаточно, чтобы все полюсы были расположены в p-плоскости слева от мнимой оси.

Метод аналого-цифровой трансформации, известный под названием метода *инвариантности импульсной характеристики*, основан на прямом применении теоремы отсчетов (теоремы Котельникова). Рассматривая импульсную характеристику аналогового фильтра-прототипа как функцию времени (сигнал), можно заменить ее последовательностью отсчетов, выбранных с достаточно малым шагом дискретизации T_d.

Результатом дискретизации импульсной характеристики аналогового фильтра будет последовательность

$$h[n]=h_a(nT_d)=\sum_{k=1}^{N} A_k e^{p_k nT_d}\cdot u[n]=\sum_{k=1}^{N} A_k r_k{}^n\cdot u[n],$$

где $r_k=e^{p_k T_d}$ – полюсы передаточной функции цифрового фильтра,

$$u[n]=\begin{cases}1, & n\ge 0,\\ 0, & n<0.\end{cases}$$

Из полученного выражения видно, что при дискретизации импульсной характеристики каузального аналогового фильтра с дробно-рациональной передаточной функцией получается сумма каузальных экспоненциальных последовательностей, следовательно, реализуемому аналоговому фильтру соответствует реализуемый цифровой фильтр. Кроме того, полюсы цифрового фильтра связаны с полюсами фильтра-прототипа соотношением

$r_k = e^{p_k T_d}$, $k = \overline{1, N}$, поэтому устойчивому аналоговому фильтру ($\mathrm{Re}\{p_k\} < 0$) соответствует устойчивый цифровой фильтр того же порядка N (так как $|r_k| = \left|e^{p_k T_d}\right| < 1$). Зная полюсы цифрового фильтра, можно сразу записать его передаточную функцию

$$H(z) = \sum_{k=1}^{N} \frac{A_k}{(1 - r_k z^{-1})},$$

и на этом аналого-цифровая трансформация заканчивается, так как зная передаточную функцию, легко составить структурную схему и разностное уравнение цифрового фильтра.

Поскольку импульсная характеристика цифрового фильтра есть продукт дискретизации импульсной характеристики аналогового фильтра, КЧХ цифрового фильтра связана с КЧХ аналогового фильтра соотношением (см. разд. 2.11)

$$H(e^{j\omega}) = H(e^{j\Omega T_d}) = \frac{1}{T_d} \sum_{k=-\infty}^{\infty} H_{\mathrm{a}}\left[j\left(\Omega + k\Omega_d\right)\right], \qquad (12.19)$$

где $\Omega_d = 2\pi / T_d$ – частота дискретизации. Очевидно, если КЧХ прототипа не финитна, а это всегда так для фильтров конечного порядка, то неизбежно наложение (суммирование) «хвостов» сдвинутых копий $H_{\mathrm{a}}(\cdot)$, и, как следствие, искажение формы получаемой КЧХ дискретного фильтра по отношению к КЧХ фильтра-прототипа. Этот эффект ограничивает практическое применение метода инвариантности импульсной характеристики в основном задачами синтеза цифровых фильтров нижних частот.

Метод билинейного преобразования основан на аппроксимации выражения (12.18), позволяющей сохранить дробную рациональность передаточной функции. Подставив разложение функции логарифма в ряд, ограниченное первым слагаемым, получим

$$p = \frac{2}{T_d}\left(\frac{z-1}{z+1} + \frac{(z-1)^3}{3(z+1)^3} + \frac{(z-1)^5}{5(z+1)^5} + \ldots \right) \approx \frac{2}{T_d}\frac{z-1}{z+1},$$

или

$$p = \frac{2}{T_d} \cdot \frac{1 - z^{-1}}{1 + z^{-1}}. \qquad (12.20)$$

Это выражение дробно-рационально относительно z^{-1}, поэтому после его подстановки в дробно-рациональную передаточную

функцию $H_{\text{a}}(p)$ аналогового прототипа получается снова дробно-рациональная, а следовательно, реализуемая передаточная функция $H(z)$ цифрового фильтра.

Выясним, как располагаются в z-плоскости полюсы передаточной функции $H(z)$, если полюсы передаточной функции прототипа $H_{\text{a}}(p)$ находятся в левой части комплексной плоскости (иными словами, является ли устойчивым цифровой фильтр, если устойчив фильтр-прототип).

Выразим на основе (12.20) комплексное переменное z через p:

$$z^{-1}pT_d + pT_d = 2 - 2z^{-1};$$

отсюда

$$z = \frac{2 + pT_d}{2 - pT_d}. \qquad (12.21)$$

Чтобы выяснить, в какое множество z-плоскости отображается мнимая ось p-плоскости (ось частоты), подставим в это выражение $j\Omega$ вместо p, тогда получим выражение для образа мнимой оси p-плоскости при отображении, описываемом выражением (12.21):

$$z = \frac{2/T_d + j\Omega}{2/T_d - j\Omega}.$$

Числитель и знаменатель этой дроби суть комплексно-сопряженные числа, поэтому модуль дроби равен единице при всех Ω. Это означает, что мнимая ось p-плоскости отображается преобразованием (12.21) на единичную окружность z-плоскости. Но переменная Ω – это частота, соответствующая описанию аналогового фильтра; роль частотной оси для цифровых цепей играет единичная окружность на z-плоскости (множество точек $e^{j\omega}$ при значениях ω, принимающих значения из интервала от $-\pi$ до π). Заменяя z на $e^{j\omega}$, получим

$$e^{j\omega} = \frac{2/T_d + j\Omega}{2/T_d - j\Omega} = \frac{1 + j\Omega T_d/2}{1 - j\Omega T_d/2};$$

тогда

$$\omega = \arg\left\{\frac{1 + j\Omega T_d/2}{1 - j\Omega T_d/2}\right\} = 2\,\text{arctg}\,\frac{\Omega T_d}{2},$$

следовательно, связь «аналоговой» и «цифровой» частот при билинейном преобразовании описывается выражениями

$$\frac{\omega}{2} = \operatorname{arctg}\frac{\Omega T_d}{2},$$

$$\frac{\Omega T_d}{2} = \operatorname{tg}\frac{\omega}{2}. \qquad (12.22)$$

Поскольку $\pi \le \omega \le \pi$, а $-\infty < \Omega < \infty$, нетрудно видеть, что вся аналоговая частотная ось (бесконечной длины) отображается на единичную окружность (длины 2π), причем это отображение *однократно* в отличие от (12.19), и вследствие этого различные участки оси Ω испытывают различное «сжатие» при отображении на ось ω (единичную окружность). Это необходимо учитывать при проектировании цифровых фильтров на этапе определения требований к частотам среза фильтров-прототипов.

Заслуживает внимания вопрос, насколько вредна нелинейная трансформация частотной оси с точки зрения задачи синтеза цифровых фильтров. Очевидно, что при проектировании фильтров с желаемыми АЧХ *кусочно-постоянного* вида указанная нелинейность трансформации частотной оси не влияет на качество цифрового фильтра, так как приводит лишь к необходимости на этапе построения аналогового фильтра-прототипа учитывать последующее изменение характерных частот фильтра (граничных частот) при билинейном преобразовании. Если же требуется построить фильтр, не являющийся типовым (ФНЧ, ФВЧ, полосовым или режекторным), то в общем случае нелинейность отображения частотной оси приводит к искажению формы АЧХ. Например, интегрирующий аналоговый фильтр имеет амплитудно-частотную характеристику гиперболического вида $\sim 1/\Omega$ и при билинейном преобразовании не приводит к интегрирующему цифровому фильтру.

Для того чтобы устойчивый аналоговый фильтр трансформировался в устойчивый же цифровой фильтр, требуется, чтобы при билинейном преобразовании левая полуплоскость p-плоскости отображалась внутрь единичной окружности. Разлагая p на мнимую и вещественную части, получим для билинейного преобразования (12.21)

$$z = \frac{2 + \operatorname{Re}\{pT_d\} + j\operatorname{Im}\{pT_d\}}{2 - \operatorname{Re}\{pT_d\} - j\operatorname{Im}\{pT_d\}}.$$

Поскольку мнимые части числителя и знаменателя одинаковы, модуль дроби будет меньше 1, если $\operatorname{Re}\{pT_d\} < 0$. Тогда любой полюс функции $H_{\text{a}}(p)$, лежащий *слева* от мнимой оси, отобража-

ется в полюс функции $H(z)$, расположенный *внутри* 1-окружности. Это означает, что устойчивый аналоговый фильтр трансформируется преобразованием (12.20) в устойчивый дискретный фильтр.

