Глава 3

Нелинейные системы

3.1 Свойства и методы анализа нелинейных систем

До сих пор мы рассматривали прохождение сигналов через различные *линейные* системы. Линейная система — это модель, в которой предполагается, что параметры системы от сигналов, воздействующих на эту цепь, не зависят. В рамках этой модели можно было эффективно использовать разложение сигнала по различным наборам функций, поскольку для линейных цепей справедлив принцип суперпозиции, позволяющий находить отклик системы на несколько сигналов как сумму откликов на каждый отдельный сигнал. В частности, откликом линейной системы на сумму гармонических воздействий всегда является набор гармонических колебаний с теми же частотами, для каждого из которых может измениться лишь амплитуда и фаза.

Во многих случаях этой моделью пользоваться нельзя, поскольку параметры элементов, входящих в рассматриваемую систему, существенным образом зависят от сигнала. Для таких систем принцип суперпозиции не справедлив. К примеру, в отклике нелинейной системы на сумму гармонических воздействий могут присутствовать гармонические колебания с частотами, которых не было в исходном сигнале. В системах, от которых требуется максимально точное воспроизведение сигнала (например, в усилителях, каналах связи, устройствах записи и хранения информации) учет нелинейностей позволяет оценить нелинейные искажения, которые являются нежелательным фактором. В то же время, многие операции, такие, как перенос спектра и генерация колебаний могут быть реализованы только в нелинейных системах. Перенос спектра является принципиально важной операцией для передачи информации, поскольку для любого канала связи существуют физические ограничения на диапазон передаваемых частот, например, для волоконно-оптического канала — это диапазон прозрачности используемого материала, для радиоканала — диапазон, в котором радиоволны не испытывают слишком большого затухания в атмосфере а приемно-передающие антенны имеют приемлемые размеры, в то время, как спектр полезного сигнала (например, человеческой речи) лежит в другом диапазоне.

При строгом рассмотрении все рассмотренные ранее элементы: активные сопротивления, катушки индуктивности и конденсаторы являются нелинейными. Вопрос состоит только в том, нужно ли учитывать их нелинейность в данной кон-

кретной задаче. Во многих случаях нелинейность проявляется при больших (по амплитуде) сигналах, в то время, как при малых (по сравнению с некоторой, характерной для данного элемента, величиной) ею можно пренебречь. Анализ нелинейных систем может быть достаточно сложен и универсального метода, подобному методу комплексных амплитуд для линейных цепей, не существует. Мы будем использовать несколько упрощений. Так, рассматриваться будут только безинерционные элементы, то есть такие, свойства которых полностью определяются их статическими характеристиками. В таких элементах изменение параметров происходит за время, много меньшее, чем характерное время изменения сигнала. Кроме того, вначале мы ограничимся рассмотрением нелинейных резистивных элементов.

Нелинейные резистивные элементы можно разделить на двуполюсные (варисторы, сопротивление которых зависит от величины протекающего тока но не зависит от его направления, и диоды, сопротивление которых существенно зависит от направления протекающего тока) и многополюсные (транзисторы, тиристоры, электронные лампы с управляющими электродами и другие).

Свойства безинерционного двуполюсного резистивного элемента определяет его вольт-амперная характеристика (ВАХ). Для пассивного (не связанного с дополнительным по отношению к сигналу источником энергии) двуполюсника ВАХ должна лежать в первом и третьем квадранте координатной плоскости (направление тока определяется приложенным напряжением) и проходить через начало координат (в отсутствие напряжения ток не течет). ВАХ, вообще говоря, может быть как монотонной, для которой производные dU/dI и dI/dU принимают только неотрицательные значения, так и немонотонной. В последнем случае различают элементы с N и S образной характеристикой (см. Рис. 3.1).

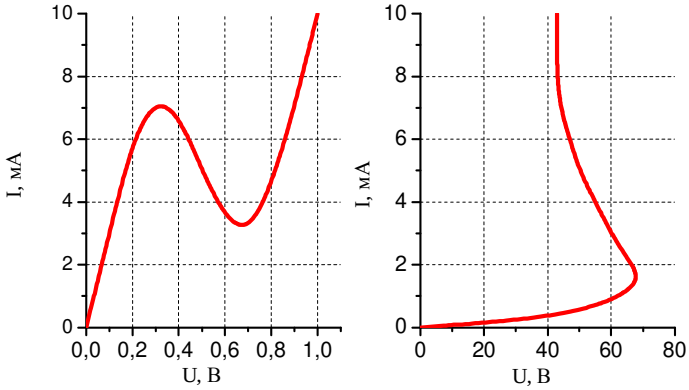


Рис. 3.1: Слева: Типичная характеристика тунельного диода (N – образная ВАХ), справа: Типичная характеристика термистора (S – образная ВАХ) .

Двуполюсники, ВАХ которых имеет участки, на которых $dI/dU < 0$ называются элементами с отрицательным сопротивлением.

В технике очень важную роль играют управляемые нелинейные элементы. Так, параметры двуполюсника могут изменятся при внешнем не электрическом воздействии: изменении температуры, давления и т.п. В этом случае каждому значению

внешнего параметра соответствует своя ВАХ. Особенно часто используются электрически управляемые элементы: транзисторы, тиристоры, электронные лампы различных видов. Поскольку такие элементы содержат дополнительные выводы для управляющего сигнала, они в общем случае относятся к четырехполюсникам (часто выводов бывает три: один является общим для управляющего и выходного сигнала). Для описания таких элементов используют семейства вольт-амперных характеристик: выходные — зависимости выходного тока от выходного напряжения при различных значениях входного напряжения или тока, входные — зависимости входного тока от входного напряжения при различных значениях выходного напряжения или тока, переходные — например, зависимость выходного напряжения от входного при различных значениях выходного тока.

3.2 Нелинейный двуполюсник

Проанализируем реакцию нелинейного двуполюсника на внешнее гармоническое воздействие.

Рассмотрим цепь, изображенную на рис. 3.2 а. Последовательно с исследуемым нелинейным элементом $R_{нелин}$ включен источник гармонического напряжения $U_{вх}$, а так же дополнительное сопротивление $R_{нагр}$ и дополнительный источник постоянного напряжения U_0. Если в рассматриваемой области ВАХ нелинейного элемента его дифференциальное сопротивление $R_{диф} \equiv dI/dU \gg R_{нагр}$ то ток, протекающий в цепи, будет определяться только нелинейным элементом, а измерить его можно, измеряя падение напряжения на $R_{нагр}$. Источник постоянного напряжения позволяет выбирать так называемую рабочую точку на ВАХ — точку с координатами (U_0, I_0) в окрестности которой будет изменяться входное напряжение и выходной ток. Пусть нелинейный элемент, входящий в рассматриваемую систему, обладает вольт-амперной характеристикой, изображенной на рис. 3.2 б. Форма тока, протекающего в цепи, будет отличаться от гармонической, и это отличие будет расти с ростом амплитуды входного напряжения.

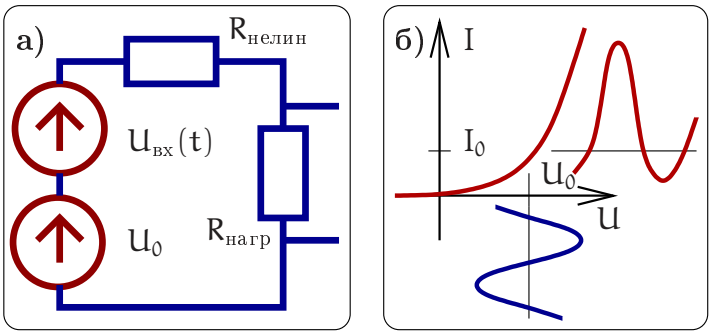


Рис. 3.2: а) Цепь, содержащая нелинейное $R_{нелин}$ и нагрузочное (линейное) сопротивление $R_{нагр}$, $R_{диф} \equiv dU_{нелин}/dI_{нелин} \gg R_{нагр}$. б) Вольт-амперная характеристика нелинейного сопротивления (на дополнительной оси, направленной вниз — пример зависимости входного напряжения от времени, на дополнительной оси направленной вправо - соответствующий ток в цепи).

Представим вольт-амперную характеристику нелинейного элемента в виде степенного многочлена

$$I = I_0 + \alpha(U - U_0) + \beta(U - U_0)^2 + \gamma(U - U_0)^3 + \ldots, \tag{3.1}$$

что фактически означает разложение функции в ряд Тейлора вблизи точки (U_0, I_0):

$$I(U) = I_0 + \left.\frac{dI}{dU}\right|_{U=U_0} (U - U_0) + \left.\frac{d^2I}{2!\,dU^2}\right|_{U=U_0} (U - U_0)^2 + \left.\frac{d^3I}{3!\,dU^3}\right|_{U=U_0} (U - U_0)^3, \tag{3.2}$$

$$\alpha = \frac{dI}{dU}, \quad \beta = \frac{d^2I}{2!\,dU^2}, \quad \gamma = \frac{d^3I}{3!\,dU^3}. \tag{3.3}$$

Число членов в разложении, определяющее точность аппроксимации характеристики, диктуется условиями решаемой задачи и величиной входного воздействия.

Определим спектральный состав тока, протекающего в нелинейной цепи (рис. 3.2) при гармоническом воздействии $U(t) - U_0 = U_1 \cos \omega t$. Подставляя выражение $U_1 \cos \omega t$ в выражение (3.1), после несложных тригонометрических преобразований найдем:

$$I(t) = I_0 + \alpha U_1 \cos \omega t + \beta U_1^2 \cos^2 \omega t + \gamma U_1^3 \cos^3 \omega t + \cdots = \tag{3.4}$$

$$= I_0 + \frac{\beta U_1^2}{2} + \left(\alpha U_1 + \gamma \frac{3U_1^3}{4}\right) \cos \omega t + \frac{\beta U_1^2}{2} \cos 2\omega t + \gamma \frac{U_1^3}{4} \cos 3\omega t + \ldots \tag{3.5}$$

Полученный результат свидетельствует о том, что ток, протекающий в нелинейной цепи при гармоническом воздействии на входе, представляет собой сумму постоянного тока и токов с частотами ω, 2ω, $3\omega, \ldots$ Величина постоянной составляющей и амплитуды всех четных гармоник тока зависят от членов разложения с четными степенями, а амплитуды основной частоты и всех нечетных гармоник определяются членами разложения с нечетными степенями.

Физический смысл нелинейного преобразования сигнала состоит в том, что изначально гармонический сигнал при взаимодействии с нелинейной системой существенно изменяет свою форму и, как следствие, отклик системы обогащается спектральными составляющими. Максимальная степень нелинейности, которую нужно учитывать, зависит от ВАХ элемента и амплитуды напряжения U. Увеличивая U можно получать гармоники более высоких степеней. Этот метод широко используется в радиофизике, нелинейной оптике и сверхвысокочастотной технике.

Пусть теперь на нелинейную систему действуют два источника гармонического напряжения с различными амплитудами и частотами: $U_1 \cos(\omega_1 t$ и $U_2 \cos(\omega_2 t$ (см. рис. 3.14, источник постоянного напряжения, задающий рабочую точку, не изображен). Для упрощения расчетов будем считать, что вольтамперную характеристику нелинейного элемента можно аппроксимировать полиномом второй степени

$$I(t) = I_0 + \alpha U(t) + \beta U^2(t), \tag{3.6}$$

Подставляя в (3.6) величину входного напряжения $U(t) = U_1 \cos(\omega_1 t) + U_2 \cos(\omega_2 t)$, найдем

$$I(t) = I_0 + \alpha\big(U_1 \cos(\omega_1 t) + U_2 \cos(\omega_2 t)\big) + \tag{3.7}$$

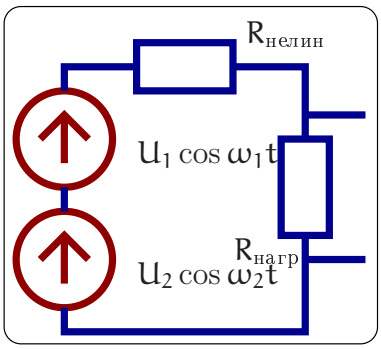


Рис. 3.3: Использование нелинейного сопротивления в схеме с двухчастотным воздействием.

$$+\beta \left(\frac{U_1^2 + U_2^2}{2} + \frac{U_1^2}{2} \cos 2\omega_1 t + \frac{U_2^2}{2} \cos 2\omega_2 t + U_1 U_2 \big[\cos(\omega_1 + \omega_2)t + \cos(\omega_1 - \omega_2)t \big] \right).$$

Таким образом, в спектре отклика нелинейной системы будут присутствовать: постоянная составляющая тока $I_0 + \beta(U_1^2 + U_2^2)/2$, составляющие тока с частотами приложенных напряжений $\sim \cos(\omega_1 t)$ и $\sim \cos(\omega_2 t)$, токи с удвоенными частотами $\sim \cos 2\omega_1 t$ и $\sim \cos 2\omega_2 t$, токи с суммарной и разностной частотами $\sim \cos(\omega_1 + \omega_2)t$ и $\sim \cos(\omega_1 - \omega_2)t$, называемые комбинационными. Устройство на основе нелинейного элемента, предназначенное для получения комбинационных частот называют смесителем.

Если один из сигналов не является гармоническим (как это обычно бывает при передаче информации), то, при взаимодействии с нелинейным элементом его спектральные составляющие будут взаимодействовать между собой, обогащая спектр протекающего тока. Поэтому при осуществлении переноса спектра обычно выбирают уровень информационного сигнала много меньшим, чем уровень второго (гармонического) сигнала. Тогда амплитуды спектральных компонент тока, появляющихся в результате такого взаимодействия (а так же гармоник спектральных компонент информационного сигнала) будут величинами второго порядка малости по сравнению с продуктами взаимодействия между информационным и гармоническим сигналом и ими можно будет пренебречь. В этом случае принято информационный сигнал называть просто сигналом, гармонический — гетеродином и говорить, что смеситель является линейным по сигналу.

Выделение возникающих в результате нелинейного взаимодействия спектральных составляющих тока или напряжения можно осуществить с использованием ранее рассмотренных линейных систем. Так, простейшие RC и RL цепочки могут служить фильтрами нижних или верхних частот. Комбинируя емкости и индуктивности можно конструировать многозвенные фильтры высокого порядка с высоким уровнем подавления нежелательных составляющих в спектре. Особую роль играют высокодобротные колебательные контуры, позволяющие выделять узкий диапазон частот.

Из всего сказанного следует, что процессы преобразования спектра сигнала, необходимые для его передачи с использованием электромагнитных волн и последующего приема могут быть осуществлены при наличии нелинейных элементов.

Частным случаем преобразования является перенос спектра сигнала из низкочастотной в высокочастотную область путем одновременного воздействия на нелинейную систему передаваемого сигнала и высокочастотного колебания, называемого несущим. Этот процесс называют модуляцией. Говоря о модуляции, мы подразумеваем медленное, по сравнению с периодом высокочастотного колебания, изменение одного из его параметров. Такими параметрами являются амплитуда, частота и фаза — см. подробное описание модулированных сигналов в разделе 1.2.2. Простейшим случаем является амплитудно-модулированный сигнал, когда передаваемая информация закодирована в медленном (по сравнению с периодом несущей) изменении (модуляцией) амплитуды. При этом спектр исходного сигнала переносится в высокочастотную область и лежит вблизи частоты несущей.

Наиболее распространенными нелинейными элементами являются устройства на основе полупроводников. Для того, что бы понять основные принципы их работы, обратимся к электронным свойствам твердых тел.

3.3 Электронный транспорт в твердых телах

Электроны в твердом теле, как известно, бывают привязанными к конкретному атому/молекуле (внутренние, или связанные электроны) и общими для всего куска вещества (внешние, или валентные). Нас интересуют последние.

В диэлектриках их почти нет, а в металлах они образуют очень плотный газ. Действительно, если обычный газ при нормальных условиях содержит $2.7 \cdot 10^{19}$ частиц на кубический сантиметр, то плотность электронного газа в металлах равна, очевидно, плотности ионов (в твердом теле!), умноженной на валентность, и для обычных металлов составляет от 10^{22} до 10^{23} штук на кубический сантиметр. Именно очень высокая плотность электронного газа и определяет его специфические квантовые свойства.

При рассмотрении электронного газа в металлах используют два подхода, до некоторой степени противоположных друг другу. В модели сильной связи в качестве отправной точки выбирается набор невзаимодействующих нейтральных атомов, а их взаимодействие рассматривается как малое возмущение. В этой модели электроны проводимости - это “почти привязанные” к атомам внешние электроны, получившие возможность скользить от атома к атому из-за того, что их внешние электронные оболочки перекрываются.

В модели слабой связи в качестве отправной точки используется газ свободных электронов, вообще не взаимодействующий с кристаллической решеткой, а взаимодействие электронов с решеткой рассматривается как возмущение.

Метод сильной связи наглядно иллюстрирует общие закономерности образования энергетических зон при сближении изолированных атомов и образовании из них кристаллической решетки без привлечения аппарата квантовой механики. Рассмотрим качественно картину возникновения энергетических зон на примере образования кристаллической решетки из изолированных атомов натрия. Электронная структура $Na^{11}(1s^2 2s^2 2p^6 3s)$: всего в атоме 11 электронов, по два электрона на $1s$ и $2s$ уровнях, 6 электронов на уровне $2p$, последний заполненный уровень в атоме натрия - $3s$, на котором находится один валентный электрон. В приближении сильной связи предполагается, что состояние электрона в кристалле

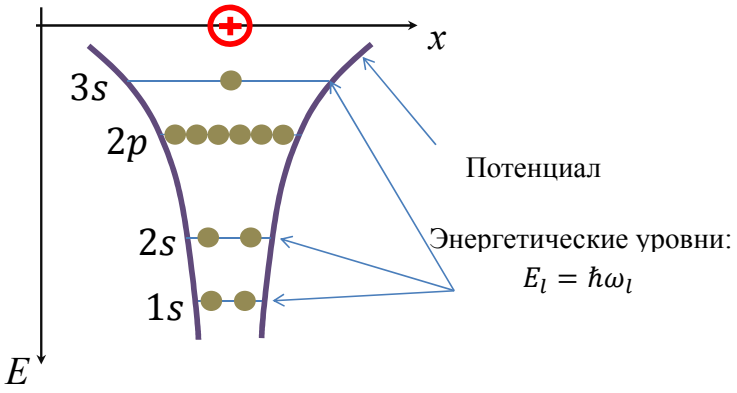


Рис. 3.4: Электронная структура атома натрия.

незначительно отличается от его состояния в изолированном атоме, поэтому будем в оценке влияния на это состояние кристаллического поля соседних атомов исходить из энергетической структуры изолированного атома. На рис. 3.4 показана схема энергетических уровней и распределение электронов на них для таких атомов.

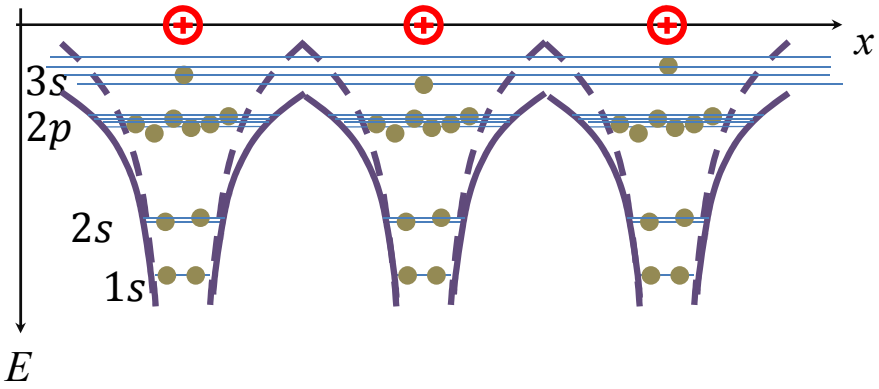


Рис. 3.5: Изменение состояний электронов при образовании кристаллической решетки.

Разрешенные уровни энергии дискретны и определяются квантовыми числами n, l, m. На каждом невырожденном по энергии уровне могут находиться с учетом спина по два электрона, а на каждом вырожденном по орбитальному квантовому числу уровне $2(2l+1)$ электронов.

Сблизим теперь эти атомы на расстояние, равное параметру кристаллической решетки (рис. 3.5). Взаимодействие с соседними атомами изменит положение энергетических уровней. В приближении сильной связи предполагается, что потенциальная энергия электрона в кристалле $U(r)$ может быть представлена суммой:

$$U(r) = U_a + \Delta U(r) \tag{3.8}$$

и считается, что возмущение слабое: $\Delta U(r) \ll U_a$.

Поскольку в кристалле каждый уровень изолированного атома повторяется N раз, то он становится N-кратно вырожденным. Электрическое поле снимает вырождение и каждый уровень расщепляется на N близко расположенных (по значениям энергии) энергетических уровней. Здесь имеется полная аналогия со связанными осцилляторами. Для двух не связанных между собой каким-либо взаимодействием совершенно одинаковых осцилляторов (математических маятников, колебательных контуров), частоты их собственных колебаний совпадают (см. рис.3.6).

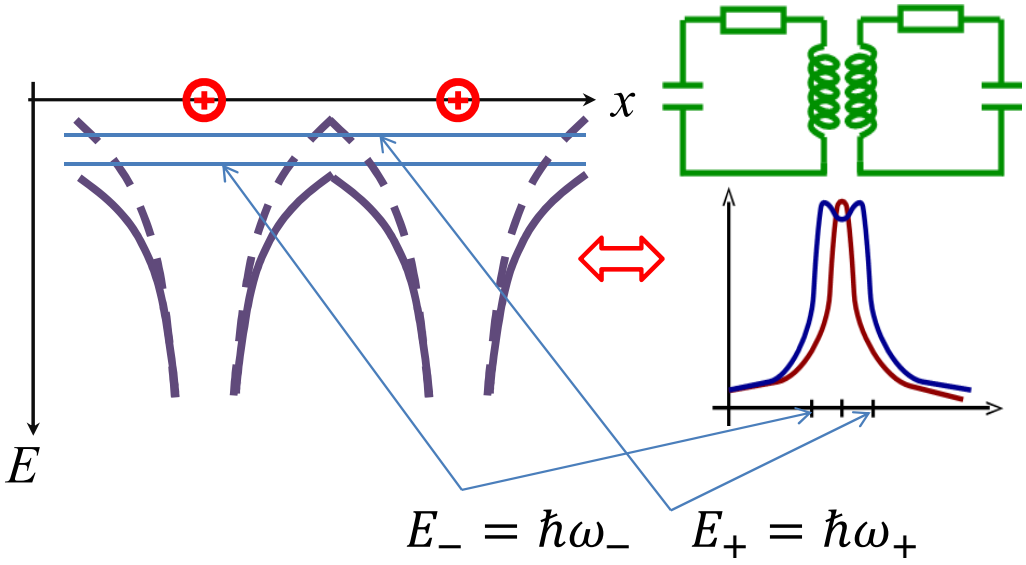


Рис. 3.6: Аналогия между расщеплением уровней и колебаниями в связанных контурах.

Взаимодействие между осцилляторами приводит к расщеплению одной частоты на две близкие частоты (при условии, что энергия взаимодействия между осцилляторами много меньше энергии собственных колебаний). Для N связанных между собой осцилляторов получим полосу из N близко расположенных частот. Аналогичный результат получается для системы взаимодействующих атомов. Число энергетических уровней, на которые расщепляется каждый энергетический уровень изолированного атома, равно числу атомов в кристалле. Величина расщепления тем больше, чем сильнее взаимодействие между атомами, т.е. чем меньше расстояние между ними. На рис. 3.7 показано схематически расщепление двух энергетических уровней атома под воздействием полей соседних атомов.

Ширина энергетической зоны обычно имеет порядок $\triangle E \simeq 1\,eV$. Ширина энергетической зоны для более высоких уровней больше, так как перекрытие электронных оболочек для этих уровней больше. Обычно энергетические зоны разделены интервалами $E_g = 0.1...10\,eV$, называемыми запрещенными зонами, но могут и перекрываться.

В реальных кристаллах содержится 10^{22} атомов. В этом случае расстояние между уровнями в зоне составляет $10^{-22} eV$, следовательно, спектр электронов в пределах энергетической зоны можно считать практически непрерывным.

Структура энергетических зон кристалла оказывает решающее влияние на величину его электропроводности. Можно оценить, каково увеличение энергии элек-

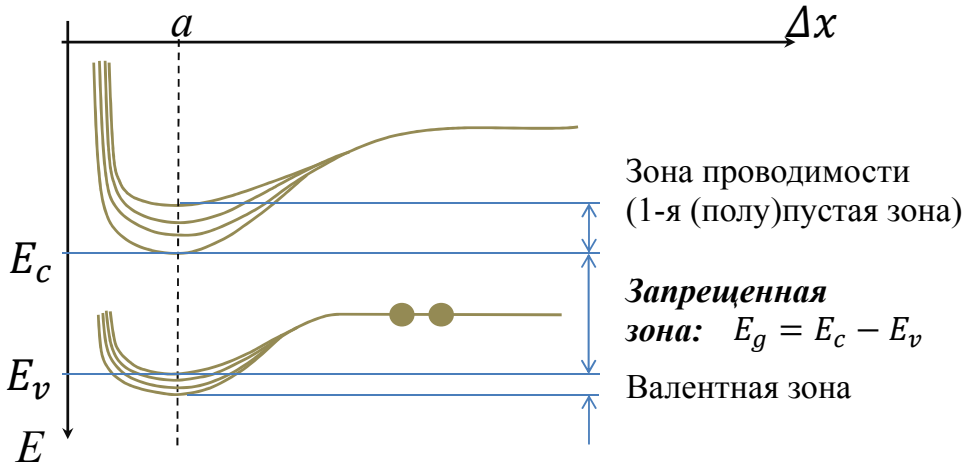


Рис. 3.7: Образование энергетических зон.

трона за счет его ускорения в электрическом поле, вызывающим электрический ток в проводниках.

В каждой энергетической зоне могут располагаться в соответствии принципом Паули не более $2(2l+1)N$ электронов - по два с противоположными спинами на каждом уровне. Число электронов в кристалле также конечно и зависит как от числа атомов N, так и от количества электронов в атоме. Поскольку электроны стремятся занять энергетические уровни с наинизшей энергией, то в кристалле нижние энергетические зоны оказываются полностью заполненными, а самые верхние заполнены либо частично, либо совершенно свободны. Частично заполненная зона образуется, например, у рассмотренного нами кристалла натрия. Этот элемент имеет полностью заполненные 1s -, 2s - и 2p -уровни, на которых располагается в общей сложности 10 электронов. В кристалле Na соответствующие 1s -, 2s - и 2p -зоны также будут полностью заполнены. Одиннадцатый валентный электрон в атоме Na располагается на 3s - уровне, на котором могут располагаться 2 электрона. Следовательно, 3s - зона кристаллического натрия будет заполнена лишь наполовину. Зонная структура металлов, подобных Na приведена на рис. 3.8, а.

Часто частично заполненная зона образуется в результате перекрытия полностью заполненной зоны со следующей совершенно свободной. Пример такой зонной структуры приведен на рис. 3.8, б для бериллия, у которого перекрываются заполненная 2s- и свободная 2p-зоны. Большую группу составляют кристаллы, у которых над целиком заполненным зонами располагаются совершенно пустые зоны, причем ширина запрещенной зоны варьируется у них от нескольких десятков электронвольт до единиц электронвольт. Типичные примеры этой группы кристаллов показаны на рис. 3.8, в, г. Это, например, углерод в модификации алмаза и кремний. Структура энергетических зон кристалла оказывает решающее влияние на величину его электропроводности. Поскольку электрический ток есть направленное движение зарядов (в металлах - электронов), то возникновение электрического тока связано с увеличением энергии электронов. Нетрудно оценить, каково увеличение энергии электрона за счет его ускорения в электрическом поле, вызывающим

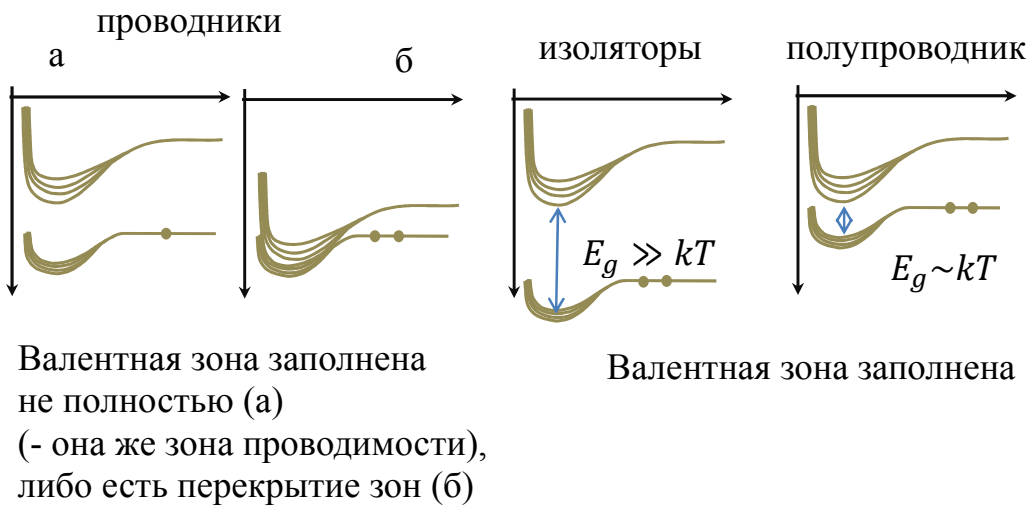


Рис. 3.8: Классификация твердых тел.

электрический ток в проводниках. Если величина напряженности электрического поля равна 10^4 В/м, то на расстоянии, равном средней длине свободного пробега электрона в кристалле, а она обычно составляет $\sim 10^{-8}$ м, электрон приобретает энергию приблизительно 10^{-4} эВ. Понятно, что эти значения энергии позволяют электрону переходить с уровня на уровень только внутри одной энергетической зоны. Для перехода между зонами необходима энергия больше ширины запрещенной зоны E_g.

Твердые тела с частично заполненными энергетическими зонами являются проводниками. Частично заполненные зоны имеют все металлы. твердые тела с полностью заполненными электронами энергетическими зонами являются непроводниками. По ширине запрещенной зоны непроводники делятся на диэлектрики и полупроводники. К диэлектрикам относят тела, имеющие относительно широкую запрещенную зону. У типичных диэлектриков $E_g > 3$ эВ. Так, у алмаза $E_g = 5,2$ эВ ; у нитрида бора $E_g = 4,6$ эВ; у Al_2O_3 $Eg = 7$ эВ. У типичных полупроводников ширина запрещенной зоны менее 3 эВ. Например, у германия $E_g = 0,66$ эВ; у кремния $E_g = 1,12$ эВ; у антимонида индия $E_g = 0,17$ эВ. Верхняя заполненная зона полупроводников и диэлектриков называется валентной зоной, следующая за ней свободная зона называется зоной проводимости. В металлах частично заполненную зону называют как валентной зоной, так и зоной проводимости.

Для технических приложений особый интерес представляют полупроводники. Концентрация носителей заряда, способных перемещаться под действием внешнего электрического поля (электронов в зоне проводимости и дырок - в валентной зоне) в полупроводниках намного меньше, чем в металлах, и ее можно изменять, изменяя таким образом проводимость. Движение носителей заряда в полупроводнике, вызванное наличием электрического поля и градиента потенциала, называют дрейфом, а созданный этими зарядами ток - дрейфовым током. Кроме того, носители заряда могут перемещаться под влиянием градиента их концентрации. Такое движение называют диффузией.

У рассмотренных нами полупроводников, называемых собственными, проводимость существенно зависит от температуры. В полупроводниковых приборах обычно используются полупроводники с примесями. Получаемый полупроводник называется несобственным (или примесным, или допированным). В качестве примесей выбирают вещества, валентность которых отличается на единицу от валентности собственного полупроводника. Примером может служить мышьяк (As, валентность - 5), который используют для допирования кремния (Si, валентность - 4). Когда такие примесные атомы замещают собой в решетке атомы исходного кристалла, 4 электрона ведут себя так же, как электроны кремния, образуя связи, а пятый оказывается слабо связанным с положительным ионом и может легко оторваться и перемещаться по кристаллу. Хотя примесные атомы - это дефекты, и кристалл с примесями идеальным не является, малая концентрация примеси (на практике - порядка 0.01%) позволяет использовать выводы, полученные нами ранее для идеального кристалла. В этом случае наличие легко ионизируемых примесных атомов означает, что внутри запрещенной зоны появляются дополнительные (примесные) уровни, расположенные вблизи дна зоны проводимости (на расстоянии $0.01 - 0.05$ эВ). Если концентрация примеси мала, что чаще всего имеет место, то ее атомы можно рассматривать как изолированные. Их энергетические уровни не расщепляются на зоны. В рассмотренном примере атомы As отдают электроны, которые переходят с примесного уровня, называемого донорным, в зону проводимости. Такой полупроводник называют полупроводником n-типа. Аналогично, допирование атомами, которые могут "забирать" электрон, образуя дырку (например, атомы бора в кремнии), приводит к появлению акцепторных уровней, расположенных вблизи потолка валентной зоны. Такие полупроводники называются полупроводниками p-типа (см. рис. 3.9).

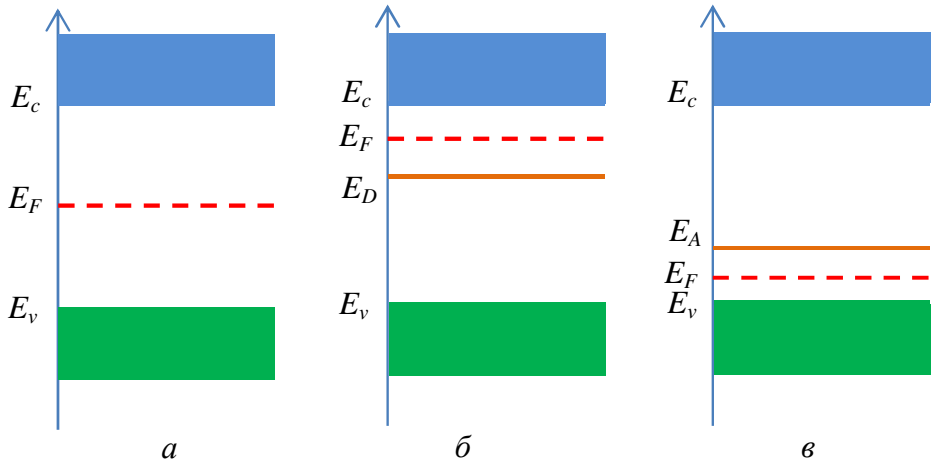


Рис. 3.9: Энергетические зоны в полупроводниках: а - собственном, б - n - типа, в - p - типа.

Если концентрация примесей в полупроводнике достаточно велика, то примесные уровни расщепляются, образуя зону, которая может пересечься с валентной

зоной. Такой полупроводник называют вырожденным.

При описании электрических свойств металлов и полупроводников важную роль играет параметр, называемый энергией Ферми (уровнем Ферми) E_F. Энергия Ферми системы невзаимодействующих фермионов — это увеличение энергии основного состояния системы при добавлении в нее еще одной частицы. Энергия Ферми может также интерпретироваться как максимальная энергия фермиона в основном состоянии при абсолютном нуле температур. Вероятность обнаружения частицы на уровне Ферми составляет 0,5 при любых температурах, уровни, лежащие ниже уровня Ферми заняты с вероятностью больше 0,5, а лежащие выше - с вероятностью, больше 0,5, свободны.

Концентрация электронов в зоне проводимости может быть записана как:

$$n = N_c e^{-(E_c - E_F)/kT}, \tag{3.9}$$

где

$$N_c = 2\left(\frac{2\pi m_n^* kT}{\hbar^2}\right)^{3/2}$$

— эффективная плотность состояний в зоне проводимости, m_n^* — эффективная масса электрона, E_c - энергия, соответствующая нижней границе зоны проводимости. Аналогичные выражения можно записать для дырок:

$$p = N_v e^{-(E_F - E_v)/kT} \tag{3.10}$$

где

$$N_v = 2\left(\frac{2\pi m_p^* kT}{\hbar^2}\right)^{3/2}$$

- эффективная плотность состояний в валентной зоне и, соответственно, , m_p^* - эффективная масса дырок а E_v - энергия, соответствующая верхней границе валентной зоны проводимости. Следует иметь ввиду, что эффективные массы носителей зарядов характеризуют их движение и не связаны напрямую с обычной массой электронов. В частности, они могут принимать отрицательные значения. Перемножая 3.10 и 3.9 , можно получить интересный результат, называемый законом действующих масс:

$$np = N_c N_v e^{-(E_c - E_v)/kT} = N_c N_v e^{-(\Delta E)/kT} \tag{3.11}$$

- в состоянии равновесия произведение концентраций носителей зарядов есть величина постоянная и не зависящая от концентрации и распределения примесей.

Так как при данной температуре количество электронов и дырок постоянно, то рекомбинация одной пары вызовет генерацию электрона и дырки в другом месте. Рекомбинация и генерация дырок и электронов в полупроводнике происходят непрерывно. Существует несколько видов рекомбинаций: межзонная, через рекомбинационные центры, поверхностная. При межзонной рекомбинации электроны из зоны проводимости непосредственно переходят в валентную зону. При этом выделяется энергия, равная ширине запрещенной зоны. Эта энергия может выделиться или в виде фотона (излучательная рекомбинация), или в виде фонона (безизлучательная рекомбинация).

В кристаллах всегда есть атомы примесей и дефекты структуры, энергетические уровни которых находятся в запрещенной зоне. Они называются рекомбинационными центрами или ловушками. Электрон из зоны проводимости может перейти на энергетический уровень ловушки, а, затем либо вернуться назад либо перейти в валентную зону. В последнем случае имеет место рекомбинация, которая носит двухступенчатый характер. Поверхностная рекомбинация обусловлена тем, что на поверхности кристалла, в силу нарушения симметрии, а так же из-за наличия загрязнений, появляются поверхностные состояния, энергетические уровни которых так же лежат в запрещенной зоне.

Если в полупроводник искусственно ввести дополнительные электроны или дырки, то возникшее электрическое поле приведет к перераспределению зарядов. Время релаксации обычно имеет порядок 10^{-12} с. При рассмотрении процессов, характерное время которых существенно больше (а это все процессы в электрических цепях, которые мы будем изучать), можно считать, что в однородном полупроводнике, независимо от характера и скорости образования носителей, не могут существовать существенные объемные заряды. Следовательно, энергия необходимая для внесения дополнительного заряда в любую область полупроводника (энергия Ферми) в отсутствие внешнего электрического поля одинакова.

В полупроводниковых приборах важную роль играют переходы - области, в которых скачкообразно меняются параметры полупроводника: тип основных носителей (электроны или дырки - pn переход), их концентрация (p^+p^- или n^+n^- переход), положение энергетических зон. Переходы между двумя полупроводниковыми материалами, имеющими различную ширину запрещенной зоны, называют гетеропереходами. Если одна из областей, образующих переход, является металлом, то такой переход называют переходом металл - полупроводник. На практике переход практически невозможно создать путем простого механического контакта двух областей с разными физическими свойствами, поскольку между поверхностями всегда оказываются загрязнения и окисные пленки.

Рассмотрим вначале контакт металл-полупроводник.

Предположим, что уровень Ферми в металле, который всегда расположен в зоне проводимости, лежит выше уровня Ферми полупроводника p-типа, с которым они образуют контакт.

Так как энергия электронов металла больше энергии носителей заряда полупроводника, то электроны будут переходить из металла в полупроводник и рекомбинировать в приграничном слое с дырками (основными носителями). В результате, у барьера появляются области объемного заряда, образованные неподвижными ионами, заряд которых не скомпенсирован носителями.

Создаваемое ими электрическое поле будет препятствовать дальнейшему движению электронов из металла в полупроводник. Так как концентрация основных носителей заряда (дырок) в приконтактном слое полупроводника понижена по сравнению с их концентрацией в его объеме, то этот слой имеет повышенное удельное сопротивление, которое будет определять сопротивление всей системы.

В состоянии равновесия уровень Ферми в системе должен быть общим, а энергетические уровни, соответствующие границам зон, в области контакта искривлены, поскольку наличие локального электрического поля изменяет энергию, которая нужна, что бы электрон перешел в зону или покинул ее (см. рис. 3.10).

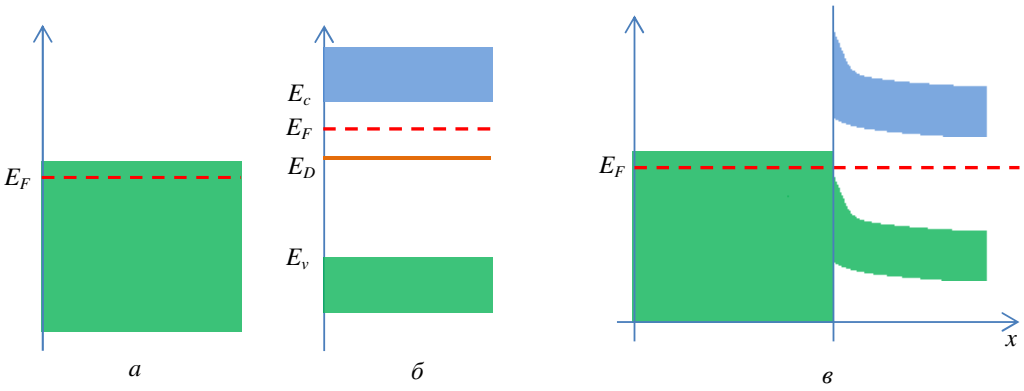


Рис. 3.10: Энергетические зоны: а - металл, б - полупроводник, в - контакт металл-полупроводник (барьер Шоттки).

Если к системе подключить внешний источник тока "плюсом" к полупроводнику, то внешнее электрическое поле уменьшит сопротивление приконтактного слоя полупроводника. Через переход потечет ток, обусловленный переходом электронов из металла в полупроводник. При обратной полярности приложенного напряжения внешнее электрическое поле увеличивает сопротивление перехода. Ток может протекать практически только за счет движения электронов в полупроводнике (неосновных носителей) и будет очень мал. Таким образом, переход между металлом и полупроводником обладает вентильными свойствами. Его называют барьером Шоттки.

Особый интерес представляет случай контакта металл-полупроводник, когда уровень Ферми металла исходно лежит ниже уровня Ферми полупроводника p-типа (или выше уровня Ферми полупроводника n-типа).При образовании контакта приконтактный слой будет не обеднен, а обогащен основными носителями и его удельное сопротивление окажется меньше, чем соответствующее сопротивление вдали от границы. Такие переходы являются основой омического, или недетектирующего, контакта. Действительно, при соединении металла с полупроводником p-типа при $\mathsf{E}_\mathsf{F}^{\mathsf{Me}} < \mathsf{E}_\mathsf{F}^{\mathsf{p}}$, электроны полупроводника перейдут в металл. В результате этого приповерхностный слой окажется обогащенным основными носителями заряда-дырками. Удельное сопротивление тонкой приконтактной области станет меньше, чем в объеме полупроводника и оказывать влияние на полное сопротивление она практически не будет. Подключение напряжения прямой или обратной полярности изменяет лишь степень обогащения приконтактной области электронами. На основе таких переходов металл-полупроводник выполняются выводы от полупроводников в полупроводниковых приборах.

3.4 Диоды и их применение

3.4.1 Принцип работы полупроводникового диода

Перейдем к рассмотрению контакта двух полупроводников. Рассмотрим переход между двумя областями полупроводника, имеющими различный тип электропроводности. Если концентрации основных носителей заряда в этих областях равны, то переход, называется симметричным. если же они существенно различаются - несимметричным. Несимметричные переходы используются чаще, поэтому будем рассматривать только их. Пусть концентрация дырок в области полупроводника с электропроводностью p-типа, много выше концентрации электронов в области n (проводимость p-области выше). Так как концентрация дырок в области p выше, чем в n-области, то дырки в результате диффузии будут переходить из p- в n-область, где вблизи границы станут рекомбинировать с электронами. Соответственно в этой области концентрация свободных электронов станет еще меньше и образуется слой нескомпенсированных положительных ионов донорной примеси. Аналогично, в n-области уход дырок приведет к образованию слоя нескомпенсированных отрицательных ионов акцепторной примеси. Возникающее локальное электрическое поле приведет к установлению равновесия. Диффузией электронов в связи с их малой концентрацией можно пренебречь. В результате, уровень Ферми частей станет одинаковым (см. рис. 3.11). В приконтактной области концентрация основных носителей становится пониженной, следовательно, эта область имеет повышенное сопротивление, которое определяет электрическое сопротивление всей системы. Некомпенсированные ионы в p-n переходе создают разность потенциа-

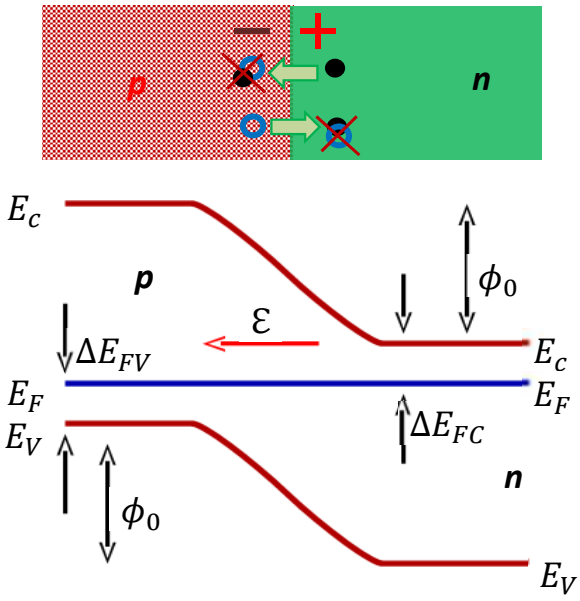


Рис. 3.11: p-n - переход в полупроводнике и его энергетическая диаграмма.

лов U_k, которую называют потенциальным барьером или контактной разностью потенциалов. Значение U_k определяется исходными положениями уровней Ферми в областях n- и p- которые, в свою очередь, зависят от концентраций примесей.

Значение U_k у германиевых полупроводниковых приборов при комнатной температуре не превышает 0,4 В; в кремниевых приборах U_k может достигать 0,7-0,8 В. Если внешний источник напряжения подключить так, что положительный контакт соединен с p-областью (говорят, что переход смещен в прямом направлении), то дополнительное внешнее электрическое поле, будет уменьшать внутреннее. Ширина приконтактной области, в которой концентрация основных носителей понижена, станет уменьшаться. При $U = U_k$ она стремится к нулю и основные носители начинают свободно диффундировать в области с противоположным типом проводимости. Через переход потечет ток, который называется прямым. Введение носителей заряда через переход в область полупроводника, где они являются неосновными носителями за счет снижения потенциального барьера называется инжекцией.

Если внешний источник напряжения подключить так, что положительный контакт соединен с n-областью (переход смещен в обратном направлении) то потенциальный барьер повышается. Движение основных носителей через переход уменьшится и при некотором значении U совсем прекратится, т. е. в этом случае электроны и дырки начнут двигаться от p-n-перехода (дефицит свободных носителей заряда в приконтактной области увеличится). При этом ток обусловлен движением только неосновных носителей. Процесс "отсоса" неосновных носителей заряда (при обратном включении напряжения) называется экстракцией.

Таким образом, p-n-переход так же обладает вентильными свойствами. При приложении напряжения, смещающего его в прямом направлении, через переход протекает электрический ток, значение которого при повышении напряжения увеличиваются по экспоненциальному закону. Изменение полярности приложенного напряжения приводит к смещению перехода в обратном направлении и его сопротивление возрастает. Через переход протекает малый ток обусловленный тепловой активацией неосновных носителей, значение которого практически не зависит от приложенного напряжения и увеличивается по экспоненциальному закону при повышении температуры.

Рассмотренная нами идеальная модель не учитывает многих особенностей реальных полупроводниковых приборов. Так в идеальном p-n-переходе обратный ток уже при сравнительно небольшом обратном напряжении не зависит от значения последнего. Однако, на практике в диодах обратный ток растет при увеличении приложенного напряжения, и может быть на 2-3 порядка выше величины, связанной с дрейфом неосновных носителей. Причиной этого являются три фактора: генерация (рождение пар "электрон-дырка") носителей непосредственно в приконтактной области, наличие канальных токов и токов утечки. Канальные токи связаны с наличием поверхностных энергетических состояний, искривляющих энергетические зоны вблизи поверхности и приводящих к появлению инверсных слоев. Эти слои называют каналами, а токи, по ним - канальными токами. Токи утечки обычно возникают из-за наличия загрязнений на поверхности полупроводника.

Наряду с электропроводностью p-n-переход имеет определенную емкость. Емкость перехода связана с наличием по обе стороны от него подвижных зарядов. Емкость p-n-перехода подразделяют на две составляющие: барьерную, отражающую перераспределение зарядов в переходе, и диффузионную, отражающую перераспределение зарядов вблизи перехода. При прямом смещении перехода в основном проявляется диффузионная емкость, при обратном - барьерная. На практике инте-

рес представляет последняя, так как при обратном смещении постоянный ток через переход мал и переход можно рассматривать как конденсатор. Следует иметь ввиду, что, поскольку запирающее напряжение влияет на ширину p-n-перехода, то при его изменении будет изменяться величина емкости, то есть такой конденсатор является нелинейным.

При приложении внешнего напряжения, превышающего определенную величину, возможно резкое уменьшение сопротивления p-n перехода, называемое пробоем. Различают три вида пробоя: туннельный, лавинный и тепловой. В основе туннельного пробоя лежит туннельный эффект: электроны и дырки могут преодолевать узкий потенциальный барьер, высота которого больше, чем их энергия.

Лавинный пробой вызывается ударной ионизацией, которая происходит тогда, когда напряженность электрического поля, вызванная обратным напряжением, достаточно велика. При этом носители заряда, движущиеся через p-n-переход, ускоряются настолько, что при соударении с атомами в зоне перехода ионизируют их. В результате появляются новые пары электрон-дырка. Вновь появившиеся носители заряда так же ускоряются электрическим полем и в свою очередь могут вызвать ионизацию следующего атома. Если процесс ударной ионизации идет лавинообразно, то по тому же закону увеличиваются количество носителей заряда и обратный ток. При лавинной ионизации ток в цепи ограничен только внешним сопротивлением.

Тепловой пробой возникает в результате разогрева перехода, когда количество Джоулевого тепла, выделяемого током в переходе, больше количества теплоты, отводимой от него. При разогреве перехода происходит интенсивная генерация электронно-дырочных пар и увеличение тока через переход. Это, в свою очередь, приводит к дальнейшему увеличению температуры. В итоге, ток через переход лавинообразно увеличивается. Следует заметить, что один вид пробоя может наступать как следствие другого вида пробоя. Пробой перехода может быть как обратимым, так и необратимым. В первом случае ток, как правило, ограничивается сопротивлением внешней цепи. Во втором случае наступает разрушение перехода из-за его перегрева.

Типичный вид зависимости тока через полупроводниковый диод от приложенного напряжения изображен на рис. 3.12. Большинство диодов имеют так называемое напряжение открывания ($\mathfrak{u}_0$), малым током при напряжении меньше которого и при обратном смещении обычно можно пренебречь. Для кремниевых диодов это напряжение обычно лежит в диапазоне 0.4 - 1 В, для германиевых - 0.2 - 0.4 В. Участок характеристики в диапазоне $\simeq \mathfrak{u}_0 - 3\mathfrak{u}_0$ хорошо аппроксимируется квадратичной параболой.

В зависимости от того, какую амплитуду имеет входной сигнал, используют различные аппроксимации вольт-амперной характеристики. При малых сигналах (амплитуда порядка $\mathfrak{u}_0$ и менее) удобно использовать аппроксимацию в виде степенного ряда. Для многих приложений достаточно ограничиться линейным и квадратичным членом и рассматривать только правую часть характеристики, полагая, что напряжение на входе - всегда положительное и изменяется в небольших пределах. На практике ко входному сигналу часто добавляют постоянную составляющую, называемую напряжением смещения, так, что бы использовался квадратичный участок характеристики диода. Такой режим используется в модуляторах и

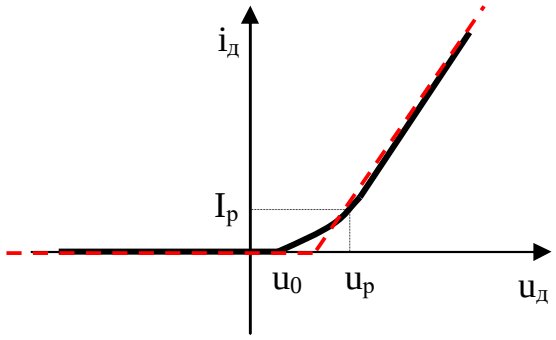


Рис. 3.12: Вольт-амперная характеристика диода (пунктир - кусочно-линейная аппроксимация)

.

квадратичных детекторах, рассматриваемых ниже. Если же напряжение на входе принимает как положительные, так и отрицательные значения и достаточно велико по сравнению с u_0, используют кусочно-линейную аппроксимацию (см. рис.3.12) при этом тонкие детали характеристики не играют роли). Такой режим используется в выпрямителях и "линейных" детекторах, так же рассматриваемых далее.

Рассмотрим несколько примеров использования нелинейных свойств диодов для модуляции и детектирования.

3.4.2 Модуляция

Модуляцией называют медленное, по сравнению с периодом несущей, изменение амплитуды (АМ), частоты (ЧМ) или фазы (ФМ). Математическое описание этих трех видов модуляции сигналов было дано в разделе 1.2.2. Теперь рассмотрим способы получения модулированного сигнала.

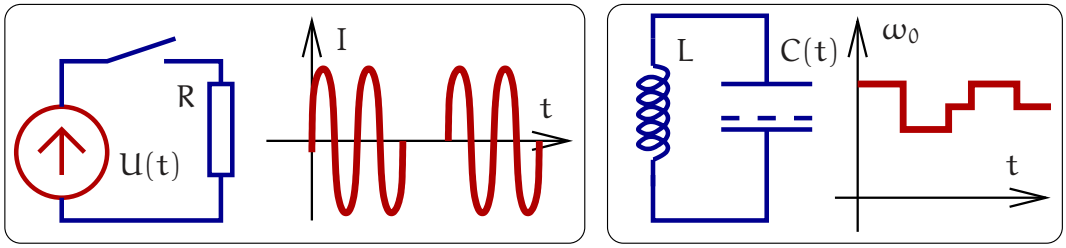


Рис. 3.13: Два примера получения импульсной модуляции амплитуды (слева) и частоты (справа).

Два простейших примера импульсной модуляции амплитуды и частоты приведены на рис. 3.13. Замыкание и размыкание ключа приводит к глубокой амплитудной модуляции (слева). На этом способе модуляции основана передача сообщений с

помощью азбуки Морзе. Справа модуляция емкости выходного контура генератора (сам генератор не показан) приводит к частотной модуляции. Из этих примеров ясно, что для получения модулированного сигнала надо управлять параметрами цепи, что возможно только с использованием *нелинейных* элементов (в нашем случае диодов).

3.4.3 Получение AM сигнала

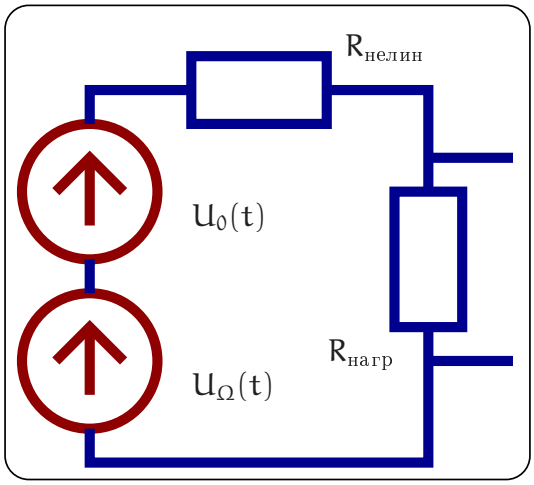


Рис. 3.14: Использование нелинейного сопротивления для получения амплитудно- модулированного сигнала.

Для модуляции иногда достаточно использования нелинейного сопротивления R, емкости C или индуктивности L. В качестве примера рассмотрим цепь, изображенную на рис. 3.14, состоящую из двух источников сигнала (несущая $V_0 = U_0 \sin \omega_0 t$ и модулирующий сигнал $V_\Omega = U_\Omega \sin \Omega t$, $\omega_0 \gg \Omega$), сопротивления нагрузки $R_{нагр}$ и нелинейного сопротивления $R_{нелин}$.

Пусть ВАХ нелинейного сопротивления описывается формулой

$$I = S_1 U + S_2 U^2 \simeq S_1 \left(U_0 \sin \omega_0 t + U_\Omega \sin \Omega t\right) + S_2 \left(U_0 \sin \omega_0 t + U_\Omega \sin \Omega t\right)^2.$$

Это приблизительно соответствует ВАХ полупроводникового диода для малых токов. Примем также, что $R_{нагр} \ll R_{нелин}$. Тогда напряжение на нагрузке будет равно

$$\begin{aligned} U_{нагр} \simeq I R_{нагр} \simeq R_{нагр} \Big(& S_1 \left[U_0 \sin \omega_0 t + U_\Omega \sin \Omega t\right] + \\ & + S_2 \left[U_0^2 \sin^2 \omega_0 t + U_\Omega^2 \sin^2 \Omega t\right] + S_2 U_0 U_\Omega \left[\cos(\omega_0 - \Omega)t - \cos(\omega_0 + \Omega)t\right] \Big) = \\ & = \mathbf{R_{нагр}} \Big(\mathbf{S_1} [\mathbf{U_0} \sin \boldsymbol{\omega_0 t} + \mathbf{S_2 U_0 U_\Omega} \left[\cos(\boldsymbol{\omega_0} - \Omega)\mathbf{t} - \cos(\boldsymbol{\omega_0} + \Omega)\mathbf{t}\right] \Big) + \\ & + R_{нагр} \Big(S_1 U_\Omega \sin \Omega t + S_2 \left[U_0^2 \sin^2 \omega_0 t + U_\Omega^2 \sin^2 \Omega t\right]. \end{aligned} \tag{3.12}$$

Напомним, что модуляция соответствует появлению в спектре выходного напряжения частот $\omega_0 \pm \Omega$. Мы видим, что такие частоты присутствуют в выходном

напряжении (выделенные члены в формуле (3.12)), что соответствует амплитудной модуляции. Правда, есть и "ненужные" нам частоты (Ω, 2Ω, $2\omega_0$). Чтобы избавится от них, надо выходной сигнал пропустить через полосовой фильтр так, чтобы остались только частоты ω_0, $\omega_0 \pm \Omega$.

Если ВАХ содержит дополнительные члены $S_3U^3 + S_4U^4 + \ldots$, то появится искажение сигнала. Подробнее:

$$\begin{aligned} S_3(V_0 + V_\Omega)^3 &\Rightarrow 3V_0V_\Omega^2 = 3U_0U_\Omega^2 \sin\omega_0 t \sin^2\Omega t \Rightarrow \\ &\Rightarrow \frac{3U_0U_\Omega^2}{4}\Big(\sin(\omega+2\Omega)t + \sin(\omega-2\Omega)t\Big), \\ S_4(V_0 + V_\Omega)^4 &\Rightarrow 4V_0V_\Omega^3 = 4U_0U_\Omega^3 \sin\omega t \sin^3\Omega t \Rightarrow \\ &\Rightarrow \frac{U_1U_2^3}{2}\Big(\cos(\omega+3\Omega)t + \cos(\omega-3\Omega)t\Big) \end{aligned}$$

. Мы видим, что присутствуют составляющие на частотах ($\omega_0 \pm 2\Omega$, $\omega_0 \pm 3\Omega$). От таких искажений с помощью полосового фильтра не избавишься, поэтому обычно стараются выбрать так рабочую точку на ВАХ, чтобы коэффициенты S_3, S_4 были достаточно малы.