Итак, установлено, что билинейное преобразование трансформирует устойчивый реализуемый аналоговый фильтр в устойчивый реализуемый цифровой фильтр. При этом вследствие однократности отображения частотной оси на 1-окружность отсутствует наложение «хвостов» КЧХ, что является достоинством билинейного преобразования. К недостаткам следует отнести то, что не сохраняются ни импульсная, ни фазочастотная характеристики фильтра (точнее говоря, импульсная и фазочастотная характеристики дискретного фильтра могут *сильно отличаться по форме* от соответствующих характеристик прототипа).

Порядок расчета цифрового фильтра методом билинейного преобразования состоит в следующем:

1) определение характерных частот ЦФ и пересчет их в частоты аналогового фильтра в соответствии с (12.22);

2) синтез аналогового фильтра, удовлетворяющего заданным условиям;

3) подстановка формулы (12.20) билинейного преобразования в выражение $H_{\text{a}}(p)$ передаточной функции фильтра-прототипа.

Реализация цифровых фильтров (и других алгоритмов цифровой обработки сигналов) возможна на различной элементной базе. Выбор конкретного воплощения алгоритма ЦОС производится разработчиком с учетом различных показателей, к которым относятся стоимость, массогабаритные характеристики, энергопотребление, быстродействие и т.п. В каждом конкретном случае один или несколько показателей играют наиболее важную роль в выборе способа реализации. Например, в системах *подвижной радиосвязи* главными показателями являются быстродействие (обработка должна выполняться в реальном времени) и массогабаритные характеристики, при этом желательно обеспечить малое энергопотребление и умеренную цену мобильной станции. Устройство обработки сигналов в таких системах работает по жестким алгоритмам, которые не изменяются в течение всего срока эксплуатации изделия; точность представления данных (разрядность) должна быть достаточна для обеспечения комфортности восприятия речи (разборчивости и возможности узнавания собеседника) и является поэтому сравнительно невысокой. В системах обработки геофизической информации, напротив, обработка в реальном времени не требуется, массогабаритные характеристики не играют доминирующей роли, однако на передний план выступают точ-

ность представления данных и гибкость реализуемых алгоритмов. Поэтому в таких системах обработка сигналов реализуется обычно на универсальных цифровых вычислительных машинах. В системах реального времени (например, в подвижной радиосвязи), когда на обработку отсчета сигнала отводится временной интервал, равный шагу дискретизации, как правило, используются специализированные цифровые устройства, называемые *цифровыми сигнальными процессорами* (*ЦСП*). Следует отметить, что в последнее время в устройствах цифровой обработки сигналов широкое распространение получили *программируемые логические интегральные схемы* (ПЛИС). ПЛИС представляет собой интегральную схему сверхвысокой степени интеграции, содержащую на кристалле порядка 1 миллиона логических вентилей, которые могут быть *программным* путем соединены в логическую структуру, реализующую заданный алгоритм цифровой обработки сигналов *аппаратным способом*. Таким образом, ПЛИС сочетают в себе преимущества аппаратной реализации алгоритмов (главное из которых – быстродействие) с достоинствами программируемых устройств.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем состоят преимущества цифровой обработки сигналов перед аналоговой?
2. В чем заключается отличие дискретных и цифровых сигналов?
3. Могут ли быть неустойчивыми трансверсальные цепи? рекурсивные цепи?
4. На чем основан синтез цифровых фильтров методом инвариантности импульсной характеристики? методом билинейного преобразования?
5. Зачем применяются окна при синтезе цифровых фильтров?

УПРАЖНЕНИЯ

1. Требуется построить «гребенку» фильтров для частотного разделения аналоговых сигналов, занимающих полосы частот (0...800 Гц), (1000...1800 Гц) и (2000...2800 Гц). Выберите частоту дискретизации и рассчитайте граничные частоты цифровых фильтров. Постройте график АЧХ фильтров (качественно).
2. На основании данных предыдущего упражнения рассчитайте граничные частоты аналоговых фильтров-прототипов для последующего синтеза ЦФ методом билинейного преобразования. Постройте качественный график АЧХ фильтров-прототипов.

13. ОСНОВЫ КРИПТОЗАЩИТЫ СООБЩЕНИЙ В СИСТЕМАХ СВЯЗИ

По мере развития и все более широкого распространения средств связи возрастает роль систем повышения информационной безопасности. Ценность информации, передаваемой по каналам и сетям связи, повышается, поэтому растет и интерес отдельных лиц и сообществ к ее несанкционированному (незаконному) получению, фальсификации, уничтожению и т.п. Растет количество «взломов» информационных систем банков, крупных фирм, военных ведомств. В отдельных случаях эти действия угрожают самому существованию человечества (например, известны случаи проникновения посторонних лиц в информационные системы, связанные с запуском стратегических ракет). Поэтому существует необходимость создания систем передачи информации, защищенных от незаконного вмешательства.

Несанкционированный сбор информации конфиденциального характера – реальность, и было бы проявлением неоправданного оптимизма не придавать ему значения. Методы сбора информации весьма разнообразны [25]; перечислим лишь некоторые из них:

- акустический контроль помещений, автомобилей, людей;
- прослушивание телефонных каналов, перехват факсовой и модемной передачи сообщений;
- перехват компьютерной информации (при передаче по сетям, путем анализа радиоизлучений компьютера, путем внедрения в локальные сети, серверы, базы данных);
- скрытая фото- и видеосъемка, визуальное наблюдение;
- кража документов и носителей информации;
- подкуп и шантаж сотрудников, знакомых, родственников интересующего лица и т.д.

В соответствии с большим разнообразием способов незаконного получения информации существует и достаточно широкий

спектр методов ее защиты – от простого экранирования линий связи до шифрования сообщений. В этом разделе будут кратко рассмотрены криптографические методы защиты информации.

13.1. ОСНОВНЫЕ ПОНЯТИЯ КРИПТОГРАФИИ

Информация, которая нуждается в защите, называется защищаемой (приватной, конфиденциальной, секретной). Принято говорить, что такая информация содержит тайну (государственную, коммерческую, врачебную, личную, тайну следствия и т.п.). Во всех случаях имеется круг законных пользователей информации. Кроме того, могут существовать люди или группы людей, которые заинтересованы в том, чтобы использовать информацию в своих целях, а значит – во вред законным пользователям. Для краткости эти незаконные пользователи называются *противниками*. Действия противника представляют *угрозу* для законного пользователя (угрозу *разглашения*, *подмены*, *имитации, уничтожения информации* и т.п.).

Возможны три подхода к защите информации [28]:

- создать абсолютно надежный канал связи, совершенно недоступный для незаконных пользователей;
- передавать информацию по общедоступному каналу, сохраняя в тайне сам факт передачи;
- передавать информацию по общедоступному каналу в такой форме, чтобы воспользоваться ею мог только адресат.

Первый способ при современном развитии техники практически нереализуем или сопряжен с чрезвычайно большими затратами.

Реализацией второго способа занимается так называемая *стеганография*. Известно немало примеров стеганографического сокрытия информации: от нанесения сообщения на обритую голову раба, которого посылали к адресату после того, как волосы отрастали, до использования симпатических чернил, которыми секретное сообщение вписывается между строк другого сообщения. В настоящее время используется, например, прием, состоящий в выборочной записи символов сообщения в младшие разряды цифрового изображения – на качество его визуального восприятия эти искажения практически не влияют, а найти «спрятанные» данные в больших массивах, которыми представляются изображения, не так легко.

Третий способ – *криптографическая* защита информации путем шифрования – в настоящее время является одним из самых

эффективных. Следует различать понятия кода и шифра[8]. Задачей криптографии является разработка методов преобразования сообщений, затрудняющих (в идеале – исключающих) извлечение противником информации из перехватываемых сообщений. По открытому каналу связи при этом передается *криптограмма* (*шифртекст*) – результат преобразования сообщения с помощью *шифра* (криптографического алгоритма). Наблюдение криптограммы для противника, не имеющего *ключа*, является бесполезным с точки зрения получения информации.

Зашифрованием называется процесс преобразования открытого сообщения в зашифрованное (криптограмму) с помощью шифра, а *расшифрованием* – обратный процесс преобразования зашифрованных (закрытых) данных в открытые с помощью шифра (*шифрование* – общее наименование для зашифрования и расшифрования). *Дешифрованием* (*вскрытием, взломом*) называют процесс преобразования зашифрованных данных в открытые при неизвестном (частично или полностью) шифре. Решением всех перечисленных задач занимается наука, называемая *криптологией*. Отрасль криптологии, занимающаяся разработкой методов шифрования, называется *криптографией*, а отрасль, занимающаяся разработкой методов взлома – *криптоанализом*.