3.4.4 Детектирование АМ сигнала

Такая же схема (см. рис. 3.15) может быть использована и для детектирования АМ сигнала.

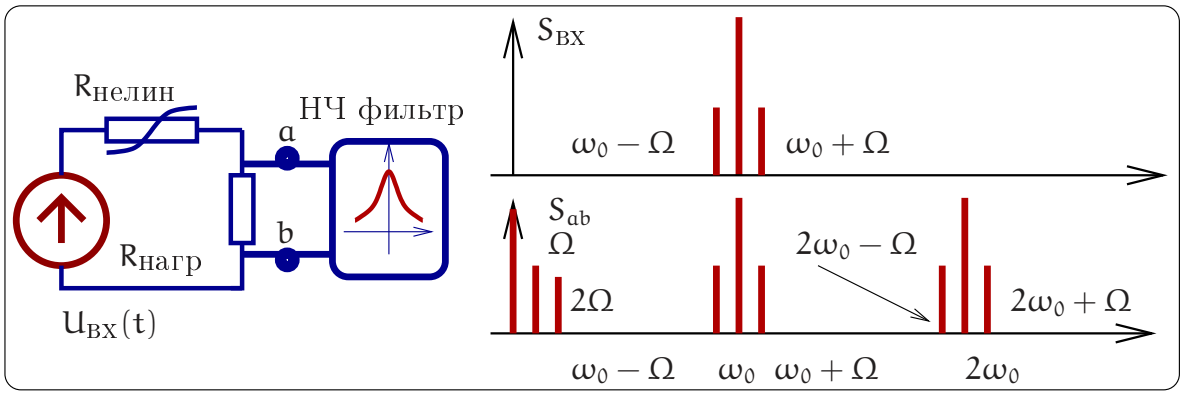


Рис. 3.15: Схема детектирования АМ сигнала (слева) и спектры входного и выходного сигналов (справа).

Пусть входное напряжение есть амплитудно-модулированный сигнал:

$$\begin{aligned} U(t) &= U_0\,(1 + m\sin\Omega t)\,\sin\omega_0 t = \\ &= U_0\left(\sin\omega_0 t + \frac{m}{2}\,[\cos(\omega_0-\Omega)t - \cos(\omega_0+\Omega)t]\right). \end{aligned}$$

Нашей задачей является выделить сигнал модуляции на частоте Ω. Пусть опять ВАХ нелинейного сопротивления описывается формулой $I = S_1U + S_2U^2$ (это приблизительно соответствует ВАХ полупроводникового диода для малых токов).

Примем также, что $R_{нагр} \ll R_{нелин}$, $m \ll 1$. Тогда для выходного напряжения получаем:

$$\begin{aligned} U_{ab}(t) &\simeq R_{нагр} I = R_{нагр}\left(S_1 U(t) + S_2 U(t)^2 + \dots\right) = \\ &= R_{нагр} S_1 U_0 \,(1 + m\sin\Omega t)\,\sin\omega_0 t + \\ &\quad + R_{нагр} S_2 U_0^2 \underbrace{(1 + m\sin\Omega t)^2}_{1+2m\sin\Omega t} \underbrace{\sin^2\omega_0 t}_{1/2} + \dots = \\ &= S_1 \dots + S_2 R_{нагр} S_2 U_0^2 \left(\frac{1}{2} + \frac{1}{2}\,2m\sin\Omega t + \dots\right). \end{aligned}$$

Мы видим, что в спектре выходного сигнала присутствует нужная нам частота Ω, сигнал на которой должен быть затем отфильтрован. После фильтра мы получим:

$$U_{ab}^{\text{после фильтра}}(t) \simeq R_{нагр} S_2 U_0^2 \times m\sin\Omega t.$$

Полезно сравнить спектры входного и выходного сигналов, приведенные на рис. 3.15 справа. Мы видим, что три частоты (ω_0, $\omega_0 \pm \Omega$) во входном напряжении, превращаются в три "набора": (0, Ω, 2Ω), (ω_0, $\omega_0 \pm \Omega$), ($2\omega_0$, $2\omega_0 \pm \Omega$). Если ВАХ диода описывается более сложной функцией, содержащей и другие члены типа $S_3U^3 + S_4U^4 + \dots$, то будут и "наборы" вида: ($3\omega_0$, $3\omega_0 \pm \Omega$, $3\omega_0 \pm 2\Omega$, $3\omega_0$, $3\omega_0 \pm 3\Omega$). Подчеркнем, что именно наличие *нелинейного* элемента приводит к такому умножению частот.

Однополупериодный выпрямитель и "линейное" детектирование

Использование квадратичной ВАХ диода вида $I = S_1U + S_2U^2$ соответствует случаю, когда входной сигнал мал и нет возможности предварительно усилить его до детектирования. В противоположном случае большого входного сигнала ВАХ диода можно аппроксимировать кусочно линейной функцией, как показано на рис. 3.16в: в прямом направлении ток пропорционален напряжению $I = U_д/R_i$ (R_i — сопротивление диода в прямом направлении), а в обратном направлении ток через диод отсутствует.

Рассмотрим схему на рис. 3.16а. Пока примем, что входное напряжение не модулировано и равно $U(t) = U_0 \cos\omega_0 t$. Выберем время релаксации RC цепочки достаточно большим: $R_{нагр}C \gg 1/\omega_0$, т.е. за период $2\pi/\omega_0$ конденсатор не успевает разрядиться. Тогда большую часть периода диод будет заперт, т.к. напряжение $U_{вых}$ в это время будет больше $U_{вх}$ и ток через диод будет отсутствовать. В это время конденсатор будет медленно разряжаться на сопротивление $R_{нагр}$. Диод будет открываться на малую часть периода, когда входное напряжение больше напряжения на конденсаторе. В это время через диод будут проходить импульсы тока, показанные на рис. 3.16г.. Время t_0 открытого состояния диода обычно измеряют в радианной мере по формуле

$$\theta = \frac{\omega_0 t_0}{2}$$

и величину θ называют углом отсечки (см. также рис. 3.16б).

Расчет, который мы здесь не приводим, дает формулу для расчета угла отсечки:

$$\tan\theta - \theta = \frac{\pi R_i}{R_{нагр}}.$$

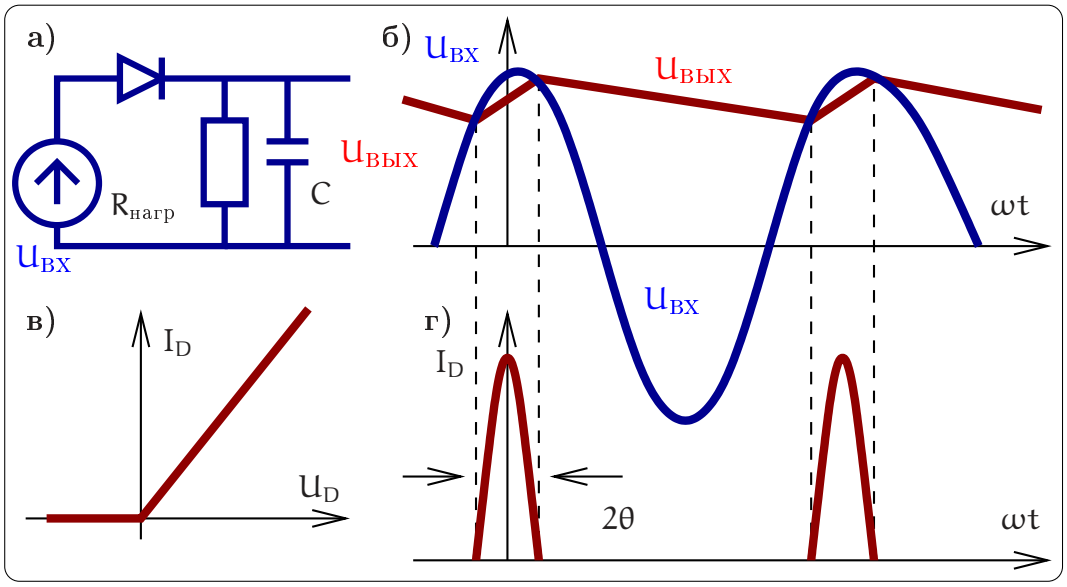


Рис. 3.16: Однополупериодное детектирование.

Для практически интересного случая, когда $R_i \ll R_{нагр}$, а следовательно, и $\theta \ll 1$, получаем асимптотику

$$\theta \simeq \sqrt[3]{\frac{3\pi R_i}{R_{нагр}}}.$$

Этой формулой обычно и пользуются на практике.

Теперь рассмотрим случай, когда входное напряжение амплитудно-модулировано:

$$U(t) = U_0 (1 + m \sin \Omega t) \sin \omega_0 t, \qquad \Omega \ll \omega_0.$$

Очевидно, что при следующих условиях

$$\omega_0 R_{нагр} C \gg 1, \qquad \Omega R_{нагр} C \ll 1 \tag{3.13}$$

будет реализовано амплитудное детектирование, т.е. в выходном сигнале присутствует постоянная составляющая и $U_{вых}(t) \sim U_0 m \sin \Omega t$. Действительно, первое неравенство в (3.13) означает, что за период $2\pi/\omega_0$ конденсатор не успевает разрядиться. А при выполнении второго неравенства в (3.13) напряжение на конденсаторе успевает изменяться с частотой модуляции Ω. Очевидно, что конденсатор C вместе с сопротивлением нагрузки образуют фильтр низких частот. Такой режим называют "линейным" детектированием, подразумевая, что используется линейная аппроксимация характеристики диода (не путать с линейностью в смысле принципа суперпозиции: линейные системы не меняют спектральный состав сигнала и для детектирования использованы быть не могут).

3.4.5 Фазовое детектирование

Для детектирования ФМ сигнала нужно опорное колебание. Пусть входное ФМ напряжение имеет вид $U_{ВХ}(t) = U_0 \cos(\omega_0 t + \phi(t))$, где в величине $\phi(t) \ll 1$ записана

информация:

$$U_{вх}(t) \;=\; U_0\,\cos(\omega t+\phi(t)) = U_0\cos\phi\,\cos\omega t - U_0\sin\phi\,\sin\omega t.$$

Далее будем считать, что $\phi(t) \ll 1$.

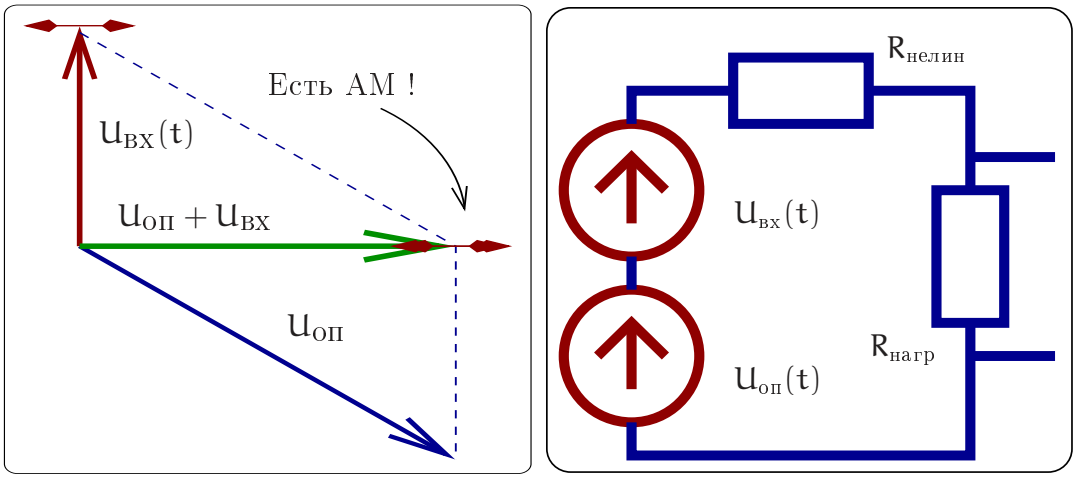


Рис. 3.17: Слева: фазовая диаграмма, показывающая, что сумма ФМ сигнала и опорного напряжения может быть АМ сигналом (при правильно подобранной фазе опорного напряжения). Справа: принципиальная схема фазового детектора.

Принцип детектирования ФМ сигнала заключается в том, чтобы до детектирования сначала превратить ФМ сигнал в АМ сигнал, который потом детектировать уже известным нам способом. Для превращения ФМ в АМ к ФМ сигналу добавляют опорное напряжение на частоте несущей. Фаза опорного напряжения должна быть выбрана оптимальным образом — это показано на фазовой диаграмме на рис. 3.17 слева. Принципиальная схема фазового детектора приведена на том же рисунке справа. Рассмотрим подробнее:

$$U(t) = U_{вх}(t) + U_{оп}(t) = \underbrace{\left(U_0\cos\phi\,\cos\omega t - U_0\sin\phi\,\sin\omega t\right)}_{U_{вх}(t)} + \underbrace{\left(-U_0\cos(\omega t) - U_1\sin\omega t\right)}_{U_{оп}(t)} =$$

$$\simeq -U_0\underbrace{\sin\phi(t)}_{\simeq\phi(t)}\sin\omega t + U_1\sin\omega t - U_0\underbrace{(1-\cos\phi)}_{\simeq\phi^2/2\ll 1}\cos\omega t \simeq$$

$$\simeq -U_1\left(1+\frac{U_0\phi(t)}{U_1}\right)\sin\omega t \quad\Rightarrow\quad \textbf{АМ сигнал.}$$

Мы видим, что эта сумма напряжений имеет вид АМ сигнала, который для детектирования можно подать на вход однополупериодного детектора, как это показано на рис. 3.17 справа. Фаза опорного напряжения определяется соотношением между U_0 и U_1.

Иногда применяют схему балансного фазового детектора, изображенную на рис. 3.18 слева. Важно, чтобы оба плеча балансного детектора были идентичны друг другу. В этой схеме на вход каждого детектора подаются напряжения

$$U_{A0} = U_{оп} - U_{вх}, \quad U_{B0} = U_{оп} + U_{вх}, \quad U_{оп} = U_1\cos(\omega t+\theta),$$

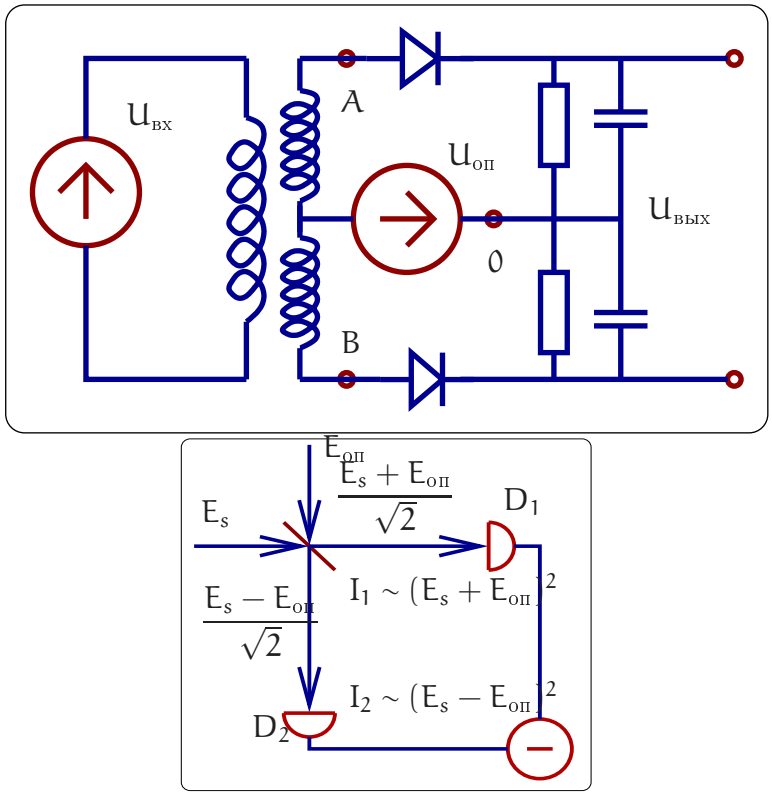


Рис. 3.18: Вверху: принципиальная схема балансного фазового детектора. Внизу: принципиальная схема балансного гомодинного детектора, применяемого в оптике.

где θ – фаза опорного колебания. Пусть детекторы квадратичные, т.е токи в детекторах пропорциональны квадрату напряжения. Тогда на выходе мы получим напряжение пропорциональное разности квадратов напряжений $U_{A0}^2 - U_{B0}^2$:

$$U_{вых} \sim (U_{оп} + U_{вх})^2 - (U_{оп} - U_{вх})^2 = 2U_{оп}U_{вх} = -U_0U_1\cos[\theta - \phi(t)] + \dots$$

$$\text{После фильтрации:} \quad -U_0U_1 \sin\phi(t), \quad \text{при } \theta = \frac{\pi}{2}$$

Меняя фазу θ можно измерять любую квадратуру, т.е. детектировать АМ-, ФМ-сигналы или сигнал, содержащий комбинацию АМ и ФМ.

Заметим, что в оптике аналогом фазового детектора является балансный гомодинный детектор, схема которого приведена на рис. 3.18 снизу.

3.4.6 Частотное детектирование

Согласно 1.51 частотно-модулированный сигнал при $\Delta\omega(t) = m\sin(\Omega t)$ представим в виде

$$U(t) = U_0 \sin(\omega_0 t - \frac{m}{\Omega}\cos(\Omega t)).$$

Он может быть преобразован в АМ сигнал пропусканием через линейную цепь, коэффициент пропускания которой имеет частотную зависимость. Например, для этого можно использовать резонансный контур, настраивая несущую частоту ω_0 на склон резонансной кривой контура (см. рис. 3.19).

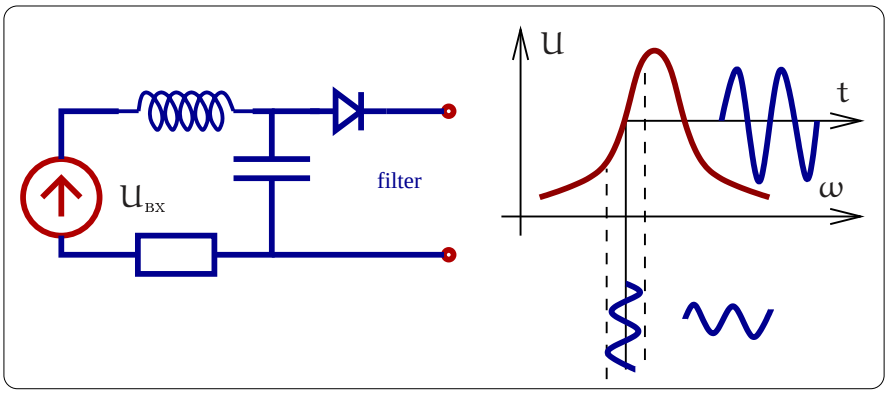


Рис. 3.19: Детектирование частотно-модулированного сигнала.

3.4.7 Синхронное детектирование

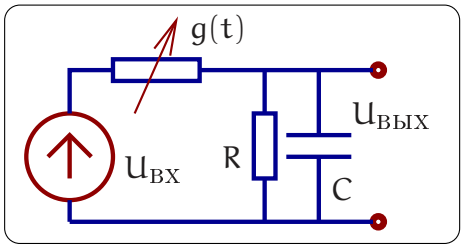


Рис. 3.20: Синхронное детектирование.

Для детектирования вместо нелинейного элемента может быть использован линейный элемент (например, сопротивление), величина которого не постоянна, а изменяется с частотой несущей. В качестве примера рассмотрим схему на рис. 3.20. Пусть

$$\begin{aligned} U_{вх}(t) &= U_m(t)\cos(\omega t + \phi(t)), \\ g(t) &= g_0 + g_1\cos(\omega t + \theta), \qquad R, \quad \frac{1}{i\omega C} \ll \frac{1}{g}. \end{aligned}$$

Примем для простоты, что проводимость $g(t)$ достаточно мала, так что большая часть входного напряжения падает на ней. Тогда ток определяется формулой

$$\begin{aligned} I(t) &\simeq g(t)U_{вх}(t) = g_0 U_m(t)\cos(\omega t + \phi(t)) + \frac{g_1 U_m(t)}{2}\cos(2\omega t + \phi(t) + \theta) + \\ &+ \underbrace{\frac{g_1 U_m(t)}{2}\cos(\phi(t) - \theta)}_{\text{н.ч. составляющая}}, \\ I_{нч} &= \frac{g_1 U_m(t)}{2}\sin\phi(t), \quad \text{при } \theta = \frac{\pi}{2}. \end{aligned}$$

Мы видим, что ток содержит медленную составляющую, что и означает детектирование. (Эта медленная составляющая отделяется от высокочастотных составляющих с помощью простейшего RC-фильтра.).

3.5 Транзисторы и их применение

Транзисторы - это полупроводниковые устройства, позволяющие с помощью тока в одной цепи, называемой входной, управлять током в другой. Транзисторы используется для усиления, генерации и преобразования электрических сигналов.

Изобретение в 1947 г. группой сотрудников из Bell Laboratories транзистора — полупроводникового прибора, способного усиливать электрические сигналы, стало отправной точкой развития твердотельной электроники, и, впоследствии, микроэлектроники, которое, в свою очередь, коренным образом изменило методы экспериментальной физики, технику и жизнь человечества в целом. Если сосчитать все транзисторы, входящие в состав интегральных микросхем, то окажется, что на каждого жителя земли их изготовлено более миллиарда!

Разновидностей транзисторов, отличающихся параметрами и технологией изготовления существует очень много, однако можно выделить два больших и важных класса: полевые и биполярные транзисторы. Принцип действия полевого транзистора был предсказан раньше: В 1925 году немецкий физик Юлиус Лилиенфельд подал первую патентную заявку на твердотельный усилитель, состоящий из слоёв металла и полупроводника. Однако, технология того времени не позволяла изготовить такое устройство. В 1946 году сотрудники Bell Laboratory обнаружили эффект управления инжекцией носителей заряда с помощью дополнительного источника тока и использовали этот принцип для создания биполярного транзистора. И, хотя в настоящее время их доля на рынке полупроводниковых приборов составляет менее 5%, мы начнем рассмотрение именно с них.

3.5.1 Биполярный транзистор

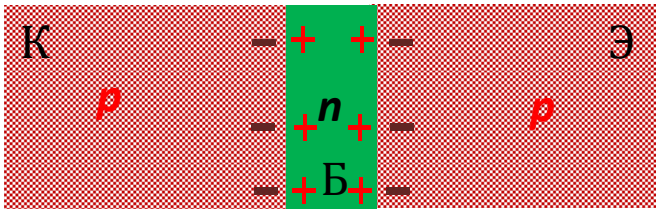


Рис. 3.21: Структура биполярного транзистора.

Упрощенная структура биполярного транзистора показана на рис. 3.21. Транзистор состоит из трех областей с различным типом проводимости. В зависимости от чередования этих областей различают pnp и npn транзисторы. Их обозначение на электрических схемах приведено на рис. 3.22. При подключении напряжений к отдельным слоям биполярного транзистора оказывается, что к одному переходу приложено прямое напряжение, к другому - обратное. При этом переход, к которому при нормальном включении приложено прямое напряжение, называют эмиттерным, а соответствующий наружный слой и вывод от него) - эмиттером (Э);

средний слой называют базой (Б). Второй переход, смещенный приложенным напряжением в обратном направлении, называют коллекторным, а соответствующий наружный слой - коллектором (К). Такие же названия используют для выводов, соединенных с соответствующими областями.

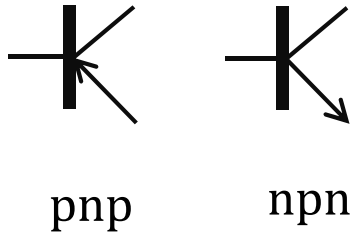


Рис. 3.22: Обозначение биполярных транзисторов на схемах Слева: pnp - транзистор, справа - npn транзистор.

У биполярных транзисторов проводимость эмиттерной и коллекторной областей много больше, чем у базовой, толщина базовой области мала по сравнению с диффузионной длиной для неосновных носителей заряда. Кроме того, удельное сопротивление области эмиттера несколько меньше, чем области коллектора.

Все положения, рассмотренные ранее для единичного p- и-перехода, справедливы для каждого из переходов транзистора. В равновесном состоянии наблюдается динамическое равновесие между потоками дырок и электронов, протекающими через каждый переход, результирующие токи равны нулю. Предположим, что к тран-

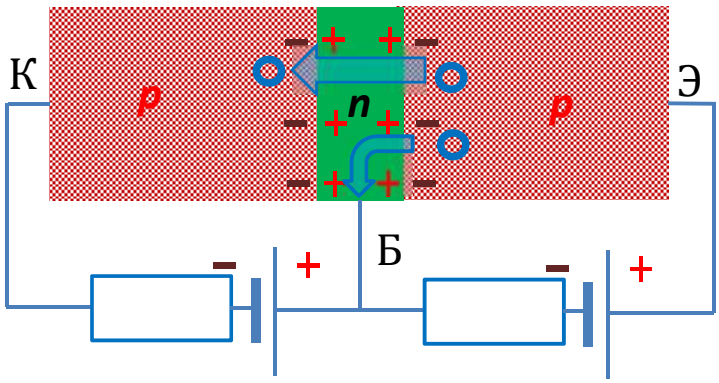


Рис. 3.23: Принцип действия биполярного транзистора.

зистору подключены источники напряжения, как показано на рис. 3.23. При этом эмиттерный переход смещается в прямом направлении, а коллекторный-в обратном. В результате снижения потенциального барьера дырки из области эмиттера диффундируют через p-n переход в область базы (инжекция дырок), а электроны - из области базы в область эмиттера. Так как удельное сопротивление базы высокое, дырочный поток преобладает над электронным, которым в первом приближении можно пренебречь.

Дырки, инжектированные в базу, создают вблизи p-n- перехода электрический заряд, который компенсируется электронами, приходящими от источника.

Приход электронов в базу из внешней цепи создает в последней электрический ток, который направлен из базы. Вследствие разности концентраций инжектированные в базу носители заряда и носители заряда, компенсировавшие их заряд и тем самым обеспечившие электронейтральность базы, движутся вглубь ее по направлению к коллектору. Если бы база была достаточно толстой, то все инжектированные носители заряда рекомбинировали бы в ней и в области, прилегающей к коллекторному переходу, их концентрация стала бы равновесной. Тогда через коллекторный переход протекал бы только малый ток, равный току обратносмещенного диода. Однако, поскольку ширина базы во много раз меньше диффузионной длины, время жизни неосновных носителей заряда в базе во много раз больше времени, необходимого для прохождения ими базы. Большинство дырок, инжектированных в нее, не успевают рекомбинировать с электронами и, попав вблизи коллекторного p-n-перехода в ускоряющее поле, втягиваются в коллектор (экстракция дырок). Электроны, число которых равно числу дырок, ушедших через коллекторный переход, в свою очередь, уходят через базовый вывод, создавая ток, направленный в базу транзистора. Таким образом, ток через базовый вывод транзистора определяют две встречено направленные составляющие тока. Если бы в базе процессы рекомбинации отсутствовали, то эти токи были бы равны между собой, а результирующий ток базы был бы равен нулю.

Подчеркнем, что данная модель биполярного транзистора является предельно упрощенной. В реальном транзисторе допирование эмиттерной области выше, чем коллекторной. Поскольку переход коллектор-база имеет большее сопротивление, на нем выделяется больше тепла и его делают большего размера, чем переход база-эмиттер. Таким образом, у реальных транзисторов коллектор и эмиттер не симметричны в отличие от нашей модели.

Ток эмиттерного перехода несколько больше тока коллекторного перехода. Относительное число неосновных носителей заряда, достигших коллекторного перехода транзистора, характеризуется коэффициентом переноса

$$k = I_K^p / I_E^p$$

где I_K^p I_E^p - токи коллекторного и эмиттерного переходов, созданные дырками. Дырки в базе являются неосновными носителями заряда и свободно проходят через запертый коллекторный p-n-переход в область коллектора. За время, определяемое постоянной времени диэлектрической релаксации τ_ε, они компенсируются электронами, создающими ток коллектора и приходящими из внешней цепи. Изменение напряжения, приложенного к эмиттерному переходу, вызывает изменение количества инжектируемых в базу неосновных носителей заряда и соответствующее изменение тока эмиттера и коллектора. Следовательно, коллекторным током можно управлять, задавая сравнительно небольшой ток в цепи эмиттер-база. Существуют различные схемы включения транзистора: схема с общим эмиттером, общим коллектором и общей базой (имеется ввиду, что один из выводов является общим для входной и выходной цепи). Рис. 3.24 иллюстрирует включение pnp транзистора по схеме с общим коллектором.

На рис. 3.25 приведен пример семейства зависимостей тока в цепи коллектор-

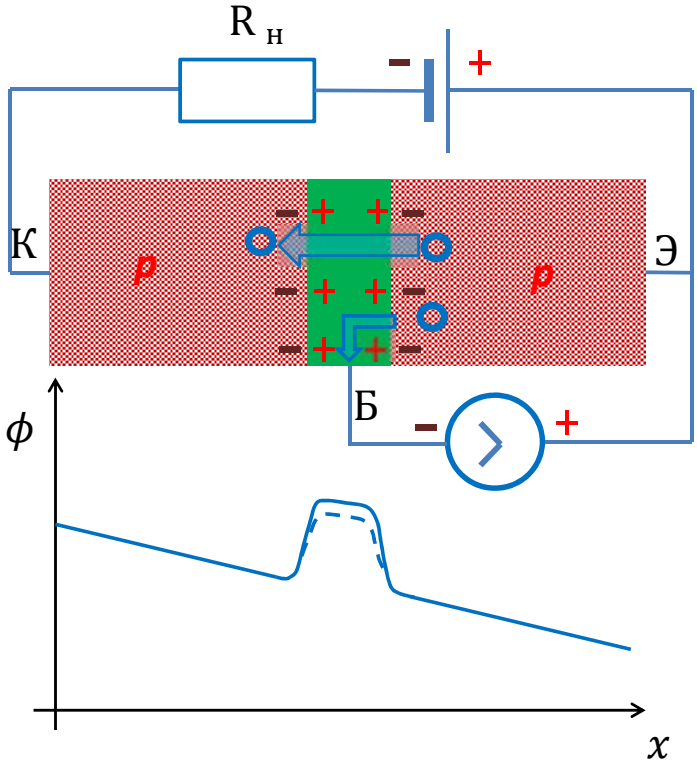


Рис. 3.24: Вверху: включение биполярного транзистора по схеме с общим эмиттером. Внизу: распределение потенциала вдоль транзистора.