Помимо получения информации из перехваченного сообщения противник может преследовать и другие цели, например, он может попытаться изменить содержание сообщения. В защищенной системе, таким образом, получатель сообщения должен иметь возможность проверить его подлинность и целостность. Защита от навязывания ложных данных называется *имитозащитой*. Для этого к криптограмме добавляется *имитовставка,* представляющая собой последовательность данных фиксированной длины, полученную по определенному алгоритму из открытых данных и ключа. Получатель зашифрованного сообщения может проверить, соответствует ли имитовставка содержанию расшифрованного сообщения.

Криптостойкостью называется характеристика шифра, определяющая его способность противостоять дешифрованию. Обычно

[8] В ранних литературных источниках понятия кода и шифра часто смешивались, однако в последнее время под кодированием понимается такое преобразование, которое преследует цели сжатия данных или повышения помехоустойчивости передачи. Коды и алгоритмы декодирования являются общеизвестными, в отличие от шифров, которые либо неизвестны противнику, либо известны лишь частично (например, шифры с открытым ключом).

количественной мерой криптостойкости считается время, необходимое для дешифрования [25].

Проблема криптозащиты существует многие сотни лет, что естественно, если вспомнить, что у любого человека есть свои секреты. Математические основы криптографии были разработаны К. Шенноном [26]. В общем случае секретная система связи может быть представлена структурной схемой, показанной на рис. 13.1.

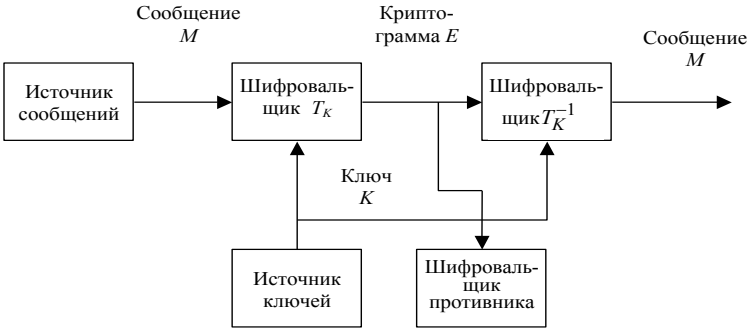


Рис. 13.1. Структура секретной системы

Сообщение M , вырабатываемое источником сообщений, поступает к шифровальщику. Шифровальщиком может быть как человек, так и устройство. От источника ключей к шифровальщику поступает ключ K . Шифровальщик на основе ключа осуществляет преобразование сообщения в криптограмму

$$E = f(M, K) .$$

Удобнее понимать это выражение не как функцию двух переменных, а как отображение

$$E = T_K\{M\},$$

при этом ключ играет роль *параметра* этого отображения. Выбранный ключ должен быть каким-то образом передан (сообщен) шифровальщику на приемной стороне системы, который при помощи ключа расшифровывает сообщение. Чтобы секретная система была пригодной для передачи информации, необходимо иметь возможность восстановить по криптограмме и ключу исходное сообщение, т.е. должно существовать (единственное) обратное отображение, так что

$$M = T_K^{-1}\{E\} = T_K^{-1}\left\{T_K\left[M\right]\right\},$$

$$T_K^{-1} T_K = I,$$

где I – тождественное отображение.

Предполагается, что множество всех сообщений конечно и сообщения $M_1, \ldots M_n$ вырабатываются источником с вероятностями $q_1, \ldots, q_n$. Аналогично источник ключей вырабатывает ключи $K_1, \ldots, K_m$ с вероятностями $p_1, \ldots, p_m$. В современной криптографии принято считать, что противнику известно множество отображений T_{K_i}, $i = \overline{1,m}$ и набор априорных вероятностей $p_1, \ldots, p_m$. Это предположение является пессимистическим, но, в конечном счете, реалистическим, так как рано или поздно всякая секретная система может быть раскрыта [26]. Если противник не сможет взломать алгоритм шифрования, даже зная, как он устроен, то тем более это окажется ему не по силам без такого знания.

В настоящее время используются следующие криптографические методы защиты [25]:

- методы замены (подстановки);
- методы на основе датчика псевдослучайных чисел;
- методы перемешивания (алгоритмические);
- системы с открытым ключом.

13.2. МЕТОДЫ ЗАМЕНЫ

Рассмотрим несколько примеров шифров замены. В примерах используются буквы русского алфавита (без буквы «ё») плюс пробел, обозначаемый символом _. Для удобства приведем русский алфавит в виде таблицы, содержащей также номера букв.

Буква	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л
№	01	02	03	04	05	06	07	08	09	10	11	12
Буква	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч
№	13	14	15	16	17	18	19	20	21	22	23	24
Буква	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	_			
№	25	26	27	28	29	30	31	32	33			

13.2.1. ШИФР ПРОСТОЙ ПОДСТАНОВКИ

В основе шифрования этим методом лежит алгебраическая операция над алфавитом сообщения, называемая *подстановкой*.

Подстановкой называется взаимно однозначное отображение конечного множества M *на себя*. Это означает, что каждому элементу множества (например, символу алфавита) ставится в соответствие один и только один элемент этого же множества. Очевидно, для задания конкретной подстановки природа элементов не имеет значения, важно лишь их количество N (называемое степенью подстановки). Можно поэтому в качестве множества рассматривать множество целых чисел $M = \{1, 2, 3, ..., N\}$.

Любая подстановка может быть описана матрицей $2\times N$, в которой первая строка содержит числа 1, 2, …, N, а вторая строка состоит из тех же чисел, расположенных в порядке, определяемом подстановкой, например, если число i после подстановки получает значение k_i, то можно записать

$$S = \begin{pmatrix} 1 & 2 & . & . & . & N \\ k_1 & k_2 & . & . & . & k_N \end{pmatrix},$$

где $k_i \in M \quad \forall i$.

Очевидно, если последовательно применить две подстановки S_1 и S_2, то их результат снова будет подстановкой, равной композиции (произведению) исходных подстановок $S = S_1 \times S_2$. Множество подстановок одинаковой степени образует *группу*[9] относительно «умножения», определенного таким образом. Отсюда следует, что для каждой подстановки существует обратная. Нейтральным элементом группы подстановок служит тривиальная подстановка, оставляющая на месте все элементы множества.

Пусть открытый текст представляет собой фразу «открытый текст», а подстановка задана таблицей

$$S = \begin{pmatrix} 1 & 2 & 3 & . & . & 32 & 33 \\ 33 & 32 & 31 & . & . & 2 & 1 \end{pmatrix}.$$

Шифртекст имеет вид «тоцреоечаоыцпо». Очевидно, что ключом к этому шифру является сама подстановка S. Количество всевозможных подстановок для алфавита из N символов равно $N!$,

[9] Определение группы см. в разд. 2.

что делает раскрытие ключа методом полного перебора проблематичным. На первый взгляд, криптостойкость такого шифра должна быть очень высокой. Однако недостатком шифра простой подстановки является то, что статистические свойства текста (частóты появления букв) при шифровании сохраняются, благодаря чему криптостойкость шифра на самом деле низка. Если в распоряжении дешифровальщика окажется достаточно длинный шифртекст (несколько десятков знаков), то шифр может быть взломан за несколько минут путем подсчета частот букв и сравнения с известными статистическими характеристиками русского (английского и т.д.) текста.

Частным случаем шифра замены является известный *шифр Цезаря*, состоящий в замене каждой буквы сообщения на букву, отстоящую от нее в алфавите на фиксированное число шагов. Например, буква «а» заменяется на «д», «б» на «е», «в» на «ж» и так далее; буквы «ю» и «я» заменяются на «в» и «г» соответственно (пробел не учитывается).

13.2.2. ШИФР ВИЖЕНЕРА

Этот шифр, опубликованный в 1586 г., определяется формулой

$$y_i = x_i + k_i (\mathrm{mod}\, N),$$

где x_i – номер буквы открытого текста в алфавите; y_i – номер буквы шифртекста; k_i – номер i-й буквы ключа (ключом является слово или любая последовательность букв длины d, повторяемая столько раз, сколько требуется для зашифрования всего сообщения). Число d называется периодом шифра Виженера. Очевидно, упомянутый выше шифр Цезаря является шифром Виженера при $d = 1$. Вариант шифра Виженера был реализован в механической шифровальной роторной машине [29], изобретенной в 20-х годах XX в.

13.2.3. ШИФРЫ БОФОРА

Эти шифры определяются формулами

$$y_i = k_i - x_i (\mathrm{mod}\, N)$$

и

$$y_i = x_i - k_i (\mathrm{mod}\, N),$$

где смысл обозначений такой же, как в шифре Виженера.