эмиттер для различных величин тока в цепи база-эмиттер. Линию, соединяющую значение тока при нулевом напряжении коллектор-эмиттер и нулевое значение тока при напряжении коллектор-эмиттер, равном напряжению питания, называют нагрузочной прямой. Очевидно, ток в цепи коллектор-эмиттер определяется точкой пересечения нагрузочной прямой и зависимости, задаваемой током база-эмиттер. Для получения максимальной амплитуды выходного сигнала и наименьших нелинейных искажений необходимо правильно выбрать режим работы усилителя, то есть подобрать значение значения постоянного тока базы и напряжения на коллекторе при нулевой амплитуде переменного сигнала на базе так, что бы рабочая точка A лежала вблизи середины нагрузочной прямой (так, чтобы $V_{КЭ}$ составляло приблизительно половину напряжения питания).

Для простого расчета электрической схемы усилителя на биполярном транзисторе используя его справочные данные, можно вначале по его выходной вольт-амперной характеристике (см. рис. 3.26) выбрать ток базы, соответствующий рабочей точке, а, затем, по входной вольт-амперной характеристике определить отвечающее этому току напряжение база-эмиттер. После этого легко рассчитать величины всех сопротивлений в схеме. При расчете схемы необходимо учесть разброс параметров транзисторов (прежде всего, коэффициента усиления) и их изменение с температурой. Для обеспечения температурной стабильности рабочей точки значение тока в цепи делителя напряжения на резисторах R_1 и R_2 обычно выби-

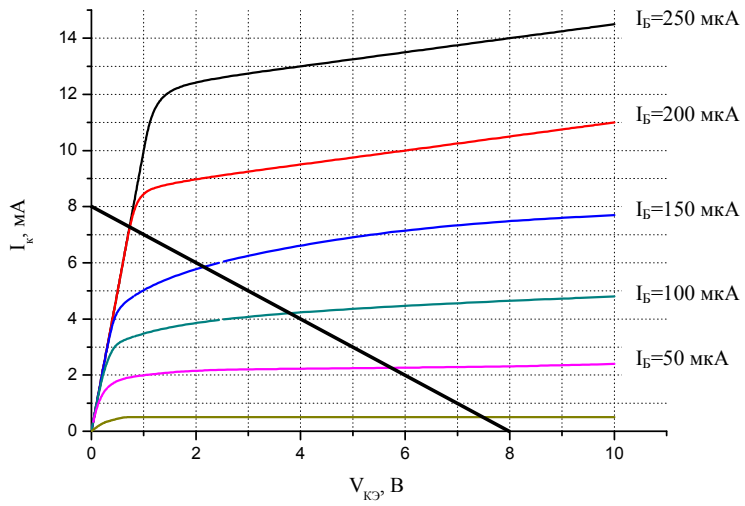


Рис. 3.25: Пример зависимостей тока в цепи коллектор-эмиттер для различных величин тока в цепи база-эмиттер.

рают в 5-10 раз больше рассчитанного тока базы, а в цепь эмиттера добавляют резистор R_3 (см.рис 3.27). Можно убедиться, что это приводит к возникновению отрицательной обратной связи по току: увеличение тока в цепи коллектор-эмиттер приводит к повышению напряжения на этом резисторе. Поскольку напряжение на базе задается по отношению к общему проводу, напряжение на самом переходе база-эмиттер при этом уменьшается, транзистор закрывается. Для того, что бы эта обратная связь не уменьшала коэффициент усиления по переменному току, параллельно с R_3 обычно включают конденсатор C. Импеданс цепи R_3C в рабочем диапазоне частот должен удовлетворять условию:

$$\frac{h_{21}}{h_{11}} \cdot \frac{R_3}{\sqrt{1+(\omega R_3 C)^2}} \ll 1 \tag{3.14}$$

Поскольку подключение источника сигналов и нагрузки к усилителю не должно изменять режим его работы по постоянному току, в схеме используют разделительные конденсаторы C_b.

Вольт-амперные характеристики транзистора, вообще говоря, являются нелинейными. Однако когда уровни сигналов невелики, то связь между входными и выходными сигналами можно с хорошей точностью считать линейной. В этом случае для расчета электронных схем можно представить транзистор в виде линейного четырехполюсника (см. рис.3.28) и описывать его набором дифференциальных параметров, связывающих малые приращения токов и напряжений на его входе и выходе.

Выбирать эти параметры можно по-разному, но для биполярных транзисторов чаще всего используют так называемые h - параметры:

$$\begin{cases} \delta u_1 = h_{11}\delta i_1 + h_{12}\delta u_2 \\ \delta i_2 = h_{21}\delta i_1 + h_{22}\delta u_2 \end{cases} \tag{3.15}$$

Здесь u_1, i_1 - напряжение и ток на входе четырехполюсника (в базовой цепи

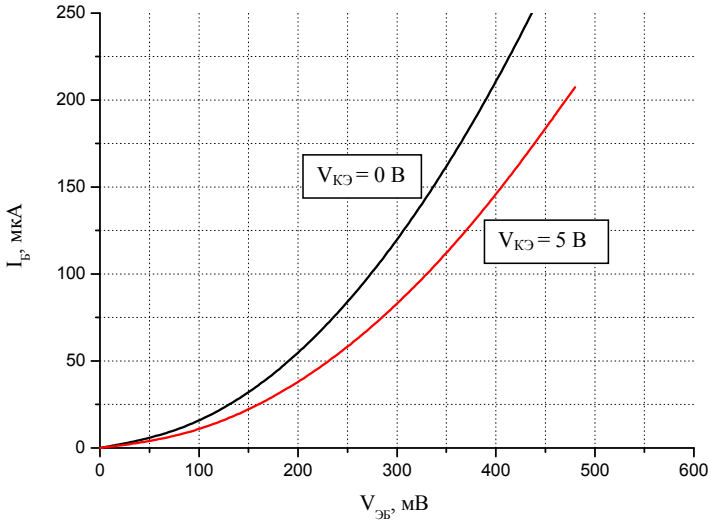


Рис. 3.26: Пример входных характеристик биполярного транзистора.

транзистора), u_2, i_2 - на его выходе. Таким образом, h_{11} - это входное сопротивление транзистора при постоянном напряжении на выходе ($\delta u_2 = 0$), h_{12} характеризует коэффициент обратной связи по напряжению при неизменном входном токе h_{21} - коэффициент передачи тока, а h_{22} - выходную проводимость. Заметим, что значения h-параметров могут различаться для разных схем включения. В справочниках часто приводят коэффициент усиления по току в схеме с общим эмиттером, который иногда обозначается $\beta \equiv h_{21Оэ}$.

3.5.2 Полевой транзистор

Работа полевых транзисторов основана на изменении электропроводности полупроводника под действием внешнего поперечного электрического поля. Наиболее распространенными являются полевые транзисторы с изолированным затвором и управляющим p-n переходом. Упрощенно устройство полевого транзистора с изолированным затвором показано на рис. 3.30. Слой полупроводника с двумя выводами, называемыми стоком и истоком, отделен от управляющего электрода, называемого затвором, тонким слоем диэлектрика. При приложении напряжения между затвором и подложкой (обычно она электрически соединена с истоком) электрическое **поле** изменяет распределение подвижных зарядов под управляющим электродом. В результате ширина области, через которую может протекать ток в цепи сток-исток (эта область называется каналом) изменяется, а, значит, изменяется ее сопротивление. Таким образом, изменяя напряжение на затворе, можно управлять током истока. Принципиальным отличием полевых транзисторов от биполярных является то, что ток в них переносится основными носителями, что исключает процессы рекомбинации и связанную с ними инерционность. Преимуществом полевого транзистора является его высокое входное сопротивление: поскольку затвор изолирован от канала, в цепи затвора протекают только токи, связные с перезарядкой емкости, образованной каналом и управляющим электродом. Заметим, что

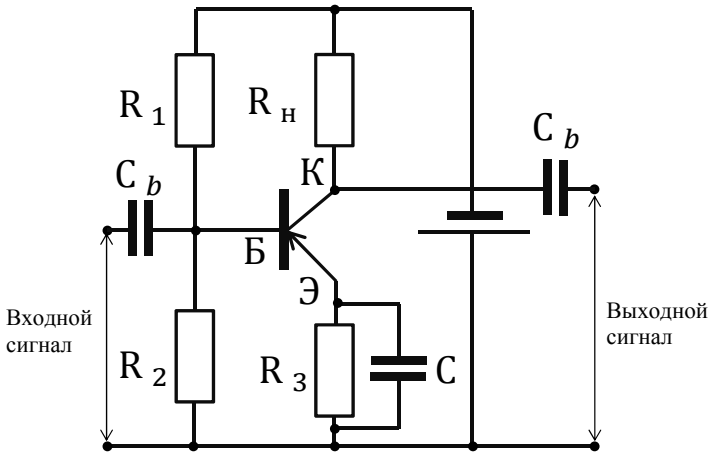


Рис. 3.27: Типовая схема усилителя сигналов на биполярном транзисторе с общим эмиттером.

Рис. 3.28: Представление биполярного транзистора как линейного четырехполюсника.

в данной, упрощенной модели не принимаются во внимания процессы на границах раздела полупроводник-диэлектрик. Кроме того, в рамках этой модели полевой транзистор выглядит симметричным относительно перестановки выводов стока и истока, в то время, как в реальных приборах такой симметрии может не быть. Мы также не рассматриваем особенности, связанные с реализацией канала: он может существовать изначально (транзисторы со встроенным каналом) или формироваться под действием поля, создаваемого потенциалом затвора (транзисторы с индуцированным каналом).

В литературе транзисторы с изолированным затвором сокращенно называют FET (от слов Field Effect Transistor), в русских справочниках встречается сокращение МОП - от слов Металл Оксид Полупроводник, поскольку, как правило, затворный электрод представляет собой металлическую пленку а изолятором служит окисный слой на ее поверхности. Полевой транзистор с p-n переходом отличается тем, что роль затвора в нем играет область полупроводника с типом проводимости, противоположной каналу (см. рис. 3.31. Принцип действия тот же, что и транзисторов с изолированным затвором: электрическое поле, создаваемое приложенным к переходу напряжением обогащает либо обедняет канал носителями заряда.

На рисунке 3.32 приведены примеры зависимостей тока стока от напряжения сток-исток для различных значений напряжения на затворе.

Простейшая схема усилителя на полевом транзисторе представлена на рис. 3.33. Сопротивление R1 обычно имеет большую величину (>100 кОм) и служит для сто-

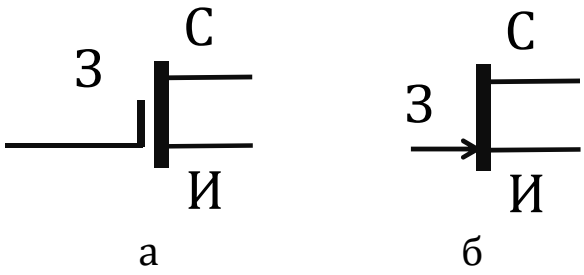


Рис. 3.29: Обозначение полевых транзисторов на электрических схемах: а - транзистор с изолированным затвором, б - транзистор с p-n переходом.

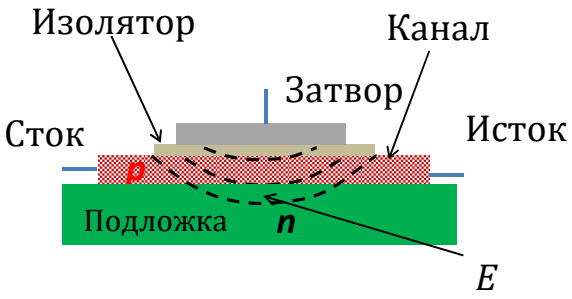


Рис. 3.30: Устройство полевого транзистора с изолированным затвором.

ка зарядов, которые могут накапливаться на затворе (предполагается, что источник сигнала может обладать высоким выходным сопротивлением, либо отделен по постоянному току от усилителя блокировочным конденсатором). В первом приближении можно рассматривать полевой транзистор как сопротивление, включенное между стоком и истоком, величина которого зависит от потенциала затвора. Важно, что при небольших изменениях тока стока (малосигнальный режим) влиянием этих изменений на напряжение на затворе можно пренебречь. Аналитически получить значение выходного напряжения можно, рассмотрев приращения тока стока $\Delta I_{вых}$ и напряжения на стоке $\Delta U_{вых}$. Считая напряжение питания постоянным, получим:

$$\Delta U_{=} = 0 = R_н \Delta I_{вых} + \Delta U_{вых} \;\Rightarrow\; \Delta I_{вых} = -\frac{\Delta U_{вых}}{R_н} \tag{3.16}$$

$$\Delta I_{вых} = \underbrace{\frac{\partial I_{вых}}{\partial U_{вх}}}_{S} \Delta U_{вх} + \underbrace{\frac{\partial I_{вых}}{\partial U_{вых}}}_{1/R_{си}} \Delta U_{вых}. \tag{3.17}$$

Величина $S = \frac{\Delta I_{си}}{\Delta U_{зи}}$ называется крутизной, $R_{си} = \frac{\Delta U_{си}}{\Delta I_{си}}$ - динамическим сопротивлением сток-исток. Подставляем (3.16) $\rightarrow$ (3.17):

$$\frac{-\Delta U_{вых}}{R_н} = S\Delta U_{вх} + \frac{\Delta U_{вых}}{R_{си}},$$

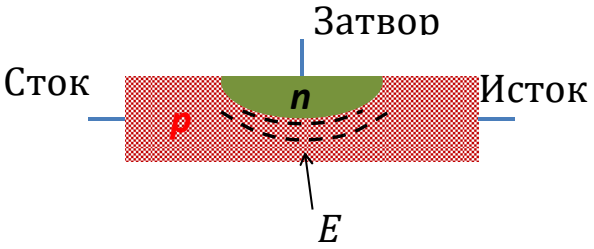


Рис. 3.31: Устройство полевого транзистора с p-n переходом.

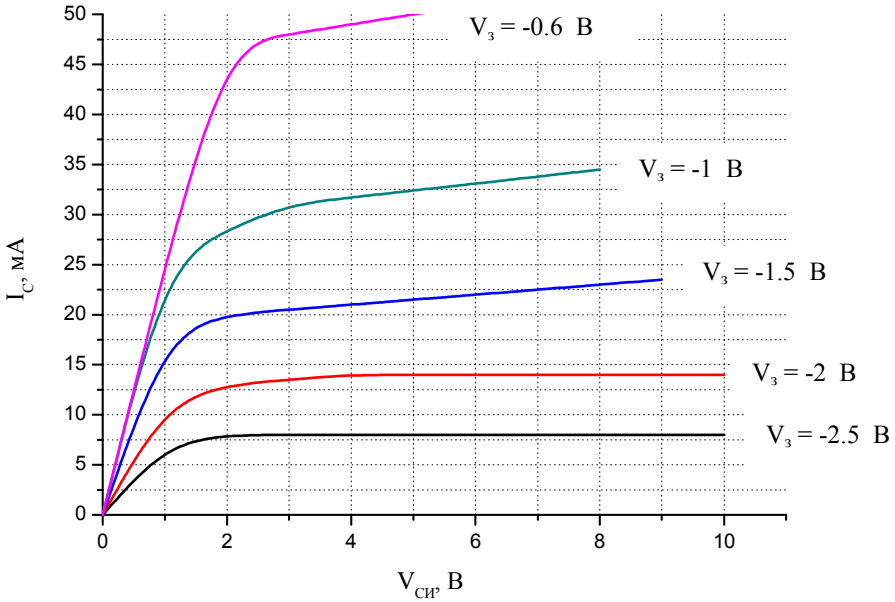


Рис. 3.32: Пример семейства выходных характеристик полевого транзистора для различных величин напряжения на затворе.

$$K_U = \frac{\Delta U_{ВЫХ}}{\Delta U_{ВХ}} = -S \cdot \frac{R_Н R_{СИ}}{R_Н + R_{СИ}} = -\underbrace{SR_{СИ}}_{\mu} \cdot \frac{R_Н}{R_Н + R_{СИ}}$$

$\mu = SR_{СИ}$. При $R_{СИ} \gg R_Н$ имеем $K_U \simeq -SR_Н$. Величины S и $R_{СИ}$ не постоянны, а зависят от выбора “рабочей точки”.

3.5.3 Эквивалентные схемы усилителей

Простейшему усилителю, изображенному на рис.3.33, можно сопоставить эквивалентную схему, в которой присутствует источник тока, величина которого определяется входным напряжением (см. рис. 3.34).

Реальные усилители, как правило, применяются для усиления сигналов: переменных токов и напряжений. В этом случае необходимо принимать во внимание наличие емкостей и индуктивностей, как специально добавляемых в схему для формирования желаемой АЧХ, так и собственных (иногда называемых "паразит-

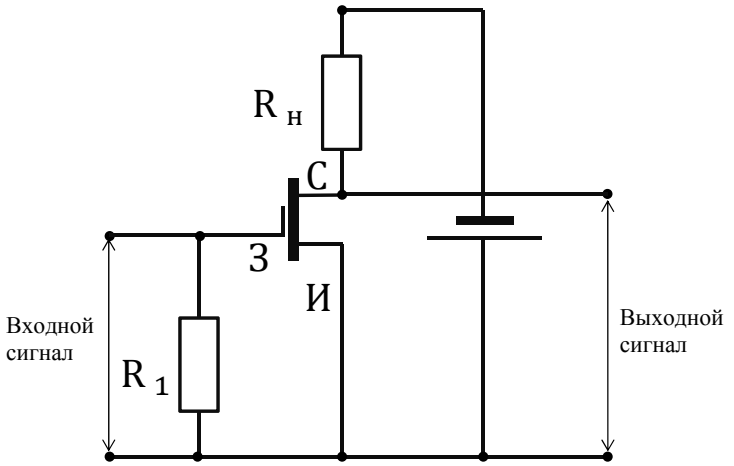


Рис. 3.33: Простейший усилитель на полевом транзисторе.

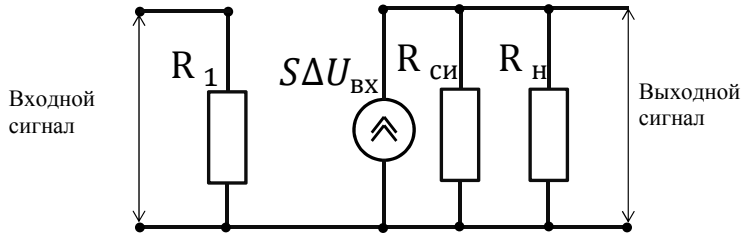


Рис. 3.34: Эквивалентная схема простейшего усилителя на полевом транзисторе.

ными"). Собственными емкостями являются емкость сток-исток и исток-затвор транзистора а так же емкости между соединительными проводниками. Проводники и канал полевого транзистора обладают так же собственной индуктивностью, однако на низких и радио частотах (до сотен мегагерц) ими можно пренебречь. Схема усилителя переменного напряжения на полевом транзисторе с корректирующими емкостями изображена на рис. 3.35., его эквивалентная схема - на рис.3.36.

Амплитудно-частотная характеристика такого усилителя имеет спад на низких и высоких частотах (см. рис. 3.37). Граничными частотами принято называть частоты, на которых коэффициент усиления по напряжению уменьшается в $\sqrt{2}$ раз. В нашей схеме они задаются цепочками $R_1C_1C_{ЗИ}$ и $R_НR_{СИ}C_{СИ}$, соответственно.

3.5.4 Обратная связь в усилителях

Обратной связью (ОС) в радиофизических системах называют передачу выходного сигнала через специальную цепь с выхода усилителя на его вход.

Введение обратной связи позволяет изменять амплитудно-частотные характеристики цепей, повышать стабильность и уменьшать нелинейные искажения. С обратной связью, используемой для повышения температурной стабильности мы

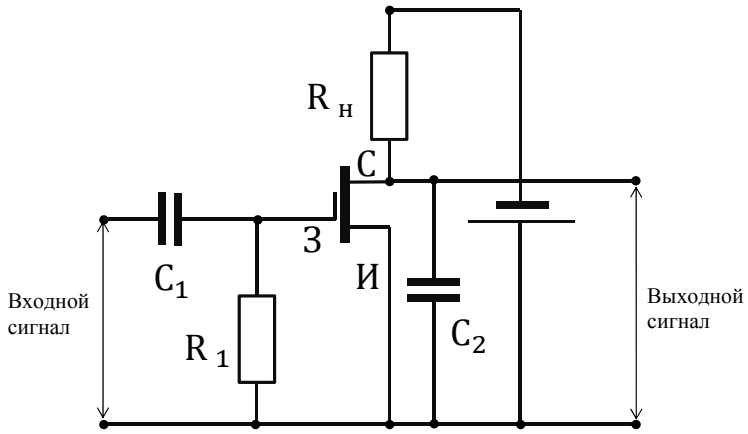


Рис. 3.35: Усилитель переменного напряжения на полевом транзисторе.

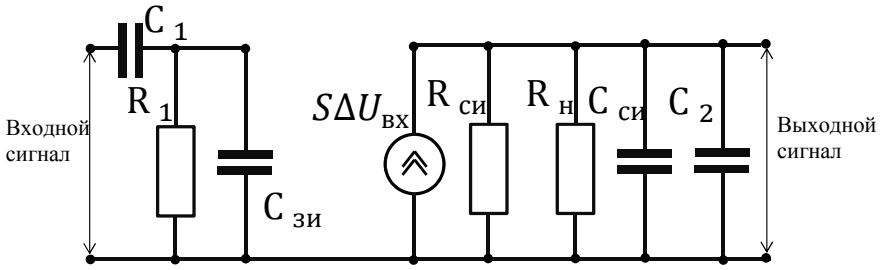


Рис. 3.36: Эквивалентная схема усилителя переменного напряжения на полевом транзисторе.

уже сталкивались при рассмотрении усилителя на биполярном транзисторе (см. рис.3.27) Кроме того, в цепях, охваченных обратной связью, при определенных условиях могут возникать автоколебания. В общем виде структурная схема усилителя с обратной связью изображена на рис.3.38. Знаком "+" обозначен узел суммирования сигнала генератора и сигнала обратной связи. Для простоты мы не будем пока рассматривать устройство сумматора и предположим, что и усилитель, и цепь обратной связи передают сигналы только в одном направлении.

В рассматриваемой схеме

$$S_{in}(\omega) = S_g(\omega) + S_\beta(\omega), \quad S_\beta(\omega) = \beta S_{out}(\omega), \tag{3.18}$$

$$S_{out}(\omega) = K(\omega)S_{in}(\omega) = K(\omega)(S_g(\omega) + \beta S_{out}(\omega)), \tag{3.19}$$

где $\beta(\omega)$ и $K(\omega)$ — коэффициенты передачи собственно усилителя и цепи обратной связи. Следовательно, комплексный коэффициент передачи линейного усилителя, охваченного обратной связью:

$$\tilde{K}_\beta(\omega) = \frac{S_{out}(\omega)}{S_g(\omega)} = \frac{\tilde{K}(\omega)}{1 - \tilde{\beta}(\omega)\tilde{K}(\omega)}. \tag{3.20}$$

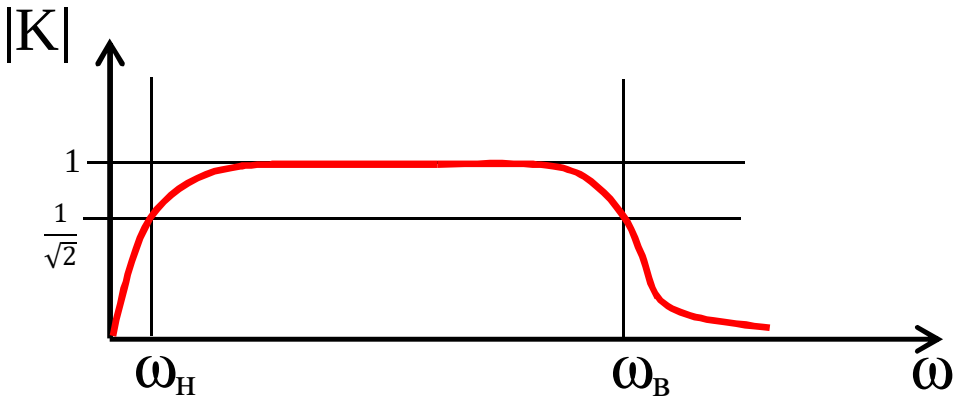


Рис. 3.37: АЧХ усилителя переменного напряжения на полевом транзисторе.

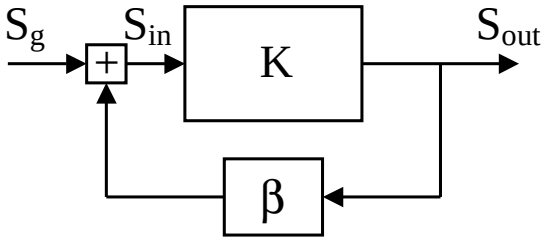


Рис. 3.38: Блок-схема усилителя с обратной связью.

Если на некоторой частоте

$$|1 - \tilde{\beta}(\omega)\tilde{K}(\omega)| > 1, \tag{3.21}$$

то обратная связь уменьшает модуль коэффициента усиления и считается отрицательной (ООС). Если выполняется обратное неравенство:

$$|1 - \tilde{\beta}(\omega)\tilde{K}(\omega)| > 1, \tag{3.22}$$

то такая обратная связь называется положительной (ПОС). Заметим, что, хотя формально введение ПОС увеличивает коэффициент усиления устройства, на практике такой способ используется очень редко, поскольку при этом снижается стабильность усилителя. Положительная обратная связь используется для создания генераторов колебаний, в том числе гармонических.

Сигнал обратной связи может быть пропорционален выходному напряжению или выходному току. В первом случае говорят об обратной связи по напряжению, во втором — по току.

Влияние отрицательной обратной связи (ООС) на параметры усилителя

Отрицательная обратная связь обладает одним замечательным свойством, широко используемым при создании стабильных усилителей с нужными характеристиками. Из соотношения (3.20) следует, что при

$$|\beta K| \gg 1 \tag{3.23}$$

$$\tilde{K}_\beta(\omega) \simeq -\frac{1}{\tilde{\beta}(\omega)}. \tag{3.24}$$

Коэффициент усиления в этом случае определяется только цепью обратной связи и не зависит от коэффициента усиления K базового усилителя. Цепь обратной связи обычно состоит из пассивных элементов (резисторов, конденсаторов, катушек индуктивности), обладающих относительно высокой стабильностью. Соответственно, стабильным будет и коэффициент усиления K_β.

Чтобы соотношение (3.23) выполнялось при малом значении β в широком диапазоне частот, базовый усилитель должен иметь в том же диапазоне частот большой коэффициент усиления. Специально разработанные для использования в качестве базовых микросхемы названы **операционными усилителями.**

3.5.5 Операционные усилители

Операционный усилитель (ОУ) - это усилитель (обычно выполненный в виде интегральной микросхемы), специально разработанный для использования в схемах с глубокой отрицательной обратной связью (ОС).

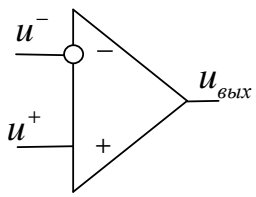


Рис. 3.39: Схематическое изображение ОУ.

Операционные усилители, как правило, имеют дифференциальный вход (см. рис.3.39), то есть усиливаемым сигналом является разность потенциалов между двумя входами ОУ. Один из входов является неинвертирующим входом. При положительном потенциале на нем выходное напряжение также положительное. Другой вход является инвертирующим. При положительном потенциале на нем выходное напряжение отрицательно. Упрощенная эквивалентная схема ОУ изображена на рис. 3.40.

Коэффициент передачи операционного усилителя без обратной связи

$$\tilde{K}_{oy}(\omega) = \frac{K_{oy}(0)}{1 + i\omega/\omega_B} = \frac{K_{oy}(0)}{1 + iK_{oy}(0)f/f_1}, \tag{3.25}$$

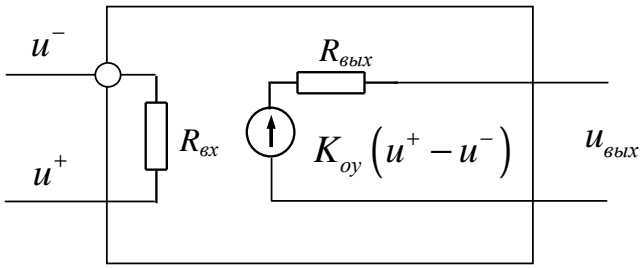


Рис. 3.40: Упрощенная эквивалентная схема ОУ

где

$$f_1 = (\omega_B/2\pi)K_{oy}(0)$$

— частота, на которой коэффициент усиления равен единице (частота "единичного усиления"). Коэффициент усиления реальных ОУ на нулевой частоте $10^4 - 10^7$. Частота "единичного усиления"современных ОУ $10^6 - 10^9$Гц. Входное сопротивление $R_{ВХ}$ ОУ, входной каскад которых выполнен на биполярных транзисторах, составляет $10^5 - 10^7$ Ом. Если же во входном каскаде использованы полевые триоды, $R_{ВХ}$ может достигать 10^{12} Ом. Выходное сопротивление лежит в диапазоне от единиц до нескольких сотен Ом.

В случае операций с единственным источником сигнала, он подключается к одному из входов. Другой вход "заземляется". Если при этом сигнал подается на неинвертирующий вход, усилитель называется неинвертирующим. В другом случае — инвертирующим.

Неинвертирующий усилитель.