В рассмотренных шифрах, очевидно, криптостойкость тем выше, чем больше длина ключа. Поэтому дальнейшим развитием шифра Виженера является шифр с автоключом, когда зашифрование текста (и расшифрование криптограммы) начинается с некоторого ключа, называемого первичным, а затем к ключу дописывается либо открытый текст, либо шифртекст.

13.3. МЕТОДЫ ШИФРОВАНИЯ НА ОСНОВЕ ДАТЧИКА ПСЕВДОСЛУЧАЙНЫХ ЧИСЕЛ

Принцип шифрования заключается в генерировании псевдослучайной последовательности, называемой *гаммой,* и наложении гаммы на открытое сообщение некоторым обратимым образом. Например, если открытые данные и гамма представлены в двоичном коде, то подходящей является операция сложения по модулю 2; если исходным является русский текст, то в качестве гаммы можно использовать последовательность независимых псевдослучайных целых чисел от 1 до 33 с равными вероятностями, а операция наложения будет заключаться в сложении номеров букв с числами гаммы по модулю 33 и т.д. Расшифрование требует повторного генерирования гаммы и применения обратной операции (в рассмотренных случаях вычитания по соответствующему модулю). Ключом в данном случае может служить некоторый параметр датчика псевдослучайных чисел.

Линейный датчик псевдослучайных чисел реализуется на цифровой вычислительной машине и описывается рекуррентной формулой

$$g_{i+1} = [Ag_i + C] \bmod M,$$

где g_i, $i = 1, 2, 3, \ldots$ – последовательность псевдослучайных чисел, g_0 – стартовое значение, A и C – некоторые константы. Обычно принимается $M = 2^b$, где b – разрядность машины.

Если гамма имеет период больший, чем длина зашифрованного сообщения, и если не известна никакая часть исходного текста, то этот шифр можно раскрыть только прямым перебором возможных ключей, поэтому криптостойкость определяется размером ключа [25].

Шифрование на основе датчиков псевдослучайных чисел наиболее часто применяется в программной реализации криптозащиты данных, так как, с одной стороны, он достаточно прост для про-

граммирования, а с другой стороны, обладает высокой криптостойкостью.

Одной из форм реализации описанного метода является метод «одноразового блокнота». Суть этого метода состоит в том, что у шифровальщиков на передающей и приемной сторонах имеются два идентичных блокнота с *одинаковой* гаммой, содержащей случайные независимые равновероятные числа. Эта гамма, являющаяся ключом, используется только *один* раз, после чего соответствующий лист блокнота вырывается и уничтожается. Если эти условия соблюдены, шифр является *абсолютно* криптостойким, т.е. не может быть взломан в принципе [29]. Однако этот метод очень сложно реализовать, так как трудно снабдить всех возможных получателей шифртекстов идентичными шифровальными блокнотами и обеспечить их однократное использование. Абсолютно стойкие шифры применяются только в системах секретной связи с небольшим объемом передаваемой информации, обычно для передачи особо важной государственной информации [27].

13.4. МЕТОДЫ ПЕРЕМЕШИВАНИЯ

Методы перемешивания основаны на работе К. Шеннона [26], в которой задачи криптологии рассматривались с информационной точки зрения. Прежде чем перейти к конкретным системам шифрования на основе перемешивания, приведем основные положения работы Шеннона.

Основные вопросы, которые представляют интерес с теоретической точки зрения, состоят в следующем. Насколько устойчива система, если шифровальщик противника располагает всеми необходимыми средствами для криптоанализа и неограниченным временем? Имеет ли криптограмма *единственное* решение (даже если оно может быть найдено за практически неприемлемое время), а если нет, то сколько решений она имеет? Какой объем шифрованного текста нужно перехватить, чтобы решение стало единственным? Существуют ли секретные системы, для которых нельзя найти единственное решение независимо от объема перехваченной криптограммы? Существуют ли системы, в которых противник вообще не получает никакой информации независимо от объема перехваченного шифртекста? Рассмотрение этих вопросов основывается на понятиях теории информации.

Пусть имеется конечное множество сообщений $M_1, \ldots M_n$ с априорными вероятностями $P(M_1)$, $P(M_2)$, …, $P(M_n)$ и эти со-

общения преобразуются в криптограммы E_1, E_2, …, E_n, так что $E = T_i\{M\}$.

После перехвата криптограммы E шифровальщик противника может вычислить условные (апостериорные) вероятности различных сообщений $P_E(M) = P(M \mid E)$. Тогда *совершенная* секретная система удовлетворяет следующему условию: для всех E все апостериорные вероятности равны априорным

$$P_E(M_i) = P(M_i) \quad \forall i \quad \forall E .$$

В этом и только в этом случае перехват шифртекста не дает противнику никакой информации.

По теореме Байеса

$$P_E(M) = \frac{P(M)P_M(E)}{P(E)}, \tag{13.1}$$

где $P(M)$ – априорная вероятность сообщения M; $P_M(E)$ – условная вероятность криптограммы E при условии, что выбрано сообщение M; $P(E)$ – вероятность получения криптограммы E; $P_E(M)$ – апостериорная вероятность сообщения M при условии, что перехвачена криптограмма E.

Из выражения (13.1) следует, что для совершенной секретности необходимо и достаточно, чтобы выполнялось равенство

$$P_M(E) = P(E)$$

для любых сообщения M и криптограммы E.

Очевидно, различных криптограмм должно быть не меньше, чем различных сообщений, так как при любом ключе сообщению M должна однозначно соответствовать криптограмма E, которой, в свою очередь, однозначно соответствует сообщение M. Кроме того, число различных ключей должно быть также не меньше, чем различных сообщений. Например, совершенная система получается, если число сообщений равно числу ключей и числу криптограмм, причем все ключи равновероятны [26].

В терминах теории информации секретная система содержит два источника неопределенности: источник сообщений[10] с энтропией

$$H(M) = -\sum_{i=1}^{n} P(M_i)\log P(M_i)$$

[10] Здесь, очевидно, имеется в виду *конечное* число сообщений.

и источник ключей с энтропией

$$H(K) = -\sum_{j=1}^{m} P(K_j) \log P(K_j).$$

Количество информации источника не может быть больше, чем $\log n$ (это количество соответствует равновероятным сообщениям). Эта информация может быть полностью скрыта, только если неопределенность ключа не меньше $\log n$. Таким образом, неопределенность источника ключей устанавливает предел – максимальное количество информации, которое может быть скрыто при помощи ключей данного источника.

Если источник порождает последовательности бесконечной длины, то никакой конечный ключ не дает совершенной секретности. Пусть длина сообщения равна N, а длина ключа M. Тогда для совершенной секретности требуется

$$N \log n \le M \log m,$$

где n и m – объемы алфавитов для сообщений и ключей соответственно.

Совершенные системы применяются на практике для засекречивания сравнительно коротких сообщений и в тех случаях, когда полной секретности придается чрезвычайное значение.

В качестве теоретической меры секретности Шеннон предложил использовать ненадежность[11]. Имеются две ненадежности: ненадежность ключа

$$H_E(K) = H(K \mid E) = -\sum_{E,K} P(E,K) \log P(K \mid E) \qquad (13.2)$$

и ненадежность сообщения

$$H_E(M) = H(M \mid E) = -\sum_{E,M} P(E,M) \log P(M \mid E), \qquad (13.3)$$

где $P(E,M)$ – совместная вероятность криптограммы E и сообщения M, $P(E,K)$ – совместная вероятность криптограммы E и ключа K, $P(M \mid E)$ и $P(K \mid E)$ – апостериорные вероятности для сообщения и ключа при условии перехвата криптограммы E.

[11] Ненадежностью называется условная энтропия (см. разд. 8.2).

Суммирование в (13.2) и (13.3) проводится по всем возможным криптограммам и ключам (соответственно по всем возможным криптограммам и сообщениям) определенной длины, поэтому ненадежности зависят от числа N перехваченных букв криптограммы. Постепенное убывание ненадежностей с ростом N соответствует увеличению сведений о ключе и сообщении, имеющихся у шифровальщика противника. Если ненадежность становится нулевой, это означает, что единственное сообщение (или единственный ключ) имеет апостериорную вероятность 1, а все остальные – нулевые апостериорные вероятности. Шеннон доказал, что в принципе можно построить систему, в которой ненадежности не стремятся к нулю при $N \to \infty$, и даже «строго идеальную систему», в которой $H_E(K) = H(K)$. Для сообщений на естественных языках характерны неравновероятность и зависимость букв исходного текста, поэтому для них построение идеальной системы хотя и возможно, но осуществить его практически очень сложно [26].

Статистическая зависимость символов (букв) естественного языка позволяет раскрывать многие типы шифров. Например, шифр Цезаря (и вообще шифры на основе подстановок) раскрывается на основе простого подсчета частот встречаемости различных букв в криптограмме и сравнения их с известными для данного языка частотами. Для некоторых шифров могут потребоваться более тонкие методы статистического анализа. В любом случае статистический анализ криптограмм строится на основе некоторых *статистик*. Под статистикой понимается функция наблюдаемой криптограммы, значение которой позволяет сделать какие-либо более или менее правдоподобные выводы относительно использованного ключа. Удачный выбор подходящих статистик определяет успех криптоаналитика при взломе шифра, многократно сужая возможный круг ключей до того предела, когда уже можно установить правильный ключ путем перебора.

Для того чтобы затруднить раскрытие шифров статистическими методами, Шеннон предложил использовать две идеи, которые он назвал «распылением» и «запутыванием». Смысл «распыления» состоит в таком преобразовании текста, при котором статистические связи между его элементами становятся трудно наблюдаемыми, хотя избыточность сообщения при этом и не изменяется. Идея «запутывания» (перемешивания) заключается в том, чтобы сделать соотношения между простыми статистиками в пространстве криптограмм и простыми подмножествами в пространстве ключей весьма сложными и беспорядочными [26]. На этих идеях основано

использование составных шифров, состоящих часто в последовательном применении простых подстановок и перестановок.

В шифрах подстановки буквы сообщения заменяются другими буквами, поэтому статистические свойства текста сохраняются, но относятся теперь к буквам другого алфавита. В шифрах перестановки буквы сообщения остаются теми же, но изменяется их расположение в тексте[12]. Комбинирование поочередно применяемых подстановок и перестановок дает возможность создавать весьма стойкие шифры.

Один из примеров реализации такого подхода – государственный стандарт шифрования США, известный под названием DES (Data Encryption Standard). Алгоритм шифрования является открытым; секретным при каждом использовании алгоритма является только ключ. Ключ длиной в 56 бит обеспечивает высокий уровень стойкости шифра, так как общее количество ключей составляет около $7{,}2 \cdot 10^{16}$. Открытый текст и криптограмма при этом являются двоичными 64-значными последовательностями. Криптоалгоритм DES представляет собой суперпозицию из 16 последовательных шифроциклов, в каждом из которых происходят подстановки и перестановки в четырехбитовых группах, при этом в каждом цикле используются 48 бит ключа, которые выбираются из полного ключа длиной в 56 бит [25]. По утверждению Национального бюро стандартов США, метод DES имеет высокую криптостойкость, которая делает его раскрытие дороже получаемой при этом прибыли, экономичен в реализации и эффективен в смысле быстродействия. Наиболее существенным недостатком считается слишком короткий ключ: для дешифрования требуется перебрать 2^{56} (или $7{,}2 \cdot 10^{16}$) ключей, что достаточно много для современной аппаратуры, но может оказаться преодолимым в ближайшем будущем. Впрочем, метод допускает простую модификацию в виде последовательного применения к сообщению нескольких циклов шифрования с различными ключами: например, уже при трех ключах для дешифрования требуется выполнение около 2^{168} (т.е. $3{,}7 \cdot 10^{50}$) операций.

Отечественный стандарт шифрования данных, предусмотренный ГОСТ 28147-89, предназначен для шифрования данных аппа-

[12] Простой пример шифра перестановки может быть реализован, например, путем записи сообщения в разграфленный прямоугольник по строкам и последующего считывания по столбцам. Один из методов шифрования основан на записи сообщения на ячейках граней кубика Рубика и считывании после заданного числа заданных трансформаций [25].

ратным или программным путем [25]. Стандарт предусматривает различные режимы шифрования, но в любом случае используется 256-разрядный двоичный ключ. Двоичные данные, подлежащие зашифрованию, разбиваются на блоки по 64 разряда, над которыми выполняются преобразования, включающие суммирования по модулям 2, 2^{32} и $2^{32}-1$, подстановки и циклические сдвиги. Стандарт обладает всеми достоинствами алгоритма DES и в то же время свободен от его недостатков, однако имеет собственный существенный недостаток – очень низкое быстродействие при программной реализации [25].

13.5. КРИПТОСИСТЕМЫ С ОТКРЫТЫМ КЛЮЧОМ

Системы с открытым ключом названы так потому, что ключ, используемый при зашифровании данных, не является секретным и может быть, например, опубликован в средствах массовой информации. Также несекретным является алгоритм зашифрования. Защита данных обеспечивается тем, что для расшифрования необходим другой (секретный) ключ, причем он не может быть определен по открытому ключу зашифрования. Алгоритмы шифрования с открытым ключом называют поэтому несимметричными алгоритмами [29]. Наиболее известен метод шифрования с открытым ключом RSA[13].

Согласно методу RSA, для генерирования ключей необходимо выполнить следующие действия [25].

1. Выбрать два больших простых числа p и q.
2. Определить их произведение $n = pq$.
3. Выбрать число d, взаимно простое с числом $(p-1)(q-1)$.
4. Определить число e, для которого выполняется условие

$$ed \bmod[(p-1)(q-1)] = 1.$$

5. Назвать открытым ключом числа e и n, а секретным ключом числа d и n.

[13] Метод назван по первым буквам фамилий авторов – Rivest, Shamir, Adleman.

Для применения полученных ключей открытое сообщение необходимо закодировать числами от 0 до $n-1$. Каждое такое число $M(i)$ зашифровывается при помощи открытого ключа по формуле

$$C(i) = \left[M(i)\right]^e \bmod(n). \qquad (13.4)$$

Для расшифрования используется формула с секретным ключом

$$M(i) = \left[C(i)\right]^d \bmod(n). \qquad (13.5)$$

Пример 13.1. Рассмотрим в качестве простого примера алгоритм RSA, основанный на очень малых числах p и q. Предположим, что шифрованию подлежит сообщение на русском языке [25]. Буквы сообщения можно представить числами от 0 до 32 (см. разд. 13.2). Тогда за n можно принять число 33, а за простые числа p и q соответственно 3 и 11. Итак, согласно описанному алгоритму:

1) выберем два простых числа $p = 3$ и $q = 11$;
2) найдем $n = 3 \cdot 11 = 33$;
3) за число d, взаимно простое с числом $(p-1)(q-1) = 20$, примем число 3;
4) соотношению $e \cdot 3 \bmod(20) = 1$ удовлетворяют числа 7, 27, 47,…; выберем $e = 7$.

Итак, открытым ключом для зашифрования является пара чисел $e = 7$ и $n = 33$, а закрытым (секретным) ключом для расшифрования – пара чисел $d = 3$ и $n = 33$.

Зашифруем слово ДОМ. Буквам Д, О и М соответствуют числа 5, 15 и 13. Используя открытый ключ, получим на основании (13.4) криптограмму, состоящую из чисел:

$$C_1 = 5^7 \bmod(33) = 78125 \bmod(33) = 14;$$

$$C_2 = 15^7 \bmod(33) = 170859375 \bmod(33) = 27;$$

$$C_3 = 13^7 \bmod(33) = 62748517 \bmod(33) = 7.$$

Для расшифрования криптограммы $\{14, 27, 7\}$ воспользуемся формулой (13.5) и секретным ключом:

$$M_1 = 14^3 \bmod(33) = 2744 \bmod(33) = 5;$$

$$M_2 = 27^3 \bmod(33) = 19683 \bmod(33) = 15;$$

$$M_3 = 7^3 \bmod(33) = 343 \bmod(33) = 13. \quad \blacktriangleleft$$

Легко видеть, что в результате расшифрования получилось исходное открытое сообщение ДОМ. Следует отметить, что на практике применяются настолько большие числа p и q, что, зная e и n (открытый ключ), невозможно найти d за приемлемое время, так как в настоящее время не только не известен достаточно эффективный (полиномиальный) алгоритм разложения большого числа на простые множители, но и сам вопрос о существовании таких алгоритмов (а следовательно, о возможности взлома систем с открытым ключом в будущем) остается открытым [25, 27]. Тем не менее нельзя исключить открытие в будущем эффективных алгоритмов определения делителей целых чисел (*факторизации*), вследствие чего метод шифрования с открытым ключом станет абсолютно бесполезным. Пока этого не произошло, метод RSA имеет важные преимущества перед другими криптосистемами, такие как очень высокая криптостойкость и простота аппаратной и программной реализации[14].

13.6. ЦИФРОВАЯ ПОДПИСЬ

Одной из важных задач, связанных с передачей документов по каналам связи или с пересылкой их на машинных носителях в электронной форме, является задача аутентификации (подтверждения подлинности). Для документа в обычной (бумажной) форме эта проблема решается за счет жесткой связи информации с носителем (бумагой). Документ в электронной форме такой связи не имеет и иметь не может.