Принципиальная схема неинвертирующего усилителя с обратной связью по напряжению изображена на рис.3.41. В этой схеме напряжение обратной связи (в комплексном представлении)

$$\tilde{U}_\beta = \tilde{U}^- = \frac{Z_1}{Z_1 + Z_2}\tilde{U}_{ВЫХ}.$$

Разность напряжений на дифференциальном входе

$$\tilde{U}^+ - \tilde{U}^- = \tilde{U}_{ВХ} + \beta\tilde{U}_{ВЫХ},$$

где

$$\tilde{\beta} = -\frac{Z_1}{Z_1 + Z_2}. \tag{3.26}$$

Выходное напряжение

$$\tilde{U}_{ВЫХ} = K_{oy}(\tilde{U}^+ - \tilde{U}^-) = K_{oy}(\tilde{U}^+ + \tilde{\beta}\tilde{U}_{ВЫХ}).$$

Следовательно, коэффициент усиления неинвертирующего усилителя

$$\tilde{K}^+ \equiv \frac{\tilde{U}_{ВЫХ}}{\tilde{U}_{ВХ}} = \frac{K_{oy}}{1 - \tilde{\beta}K_{oy}} \simeq -\frac{1}{\tilde{\beta}} = 1 + \frac{Z_2}{Z_1}. \tag{3.27}$$

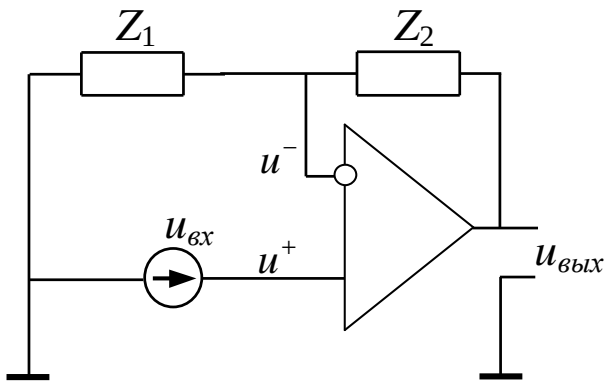


Рис. 3.41: Неинвертирующй усилитель на ОУ

Инвертирующий усилитель.

Принципиальная схема инвертирующего усилителя изображена на рис.3.42. Здесь напряжение на инвертирующем входе

$$\tilde{U}^- = \tilde{U}_{ВХ} - \tilde{I}_{ВХ} Z_1,$$

а входной ток

$$\tilde{I}_{ВХ} = \frac{\tilde{U}_{ВХ} - \tilde{U}_{ВЫХ}}{Z_1 + Z_2}. \tag{3.28}$$

(Входным током операционного усилителя $\tilde{U}^-/R_{ВХ}$ при $|Z_2| \ll R_{ВХ}$ можно пренебречь.)
Выходное напряжение

$$\tilde{U}_{ВЫХ} = K_{оу}(\tilde{U}^+ - \tilde{U}^-) = -K_{оу}\tilde{U}^- = -K_{оу}\left(\tilde{U}_{ВХ}\frac{Z_2}{Z_1 + Z_2} - \tilde{\beta}\tilde{U}_{ВЫХ}\right).$$

Следовательно, коэффициент усиления инвертирующего усилителя

$$\tilde{K}^- = \frac{-K_{оу}}{1 - \tilde{\beta}K_{оу}}\frac{Z_2}{Z_1 + Z_2} \tag{3.29}$$

по модулю в $|Z_2/(Z_1 + Z_2)|$ раз меньше K^+. Это связано с тем, что входное напряжение поступает на инвертирующий вход ОУ ослабленным в это число раз.
При $|\tilde{\beta}K_{оу}| \gg 1$ из (3.29) получим

$$\tilde{K}^- = -\frac{Z_2}{Z_1}. \tag{3.30}$$

При большом усилении различие между модулями K^+ и K^- пренебрежимо мало. Однако инвертирущие и неинвертирующие усилители существенно различаются по

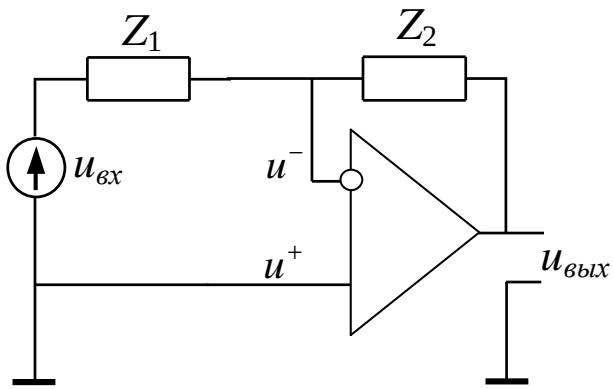


Рис. 3.42: Инвертирующй усилитель на ОУ

входному сопротивлению.

Входное сопротивление неинвертирующего усилителя.
По определению

$$Z_{\text{ВХ}} = \frac{\tilde{U}_{\text{ВХ}}}{\tilde{I}_{\text{ВХ}}}.$$

Поскольку в этой схеме $\tilde{U}_{\text{ВХ}} = \tilde{U}^{+}$ и $K_{\text{оу}}(\tilde{U}^{=} - \tilde{U}^{-}) = \tilde{K}^{+}\tilde{U}^{+}$, то

$$Z^{+}{}_{\text{ВХ}} = \frac{K_{\text{оу}}}{\tilde{K}^{+}} \frac{(\tilde{U}^{+} - \tilde{U}^{-})}{\tilde{I}_{\text{ВХ}}} = \frac{K_{\text{оу}}}{\tilde{K}^{+}} R_{\text{ВХ}} > R_{\text{ВХ}}. \tag{3.31}$$

Входное сопротивление неинвертирующего усилителя больше входного сопротивления ОУ в $K_{\text{оу}}/\tilde{K}^{+}$ раз.

Входное сопротивление инвертирующего усилителя.
Из соотношений (3.28), (3.30)

$$\tilde{I}_{\text{ВХ}} = \frac{\tilde{U}_{\text{ВХ}} - \tilde{K}^{-}\tilde{U}_{\text{ВХ}}}{Z_1 + Z_2} = \frac{\tilde{U}_{\text{ВХ}}}{Z_1}.$$

Следовательно, входное сопротивление инвертирующего усилителя

$$Z^{-}_{\text{ВХ}} = Z_1. \tag{3.32}$$

Оно в $K^{-}/K_{\text{оу}}$ раз меньше входного сопротивления $R_{\text{ВХ}}$ ОУ.

Влияние обратной связи на выходное сопротивление усилителя.
По определению

$$Z_{\text{ВЫХ}} = \frac{\tilde{U}_{\text{ВЫХ}}}{\tilde{I}_{\text{ВЫХ}}}.$$

Измерять выходное сопротивление можно различными способами. Например, в соответствии с определением для измерения $Z_{\text{ВЫХ}}$ можно подать на выход усилителя

напряжение и измерить создаваемый им ток. Входное напряжение при этом должно быть равно 0. В этом измерении схемы инвертирующего и неинвертирующего усилителя будут идентичными. Следовательно, одинаковыми будут и их выходные сопротивления. При таком измерении выходной ток будет равен

$$\tilde{I}_{\text{вых}} = \frac{\tilde{U}_{\text{вых}} - K_{\text{оу}}\tilde{U}^-}{R_{\text{вых}}}.$$

(Током через цепь обратной связи пренебрегаем.) Когда $\tilde{U}_{\text{вх}} = 0$, $\tilde{U}^- = -\tilde{\beta}\tilde{U}_{\text{вых}}$. Из этих соотношений получим

$$Z_{\text{вых}} = \frac{R_{\text{вых}}}{1 - \tilde{\beta}K_{\text{оу}}} \simeq R_{\text{вых}}\frac{\tilde{K}^+}{K_{\text{оу}}}.$$

Это подтверждает общую закономерность: отрицательная обратная связь по напряжению уменьшает выходное сопротивление усилителя (связь по напряжению — связь, при которой сигнал обратной связи снимается с нагрузки усилителя). Также известно, что отрицательная обратная связь по току, т.е. когда напряжение обратной связи берется с сопротивления, включенного последовательно с выходным $R_{\text{вых}}$ и с нагрузкой, увеличивает выходное сопротивление.

3.5.6 Полосовые усилители на ОУ.

Использование в обратной связи частотно-зависимых цепей позволяет сформировать желаемую зависимость коэффициента усиления от частоты. Простейшим примером является полосовой усилитель.

Принципиальная схема неинвертирующего полосового усилителя на ОУ изображена на рис.3.43. В этой схеме

$$Z_1 = R_1 + \frac{1}{i\omega C_1}, \quad Z_2 = \frac{R_2}{1 + i\omega R_2 C_2}.$$

Соответственно, коэффициент усиления напряжения

$$\tilde{K}^+ = 1 + \frac{Z_2}{Z_1} \simeq \frac{Z_2}{Z_1} = K_0\frac{1}{1 + i\omega R_2 C_2}\frac{i\omega R_1 C_1}{1 + i\omega R_1 C_1}, \tag{3.33}$$

где $K_0 = R_2/R_1$. Если $1/R_2C_2 \gg 1/R_1C_1$, то в области низких частот определяющим будет последний сомножитель в (3.33), а в области высоких частот — второй. Соответственно, нижняя частота такого усилителя будет равна

$$\omega_{\text{н}} = \frac{1}{R_1 C_1}, \quad \text{а верхняя } \omega_{\text{в}} = \frac{1}{R_2 C_2}. \tag{3.34}$$

Принципиальная схема инвертирующего УНЧ на ОУ изображена на рис.3.44. И в этом случае коэффициент усиления с точностью до знака определяется правой частью соотношения (3.33).

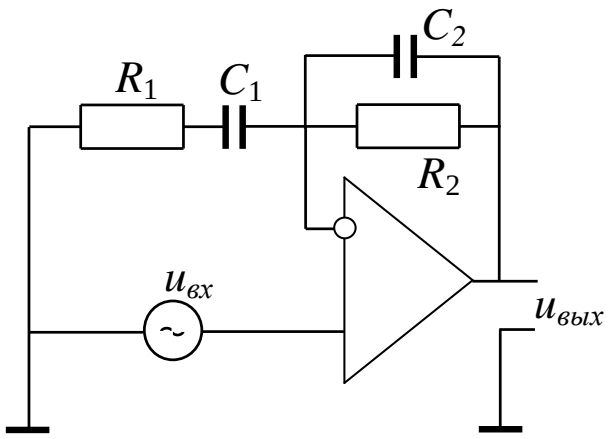


Рис. 3.43: Неинвертирующй полосовой усилитель

3.6 Генераторы.

В радиофизических системах часто бывают нужны источники периодических колебаний. Устройства, преобразующие энергию источника (обычно — источника постоянного тока) в энергию колебаний называются автогенераторами (или просто генераторами). Автоколебания в физических системах возникают при наличии в системе положительной обратной связи и притока энергии. Генератор может представлять собой усилитель, выход которого соединен со входом через цепь положительной обратной связи (см. рис. 3.45).

Качественно процесс возникновения автоколебаний можно представить себе следующим образом: предположим, что какая-либо спектральная компонента флуктуаций на входе усилителя после прохождения через усилитель, цепь положительной обратной связи и сложения с этими флуктуациями на входе увеличивается по амплитуде. Важно, что флуктуации присутствуют во всех активных элементах усилителей (транзисторах, электронных лампах) — см. раздел “Шумы”. Поэтому очевидно, что колебания в таких условиях возникнут неизбежно и их амплитуда будет возрастать. Ограничение амплитуды происходит из-за того, что у любого активного элемента (транзистора, электронной лампы), входящего в состав усилителя, коэффициент усиления при увеличении амплитуды сигнала рано или поздно начинает уменьшаться. В установлении постоянной амплитуды колебаний нелинейность играет принципиальную роль.

Пусть комплексный коэффициент усиления усилителя есть $K(\omega)$, а коэффициент передачи обратной связи — $\beta(\omega)$. Для возникновения автоколебаний напряжение $u_\beta(\omega)$ на “выходе” обратной связи (см. рис. 3.45) должно быть больше или равно входному напряжению и совпадать с ним по фазе. Тогда условие существования *стационарных* гармонических колебаний на частоте ω_0 имеет вид:

$$|K(\omega_0)\beta(\omega_0)| = 1\,, \tag{3.35}$$

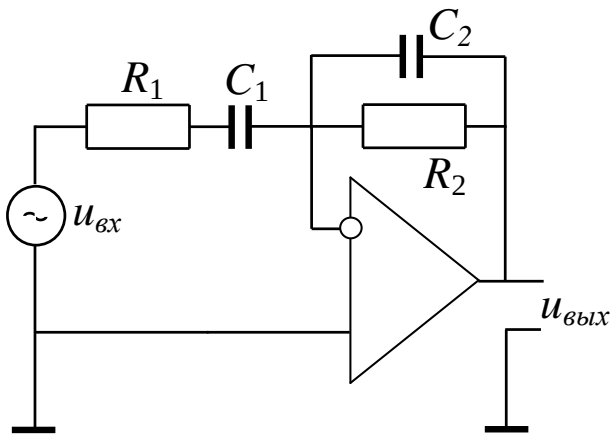


Рис. 3.44: Инвертирующй полосовой усилитель

$$\arg(K(\omega_0) + \arg(\beta(\omega_0) = 2\pi n, \quad n = 0, 1, 2, \ldots \tag{3.36}$$

(3.35) **и** (3.36) называют условиями баланса амплитуд и баланса фаз, соответственно.

3.6.1 Стационарные автоколебания

Возникновение автоколебаний в генераторе (линейное приближение)

Рассмотрим процесс генерации гармонических колебаний на примере автогенератора на полевом транзисторе с колебательным контуром в цепи затвора и индуктивной обратной связью (см. рис.3.46).

Автогенераторы, в составе которых присутствует колебательный контур с достаточно высокой добротностью (LC-контур, объемный резонатор), а активный элемент служит для компенсации потерь энергии в этом контуре, иногда называют автогенераторами Томсоновского типа.

Будем считать для начала, что напряжение на затворе меняется в небольших пределах, так, что полевой транзистор может рассматриваться как линейный управляемый источник тока:

$$i_c = Su_З \tag{3.37}$$

здесь i_c — ток стока транзистора, $u_З$ — напряжение на затворе, S — крутизна. Сопротивление R_2 и емкость C_2 в схеме на рис. 3.46 служат для выбора рабочей точки. На частотах работы генератора импеданс цепи R_2C_2 мал и его влиянием на усиление можно пренебречь. Тогда дифференциальное уравнение для колебаний в контуре RLC можно записать в виде:

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = \pm M\frac{di_c}{dt}, \tag{3.38}$$

здесь q — заряд на конденсаторе, M — коэффициент взаимоиндукции (знак зависит от направления витков катушки связи по отношению к виткам L, которое

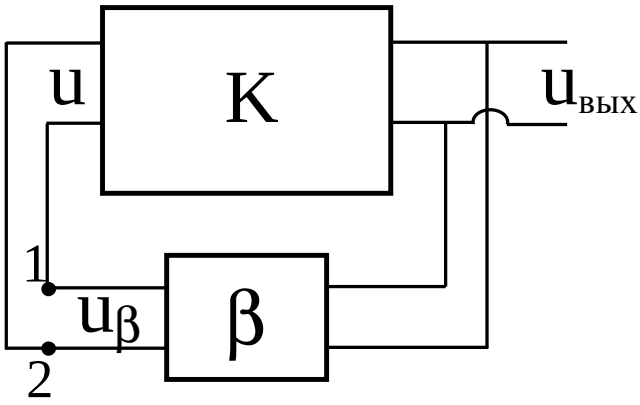


Рис. 3.45: Блок-схема генератора, содержащего усилитель и положительную обратную связь.

должно обеспечивать положительную обратную связь). Тогда:

$$M\frac{di_c}{dt} = MS\frac{du_з}{dt}. \tag{3.39}$$

Поскольку $u_з = q/C$, уравнение (3.38) можно переписать в виде:

$$L\frac{d^2q}{dt^2} + \left(R - \frac{MS}{C}\right)\frac{dq}{dt} + \frac{q}{C} = 0 \tag{3.40}$$

или

$$\frac{d^2q}{dt^2} + 2\delta\frac{dq}{dt} + \omega_0^2 q = 0, \tag{3.41}$$

где

$$2\delta = \left(R - \frac{MS}{C}\right), \quad \omega_0^2 = \frac{1}{LC}. \tag{3.42}$$

Уравнение (3.41) описывает малые колебания в контуре генератора. Величину MS/C можно считать отрицательным сопротивлением, которое вносит в контур активный элемент. Стационарные колебания будут наблюдаться при условии

$$\frac{MS}{C} = R. \tag{3.43}$$

Напомним еще раз, что крутизна S зависит от напряжения на затворе. Именно эта зависимость и будет определять стационарную амплитуду колебаний (см. детали ниже в разделе 3.6.1).

Существует множество различных схем генераторов. Упомянем одну из них, называемую индуктивной трехточкой — ее схема приведена на рис 3.47. Обратная связь формируется за счет того, что ток истока транзистора протекает через часть витков катушки контура LC, конденсаторы $C_{р1}, C_{р2}$ разделяют цепи постоянного и переменного тока, резистор $R_З$ задает режим транзистора по постоянному току.

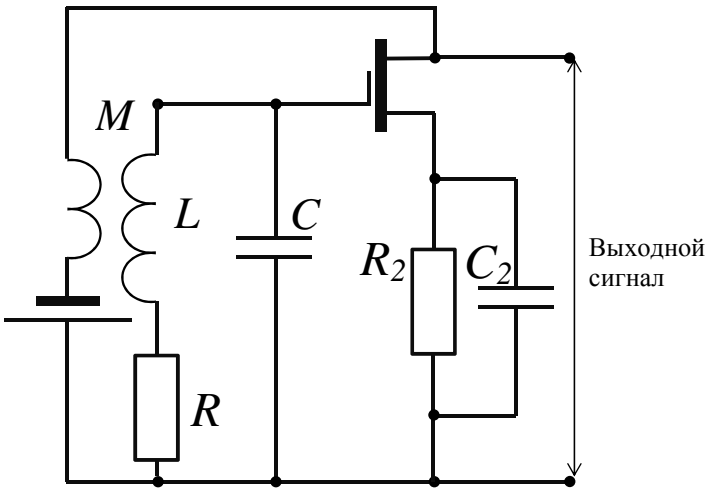


Рис. 3.46: Автогенератор на полевом транзисторе с колебательным контуром в цепи затвора и индуктивной обратной связью.

Амплитуда автоколебаний

Очевидно, что условие (3.42) на практике никогда не выполняется строго. Таким образом, в линейном приближении автоколебания оказываются не устойчивыми: при $\frac{MS}{C} < R$ они будут затухать, а при $\frac{MS}{C} > R$ — неограниченно нарастать.

Для установления стационарных колебаний в автогенераторах принципиальную роль играет нелинейность усилителя, т.е. нелинейная зависимость выходного напряжения от входного. Обычно роль этой нелинейности играет зависимость коэффициента усиления активного элемента от амплитуды входного напряжения u.

Будем считать, что коэффициент усиления не зависит от частоты и в некотором диапазоне значений напряжения u на входе активного элемента можно представить, как:

$$u_{вых} = K_1 u + K_2 u^2 + K_3 u^3 + K_4 u^4 + K_5 u^5 + \ldots \tag{3.44}$$

Предположим, что колебания на входе имеют вид $u(t) = A_0 \cos \omega t$.

Строго говоря, из-за нелинейности в автогенераторе колебания не являются строго гармоническими, но, при определенных условиях, амплитуда высших гармоник может быть много меньше, чем амплитуда колебаний на основной частоте.

Заметим, что отбрасывать все нелинейные члены в (3.44) нельзя, поскольку амплитуда выходного напряжения на частоте ω зависит не только от K_1, но и от остальных нечетных членов в разложении, в то время, как четные члены вклада в амплитуду колебаний на основной частоте не дадут, например, член $K_2 u^2$ дает вклад в постоянную составляющую и во вторую гармонику: $\cos^2 \omega t = \frac{1}{2}(1 - \cos 2\omega t)$.

Подставляя $u = A_0 \cos \omega t$ в (3.44) и выбирая только коэффициенты при $\cos \omega t$, получаем:

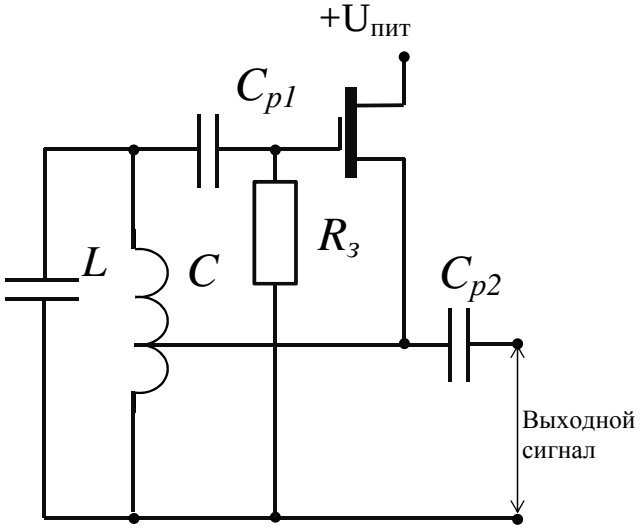


Рис. 3.47: Автогенератор на полевом транзисторе по схеме "индуктивной трехточки".

$$A_{вых}(\omega) = A_0(K_1 + \frac{3}{4}K_3A_0^2 + \frac{5}{8}K_5A_0^4). \tag{3.45}$$

Второе слагаемое дает кубическая нелинейность, третье - нелинейность пятой степени (см. 3.5).

Здесь мы ограничились пятью первыми членами в разложении — этого достаточно, чтобы получить качественно правильный результат, и не учитывали фазовые соотношения, считая, что, для основной частоты условие баланса фаз (3.36) выполнено.

Обозначим

$$\overline{K(A)} \equiv K_1 + \frac{3}{4}K_3A^2 + \frac{5}{8}K_5u^4 \tag{3.46}$$

и подставим его в условие баланса амплитуд (3.35):

$$A|\overline{K(A)}\beta| = A. \tag{3.47}$$

Заметим, что это условие эквивалентно условию (3.42) для LC-генератора. Из уравнения (3.47) можно найти значение амплитуды установившихся колебаний. Одно из решений этого уравнения, $A = 0$, соответствует состоянию покоя. Оно может быть устойчивым или неустойчивым. Другие значения стационарной амплитуды следуют из уравнения

$$|\overline{K(A)}\beta| = 1. \tag{3.48}$$

Мягкий режим возбуждения

Рассмотрим сначала случай $K_3 < 0, K_5 < 0$. Для этого случая графическое решение уравнения (3.48) приведено на рис. 3.48. Очевидно, что состояние с $A_0 = 0$

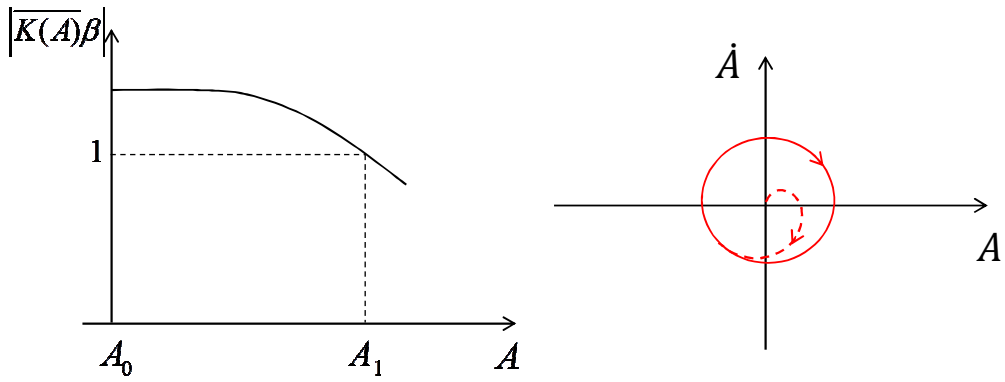


Рис. 3.48: Слева: зависимость $|\overline{K(A)}\beta$ при мягком режиме возбуждения. Справа: фазовый портрет автоколебательной системы томсоновского типа при мягком режиме возбуждения.

неустойчивое. Поскольку $|\overline{K(0)}\beta| > 1$, то любое, сколь угодно малое отклонение амплитуды δA от нуля приведет к ее нарастанию. Амплитуда будет расти до некоторого значения A_1.

Покажем, что состояние с амплитудой A_1 устойчивое. Допустим, что $A = A_1 + \delta A$. Если $\delta A > 0$, то $|\overline{K(A)}\beta| < 1$ и амплитуда будет уменьшаться. Если же $\delta A < 0$, то $|\overline{K(A)}\beta| > 1$, и амплитуда будет расти. В обоих случаях значение амплитуды возвращается к ее стационарному значению A_1.

В данном случае для возбуждения автоколебаний из состояния покоя достаточно сколь угодно малое отклонение от состояния покоя. Поскольку в электрических цепях всегда присутствуют флуктуации, это означает, что колебания в таком генераторе будут возникать неизбежно, как только на него подано питание. Подобный режим возбуждения колебаний называют **мягким**.

Жесткий режим возбуждения

Теперь рассмотрим случай $K_3 > 0, K_5 < 0$. Для этого случая графическое решение уравнения (3.48) приведено на рис. 3.49. Очевидно, что возможны три решения: $A_0 = 0$, A_2 и A_3. Состояние с $A_0 = 0$ устойчивое, поскольку $|\overline{K(A)}\beta| < 1$, пока возмущение амплитуды не превышает A_2.

Состояние A_2 неустойчиво. Действительно, если $A < A_2$, то $|\overline{K(A)}\beta| < 1$ и в дальнейшем амплитуда уменьшится до нуля. Если же $A > A_2$, то амплитуда будет нарастать до стационарного значения A_3.

Следовательно, для возбуждения автоколебаний из состояния покоя необходимо достаточное начальное возмущение. Иногда (но не всегда) роль такого возмущения может играть переходный процесс при включении питания. Подобное возбуждение автоколебаний называют **жестким**.

Стационарное состояние с амплитудой A_3 устойчиво — это можно показать аналогично тому, как была доказана устойчивость A_1 в режиме мягкого возбуждения колебаний.

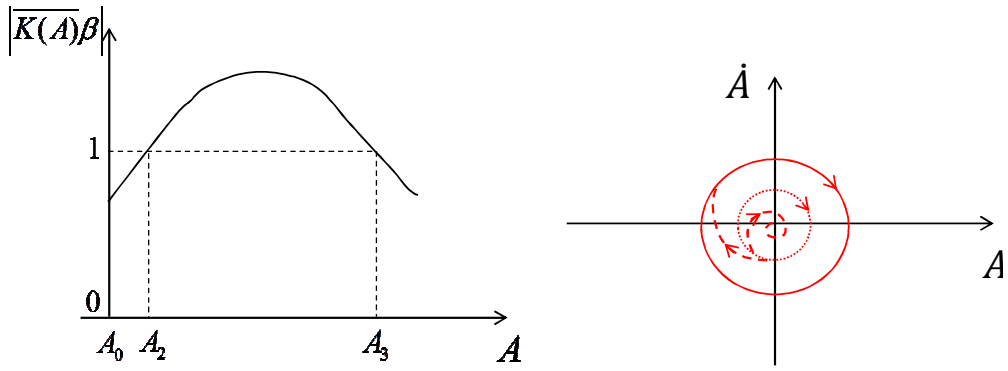


Рис. 3.49: Слева: зависимость $|\overline{K(A)}\beta$ при жестком режиме возбуждения. Справа: фазовый портрет автоколебательной системы томсоновского типа при жестком режиме возбуждения.

Наглядным представлением процессов, происходящих в автоколебательной системе является *фазовый портрет* — график зависимости производной какой-либо координаты от самой координаты, где время выступает в качестве параметра. Если в качестве переменной взять напряжение на выходе автогенератора Томсоновского типа с мягким режимом возбуждения колебаний, то фазовый портрет установившегося режима будет представлять собой движение по окружности, а переходные процессы — спирали, которые сходятся к этой окружности (см.рис. 3.48).

В случае жесткого режима возбуждения колебаний существует два ненулевых решения уравнения (3.48), каждому из которых соответствуют гармонические колебания на выходе. На рис. 3.49 им соответствуют окружности, однако одно из них (окружность, изображенная пунктиром) будет неустойчивым, соответственно, сколь угодно малые флуктуации приведут к тому, что система из этого режима перейдет либо к устойчивому циклу, либо в точку с координатами $(0,0)$ — колебания прекратятся.

3.6.2 RC-генератор

Поскольку индуктивность катушки зависит от ее размеров, сделать компактный генератор Томсоновского типа для низких ($<$10 кГц) частот оказывается сложно. Чаще в цепи обратной связи таких генераторов используют частотно-селективные цепи, содержащие только сопротивления и конденсаторы. Примером такого генератора является генератор на операционном усилителе с цепью Вина (см. рис. 3.50).

Генератор представляет собой неинвертирующий усилитель на ОУ выход которого соединён со входом цепью $R_2C_2R_1C_1$, называемой цепью Вина. Коэффициент передачи этой цепи:

$$K(\omega) = \frac{1}{\left(1 + \frac{C_2}{C_1} + \frac{R_1}{R_2}\right) + i\left(\frac{\omega^2 R_1 C_1 R_2 C_2 - 1}{\omega C_1 R_2}\right)}. \tag{3.49}$$

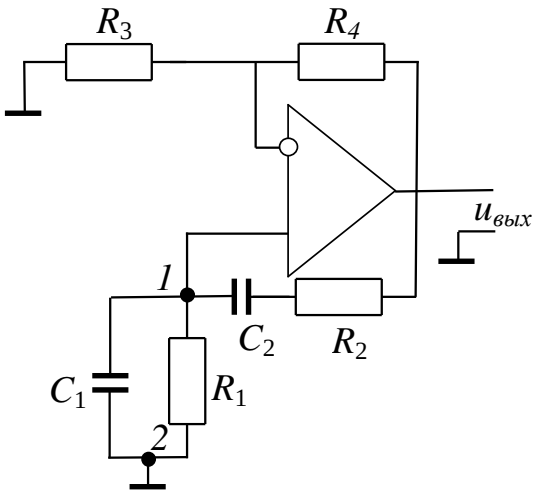


Рис. 3.50: RC генератор с цепочкой Вина.

Чаще всего используется симметричная цепь Вина: $C_1 = C_2 = C$, $R_1 = R_2 = R$. В этом случае:

$$K(\omega) = \frac{1}{\left(3 + i\frac{\omega^2 R^2 C^2 R_2}{\omega RC}\right)}, \tag{3.50}$$

$$|K(\omega)| = \frac{\omega RC}{\sqrt{\omega^4 R^4 C^4 + 7\omega^2 R^2 C^2 + 1}}, \tag{3.51}$$

$$\varphi_K = \operatorname{arctg}\frac{1 - \omega R^2 C^2}{3\omega C_1 R_1}. \tag{3.52}$$

Используя условия (3.35) и (3.36) получаем, что стационарные колебания могут существовать на частоте $\omega_0 = 1/RC$, на которой $\varphi_K = 0$ (считаем, что операционный усилитель является идеальным и в рассматриваемом диапазоне частот не вносит сдвига фазы). На этой частоте $|K(\omega)| = 1/3$, следовательно, коэффициент усиления усилителя $K_0 = 1 + R_4/R_3$ должен быть ≥ 3. Необходимо отметить, что, в отличие от генераторов Томсоновского типа, колебания на выходе которых по форме очень близки к гармоническим, RC-генераторы без дополнительных цепей стабилизации могут генерировать колебания, форма которых может сильно отличаться от гармонической и приближаться к прямоугольной. (см.рис. 3.51).

Качественно это можно объяснить следующим образом: высокодобротный колебательный контур является эффективным фильтром для гармоник основной частоты в то время, как АЧХ цепи Вина (3.51) пологая. Для того, чтобы получить гармонические колебания с помощью генератора, изображенного на рис. 3.50, необходимо с помощью цепи R_4R_3 устанавливать коэффициент усиления K_0 как можно ближе к пороговому значению $K_0 = 3$. В случае же, когда коэффициент усиления выбран с “запасом”, т.е. $K_0 \gg 3$, форма колебаний будет подобна изображенной на рис. 3.51.

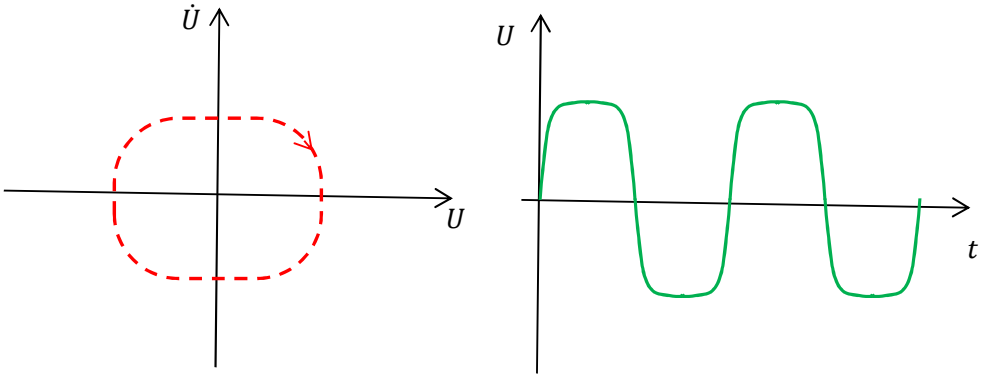


Рис. 3.51: Слева: пример фазового портрета RC генератора, справа: соответствующая форма колебаний на выходе.

3.6.3 Релаксационный генератор

Периодические незатухающие колебания, сильно отличающиеся по форме от гармонических, называются релаксационными. Происхождение названия связанно с тем, что простейшим источником таких колебаний может быть схема, содержащая бистабильный элемент (устройство с двумя устойчивыми состояниями) и энергоёмкий элемент (например, конденсатор). Процесс генерации колебаний заключается в чередовании быстрых (скачкообразных) изменений состояния бистабильного элемента в промежутке между которыми происходит медленное изменение энергии (релаксация) запасенной в энергоёмком элементе. Самым простым примером релаксационного генератора может быть RC-генератор с газоразрядной лампой (см. рис. 3.52).

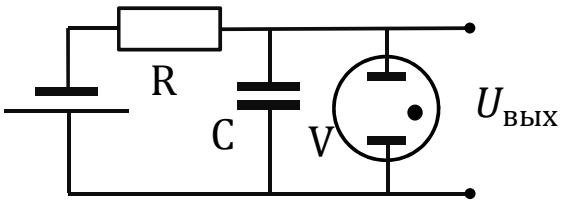


Рис. 3.52: Релаксационный генератор с газоразрядной лампой.

Сопротивление газоразрядной лампы в отсутствие разряда R_0 много больше, чем при его наличии $R_р$, причем зажигание разряда происходит при напряжении $U_З$ которое выше, чем напряжение $U_Г$ при котором он гаснет. Будем считать, что зажигание и прекращение разряда происходит за время существенно меньшее, чем постоянная времени цепочки RC, $R_0 \gg R$, $R_р \ll R$. Тогда при включении питания конденсатор будет заряжаться через R до тех пор, пока напряжение на нем не достигнет $U_З$, при котором зажжется разряд в лампе V. После этого конденсатор быстро разрядится через лампу. При напряжении на ней равном $U_Г$ разряд по-

гаснет и процесс повторится. Период колебаний, форма которых будет близка к пилообразной (см. рис. 3.53), равен:

$$T \simeq RC\frac{E - U_з}{E + U_з}. \tag{3.53}$$

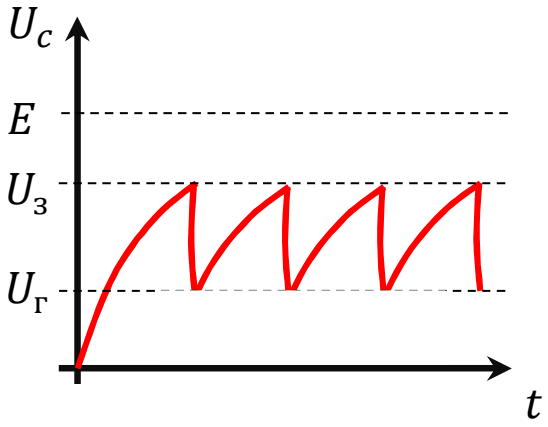


Рис. 3.53: Напряжение на выходе релаксационного генератора.

3.6.4 Мультивибратор на операционном усилителе

На практике бывает необходимо иметь источники периодических колебаний заданной формы. Это может быть последовательность прямоугольных импульсов (меандр), пилообразные колебания и любые другие.

Простейший генератор прямоугольных импульсов на операционном усилителе (ОУ) изображен на рис.3.54.

Предположим, что после включения напряжение на неинвертирующем входе оказалось больше, чем на инвертирующем (принципиального значения это предположение не имеет). Тогда на выходе ОУ быстро установится максимальное положительное напряжение (можно легко проверить, что для высоких частот, соответствующей этому переходному процессу, коэффициент передачи цепи, образованной R и C стремится к нулю, соответственно, усиление максимально). После этого, конденсатор начнет заряжаться через резистор R и, в какой-то момент, напряжение на инвертирующем входе станет больше, чем на неинвертирующем (задается делителем R_2, R_1). В этот момент напряжение на выходе ОУ быстро изменит знак (станет максимальным отрицательным), конденсатор начнет перезаряжаться, процесс будет повторяться циклически. Период колебаний равен:

$$T = 2RC\ln(\frac{2R_2}{R_1} + 1).$$

3.6.5 Кварцевый генератор

Использование высокодобротного колебательного контура в автогенераторе позволяет получить колебания, по своей форме очень близкие к гармоническим. В

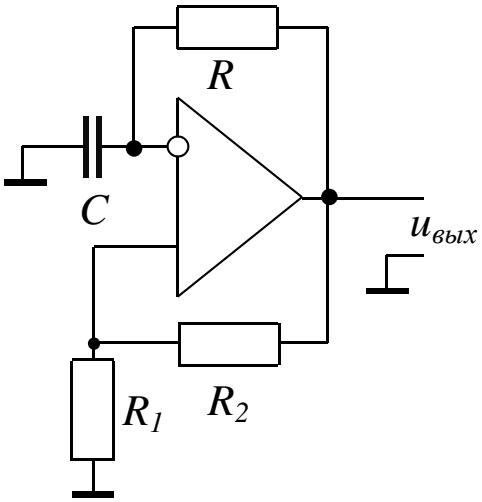


Рис. 3.54: Мультивибратор на операционном усилителе.

радиочастотном диапазоне (100 кГц – 100 МГц) контуры, состоящие из катушки и конденсатора, могут иметь добротность $Q = 50 \div 300$. Для обеспечения высокой стабильности частоты подбирают конденсаторы с заполнением, диэлектрическая проницаемость которого увеличивается с температурой. Это позволяет, до некоторой степени, компенсировать изменение индуктивности, которое, чаще всего, связано с изменением геометрических размеров катушки. Более высокую

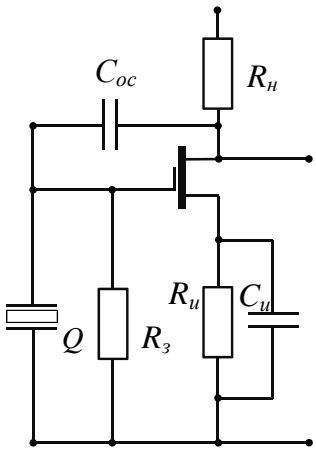


Рис. 3.55: Автогенератор с кварцевым резонатором.

эквивалентную добротность и стабильность частоты имеют генераторы с электромеханическим резонатором. Обычно в качестве такого резонатора используется пластина из пьезоэлектрика, чаще всего — кристаллического кварца (см. раздел 2.2.2), отчего такие генераторы называют кварцевыми. Относительная нестабиль-

ность частоты такого генератора может не превышать $\frac{\Delta\omega}{\omega} < 10^{-9}$ за сутки. Схема простейшего кварцевого генератора изображена на рис. 3.55.

Глава 4

Шумы в радиофизических системах.

4.1 Случайные процессы

В любых радиофизических системах кроме детерминированных сигналов — токов и напряжений, которые можно рассчитать, зная параметры системы и задавая внешнее воздействие на нее (сигнал на входе), присутствуют их случайные изменения (флуктуации), причиной которых могут быть самые различные физические явления. Такие изменения мы будем называть шумами. Шумы могут быть связаны как с помехами, попадающими на вход одновременно с входным сигналом, так и возникать в элементах самой системы. Физические механизмы возникновения шумов мы рассмотрим позднее, начнем со способов их описания. С математической точки зрения шум — это случайный процесс, обозначим его $\xi(t)$. В результате регистрации случайного процесса получают некоторую случайную запись $x(t)$, называемую *реализацией случайного процесса*. При повторении регистрации при одних и тех же начальных условиях получаются различные реализации $x(t)$. Таким образом, каждому моменту времени t соответствует бесконечное множество возможных значений x. В этом смысле $\xi(t)$ - это совокупность всех возможных реализаций $x(t)$. Каждое из значений x встречается с некоторой вероятностью. Вероятность значения x_2 в момент t_2 может зависеть от значения x_1 в предшествующий момент t_1. Для описания шумов используются статистические характеристики: среднее значение, дисперсия, корреляционная функция, спектральная плотность мощности шума и другие. Характеристики шума в общем случае тоже зависят от времени.

4.1.1 Основные характеристики случайных процессов.

Среднее значение (математическое ожидание) случайного процесса $\langle\xi(t)\rangle$ (напоминаем, угловые скобки означают усреднение по всем возможным реализациям).

Поскольку под шумами мы понимаем флуктуации — отклонения от некоторого среднего значения, то среднее значение шумового тока или напряжения будем считать равным нулю:

$$\xi_{noise}(t) = \xi(t) - \langle\xi(t)\rangle$$

$$\langle \xi_{noise}(t) \rangle = 0.$$

Интенсивность флуктуаций можно характеризовать дисперсией:

$$\sigma(t) = \langle \xi^2(t) \rangle - \langle \xi(t) \rangle^2 = \langle \xi^2_{noise}(t) \rangle. \quad (4.1)$$

Поскольку мощность, выделяемая на сопротивлении, пропорциональна квадрату тока через него (или квадрату напряжения на нем), дисперсия шумового тока или напряжения пропорциональна средней мощности флуктуаций. Если в цепи присутствуют два источника шума, то суммарная дисперсия:

$$\sigma_\sum(t) = \langle (\xi_1(t) + \xi_2(t))^2 \rangle = \langle \xi_1^2(t) \rangle + \langle \xi_2^2(t) \rangle + 2\langle (\xi_1(t)(\xi_2(t) \rangle. \quad (4.2)$$

Если эти два источника статистически независимы, то последний член равен нулю, таким образом, общая мощность шума от нескольких независимых источников равна сумме их мощностей. Мерой статистической зависимости шумовых процессов является корреляционная функция:

$$B(\xi_1(t_1), \xi_2(t_2)) = \langle \xi_1(t_1)\xi_2(t_2) \rangle.$$

Часто важное значение имеет то, насколько статистически зависимы (коррелированы) значения одного и того же случайного процесса в различные моменты времени. В этом случае используется автокорреляционная функция (приставку "авто" часто опускают).

4.1.2 Характеристики стационарного шума.

Существует практически важный вид шумов, статистические характеристики которых от времени не зависят. Таковыми обычно можно считать шумы приборов в установившемся (стационарном) режиме, когда их параметры не меняются со временем. Подобные шумы называют стационарными. Далее мы будем рассматривать именно такие шумы.

До сих пор статистические характеристики определялись нами как средние по реализациям. На практике осуществить такое усреднение сложно, поскольку оно предполагает, что, всякий раз, в точности воспроизводятся начальные условия. Эргодическая гипотеза (одно из важнейших положений, лежащих в основе термодинамики) состоит в том, что, вместо усреднения по реализациям для стационарных случайных процессов можно использовать усреднение по времени одной, достаточно продолжительной реализации. Пояснить это можно на примере игрального кубика. Для определения среднего числа, выпадающего на кубике, надо взять как можно больше (допустим, миллион) одинаковых кубиков (ансамбль), бросить их и усреднить полученный результат. Это есть усреднение по ансамблю. Очевидное неудобство такого усреднения — надо где-то найти этот миллион одинаковых кубиков. Вроде бы очевидно, что можно бросить один и тот же кубик миллион раз и усреднить — это пример усреднения по времени. Но в этом случае условия во время бросания кубика от раза к разу могут меняться (нестационарный случайный процесс) и мы можем получить *разные* результаты в первом и во втором случаях.

И только для *стационарного* случайного процесса усреднение по ансамблю и по времени должно дать одинаковые результаты. Усреднение одной реализации по времени мы будем обозначать чертой сверху: $\overline{x(t)}$.

Для стационарного процесса среднее значение и дисперсия не зависят от времени. Дисперсия стационарного шума:

$$\sigma = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt = \overline{\Delta x^2}. \tag{4.3}$$

Подчеркнем, что время усреднения должно быть достаточно большим (много больше времени корреляции — см. ниже).

Автокорреляционная функция стационарного процесса зависит только от разности времен $\tau = t_2 - t_1$:

$$B(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)dt. \tag{4.4}$$

Функция $B(\tau)$ характеризуют связь (корреляцию) между значениями $x(t)$, разделенными промежутком времени τ. Чем медленнее, плавнее изменяется $x(t)$, тем медленнее спадает зависимость автокорреляционной функции от τ. Можно сказать, что функция автокорреляции определяет "память", или "последействие".

При увеличении времени τ автокорреляционная функция шума стремится к нулю: $\lim_{\tau\to\infty} B(\tau) = 0$. *Временем корреляции* τ^* для стационарного шума будем считать время, за которое корреляционная функция спадает в e раз.

На практике при использовании эргодической гипотезы время усреднения T велико, но конечно, поэтому при выполнении усреднения важно следить, чтобы выполнялось условие $T \gg \tau^*$.

Функция $B(\tau)$ — четная. Обратим внимание, что, при $\tau = 0$ $B(\tau)$ принимает наибольшее значение равное дисперсии:

$$B(0) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+0)dt = \sigma. \tag{4.5}$$

4.1.3 Гауссовы шумы.

В радиофизике (как и вообще в природе) наиболее часто встречаются *нормальные* случайные процессы. Функция распределения вероятности для них описывается распределением Гаусса:

$$P(a \le x(t) \le b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-x(t)^2/2\sigma^2} dx.$$

здесь P — вероятность того, что в момент времени t значение $x(t)$ лежит в интервале от a до b. Центральная предельная теорема математической статистики утверждает, что случайный процесс, который является независимой суммой

большого числа N случайных процессов одной природы, описывается функцией распределения вероятности, которая асимптотически стремится к гауссовому распределению при стремлении $N \to \infty$. Классический пример — максвелловское распределение скоростей молекул в идеальном газе. Нормальные случайные процессы так же называют *гауссовскими*. Очевидно, что вероятность обнаружить значение $|x(t)| \gg \sqrt{\sigma}$ для такого процесса мала.

В дальнейшем мы будем предполагать, что рассматриваемые шумы — гауссовы. Заметим, однако, что при прохождении шумов через нелинейные системы характер распределения может существенно изменяться.

4.1.4 Спектральная плотность мощности шума

Традиционное определение спектральной плотности

Аналогично детерминированным сигналам шумы удобно рассматривать в спектральном представлении. Однако, фурье-образ, вычисленный для флуктуаций сам будет случайной функцией частоты. Физический смысл имеет **энергетический** спектр. Мы будем определять **спектральную плотность мощности флуктуаций** на частоте ω как дисперсию шума в единичной полосе частот возле частоты ω. Пояснить это можно, описав следующую процедуру измерений: предположим, что источник шумового напряжения действует на входе идеального фильтра, коэффициент передачи которого равен единице в полосе $\omega \div \omega + \Delta\omega$ и нулю вне этой полосы (рис. 4.1).

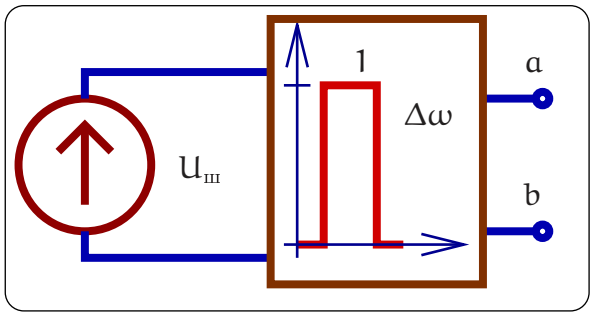


Рис. 4.1: К определению спектральной плотности.

На выходе фильтра мы получим реализацию случайного процесса. Используя эргодическую гипотезу и записывая реализации случайного напряжения на выходе фильтра в течение достаточно долгого времени мы можем измерить дисперсию случайного напряжения Δu^2. Эта дисперсия будет пропорциональна полосе фильтра $\Delta\omega$:

$$\Delta u^2 \equiv \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} u_{ab}(t)^2 \, dt \simeq S_u(\omega) \times 2 \times \frac{\Delta\omega}{2\pi} \tag{4.6}$$

Если выбрать полосу фильтра $\Delta\omega$ достаточно малой (так, что бы результат этого измерения не изменялся при смещении центральной частоты фильтра на $d\omega \simeq \Delta\omega$), то, очевидно, коэффициент пропорциональности $S_u(\omega)$ не зависит от частоты в пределах $\omega \div \omega + \Delta\omega$ и характеризует флуктуации на частотах, проходящих через фильтр. Величину $S_u(\omega)$ мы и назовем спектральной плотностью.

Отсюда сразу следует физический смысл спектральной плотности: это среднеквадратичное напряжение шума, генерируемое в *единичной* спектральной полосе частот. Другими словами, спектральная плотность мощности шума показывает, какая энергия в среднем за единицу времени приходится на единичный частотный интервал вблизи частоты ω, или какова *мощность* шума на частоте ω.

Размерность спектральной плотности мощности шума: если имеются ввиду флуктуации напряжения, то $[S(\omega)] = В^2/Гц$.; в случае флуктуаций тока $[S(\omega)] = А^2/Гц$.

Очевидно, что полная дисперсия будет определяться интегралом:

$$\sigma \;=\; \int_{-\infty}^{\infty} S(\omega)\,\frac{d\omega}{2\pi} \tag{4.7}$$

Формально величина $S_u(\omega)$ определена как для положительных, так и отрицательных частот — так называемое двустороннее определение спектральной плотности. С "двусторонним" определением связано появление множителя 2 в (4.6). Появление отрицательных частот при математическом переходе к спектральному представлению подробно разъяснено ранее — см. 1.11. Физический смысл имеет лишь *односторонне* определенная спектральная плотность $S_u^+(\omega)$, определенная только для положительных частот. Для $\omega > 0$ эти величины связаны очевидным равенством:

$$S_u^+(\omega) = 2S_u(\omega).$$

Дисперсия шума в произвольной полосе частот $\omega_1 \div \omega_2$ может быть выражена через интегралы от односторонней или двусторонней спектральной плотности:

$$\sigma \;=\; \int_{\omega_1}^{\omega_2} S^+(\omega)\,\frac{d\omega}{2\pi} = \int_{-\omega_2}^{-\omega_1} S(\omega)\,\frac{d\omega}{2\pi} + \int_{\omega_1}^{\omega_2} S(\omega)\,\frac{d\omega}{2\pi}. \tag{4.8}$$

Пояснения к определению спектральной плотности (4.6)

Применим преобразование Фурье к случайной величине $u(t)$ и выразим напряжение на выходе узкополосного фильтра (рис. 4.1):

$$u(\omega) = \int_{-\infty}^{\infty} u(t)\,e^{-i\omega t}\,dt, \tag{4.9}$$

$$u_{ab} = \int_{\Delta\omega} u(\omega)\,e^{-i\omega t}\,\frac{d\omega}{2\pi} \tag{4.10}$$

Среднее от этого напряжения будет равно нулю, а среднекрадратичное отклонение равно

$$\overline{u_{ab}^2} = \int_{\Delta\omega}\int_{\Delta\omega} \overline{u(\omega)u^*(\omega')}\,e^{-i(\omega-\omega')t}\,\frac{d\omega\,d\omega'}{(2\pi)^2}. \tag{4.11}$$

Сопоставляя это выражение и определение (4.6) спектральной плотности мы приходим к выводу, что должно выполняться равенство:

$$\overline{u(\omega)\,u(\omega')^*} \;=\; 2\pi \times \delta(\omega-\omega') \times 2S(\omega).$$

Чтобы доказать это, выпишем

$$\begin{aligned}\overline{u(\omega)\,u(\omega')^*} &= \int_{-\infty}^{\infty} \overline{u(t)u(t')}\, e^{-i(\omega t-\omega' t')}\, dt\, dt' = \\ &= \int_{-\infty}^{\infty} B(t-t')\, e^{-i(\omega t-\omega' t')}\, dt\, dt'.\end{aligned}$$

Это выражение можно упростить, сняв одно интегрирование. Для этого сделаем замену переменных:

$$\begin{aligned}\tau &= t-t', \quad \tau' = \frac{t+t'}{2}, \\ \omega t-\omega' t' &= (\omega+\omega')\frac{\tau}{2}+(\omega-\omega')\tau'\end{aligned}$$

и проинтегрируем по τ', принимая во внимание известное представление дельта-функции:

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ipx}\, dp. \tag{4.12}$$

В итоге получаем:

$$\overline{u(\omega)\,u(\omega')^*} = 2\pi\,\delta(\omega-\omega')\; S_u\left(\frac{\omega+\omega'}{2}\right) = 2\pi\,\delta(\omega-\omega')\; S_u(\omega), \tag{4.13}$$

$$S_u(\omega) = \int_{-\infty}^{\infty} B(\tau)\, e^{i\omega\tau}\, d\tau. \tag{4.14}$$

Нетрудно убедиться, что величина $S_u(\omega)$ есть спектральная плотность шума u в (4.6). Смысл равенства (4.13) заключается в том, что спектральные гармоники $u(\omega)$ и $u(\omega')$ статистически не зависимы, а средний квадрат одной гармоники определяется спектральной плотностью $S_u(\omega)$.

Заметим, что формула (4.14), является частью теоремы Винера-Хинчина, которая ниже будет доказана более простым способом.

Другое определение спектральной плотности

К определению спектральной плотности мощности можно подойти иначе. Будем считать, что $u(t)$ - это запись флуктуаций напряжения и вычислим среднюю по времени мощность шума:

$$\overline{P}^T = \frac{\int\limits_{-T/2}^{T/2} (x(t))^2 dt}{T} = \frac{1}{2\pi}\frac{\int\limits_{-\infty}^{\infty} |u(\omega)|^2 d\omega}{T},$$

где $u(\omega)$ обозначено обычное фурье-преобразование реализации $x(t)$ случайного процесса длительностью T:

$$u(\omega) = \int_{-T/2}^{T/2} u(t)\, e^{-i\omega t}\, dt \tag{4.15}$$

и использовано равенство Парсеваля. Следовательно, “радиофизическую” спектральную плотность средней мощности шума можно определить как

$$S_{rf}(\omega) = \lim_{T\to\infty} \frac{|u(\omega)|^2}{T}. \tag{4.16}$$

Нетрудно убедиться, что эти два определения имеют одинаковый физический смысл: спектральная плотность мощности шума показывает, какая энергия в среднем за единицу времени приходится на единичный частотный интервал вблизи частоты ω, или, другими словами, какова мощность шума на частоте ω.

Как уже говорилось, функция $S(\omega)$ — четная и определена в полосе частот от $-\infty$ до ∞. Поскольку очевидный физический смысл имеют положительные частоты, удобно использовать односторонне определенную спектральную плотность мощности шума:

$$S^+(\omega) = 2S(\omega).$$

Спектральная плотность $S^+(\omega)$ определена в полосе $[0 \div \infty)$. В технической литературе часто символом S обозначают именно одностороннюю спектральную плотность.

4.1.5 Теорема Винера - Хинчина

Теорема Винера - Хинчина утверждает, что спектральная плотность мощности стационарного шума представляет собой Фурье-образ автокорреляционной функции для этого шума. Докажем это утверждение. По определению:

$$B(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)dt. \tag{4.17}$$

Представив функцию $x(t+\tau)$ в виде интеграла Фурье:

$$x(t+\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(\omega)e^{i\omega(t+\tau)}\, d\omega, \tag{4.18}$$

получим

$$B(\tau) = \frac{1}{2\pi} \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} x(\omega)e^{i\omega\tau} \left(\int_{-T/2}^{T/2} e^{i\omega t}x(t)dt \right) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\lim_{T\to\infty} \frac{1}{T}|x(\omega)|^2 \right] e^{i\omega\tau}d\omega.$$

Следовательно,

$$B(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{rf}(\omega)e^{i\omega\tau}d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)e^{i\omega\tau}d\omega \tag{4.19}$$

(Определения спектральной плотности 4.6 и 4.16 эквивалентны, индекс rf в дальнейшем использовать не будем). Соответственно, справедливо и обратное:

$$S(\omega) = \int_{-\infty}^{\infty} B(\tau) e^{-i\omega\tau} d\tau \tag{4.20}$$

— утверждение теоремы доказано.

Для одностороннего определения спектральной плотности теорема Винера-Хинчина может быть записана в виде (пределы интегрирования и по ω, и по τ выбираются от 0 до ∞) :

$$S^+(\omega) = 4\int_0^{\infty} B(\tau)\ \cos(\omega\tau)\ d\tau \tag{4.21}$$

$$B(\tau) = \int_0^{\infty} S^+(\omega)\ \cos(\omega\tau)\ \frac{d\omega}{2\pi} \tag{4.22}$$

При измерении спектральной плотности часто используются электронные приборы, называемые анализаторами спектра. В составе такого прибора может быть набор фильтров, настроенных на разные частоты, либо один фильтр, который перестраивается по частоте. Для упрощения используются гетеродинирование - перенос спектра в удобную для анализа область частот. Это “традиционный” способ по формуле (4.6), но его используют сейчас все реже, в основном для высоких частот (на которых оцифровка невозможна).

Для относительно низких частот ($\omega < 10 \ldots 100$ ГГц) все чаще используют второй способ измерения спектральной плотности по формуле (4.16). Это связано с возможностью оцифровки сигнала и вычисления $S_T(\omega)$ на компьютере через быстрое преобразование Фурье— это оказывается проще. Кроме того, при вычислении Фурье-образа от сигнала можно получить комплексную величину, которая несет информацию не только о спектральной плотности мощности, но и о фазе каждой компоненты спектра. Приборы, реализующие эту возможность, называются векторными анализаторами.

4.1.6 Белый шум

Шум, спектральная плотность которого не зависит от частоты, принято называть белым: $S(\omega) = const$. Из соотношения 4.19 следует, что функцией автокорреляции белого шума является δ - функция. По этой причине белый шум называют δ - коррелированным процессом у которого время корреляции формально равно нулю. Строго говоря, постоянство энергетического спектра на всех частотах не может иметь места в действительности, поскольку полная мощность такого процесса была бы равна бесконечности. Таким образом, белый шум является абстрактной математической моделью. В природе таких процессов не существует. Однако, это не мешает приближенно заменять реальные достаточно широкополосные случайные процессы белым шумом тогда, когда полоса пропускания цепи, на которую воздействует случайный сигнал, оказывается существенно уже ширины спектра

шума, например, время корреляции тепловых шумов, вызванных случайным движением электронов в проводнике, весьма мало: $\tau_{th} < 10^{-15}$ сек[1] а соответствующая верхняя частота $f_B > 10^{15}$ Гц.

4.1.7 Преобразование шумов в линейных цепях.

В силу принципа суперпозиции любые сигналы, в том числе и шумовые, проходят через линейные цепи, не влияя друг на друга. Поскольку спектральная плотность мощности шума пропорциональна квадрату шумового тока (напряжения), справедливо следующее соотношение:

$$S_{вых}(\omega) = |K(\omega)|^2 S_{вх}(\omega), \tag{4.23}$$

здесь $K(\omega)$ — комплексный коэффициент передачи линейной цепи. Действительно, из определения (4.16) спектральной плотности следует

$$S_{вых}(\omega) = \lim_{T\to\infty} \frac{|u_{вых}|^2}{T} = |K(\omega)|^2 \lim_{T\to\infty} \frac{|u_{вх}|^2}{T} = |K(\omega)|^2 S_{вх}(\omega) \tag{4.24}$$

Из формулы (4.23) следует один из возможных способов определения модуля коэффициента передачи линейных цепей, широко используемый на практике: если на вход цепи подать белый шум, зависимость спектральной плотности от частоты на выходе системы будет определяться квадратом коэффициента передачи.