При передаче секретных документов (военного или дипломатического содержания) весьма вероятным является перехват документа противником, который может либо создать и переслать вместо него подложный документ, либо изменить содержимое документа законного источника. Чтобы обнаружить подмену, в зашифрованное сообщение встраивается так называемая имитовставка, которая представляет собой последовательность, полученную по определенному алгоритму на основе всего текста открытого сообщения. Получатель после расшифрования криптограммы подвергает полученный открытый текст повторной обработке тем же алгоритмом и сравнивает полученную имитовставку с принятой. Если совпадения нет, принятое сообщение считается ложным.

[14] Кроме метода RSA, к криптосистемам с открытым ключом относятся системы Эль-Гамаля и Мак-Элиса [25].

Совершенно другая ситуация может иметь место при обмене информацией коммерческого характера. Здесь возможна ситуация, когда партнеры по информационному обмену не доверяют друг другу и являются в каком-то смысле противниками. Один из партнеров может изготовить документ, зашифровать его и заявить, что получил его от партнера.

Для такого случая необходимо применять схему шифрования следующего вида. Передающий абонент зашифровывает подпись с помощью своего секретного ключа, а получатель расшифровывает ее своим несекретным ключом. Несекретный ключ может представлять собой набор проверочных соотношений, позволяющих установить подлинность подписи, но не восстановить секретный ключ. Таким образом, никто, кроме законного автора документа, не в состоянии сгенерировать правильную подпись.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В чем заключается проблема защиты информации?
2. Что такое совершенная секретная система?
3. В чем отличие криптографии и стеганографии?
4. Чем отличаются шифры от кодов?
5. В чем состоят задачи криптографии и криптоанализа?
6. Что такое система с открытым ключом? На чем основана ее криптостойкость?
7. Что такое цифровая подпись?

УПРАЖНЕНИЯ

1. Попросите партнера зашифровать сообщение длиной 80…100 букв шифром простой подстановки. Расшифруйте криптограмму, воспользовавшись оценками вероятностей букв русского текста (для оценивания можно использовать любой неспециальный текст, например газетную статью).
2. Зашифруйте короткое (10…20 букв) сообщение при помощи гаммы. В качестве гаммы используйте некоторый текст (например, фразу, строку из стихотворения, пословицу и т.п.). Предложите партнеру расшифровать текст, передав ему гамму.
3. Зашифруйте короткое слово (3…5 букв) при помощи алгоритма RSA, воспользовавшись данными примера 13.1. Расшифруйте полученный шифртекст.

14. ЭФФЕКТИВНОСТЬ И ОПТИМИЗАЦИЯ СИСТЕМ СВЯЗИ

Современные системы связи (системы передачи информации, телекоммуникационные системы) представляют собой высокотехнологичную инфраструктуру, во многом определяющую облик цивилизации. Создание средств связи, их эксплуатация, разработка новых принципов и технологических решений в условиях жесткой конкурентной борьбы предъявляют высокие требования к качеству принимаемых решений. Специалисту совершенно необходимо знать, по каким параметрам следует сравнивать между собой различные устройства и системы связи, как обеспечить выполнение предъявляемых требований по их эффективности, помехоустойчивости, скрытности, электромагнитной совместимости, информационной защищенности и т.д. Подробное изучение этих вопросов является предметом специальных дисциплин. В настоящем разделе дается лишь краткое изложение основных понятий, связанных с эффективностью систем связи и их оптимизацией.

14.1. ОСНОВНЫЕ ПОКАЗАТЕЛИ ЭФФЕКТИВНОСТИ

Назначение любой системы связи заключается в передаче информации; чем быстрее и точнее передается информация, тем лучше система. Поэтому основными показателями качества (эффективности) систем связи являются скорость передачи и верность (достоверность), под которой понимается степень соответствия принятого сообщения переданному. Конкретная количественная мера верности выбирается в зависимости от характера сообщения [22]. В дискретных системах связи мерой верности может служить, например, средняя вероятность ошибочного решения при приеме, в

аналоговых системах – среднеквадратическое отклонение принятого сообщения от переданного.

Скорость передачи информации I' измеряется в битах в секунду; ее не следует путать с технической скоростью, измеряемой в бодах (см. разд. 1). Предельная скорость передачи информации по данному каналу называется его пропускной способностью C. Однако этот предел лишь показывает потенциальные возможности, для приближения к которым могут потребоваться непомерные затраты (например, очень сложные и дорогие кодеры и декодеры, очень большое время кодирования-декодирования и т.п.). Степень использования пропускной способности канала характеризуют относительным показателем – информационной эффективностью $\eta = I'/C$.

Ресурсы, которыми располагает разработчик, всегда ограничены. Например, могут быть ограничены стоимость устройств, эксплуатационные расходы, максимальное время задержки получения сообщений, полоса частот, занимаемая системой радиосвязи, допустимый уровень электромагнитных излучений вне этой полосы, уровень скрытности связи, степень защищенности (время, необходимое для «взлома» криптограммы), массогабаритные характеристики и др. Таким образом, поиск решения при проектировании системы связи имеет характер задачи *оптимизации с ограничениями*.

Обозначая различные количественные показатели эффективности (качества) через $k_1, k_2, ..., k_m$, получаем вектор $\mathbf{K}$, который характеризует качество системы. Множество векторов пространства размерности выше 1 не является естественно упорядоченным, как одномерное пространство (например, числовая прямая); сравнение двух систем между собой по *векторному* показателю не позволяет, как правило, выбрать *безусловно лучшее* решение. Формулирование задачи выбора оптимального решения всегда предполагает задание *скалярного* показателя, такого, что искомому оптимальному решению соответствует его максимум (или минимум); этот скалярный показатель называется *целевой функцией*. Иногда за скалярный показатель качества можно принять линейную комбинацию компонент вектора $\mathbf{K}$, но тогда встает вопрос о назначении весовых коэффициентов. В большинстве случаев более оправданным является подход, когда максимизируется одна компонента вектора (например, скорость передачи информации) при ограничениях в форме неравенств, накладываемых на остальные компоненты (например, средняя вероятность ошибки при приеме двоичного символа не более 0,001; задержка не более 0,1 с; потребляемая мощность не более 0,5 Вт и т.п.).

Аргументами целевой функции служат, вообще говоря, все величины, так или иначе влияющие на значение целевой функции. Поскольку таких факторов слишком много и учесть их влияние точно не представляется возможным, неизбежно упрощение модели и сведéние всего множества влияющих величин к нескольким наиболее существенным. Тем не менее, необходимо рассматривать систему связи в виде целостного объекта, учитывая ее состав, взаимодействие ее частей друг с другом, влияние на нее окружающей среды, взаимодействие системы и пользователей, историю и перспективы развития систем данного класса и близких классов и т.д. Такой подход называется *системным* [12].

14.2. ОПТИМИЗАЦИЯ СИСТЕМ СВЯЗИ

Действие системы связи можно описать операторным уравнением, выражающим различные этапы формирования, передачи и приема сигналов. Рассматривая упрощенную систему связи (см. рис. 1.2), можно описать преобразование сообщения a в модулированный сигнал $u(t)$ операторным выражением $u(t) = M\{a, s(t)\}$, где $s(t)$ – колебание-переносчик. Преобразование модулированного сигнала на выходе передатчика в наблюдаемое колебание на входе приемника можно представить выражением $z(t) = T\{u(t), \eta(t)\}$, где $\eta(t)$ – помеха в канале. Аналогично, преобразование наблюдаемого колебания в оценку $\hat{a}$ сообщения можно описать выражением $\hat{a} = D\{z(t)\}$. Объединяя эти выражения, получим операторное уравнение системы связи

$$\hat{a} = D\{T\{M\{a, s(t)\}, \eta(t)\}\}. \qquad (14.1)$$

Задача оптимального проектирования системы связи заключается в том, чтобы обеспечить наилучшее качество в смысле выбранного критерия. Критерий должен соответствовать назначению системы, быть достаточно простым и зависеть от величин, которыми можно *управлять* в процессе проектирования.

Полная задача оптимизации проектирования системы согласно (14.1) может быть разложена на подзадачи, связанные с оптимизацией на уровне отдельных операторов, входящих в (14.1). В некоторых случаях, когда задана линия связи [12], оператор $z(t) = T\{u(t), \eta(t)\}$, описывающий взаимодействие сигнала с помехой, считается неуправляемым, тогда оптимизация осуществляется

выбором операторов $M\{\cdot\}$ и $D\{\cdot\}$. К оператору $M\{\cdot\}$ относятся выбор переносчика, метода модуляции, параметров передатчика и т.п. Выбор оператора $D\{\cdot\}$ заключается в разработке и построении оптимального демодулятора, устройств предварительной обработки (фильтрации) и т.п.

Однако нужно иметь в виду, что сама *декомпозиция* задачи оптимизации системы на подзадачи согласно уравнению (14.1) может увести от оптимального решения к квазиоптимальным: решения каждой подзадачи могут быть наилучшими, а система в целом может быть далека от оптимума. Поэтому при выборе структуры системы и отдельных подсистем и устройств следует придерживаться системного подхода: при выборе переносчика учитывать свойства линии связи, включая действующие в ней помехи, а также особенности приема (например, стационарный или мобильный приемник), и т.д. Уравнение (14.1) в настоящее время нельзя использовать непосредственно для строгой оптимизации системы, но оно показывает внутреннее единство задачи оптимизации.

14.3. ПРЕДЕЛЬНЫЕ ВОЗМОЖНОСТИ СИСТЕМ ПЕРЕДАЧИ ДИСКРЕТНЫХ СООБЩЕНИЙ

На первом этапе проектирования системы необходимо определить принципиальную возможность передавать информацию с заданной скоростью и заданной достоверностью. При этом из всех возможных помех принимается во внимание только внутренний шум приемника и считается, что кодер и декодер могут быть сколь угодно сложными. Тогда если $H' < C$, где H' – производительность источника, а C – пропускная способность канала, то за счет кодирования можно добиться сколь угодно малой вероятности ошибки $p_{\text{ош}} \to 0$, при этом техническая скорость $1/T_{\text{п}} = C$.

Для предварительной оценки затрат энергии и полосы частот, которыми определяется пропускная способность дискретного канала, удобно использовать относительные характеристики [12]: удельный расход энергии на один двоичный символ $\beta_E = P_{\text{с}} T_{\text{п}} / N_0 = E_{\text{с}} / N_0$ и удельный расход полосы $\beta_f = F_{\text{с}} T_{\text{п}}$, где $P_{\text{с}}$ – мощность сигнала, $T_{\text{п}}$ – длительность сигнала (посылки), $E_{\text{с}}$ – энергия сигнала, $F_{\text{с}}$ – ширина спектра сигнала, N_0 – спектральная плотность мощности (белого) шума. Положим, что для передачи дискретных сообщений используется непрерывный гауссовский

канал связи (см. разд. 8.7), пропускная способность которого определяется выражением (8.22)

$$C = F_{\text{к}} \log\left(1 + \frac{P_{\text{с}}}{P_{\text{ш}}}\right) = F_{\text{к}} \log\left(1 + \frac{P_{\text{с}}}{F_{\text{к}} N_0}\right).$$

Считая, что $F_{\text{с}} = F_{\text{к}}$, $1/T_{\text{п}} = C$, и учитывая $\beta_E / T_{\text{п}} = P_{\text{с}} / N_0$, можно записать $\dfrac{1}{F_{\text{с}} T_{\text{п}}} = \log\left(1 + \dfrac{\beta_E}{T_{\text{п}} F_{\text{с}}}\right)$, откуда $\dfrac{1}{\beta_f} = \log\left(1 + \beta_E \beta_f\right)$ и окончательно получаем

$$\beta_E = \beta_f \left(2^{1/\beta_f} - 1\right). \qquad (14.2)$$

Уравнение (14.2) определяет возможности обмена удельных расходов полосы и энергии. На рис. 14.1 показана диаграмма обмена, построенная согласно (14.2). Эта диаграмма позволяет оценить совместную реализуемость предполагаемой скорости передачи информации и энергетических затрат (ее называют пределом Шеннона [12]). Очевидно, реализуемы только те сочетания, которые изображаются на графике точками, лежащими выше кривой. Если ресурс канала изображается точкой, лежащей ниже кривой, реализовать систему с такими характеристиками невозможно.

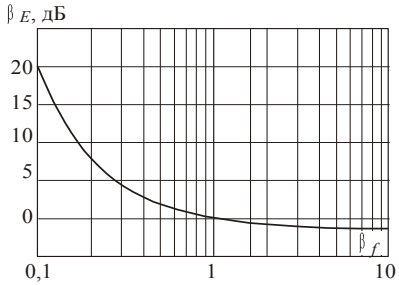


Рис. 14.1. Диаграмма обмена при передаче дискретных сообщений

При $\beta_f \to \infty$ график асимптотически стремится к величине $\ln 2 \approx 0{,}693$ (что составляет $-1{,}6$ дБ). При $\beta_f = 1$ удельный расход энергии равен 1 (0 дБ). Таким образом, ясно, что в канале с АБГШ увеличение удельного расхода полосы выше единицы приводит к сравнительно небольшому сокращению удельного расхода энергии. В то же время уменьшение β_f до значений, заметно меньших 1, приводит к сильному возрастанию удельных энергетических затрат.

14.4. ПРЕДЕЛЬНЫЕ ВОЗМОЖНОСТИ СИСТЕМ ПЕРЕДАЧИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Предположим, что каждый отсчет непрерывного сообщения передается при помощи ИКМ последовательностью двоичных символов; при этом точность передачи определяется разрядностью кода. Пусть скорость следования двоичных символов H'_ε не превышает пропускной способности канала C, тогда согласно теореме Шеннона, применяя достаточно сложные способы кодирования и декодирования, можно обеспечить сколь угодно низкую вероятность ошибки. Мы рассматриваем предельные возможности, поэтому считаем, что вероятность ошибки равна нулю, тогда точность воспроизведения сообщения определяется только скоростью следования двоичных символов. (Заметим, что влияние помех в реальном канале выражается в снижении скорости передачи информации и, следовательно, в снижении точности представления сообщения.)

Минимальная скорость следования двоичных символов, при которой можно представить сообщение с заданной точностью, равна эпсилон-энтропии источника. Предположим, что сообщение представляет собой стационарный гауссовский процесс со спектральной плотностью мощности, постоянной в полосе частот $\left(-F_\text{в}, F_\text{в}\right)$; такое сообщение можно точно восстановить по его отсчетам, взятым с шагом $T_d = 1/\left(2F_\text{в}\right)$.

Введем в качестве меры точности воспроизведения сообщения относительный средний квадрат $\delta^2 = P_\varepsilon / P_\text{с}$, или *отношение шум/сигнал* на выходе приемника. Для гауссовского источника эпсилон-энтропия равна (8.14)

$$H_\varepsilon\left(X\right) = \frac{1}{2}\log\frac{P_\text{с}}{P_\text{ш}},$$

поэтому с учетом скорости следования отсчетов производительность

$$H'_\varepsilon = -F_\text{в}\log\delta^2.$$

Считая $H'_\varepsilon = C$ и учитывая (8.19), запишем для канала с АБГШ

$$-F_\text{в}\log\delta^2 = F_\text{к}\log\left(1+\frac{P_\text{с}}{N_0F_\text{к}}\right). \qquad (14.3)$$

Для предварительной оценки затрат при передаче непрерывных сообщений используют относительные характеристики [12]: удельный расход мощности $\beta_P = P_{\text{с}} / \left(N_0 F_{\text{в}}\right)$ и удельный расход полосы $\beta_f = F_{\text{с}} / F_{\text{в}}$. Положим $F_{\text{с}} = F_{\text{к}}$, поделим обе части (14.3) на $F_{\text{в}}$ и учтем, что $P_{\text{с}} / \left(N_0 F_{\text{с}}\right) = \beta_P / \beta_f$. Тогда

$$\log \frac{1}{\delta^2} = \beta_f \log\left(1 + \frac{\beta_P}{\beta_f}\right),$$

откуда следует $\dfrac{1}{\delta^2} = \left(1 + \dfrac{\beta_P}{\beta_f}\right)^{\beta_f}$, или

$$\beta_P = \beta_f \left[\left(\frac{1}{\delta^2}\right)^{\frac{1}{\beta_f}} - 1\right]. \qquad (14.4)$$

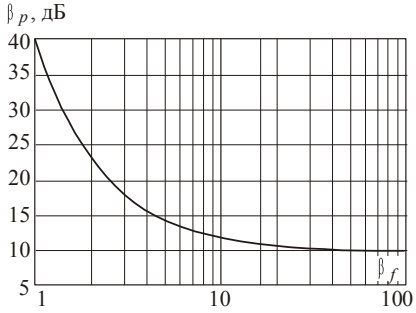


Рис. 14.2. Диаграмма обмена при передаче непрерывных сообщений

На рис. 14.2 показана диаграмма обмена, построенная согласно (14.4) при $\delta^2 = 10^{-4}$. Реализуемы только те сочетания параметров β_P и β_f, которые изображаются на графике точками, лежащими выше кривой. Чем ближе к кривой изображающая точка, тем лучше используются ресурсы канала. Подробное обсуждение диаграмм обмена для различных методов модуляции и кодирования см., например, в [12].