Дисперсия флуктуаций на выходе цепи

$$\sigma_{вых} = \int_{-\infty}^{\infty} S_{вх}(\omega)|K(\omega)|^2 \frac{d\omega}{2\pi} = \int_{0}^{\infty} S^{+}_{вх}(\omega)|K(\omega)|^2 \frac{d\omega}{2\pi}.$$

Не следует забывать, что спектральная плотность мощности флуктуаций может быть определена как для напряжения, так и для тока. Очевидно, что спектральная плотность мощности флуктуаций напряжения $S_u(\omega)$ на комплексном сопротивлении $Z(\omega)$ связана со со спектральной плотностью $S_i(\omega)$ протекающего тока как

$$S_u(\omega) = |Z(\omega)|^2 S_i(\omega). \tag{4.25}$$

Доказывается это равенство опять через формулу (4.16).

[1] Время корреляции в проводнике определяется так называемым максвелловским временем релаксации, оценку которого можно вывести из следующих соображений. Мысленно выделим в проводнике цилиндр площади S и высоты d, который будем считать RC-цепочкой с емкостью $C = \epsilon_0 S/d$ и сопротивлением $R = \rho d/S$, где ϵ_0 — диэлектрическая постоянная, а ρ — удельное сопротивление. Очевидно, что время релаксации $RC = \epsilon_0\rho$ такой цепочки не зависит от геометрии и определяется только удельным сопротивлением. Скажем, для нихрома $RC \simeq 10^{-17}$ сек, в полупроводниках на несколько порядков больше — $RC \simeq 10^{-15} \ldots 10^{-13}$ сек. Поэтому на частотах $\omega \ll 1/\tau_{th}$ такой шум можно считать белым.

4.2 Основные источники шумов

4.2.1 Тепловой шум

Тепловой шум вызывают флуктуации объемной плотности электрического заряда в проводниках, возникающие благодаря хаотическому тепловому движению носителей заряда. Несмотря на электрическую нейтральность проводника в целом, внутри него возникают переменные во времени электромагнитные поля, а на поверхности — случайная разность потенциалов. Спектр шумового напряжения оказывается очень широким из-за высокой концентрации заряженных частиц и большой средней тепловой скорости их движения. При комнатной температуре и на частотах до $\simeq 10^{12}$ Гц его можно считать белым.

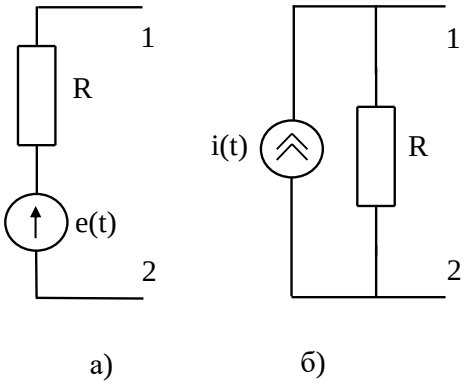


Рис. 4.2: Представление источника теплового шума в виде: а - генератора напряжения, б - генератора тока.

Для расчета шума в цепи, содержащей проводник (сопротивление), последний можно представить эквивалентной схемой, состоящей либо из источника шумового напряжения $u(t)$ последовательно с сопротивлением R (рис.4.2 а), либо из источника тока $i(t)$ с параллельным сопротивлением (рис.4.2 б) (теорема об эквивалентном генераторе). Спектральные плотности мощности флуктуаций тока и напряжения в этих схемах связаны соотношением

$$S_i = \frac{S_u}{R^2}.$$

(Обратим внимание на то, что эти схемы эквивалентны только по отношению к внешней для них цепи.)

Найдем спектральную плотность теплового шума. Для этого рассмотрим RC цепь (см. рис. 4.3). Будем считать, что резистор R поддерживается при температуре T и является источником белого шума с неизвестной пока, *но постоянной* спектральной плотностью S_0. Шум резистора создает на обкладках конденсатора некоторое шумовое напряжение, спектральную плотность которого рассчитываем

в соответствии с (4.23)

$$S_c = |K|^2 S_0 = \frac{S_0}{1 + (\omega RC)^2}$$

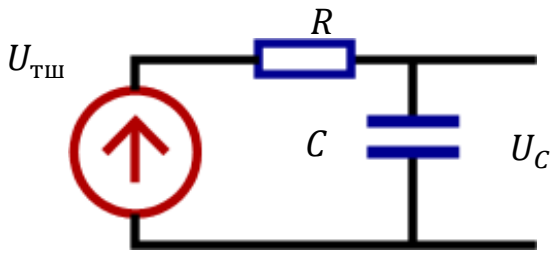


Рис. 4.3: К расчету спектральной плотности теплового шума.

Проинтегрировав это соотношение по всем частотам, найдем полную величину среднеквадратичного напряжения на конденсаторе и его среднюю энергию W_C:

$$\overline{\Delta u_c^2} = \frac{S_0}{2\pi} \int_{-\infty}^{\infty} \frac{1}{1 + (\omega RC)^2} d\omega = \frac{S_0}{2RC}, \tag{4.26}$$

$$W_C = \frac{C\overline{\Delta u_c^2}}{2} = \frac{S_0}{4R}. \tag{4.27}$$

Но, по теореме о равнораспределении энергии, в любой физической системе с одной степенью свободы и любым механизмом флуктуаций средняя энергия флуктуаций равна $kT/2$, где k — постоянная Больцмана. Следовательно,

$$\frac{CS_0}{4RC} = \frac{kT}{2}, \quad \Rightarrow \quad S_0 = 2kTR. \tag{4.28}$$

Односторонне определенная спектральная плотность тепловых шумов сопротивления в два раза больше:

$$S_0^+ = 4kTR \tag{4.29}$$

Это соотношение носит название **формулы Найквиста**. Закон о равномерном распределении энергии по степеням свободы справедлив при таких частотах и температурах, когда квантовомеханические эффекты несущественны, т. е. величина кванта энергии мала по сравнению с kT. При комнатной температуре условие выполняется для частот $\omega/2\pi < 10^{12}$Гц.

Квантовая механика позволяет получить более общее соотношение:

$$S_0 = 2R\left(\frac{\hbar\omega}{2} + \frac{\hbar\omega}{e^{\frac{\hbar\omega}{kT}} - 1}\right), \tag{4.30}$$

которое переходит в (4.28) при $\hbar\omega < kT$.

Несмотря на кажущуюся малость тепловых шумов, они в ряде случаев оказываются решающим фактором, ограничивающим реальную чувствительность электронной аппаратуры и предельную точность физических измерений. Заметим, что

в соответствии с формулой Найквиста, величина теплового шума определяется только активным сопротивлением цепи, реактивные элементы не генерируют тепловой шум. Для доказательства этого утверждения снова рассмотрим цепь рис. 4.3. Шумовой ток, генерируемый резистором, не может нагревать конденсатор, так как сдвиг фаз между током и напряжением на конденсаторе равен $\pi/2$. Допустим, что и конденсатор генерирует тепловой шум. Тогда шумовой ток, генерируемый конденсатором, должен нагревать резистор, поскольку сдвиг фаз между током и напряжением на резисторе равен 0. В результате температура резистора будет повышаться. Получается, что тепло от конденсатора с температурой T передается резистору с более высокой температурой. Но такой процесс противоречит второму началу термодинамики. Следовательно, он невозможен.

Заметим, что в реальном конденсаторе могут присутствовать потери (например, за счет возникновения токов утечки в диэлектрике между его пластинами). Это эквивалентно включению активного сопротивления параллельно с идеальной емкостью, следовательно, в такой системе появятся тепловые шумы. Таким образом, полученная нами формула Найквиста является частным случаем важного физического закона, который формулируется в виде **флуктуационно-диссипационной теоремы**, применимой к любым, а не только электрическим системам. Она утверждает, что чем больше связь системы с окружающим миром (больше диссипация), тем больше в ней тепловые флуктуации. Можно качественно сформулировать это иначе: чем интенсивнее энергия может уходить из системы в термостат (за счет диссипации), тем больше флуктуационное воздействие термостата на систему.

4.2.2 Дробовой шум

Причиной возникновения дробового шума является дискретный характер носителей заряда. В отличие от теплового шума, дробовой шум появляется при наличии транспорта частиц, то есть в неравновесной системе. Дробовой шум тока возникает при протекании тока через потенциальный барьер (например, в электронных лампах, полупроводниковых контактах).

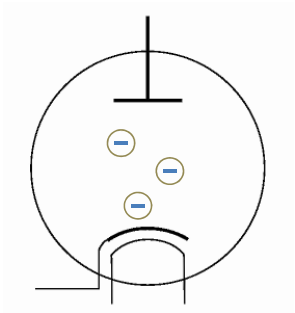


Рис. 4.4: Ламповый диод - источник дробового шума

Рассмотрим ток через вакуумный диод (рис.4.4). Ток в нем создается электронами, вылетающими из нагретого катода. Для каждого отдельного электрона у

поверхности катода вероятность вылета мала. Можно показать, что, в этом случае, число электронов, вылетающих в единицу времени, является случайной величиной, описываемой распределением Пуассона. Следовательно, ее дисперсия, так же, как и среднее значение, пропорциональны этой вероятности. Будем считать, что диод

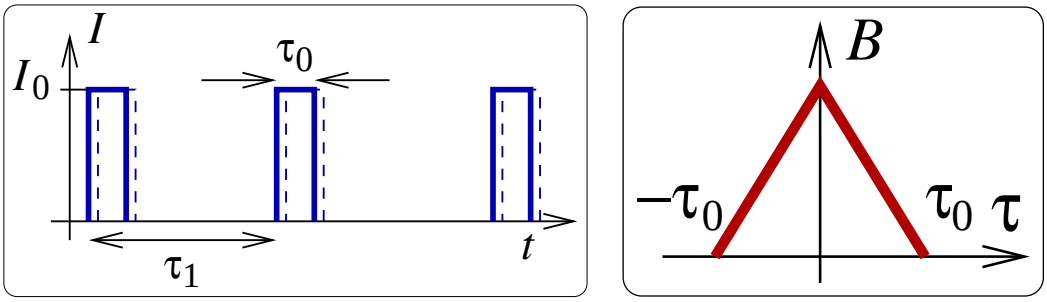


Рис. 4.5: Случайные импульсы тока в анодной цепи диода и корреляционная функция $B(\tau)$ (4.32). τ_1 — среднее время между импульсами (оно меняется от импульса к импульсу).

работает в режиме насыщения: все вылетевшие электроны достигают анода. Характерное время пролета τ_0 очень мало (для обычных ламповых диодов порядка 10^{-9} с). Тогда ток в цепи анода представляет собой последовательность коротких импульсов, которые практически не перекрываются (рис. 4.5). Время между импульсами случайно, но есть среднее время τ_1, которое характеризует интенсивность появления импульсов.

Эта модель случайных коротких импульсов оказывается верной и в более общем случае при протекании носителей через какой-либо барьер. Наличие барьера является необходимым для возникновения дробового шума.

Вернемся к нашей модели на рис. 4.5.

Будем рассматривать область частот, в которой $\tau_1 = 2\pi/\omega \gg \tau_0$. этом случае вероятность перекрытия импульсов мала. Найдем функцию корреляции для тока I в анодной цепи:

$$\overline{I(t)I(t-\tau)} = B(\tau) = \frac{1}{T}\int_{-T/2}^{T/2} I(t)I(t-\tau)dt = \frac{1}{T}\times\frac{T}{\tau_1}\int_{-\tau_0}^{\tau_0} I(t)I(t-\tau)dt. \qquad (4.31)$$

Тогда

$$B(\tau) = \begin{cases} \frac{I_0^2}{\tau_1}(\tau_0 - |\tau|), & \text{если } |\tau| < \tau_0, \\ 0, & \text{если } |\tau| \geq \tau_0 \end{cases} \qquad (4.32)$$

Нас интересуют частоты $\omega \ll 1/\tau_0$, поэтому корреляционную функцию на рис. 4.5) можно аппроксимировать дельта-функцией $B(\tau) = A\delta(\tau)$ (A — постоянная). Вычисляем постоянную перед дельта-функцией, используя (4.32) :

$$A = \int_{-\tau_0}^{\tau_0} B(\tau')d\tau' = \frac{I_0^2\tau_0^2}{\tau_1} = Ie, \quad \Rightarrow B(\tau) \simeq \bar{I}e\,\delta(\tau). \qquad (4.33)$$

Здесь $I_0 = e/\tau_0$ — ток в каждом импульсе, а средний ток $\bar{I} = I_0\tau_0/\tau_1 = e/\tau_1$.

Рассчитаем спектральную плотность мощности флуктуаций тока:

$$S^{+}_{I,др}(\omega) = 4\int_0^\infty B(\tau)\cos\omega\tau d\tau = 4 \times \frac{1}{2} \times \frac{e^2}{\tau_1} = 2e\bar{I}.$$

Здесь множитель $1/2$ появляется при интегрировании дельта-функции от нуля до бесконечности (в бесконечных пределах был бы множитель 1).

Мы доказали **теорему Шоттки**:

$$S^{+}_{I,др}(\omega) = 2e\bar{I} \tag{4.34}$$

Реально, формулой Шоттки в радиофизических системах можно пользоваться до частот порядка $10^7 \div 10^9$ Гц. Если при протекании тока через потенциальный барьер около барьера образуется объемный заряд, как например у катода вакуумного диода, дробовые флуктуации сглаживаются.

4.2.3 Генерационно-рекомбинационный шум

Образование и рекомбинация пар электрон-дырка в полупроводниках носит случайный характер. При флуктуациях числа свободных носителей заряда, очевидно, флуктуирует проводимость полупроводника. В результате, при протекании постоянного тока I_0 через полупроводник будет появляться флуктуирующая составляющая. Будем использовать двухуровневую модель: считаем, что генерация носителей происходит за счет перехода электронов из валентной зоны в зону проводимости. Пусть N_0 - среднее число носителей в зоне проводимости, $\overline{\delta N^2}$ - его дисперсия, τ_N - характерное время жизни этих носителей. Тогда автокорреляционная функция для флуктуаций числа носителей:

$$B_N(0) \equiv \overline{\delta N^2}$$

а для $t \neq 0$

$$B_N(\tau) \equiv \overline{\delta N^2}\exp(-\tau/\tau_N).$$

Спектральная плотность этих флуктуаций

$$S^{+}_{N,\Gamma\text{-}P}(\omega) = 4\int_0^\infty B_N(\tau)\cos(\omega\tau)d\tau = 4\overline{\delta N^2}\int_0^\infty \exp(-\tau/\tau_N)\cos(\omega\tau)d\tau =$$
$$4\overline{\delta N^2}\frac{\exp(-\tau/\tau_N)(-\frac{1}{\tau_N}(\cos(\omega\tau)+\omega\sin(\omega\tau))}{(-\frac{1}{\tau_N})^2+\omega^2}\Bigg|_0^\infty = \frac{4\overline{\delta N^2}\tau_N}{1+(\omega\tau_N)^2}. \tag{4.35}$$

Спектральная плотность флуктуаций тока, протекающего через такой полупроводник:

$$S^{+}_{I,\Gamma\text{-}P}(\omega) = I_0^2\frac{\overline{\delta N^2}}{N_0^2}\frac{4\tau_N}{1+(\omega\tau_N)^2}. \tag{4.36}$$

Для типичных полупроводников генерационно-рекомбинационный шум наблюдается на частотах до нескольких гигагерц.

4.2.4 Фликер-шум

Он называется так же шумом мерцания, избыточным шумом или шумом "1/f". Это шум, плотность мощности которого, ниже некоторой граничной частоты, растет с уменьшением частоты. Фликер - шум имеет наиболее существенное значение в области низких (как правило < 10 Гц), частот. Впервые он был обнаружен при исследовании спектра флуктуаций тока, протекающего через тонкую пленку, в дальнейшем такая зависимость спектральной плотности от частоты обнаруживалась в самых различных системах, включая не только электронные. Например, подобная зависимость может наблюдаться при подсчете числа автомобилей, проезжающих через перекресток в единицу времени. Спектральная плотность для этого вида флуктуаций выражается формулой вида

$$S = S_0 \frac{1}{f^\alpha}$$

где $\alpha = 0.8...1.4$. В течение длительного времени предпринимались попытки найти некий универсальный механизм, отвечающий за возникновение таких шумов, в настоящее время доказано, что источниками фликер-шума в электронных устройствах являются медленные изменения свойств материалов устройств. Во многих случаях конкретные источники таких изменений могут быть определены. К ним относятся колебания конфигураций дефектов в металлах, изменения заселенности ловушек в полупроводниках, колебания доменных структур в магнитных материалах. Примером физической модели, в рамках которой можно получить зависимость спектральной плотности флуктуаций от частоты вида $1/f$ может быть использованная выше двухуровневая модель, в которой существует не один, а много механизмов перехода между уровнями, каждому из которых соответствует свое время жизни τ. Если распределение числа носителей заряда по временам жизни описывается некоторой функцией $D(\tau)$ в диапазоне $\tau_1 \leq \tau \leq \tau_2$, то суммарная спектральная плотность

$$S(\omega) \sim \int_{\tau_1}^{\tau_2} \frac{\tau}{\omega^2\tau^2 + 1} D(\tau) d\tau$$

Если $D(\tau) \sim \tau^{-1}$, то

$$S(\omega) \sim \omega^{-1}$$

в диапазоне частот $\tau_2^{-1} \ll \omega \ll \tau_1^{-1}$.

4.3 Шумы в усилителях сигналов

Любой усилитель добавляет к усиливаемому сигналу свои собственные шумы. Их источниками могут быть тепловые, дробовые, генерационно-рекомбинационные, фликерные и иные флуктуации в элементах усилителя. Шум на выходе усилителя создается шумами на входе усилителя и шумами в самом усилителе. Можно заменить источники шума внутри усилителя эквивалентными источниками шумового тока и напряжения на его входе: генератора тока $I_{фл}$ параллельного входному сопротивлению $R_{вх}$ рис. 4.6 и источника напряжения $U_{фл}$, подключенного ко входу

как показано на том же рисунке. При этом сам усилитель считают не шумящим и имеющим бесконечное входное сопротивление (идеальным).

Напряжение на входе идеальной части усилителя, создаваемое источником $U_{фл}$, не зависит от сопротивления во входной цепи. Напряжение, создаваемое током $I_{фл}$ зависит от сопротивлений $R_{вх}$ и сопротивления источника сигнала R_g. Отсюда, в частности, следует, что, в общем случае, невозможно заменить два эквивалентных источника каким-либо одним. Источник сигнала может быть одновременно и источником шумов. На рис. 4.6 он представлен эквивалентной схемой, состоящей из I_g и R_g.

Поставим следующий мысленный эксперимент: подключим к нашему усилителю сопротивление R_g и будем постепенно нагревать его, начиная с нулевой температуры. При некоторой температуре $T = T_g$ спектральная плотность мощности шумов на выходе усилителя увеличится вдвое по сравнению с $T = 0$. Это означает, что тепловые шумы, создаваемые этим сопротивлением и собственные шумы усилителя равны. Повторяя этот эксперимент с различными сопротивлениями, найдем такую величину $R_g = R_{g,opt}$, при которой T_g будет минимальной. Такое сопротивление называют **согласованным с усилителем по шумам**, а температуру $T^* = T_{g,min}$ - **эффективной шумовой температурой усилителя**. Пояснить это можно следующим образом: если усилитель имеет эффективную шумовую температуру, равную 40 K, то, при подключении к его входу согласованного по шумам сопротивления, находящегося при температуре 300 K, шумы на выходе усилителя будут такими же, как если бы он был идеальным (не шумящим), а сопротивление находилось при физической температуре 340 K. На этом основан широко используемый на практике способ определения эффективной шумовой температуры усилителей: ко входу усилителя подключается согласованный источник шумов, спектральная плотность которых может быть легко определена. Это может быть сопротивление, которое нагревается или охлаждается, либо ламповый диод, работающий в режиме насыщения. Шумы источника соответствуют его температуре (в первом случае это физическая температура, во втором - рассчитанная по формуле $T_g = eI_0/2k$). Измеряют мощность шумов на выходе усилителя при T_{g1}, а, затем, повышают температуру источника до T_{g2} так, что бы мощность шумов на выходе увеличилась в 2 раза. Тогда, как легко показать, эффективная шумовая

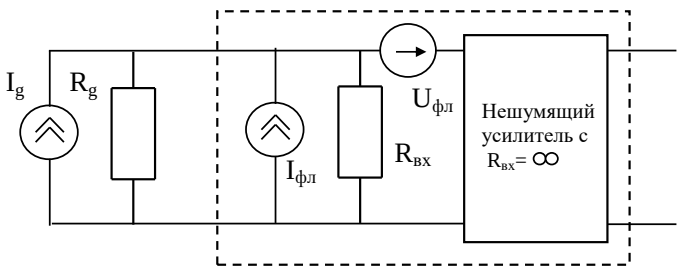


Рис. 4.6: Эквивалентная шумовая схема усилителя

температура усилителя:

$$T^* = T_{g2} - 2T_{g1}.$$

Применительно к эквивалентной схеме на рис.4.6 можно рассчитать, что согласованное по шумам сопротивление источника равно:

$$R_{g,opt} = \frac{R_{BX}}{\sqrt{1 + R_{BX}^2 S_J/S_u}}, \tag{4.37}$$

при этом

$$T^* = \frac{S_u}{kR_{BX}}\left(1 + \sqrt{1 + \frac{R_{BX}^2 S_J}{S_u}}\right). \tag{4.38}$$

Возможность выделения полезной информации из сигнала зависит от отношения его мощности к мощности шумов в полосе частот, которую этот сигнал занимает W_s/W_n. Часто его называют "отношение сигнал/шум" и обозначают s/n (signal-to-noise). Для того, чтобы на выходе усилителя это отношение было максимальным, источник сигнала должен быть согласован с усилителем по шумам.

Величина T^* зависит от отношения $R_{BX}^2 S_J/S_u$. В усилителях с большим входным сопротивлением оно может быть много больше единицы. В этом случае согласно (4.37) и (4.38)

$$R_{BX}^2 S_J/S_u \gg 1, \quad \Rightarrow \quad R_g = \sqrt{S_u/S_J}, \qquad T^* = \frac{1}{k}\sqrt{S_u S_J}. \tag{4.39}$$

На практике в цепях, предназначенных для обработки слабых сигналов чаще применяют усилители с малым входным сопротивлением: $R_{BX}^2 \ll S_u/S_J$. В этом случае:

$$R_{BX}^2 S_J/S_u \ll 1, \quad \Rightarrow \quad T^* = \frac{S_u}{kR_{BX}}, \tag{4.40}$$

а условие согласования по шумам примет простой вид:

$$R_g = R_{BX}. \tag{4.41}$$

Этой формулой удобно пользоваться, поскольку не нужно знать параметры эквивалентных источников, кроме того, если в качестве источника сигнала используется длинная линия с волновым сопротивлением $\rho = R_g$, то это условие в точности совпадает с условием оптимальной (без отражения) передачи сигнала из линии в усилитель.

Квантовый предел шумовой температуры. Согласно квантовой теории произведение спектральных плотностей

$$S_J(\omega)S_u(\omega) \geq \frac{(\hbar\omega)^2}{4}.$$

Отсюда и из (4.39) следует

$$T^* \geq \frac{\hbar\omega}{2k}.$$

Коэффициентом шума усилителя называется отношение

$$F = \frac{(s/n)_{вх}}{(s/n)_{вых}}.$$

здесь под шумом на входе понимается тепловой шум параллельно включенных сопротивлений генератора и входа усилителя, а под отношением сигнал/шум подразумевается отношение средней (за время действия) мощности сигнала к средней мощности шума в эффективной полосе частот сигнала Δf.

Коэффициент шума реального усилителя всегда больше единицы. Часто его выражают в децибелах:

$$F_{dB} = 10 \log F.$$

Применительно к схеме рис.4.6 это соотношение можно представить в виде

$$F = \frac{S_U + (S_J + S_{I_g})R_g^2}{S_{I_g} R_g^2}. \tag{4.42}$$

Шумовая температура и коэффициент шума связаны соотношением

$$T^* = T_0(F - 1), \tag{4.43}$$

здесь T_0 — физическая температура усилителя.

Коэффициент шума двух последовательно включенных усилителей при оптимальных условиях равен

$$F = F_1 + \frac{F_2 - 1}{|K_1(\omega)|^2}.$$

Здесь F_1, F_2 — коэффициенты шума первого и второго усилителей соответственно, $K_1(\omega)$ — коэффициент усиления первого усилителя. Влияние шума второго усилителя на общий коэффициент шума в $|K_1(\omega)|^2$ раз меньше, чем влияние шума первого. Таким образом, наибольший вклад в шумы любого приемного устройства дает его первый каскад. Именно он определяет предельную чувствительность, поскольку суммарный коэффициент усиления можно получить практически любой за счет увеличения числа каскадов. Современные малошумящие усилители на полевых транзисторах позволяют получить $F < 0.5dB$ на частотах до нескольких десятков гигагерц. Это соответствует ухудшению отношения сигнал/шум на величину порядка 10%.

Глава 5

Цифровые системы

5.1 Элементы теории информации

5.1.1 Аналоговый, дискретный и цифровой сигнал

При анализе и обработке сигналов их разделяют по следующим типам: аналоговый, дискретный и цифровой. Аналоговый сигнал является непрерывной функцией непрерывного аргумента. Как правило это функция времени, принимающая бесконечное множество мгновенных значений. Радиофизика “доцифровой эры” имела дело в основном с аналоговыми сигналами.

Дискретный (discrete) сигнал по своим значениям также является непрерывной функцией, но определенной только для дискретных значений аргумента. Это означает, что сигнал задается дискретной последовательностью отсчетов (samples) $u(n\Delta)$, где Δ — временной интервал между отсчетами (интервал или шаг дискретизации), $n = 0, 1, 2, ..., N$. Величина, обратная шагу дискретизации $f = 1/\Delta$, называется частотой дискретизации. Если сигнал получен дискретизацией аналогового сигнала, то он представляет собой последовательность отсчетов, значения которых в точности равны мгновенным значениям исходного сигнала в моменты времени $n\Delta$. Пример дискретизации аналогового сигнала (рис. 5.1а) представлен на рис. 5.1б. При равномерной дискретизации (Δ — постоянная) дискретный сигнал можно описывать сокращенным обозначением u_n или $u[n]$. При неравномерной дискретизации сигнала обозначения дискретных последовательностей обычно заключаются в фигурные скобки — $\{u(t_i)\}$, а значения отсчетов приводятся в виде таблиц с указанием значений координат t_i.

Цифровой (digital) сигнал дискретен как по своим значениям, так и по аргументу. Набор возможных значений задается заранее, как показано на рис. 5.1в. Такое округление принято называть квантованием сигнала по уровню (не путать с процедурой квантования в квантовой механике). Цифровой сигнал, как правило, представляет собой дискретный ряд $u_{k,n} = f_k(n\Delta)$ при $\Delta = \text{const}$. В общем случае интервалы квантования по амплитуде и времени могут быть как с равномерным распределением, так и с неравномерным, например, логарифмическим, а значения сигнала могут быть заданы в виде таблицы для произвольных значений аргумента.

Большинство сигналов, с которыми имеет дело радиофизика, являются аналоговыми по своей природе. Их можно передавать и обрабатывать в аналоговом виде,

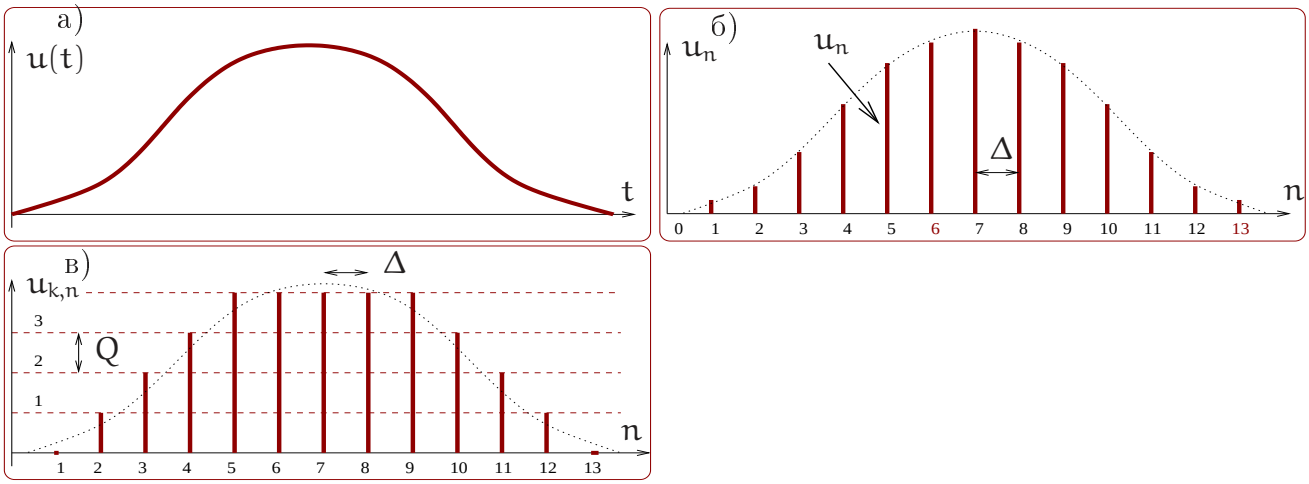


Рис. 5.1: а) Аналоговый сигнал. б) Дискретный сигнал — набор непрерывных отсчетов функции в дискретные моменты времени. в) Цифровой сигнал — набор дискретных отсчетов функции в дискретные моменты времени (в)

однако в последние десятилетия аналоговые сигналы при регистрации или измерении практически всегда преобразовываются в цифровые сигналы. Это связано с тем, что обработка цифровых сигналов значительно эффективнее, чем обработка аналоговых. Возникающие при таком преобразовании ошибки округления называются шумами (или ошибками) квантования. Следует отметить, что существуют и сигналы, которые изначально относятся к классу цифровых, например, отсчеты детекторов, регистрирующих отдельные оптические кванты или заряженные частицы.