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите основные показатели эффективности систем связи.
2. Можно ли оптимизировать систему связи по векторному критерию качества?
3. Какие показатели обычно учитываются в виде ограничений?
4. Что такое системный подход?

УПРАЖНЕНИЯ

1. Постройте диаграмму обмена согласно (14.4) для $\delta^2 = 10^{-6}$. Сравните с рис. 14.2.

2. Постройте графики зависимости отношения сигнал/шум $1/\delta^2$ от удельных затрат мощности β_P при нескольких заданных значениях удельных затрат полосы β_f. Объясните поведение кривых.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Владимиров В.С.* Обобщенные функции в математической физике. – М.: Наука, 1976. – 280 с.
2. *Френкс Л.* Теория сигналов. – М.: Сов. радио, 1974. – 344 с.
3. *Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа. – М.: Наука, 1972. – 496 с.
4. *Ратынский М.В.* Основы сотовой связи. – М.: Радио и связь, 2000. – 248 с.
5. *Оппенгейм А.В., Шафер Р.В.* Цифровая обработка сигналов. – М.: Связь, 1979. – 416 с.
6. *Вентцель Е.С.* Теория вероятностей. – М.: Наука, 1969. – 576 с.
7. *Феллер В.* Введение в теорию вероятностей и ее приложения. – М.: Мир, 1984. – Т. 1. – 528 с., Т. 2. – 738 с.
8. *Баскаков С.И.* Радиотехнические цепи и сигналы. – М.: Высш. шк., 2002.
9. *Боголюбов А.Н.* Математики и механики: Биограф. справ. – Киев: Наукова думка, 1983. – 638 с.
10. *Теория* электрической связи: Учебник для вузов / Под ред. Д.Д. Кловского. – М.: Радио и связь, 1999. – 432 с.
11. *Прокис Дж.* Цифровая связь. – М.: Радио и связь, 2000. – 800 с.
12. *Радиотехнические* системы передачи информации: Учеб. пособие для вузов / Под ред. В.В. Калмыкова. – М.: Радио и связь, 1990. – 304 с.
13. *Андреев В.С.* Теория нелинейных электрических цепей. – М.: Связь, 1972. – 328 с.
14. *Сиберт У.* Цепи, сигналы, системы: В 2 ч. Ч. 1. – М.: Мир, 1988 – 336 с.
15. *Сиберт У.* Цепи, сигналы, системы: В 2 ч. Ч. 2. – М.: Мир, 1988 – 360 с.
16. *Хэмминг Р.В.* Теория кодирования и теория информации. – М.: Радио и связь, 1983. – 176 с.
17. *Васюков В.Н.* Введение в теорию сигналов: Учеб. пособие. – Новосибирск: Изд-во НГТУ, 2002. – 92 с.

18. *Тихонов В.И.* Статистическая радиотехника. – М.: Радио и связь,1982. – 624 с.

19. *Васюков В.Н.* Цифровая обработка сигналов и сигнальные процессоры в системах подвижной радиосвязи: Учебник. – Новосибирск: Изд-во НГТУ, 2003. – 292 с.

20. *Сейдж Э., Мелс Дж.* Теория оценивания и ее применение в связи и управлении. – М.: Связь, 1976. – 496 с.

21. *Назаров М.В., Кувшинов Б.И., Попов О.В.* Теория передачи сигналов. – М.: Связь, 1970. – 368 с.

22. *Теория* передачи сигналов: Учеб. для вузов / А.Г. Зюко, Д.Д. Кловский, М.В. Назаров, Л.М. Финк – М.: Связь, 1980. – 288 с.

23. *Финк Л.М.* Теория передачи дискретных сообщений. – М.: Сов. радио, 1970. – 728 с.

24. *Гоноровский И.С.* Радиотехнические цепи и сигналы. – М.: Радио и связь, 1986. – 512 с.

25. *Петраков А.В.* Основы практической защиты информации. – М.: Радио и связь, 2000. – 368 с.

26. *Шеннон К.* Теория связи в секретных системах / К. Шеннон Работы по теории информации и кибернетике. – М.: ИЛ, 1963. – 830 с.

27. *Введение* в криптографию / Под общ. ред. В. В. Ященко. – М.: МЦНМО, 2000. – 272 с.

28. *Ященко В.В.* Основные понятия криптографии // Математическое просвещение. Сер. 3. – Вып. 2. – 1998. – С. 53–70.

29. *Брассар Ж.* Современная криптология. – М.: Изд.-полиграф. фирма ПОЛИМЕД, 1999. – 176 с.

30. *Феер К.* Беспроводная цифровая связь Методы модуляции и расширения спектра: Пер. с англ. – М.: Радио и связь, 2000. – 520 с.

31. *Скляр Б.* Цифровая связь. – М.: Вильямс, 2003. – 1104 с.

ПРИНЯТЫЕ ОБОЗНАЧЕНИЯ

$x(t)$, $s(t)$ – аналоговый сигнал, колебание

$b(t)$ – первичный сигнал

f – частота гармонического колебания

ω – круговая частота

$\mathbb{R}$ – поле вещественных (действительных) чисел

$\mathbb{C}$ – поле комплексных чисел

$\mathbb{F}$ – произвольное поле

$T_{\text{с}}$ – длительность сигнала

$F_{\text{с}}$ – полоса частот (ширина спектра) сигнала

$D_{\text{с}}$ – динамический диапазон сигнала

$V_{\text{с}}$ – объем сигнала

$T_{\text{к}}$ – время действия канала;

$F_{\text{к}}$ – полоса пропускания канала;

$D_{\text{к}}$ – динамический диапазон канала

$V_{\text{к}}$ – емкость канала

$\delta(t)$ – дельта-функция Дирака

$\sigma(t)$ – функция включения Хевисайда

$\delta[n]$ – дельта-последовательность

$u[n]$ – единичная ступенчатая последовательность

$\forall$ – квантор всеобщности (читается «для всех»)

$\exists$ – квантор существования (читается «существует»)

$\{\cdot\}$ – множество элементов, указанных в скобках

$\in$ – знак принадлежности элемента множеству

$T\{\cdot\}$ – произвольное преобразование

$\mathbb{L}\{\cdot\}$ – произвольный оператор

$x(t) \Leftrightarrow X(f)$ – функции времени и частоты, связанные парой преобразований Фурье

L – пространство аналоговых сигналов, определенных на всей временной оси

$L_2(T)$ – пространство аналоговых сигналов ограниченной энергии, определенных на интервале времени T

$L_2(F)$ – пространство аналоговых сигналов ограниченной энергии, имеющих ограниченную полосу частот F

l_2 – пространство последовательностей ограниченной энергии

(x, y) – скалярное произведение сигналов (векторов)

$*$	– свертка функций $x(\cdot)$ и $h(\cdot)$
$x \otimes h$	– дискретная свертка последовательностей $x[\cdot]$ и $h[\cdot]$
$\oplus$	– прямая сумма пространств
$\|x\|_2$	– норма сигнала, соответствующая пространству L_2 или l_2
$\mathbf{P}\ \cdot$	– вероятность случайного события, указанного в фигурных скобках
$\mathbf{E}\ \cdot$	– обозначение усреднения по ансамблю
$\overline{(\cdot)}$	– обозначение усреднения по ансамблю
$\langle\cdot\rangle$	– обозначение усреднения по времени
$H(\mathrm{A})$	– энтропия источника сообщений A
$I(\mathrm{A}, \mathrm{B})$	– взаимная информация источников A и B

ПРЕДМЕТНО-ИМЕННОЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

УЧЕБНОЕ ИЗДАНИЕ

Василий Николаевич Васюков

ТЕОРИЯ ЭЛЕКТРИЧЕСКОЙ СВЯЗИ

Учебник

Редактор *И.Л. Кескевич*
Корректор *Л.Н. Ветчакова*
Обложка *А.В. Волошина*
Компьютерная верстка *В.Ф. Ноздрева*

Подписано в печать 20.10.05
Формат 60×90 1/16. Бумага офсетная
Уч.-изд. л. 29,4. Печ. л. 24,5
Тираж 500 экз. Заказ № 1187

Лицензия на издательскую деятельность
Серия ИД № 04303 от 20.03.01
Издательство Новосибирского государственного
технического университета
630092, г. Новосибирск, пр. К. Маркса, 20.
Тел. (383-2) 46-31-87
E-mail: office@publish.nstu.ru

Отпечатано в типографии
Новосибирского государственного технического университета
630092, г. Новосибирск, пр. К. Маркса, 20