5.1.2 Теорема Котельникова

Естественно возникает вопрос: как правильно задать интервал дискретизации Δ при дискретизации аналогового сигнала? Правильно в том смысле, что этот интервал должен быть не слишком мал (чтобы не увеличивать чрезмерно число отсчетов), но и не слишком велик, иначе дискретный сигнал будет сильно отличаться от исходного аналогового. Проблему выбора интервала дискретизации Δ решает теорема Котельникова (называемая также *теоремой отсчетов*): Если спектр сигнала $u(t)$ *ограничен* и верхняя частота спектра сигнала *меньше* $f_c = \frac{1}{2\Delta}$, то по дискретному набору $u_n = u(n\Delta)$ можно точно восстановить исходный сигнал по формуле:

$$u(t) = \sum_{n=-\infty}^{\infty} u_n \frac{\sin\left[2\pi f_c(t - n\Delta)\right]}{2\pi f_c(t - n\Delta)}, \tag{5.1}$$

f_c — частота Найквиста, размерность f_c — Гц.

Доказательство: пусть

$$H(\omega) = \int_{-\infty}^{\infty} h(t)\, e^{-2\pi i\omega t} dt, \tag{5.2}$$

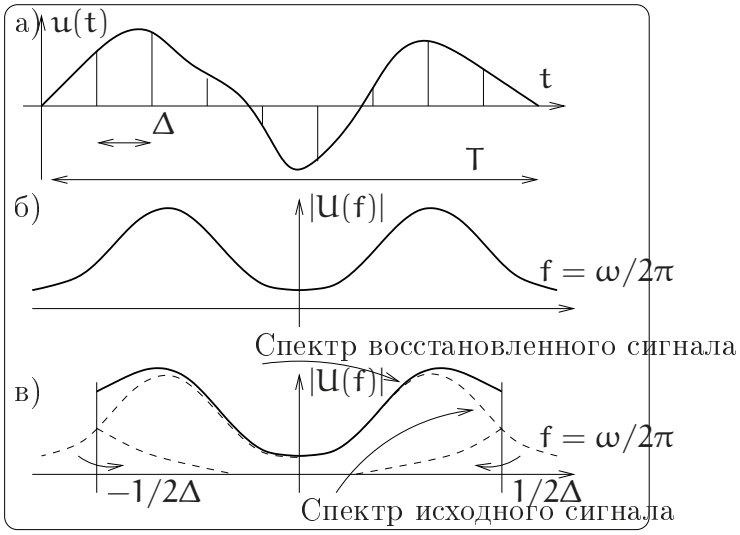


Рис. 5.2: а) Аналоговый сигнал $u(t)$, определенный на конечном интервале времени T, и его отсчеты. б) Фурье преобразование (спектр) аналогового сигнала, не ограниченное по частоте. в) Фурье преобразование (спектр) сигнала, восстановленного из дискретного по формуле (5.1).

$$h(t) = \int_{-\infty}^{\infty} H(\omega)\, e^{2\pi i\omega t} d\omega. \tag{5.3}$$

Введем функцию:

$$G(\omega) = \begin{cases} H(\omega) \text{ если } |\omega| < \Omega, \\ 0 \text{ если } \Omega \le |\omega| \end{cases} . \tag{5.4}$$

и продолжим ее периодически с периодом $1/T$ на интервал $-\infty \ldots \infty$. Тогда формально $G(\omega)$, как периодическую функцию, можно разложить в *ряд* Фурье:

$$G(\omega) = \sum C_n e^{-\pi i n\omega 2T}, \tag{5.5}$$

где

$$C_n = T \int_{-\frac{1}{2T}}^{\frac{1}{2T}} G(\omega)\, e^{\pi i n\omega 2T} d\omega. \tag{5.6}$$

Очевидно, коэффициенты в разложении $C_n = Th(nT)$.

Тогда

$$h(t) = \int_{-\infty}^{\infty} H(\omega)\, e^{2\pi i\omega t} d\omega = \int_{-\Omega}^{\Omega} G(\omega)\, e^{2\pi i\omega t} d\omega = \int_{-\Omega}^{\Omega} \sum C_n e^{-\pi i n\omega 2T} e^{2\pi i\omega t} d\omega =$$

$$= \sum C_n \int_{-\Omega}^{\Omega} e^{-2\pi i\omega(t-nT)} d\omega = \sum C_n \frac{\sin(2\pi\Omega(t-nT))}{\pi(t-nT)} = \tag{5.7}$$

$$= 2T\Omega \sum h(nT) \frac{\sin(2\pi\Omega(t-nT))}{\pi\Omega(t-nT)} = \sum h_n \frac{\sin(2\pi\Omega(t-nT))}{\pi\Omega(t-nT)}. \tag{5.8}$$

Заметим, что ограничение частоты спектра сигнала формально является довольно сильным ограничением. Например, функция, ограниченная по времени интервалом T теоретически имеет бесконечный спектр. На практике же можно выбрать "наивысшую" частоту спектра f_c так, чтобы "хвосты" спектра (содержащие частоты выше f_c) содержали достаточно малую долю энергии сигнала. Но надо помнить, что в этом случае спектр сигнала восстановленного по формуле Котельникова (5.1) отличается от спектра исходного (аналогового) сигнала. А именно, спектральные компоненты вне полосы $-f_c < f < f_c$ не отсекаются, а переносятся внутрь рабочей полосы — это свойство процедуры дискретизации сигнала (в англоязычной литературе называется aliasing). Рис. 5.2 иллюстрирует это.

Условие ограниченности полосы частот связано с энергетическими соображениями при передаче сигнала. В реальных системах это обычно реализуется введением селективных устройств — фильтров нижних частот, речь о которых пойдет в следующем разделе.

В качестве примера рассмотрим *формально неограниченный* спектр прямоугольного сигнала амплитуды u_0 и длительности τ, рассмотренный в разделе 1.2.1 (см. формулу (1.26)):

$$U(\omega) = u_0 \tau \operatorname{sinc} \frac{\omega\tau}{2}. \tag{5.9}$$

Рассмотрим несколько вариантов выбора высшей частоты f_c. Мы уже говорили, что внутри интервала частот от нуля до первого нуля функции $U(\omega)$ сосредоточено около 90% энергии сигнала. Это соответствует частоте $\omega_c = 2\pi/\tau$ или $f_c = 1/\tau$. Тогда по теореме Котельникова мы определяем интервал дискретизации $\Delta = \tau/2$. Таким образом мы получили всего 3 (!) отсчета ($t = -\tau/2,\ 0,\ \tau/2$). В качестве другого варианта выберем граничную частоту $\omega_c = 8\pi/\tau$ или $f_c = 8/\tau$ (это соответствует седьмому нулю Фурье образа $U(\omega)$). При этом интервал дискретизации равен $\Delta = \tau/8$ (9 отсчетов). На рис. 5.3 приведены графики восстановленного по теореме Котельникова сигнала для этих двух случаев. Восстановленные сигналы значимо отличаются от исходного прямоугольного сигнала.

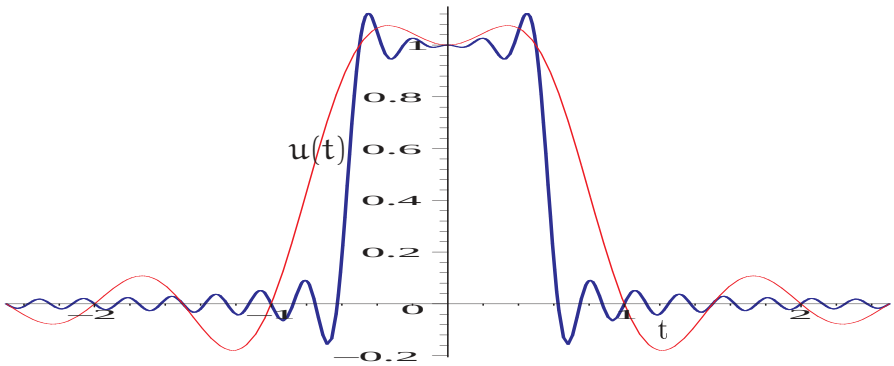


Рис. 5.3: Прямоугольный сигнал единичной амплитуды и длительности ($u_0 = 1$ В, $\tau = 1$ с) был дискретизирован с двумя различными интервалами дискретизации Δ. Приведены графики восстановленных по дискретным отсчетам по формуле (5.1) для $\Delta = \tau/2$ (тонкая линия) и $\Delta = \tau/8$ (толстая линия).

В западной литературе теорема Котельникова носит название теоремы отсчетов (sampling theorem). В математике она была известна до Котельникова (отсюда и

частота f_c "имени" Найквиста), однако в методы радиофизического анализа ее внес именно Котельников.

5.1.3 Дискретное преобразование Фурье

Перейдем теперь к рассмотрению Фурье преобразования дискретного сигнала. Рассмотрим функцию $u(t)$, которая задана конечным числом N своих отсчетов с интервалом дискретизации Δ:

$$u_k \equiv u(t_k), \quad t_k \equiv k\Delta, \quad k = 0, 1, 2 \ldots N-1\,. \tag{5.10}$$

Для простоты будем считать, что N четно. Если функция $u(t)$ определена на конечном интервале T, то считаем, что этот интервал содержит все N отсчетов. Если же $u(t)$ не ограничена по времени, то полагаем, что отсчеты покрывают область, где функция достаточно велика (по сравнению с отбрасываемыми "хвостами").

Если входной сигнал полностью определен N числами, логично предположить, что его Фурье образ будет содержать всю информацию о нем, если в этом образе так же N чисел. Поэтому не будем строить преобразование Фурье $U(f)$ для всех частот[1] f, лежащих в интервале $-f_c < f < f_c$, а ограничимся только набором частот

$$f_n \equiv \frac{n}{N\Delta}, \quad n = -\frac{N}{2} \ldots \frac{N}{2}, \qquad \left(\omega_n \equiv \frac{2\pi n}{N\Delta}\right). \tag{5.11}$$

Крайние значения n в (5.11) точно соответствуют частотам Найквиста. Внимательный читатель заметит, что в (5.11) n пробегает $N+1$, а не N значений. Оказывается, что для двух крайних значений n при дискретном преобразовании Фурье мы получим зависимые значения (они просто равны), тогда как остальные оказываются независимыми. Это уменьшает число независимых отсчетов до N. Заметим, что введение дискретного набора частот эквивалентно замене сигнала (5.10) периодической последовательностью таких сигналов с периодом $N\Delta$.

Теперь преобразуем интеграл Фурье в сумму дискретных значений:

$$U(f_n) = \int_{-\infty}^{\infty} u(t)\, e^{2\pi i f_n t}\, dt \simeq \sum_{k=0}^{N-1} u_k\, e^{2\pi i f_n t_k}\, \Delta \tag{5.12}$$

Далее, учтя формулы (5.10, 5.11), можно записать

$$U(f_n) = \Delta \sum_{k=0}^{N-1} u_k\, e^{2\pi i\, kn/N} \tag{5.13}$$

и ввести обозначение U_n:

$$U_n = \sum_{k=0}^{N-1} u_k\, e^{2\pi i\, kn/N}\,. \tag{5.14}$$

[1] В дискретном преобразовании Фурье традиционно принято использовать частоты f, измеряемые в герцах. Они соотносятся с угловыми частотами ω (они измеряются в рад/сек) по известной формуле $f = \omega/2\pi$.

Преобразование (5.14) называют *дискретным преобразованием Фурье* для N отсчетов u_k. При таком обозначении дискретное преобразование Фурье преобразует N чисел u_k (в общем случае они могут быть и комплексными) в N комплексных чисел U_n. Важно, что преобразование (5.14) не зависит от каких-либо размерных параметров, например, от интервала дискретизации Δ. Связь между обычным и дискретным преобразованием Фурье следует из (5.13, 5.14):

$$U(f_n) \simeq U_n \Delta, \tag{5.15}$$

где f_n задается (5.11).

До сих пор мы считали, что индекс n в (5.14) пробегает значения от $-N/2$ до $N/2$. Однако нетрудно заметить, что преобразование (5.14) является периодическим по n с периодом N: $U_{-n} = U_{N-n}$. Учитывая это обстоятельство, можно положить, что индекс n в U_n пробегает значения от 0 до $N-1$ (это полный период). В этом случае оба индекса k и n изменяются в одинаковых пределах (от 0 до $N-1$). При этом надо помнить, что нулевая частота соответствует $n = 0$, положительные частоты $0 < f < f_c$ соответствуют значениям $1 \le n \le N/2-1$, а отрицательные частоты $-f_c < f < 0$ соответствуют $N/2+1 \le n \le N-1$. Значение $n = N/2$ соответствует одновременно и $f = f_c$, и $f = -f_c$.

Обратное дискретное Фурье преобразование (из набора U_n получить набор u_k) задается соотношением

$$u_k = \frac{1}{N} \sum_{n=0}^{N-1} U_n \, e^{-2\pi i \, kn/N}. \tag{5.16}$$

Формулы (5.14 и 5.16) отличаются друг от друга только знаком в показателе экспоненты и множителем $1/N$. Это означает, что численные процедуры для прямого ДПФ могут быть легко модифицированы и для обратного ДПФ.

В целом свойства ДПФ аналогичны свойствам непрерывного преобразования Фурье, однако дискретный характер сигнала привносит некоторую специфику. Не останавливаясь на этом подробно, приведем лишь для справки дискретный аналог равенства Парсеваля (1.38):

$$\sum_{k=0}^{N-1} |u_k|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |U_n|^2. \tag{5.17}$$

Быстрое преобразование Фурье (БПФ)

Практически всегда при обработке сигнала требуется выполнить преобразование Фурье. При этом весьма актуально сокращение объема вычислительных операций. Процедура быстрого преобразования Фурье позволяет сократить количество вычислений, необходимое для дискретного преобразования Фурье. Перепишем формулу (5.14) для дискретного преобразования Фурье в виде:

$$U_n = \sum_{k=0}^{N-1} W^{nk} u_k, \quad W \equiv e^{2\pi i/N} \, . \tag{5.18}$$

Это эквивалентно тому, что матрица, чей элемент (n,k) есть комплексная константа W, возведенная в степень $n \times k$, умножается на вектор u_k. Очевидно, что для вычисления одного элемента U_n потребуется N операций комплексного умножения, а для вычисления всех элементов U_n — N^2 операций (плюс еще меньшее количество операций для генерации коэффициентов W^{nk}). Таким образом, для реализации ДПФ требуется $\mathcal{O}(N^2)$ операций. Алгоритм быстрого преобразования Фурье (БПФ) выгодно отличается от ДПФ тем, что для решения той же задачи ему требуется всего лишь $\mathcal{O}(N \log_2 N)$ операций. Разница между $\mathcal{O}(N^2)$ и $\mathcal{O}(N \log_2 N)$ огромна, например, при $N = 10^6$ БПФ дает выигрыш в $\approx 5 \times 10^4$ раз! Алгоритм БПФ стал широко известен в середине 60-х после работ Кули и Тьки (J.W.Cooley, J.W.Tukey), однако позже выяснилось, что подобные методы были независимо и раньше открыты десятком других исследователей, начиная с Гаусса (1805 год).

Сначала покажем, что дискретное преобразование Фурье длины N может быть записано как сумма двух преобразований длины $N/2$ (естественно, если N четно). Это доказывает лемма Даниельсона и Ланца (Danielson and Lanczos)). Действительно,

$$U_n = \sum_{k=0}^{N-1} u_k\, e^{2\pi i\, kn/N} = \sum_{k=0}^{N/2-1} u_{2k}\, e^{2\pi i\,(2k)n/N} + \sum_{k=0}^{N/2-1} u_{2k+1}\, e^{2\pi i\,(2k+1)n/N} =$$
$$= \sum_{k=0}^{N/2-1} u_{2k}\, e^{2\pi i\, kn/(N/2)} + e^{2\pi i\, n/N} \sum_{k=0}^{N/2-1} u_{2k+1}\, e^{2\pi i\, kn/(N/2)} = U_n^{\text{чет}}(N/2) + W^n U_n^{\text{нечет}}(N/2)\ . \tag{5.19}$$

Здесь $U_n^{\text{чет}}(N/2)$ обозначает дискретное преобразование Фурье длины $N/2$, образованное только четными членами u_k, а $U_n^{\text{нечет}}(N/2)$ соответствует такому же преобразованию, но из нечетных членов u_k. Очевидно, что преобразование Фурье, например, $U_n^{\text{чет}}(N/2)$ дает лишь $N/2$ спектральных коэффициентов, поэтому непосредственно использовать формулу (5.19) можно только при $0 \le n \le (N/2) - 1$. Для остальных n (т.е. $(N/2) - 1 \le n \le N - 1$) следует воспользоваться периодичностью спектра дискретного сигнала: $U_n^{\text{чет}}(N/2) = U_{n+N/2}^{\text{чет}}(N/2)$ (аналогично и для Фурье преобразования нечетных членов $U_n^{\text{нечет}}(N/2) = U_{n+N/2}^{\text{нечет}}(N/2)$).

Оценим количество операций, необходимое для вычисления Фурье преобразования с использованием (5.19). Каждое преобразование размерности $N/2$ требует $N^2/4$ операций. Кроме того, умножение на экспоненциальный множитель W добавляет еще $N/2$ операций. Таким образом получаем $2\,N^2/4 + N/2 = N(N+1)/2$ операций — это почти вдвое меньше, чем при вычислении Фурье преобразования прямым образом, требующим N^2 операций.

Замечательно, что свойство (5.19) можно применять *рекурсивно*. Если вычисление U_n может быть сведено в вычислению $U_n^{\text{чет}}$ и $U_n^{\text{нечет}}$ (длины $N/2$), то в свою очередь, например, $U_n^{\text{чет}}$ может быть сведено к сумме вычислений $U_n^{\text{чет-чет}}$ и $U_n^{\text{чет-нечет}}$ длины $N/4$. Таким же образом можно и дальше продолжать рекурсию:

$$U_n = U_n^{\text{чет}}(N/2) + W^n U_n^{\text{нечет}}(N/2) = \tag{5.20}$$
$$= \left[U_n^{\text{чет-чет}}(N/4) + W^n U_n^{\text{чет-нечет}}(N/4)\right] + W^n \left[U_n^{\text{нечет-чет}}(N/4) + W^n U_n^{\text{нечет-нечет}}(N/4)\right] = \ldots$$

Остановимся на самом простом случае, когда N есть степень 2, т.е. $N = 2^m$ (m — целое). На практике рекомендуют использовать БПФ только с такими N (хотя возможны и другие варианты). Если число $N \neq 2^m$, то его надо увеличить до ближайшей степени двойки, заполнив добавленные позиции нулями. Очевидно, что в этом случае ($N = 2^m$) мы можем продолжать рекурсию по формуле (5.19), уменьшая N вплоть до единицы. Это означает, что, продолжая рекурсию в духе (5.20), в конце концов мы получим члены, содержащие одно-точечное (т.е. длины 1) преобразование:

$$U_n^{\text{нечет-чет-чет-...- нечет-чет-нечет}}(1) = u_k \quad \text{для некоторого индекса } k \tag{5.21}$$

Теперь надо выяснить какой конкретной комбинации (чет) и (нечет) соответствует u_k в выражении (5.21). Кратко ответ можно сформулировать просто — надо произвести обращение (реверсию) битов в комбинации (нечет-чет-чет-...- нечет-чет-нечет) и это число в *двоичной* системе будет равно числу k в (5.21). Для этого надо сначала записать комбинацию (нечет-чет-чет-...- нечет-чет-нечет) в двоичной системе, присвоив значения чет$= 0$, нечет$= 1$. Например, комбинация в (5.21) запишется в виде $(100\ldots101)$. Для обращения (инверсии) надо просто заменить порядок следования нулей и единиц на обратный (т.е. переписать число справа налево), в нашем примере получится число $(101\ldots001)$. Это число и будет равно числу k в (5.21), записанному в двоичной системе.

Дальнейшее почти очевидно. Мы можем выбрать два соответствующих одно-точечных преобразования вида (5.21), образующих 2-точечное преобразование. Таких пар будет $N/2$. Далее собираем из 2-точечных преобразований 4-точечные и так далее, пока не получим две половинки полного преобразования в соответствии с формулой (5.20). Каждая такая комбинация требует N операций, а количество комбинаций есть $\log_2 N$, поэтому весь алгоритм требует порядка $N \log_2 N$ операций (мы считаем, что операция сортировки при обращении битов требует меньшее число операций).

Мы только коснулись методов описания и обработки дискретного сигнала, более детальную информацию можно найти в специальной литературе, например, в [4].

5.1.4 Количество информации

Рассмотрим передачу дискретного сообщения с помощью отдельных импульсов. Пусть амплитуда импульсов может принимать m значений. Число m называют числом градаций. Пусть все сообщение состоит из n импульсов. Тогда полное количество комбинаций элементов равно

$$N = m^n.$$

Нас интересует величина I, которую можно назвать количеством информации. Очевидно, величина N не может быть количеством информации, поскольку интуитивно ясно, что должно выполняться условие аддитивности, т.е. I должно быть пропорционально количеству импульсов: $I \sim n$ (как стоимость телеграммы). С другой стороны ясно, что количество информации должно зависеть от полного

количества комбинаций $I = f(N)$. Выпишем условие аддитивности и дифференциал N:

$$\begin{aligned} df &= K\,dn, \\ dN &= N \ln m\,dn, \end{aligned}$$

где K — постоянная. Мы видим, что условие аддитивности будет выполняться, если функцию $f(N)$ принять в виде:

$$I = f(N) = K_1 \ln N = \log_a N.$$

Осталось определить постоянную a. Чтобы это сделать выберем $m = 2$, $n = 1$ и назовем такое сообщение ("0" или "1") единицей информации или битом. Тогда

$$1 = \log_a(2^1), \quad \Rightarrow \quad a = 2.$$

В результате получаем формулу для количества информации[2]:

$$I = \log_2 N = n \log_2 m. \tag{5.22}$$

5.1.5 Передача информации через канал связи

Посмотрим на передачу сообщения несколько иначе. Пусть сообщение представляет собой функцию $u(t)$, которая имеет спектр, ограниченный частотой F_0. Тогда по теореме Котельникова такое сообщение может быть передано с помощью набора импульсов (отсчетов), разнесенных на время $\Delta t = 1/(2F_0)$. Таким образом за время t будет передано $2F_0 t$ импульсов. Если число градаций каждого импульса равно m, то нетрудно посчитать количество информации $I(t)$, переданное за время t, и скорость R передачи информации

$$\begin{aligned} I &= 2F_0 t \log_2 m, \\ R &= \frac{dI}{dt} = 2F_0 \log_2 m. \end{aligned} \tag{5.23}$$

Из вида приведенных формул сразу следует, что для увеличения скорости передачи следует увеличивать число градаций m. Очевидно, что число градаций не может быть бесконечно из-за наличия шумов. Шеннон в 1948 г. показал, что максимальные значения количества информации I и пропускной способности R канала определяются формулами:

$$I = F_0 t \log_2 \left(1 + \frac{W_s}{W_n}\right), \tag{5.24}$$

$$R = \frac{dI}{dt} = F_0 \log_2 \left(1 + \frac{W_s}{W_n}\right), \tag{5.25}$$

где W_s — мощность сигнала, а W_n — мощность шума. Величину I, определенную в (5.24), называют еще объемом сигнала.

[2] Иногда в теоретических работах используется единица "нат": в этом случае $a = e$ и формула для количества информации I' записывается в виде $I' = n \ln m$. Мы такие экзотические единицы употреблять не будем.

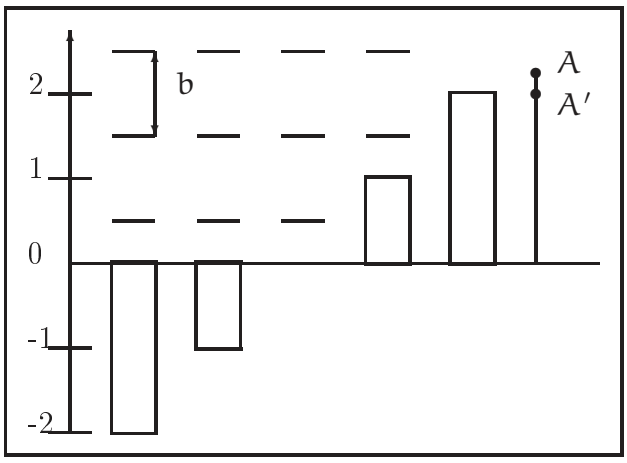


Рис. 5.4: Дискретизация сигнала с шагом b. Отрезок AA' — ошибка квантования.

Естественно, чтобы такой сигнал прошел через канал связи, последний должен обладать соответствующими характеристиками:

- полоса частот F_k, пропускаемых каналом, должна быть достаточно велика: $F_k > F_0$;

- время связи t_k через канал должно быть также достаточно велико: $t_k > t$;

- превышение сигнала над шумом в канале $H_k = \log_2\left(1 + \frac{W_s}{W_n}\right)$ должно быть также больше соответствующей величины H сигнала: $H_k > H$.

Величину $F_k t_k H_k$ называют информационной емкостью канала.

5.1.6 Шумы квантования

Рассмотрим передачу сигнала импульсами, имеющими число градаций m. Пусть обычные шумы отсутствуют. Тогда останутся шумы квантования, связанные с градацией сигнала, т.е. с тем, что передается не “истинная” амплитуда импульса, а амплитуда, кратная шагу градаций. Пусть шаг градации равен b. Если амплитуда сигнала может принимать равновероятные значения в пределах шага b, то заменяя ее дискретным значением, мы допускаем ошибку, среднее которой равно нулю, а дисперсия равна

$$\sigma^2 = \int_{-b/2}^{b/2} \frac{1}{b} x^2\, dx = \frac{b^2}{12}.$$

Здесь принято, что плотность вероятности в пределах шага квантования равна $1/b$.

Если сигнал имеет m градаций, то средний квадрат напряжения сигнального импульса равен:

$$U_s^2 = \frac{m^2 b^2}{12}.$$

Поскольку $U_s^2 = \sigma^2 + U_{qs}^2$, где U_{qs}^2 — средний квадрат квантованного сигнала, то

$$U_{qs}^2 = (m^2 - 1)\frac{b^2}{12} = (m^2 - 1)\sigma^2.$$

Таблица 5.1: Характеристики различных каналов передачи информации

	F_0	m	$\log_2 m$	$R = I/t$
Телеграф	$4 \cdot 10^2$ Гц	2	1	$8 \cdot 10^2$
Телефон	$4 \cdot 10^3$ Гц	128	7	$6 \cdot 10^4$
Телевидение (стандартной четкости, без сжатия)	$6 \cdot 10^6$ Гц	30	~ 5	$6 \cdot 10^7$

Выражаем отсюда m и подставляем в формулы (5.23) для количества и скорости передачи информации:

$$I = F_0 t \log_2 \left(1 + \frac{U_{qs}^2}{\sigma^2}\right), \tag{5.26}$$

$$R = \frac{dI}{dt} = F_0 \log_2 \left(1 + \frac{U_{qs}^2}{\sigma^2}\right). \tag{5.27}$$

Сравнивая эти формулы с формулами (5.24, 5.25), мы видим, что они похожи. Действительно, если выразить квадрат сигнала U_s^2 через мощность сигнала W_s, а квадрат дисперсии ошибки дикретизации — через мощность шума дискретизации W_{qn}, то формально получим совпадение с формулами (5.24, 5.25). Подчеркнем, что сходство это формальное, поскольку в формулах Шеннона имеются в виду физические шумы. Однако, если физические шумы, сопровождающие сигнал или добавляющиеся в канале, будут превосходить шумы квантования, то очевидно, что формулы (5.26, 5.27) должны перейти в (5.24, 5.25).

Подчеркнем, что эти рассуждения не являются выводом (5.24, 5.25), а приведены здесь в качестве иллюстрации.

5.1.7 Различные каналы передачи информации

В таблице 5.1 приведены различные параметры трех традиционных каналов связи. Интересно заметить, что через зрение человек получает $2 \cdot 10^4$ бит/сек. Это много меньше, чем скорость передачи информации по телевидению. Дело в том, что зрение устроено значительно более экономно: в качестве информации записывается не каждый кадр, а лишь *изменение* картинки. При этом мозг должен помнить всю текущую картинку (в оперативной памяти). Аналогичный подход используется в алгоритмах сжатия видеоинформации mpeg и им подобных.

Среди современных каналов информации отметим три: СВЧ кабель, витая пара и оптический волновод.

Полоса частот, передаваемых по СВЧ кабелю составляет $F_0 \simeq 10^{10}$ Гц. Это означает, что по СВЧ кабелю можно передавать около 1000 телевизионных каналов или $2,5 \cdot 10^6$ телефонных каналов.

СВЧ кабель постепенно вытесняется витой парой, которая при сравнимых параметрах значительно дешевле и удобнее особенно при использовании в локальных сетях. Скорость передачи информации по витой паре достигает 1000 Мбит/сек.

Значительно более широкая полоса частот может передаваться по оптическому кабелю: $F_0 \simeq 10^{14}$ Гц. Скорость передачи информации составляет до 40 Тбит/сек (в лабораторных условиях продемонстрировано ~ 200 Тбит/сек). Оптический кабель представляет собой совокупность световодов. Каждый световод - диэлектрический волновод из плавленого кварца, по которому свет распространяется благодаря эффекту полного внутреннего отражения. Толщина сердцевины (имеющей больший показатель преломления) в волноводах, рассчитанных на длину волны $\lambda \sim 1.55$ мкм имеет диаметр около 10 мкм, оболочка — 125 мкм. Затухание составляет $0,2$ дБ/км (интенсивность уменьшается в e раз на расстоянии 30 км). Чтобы компенсировать затухание в оптических кабелях, имеющих длину более нескольких десятков километров, используют промежуточные оптические усилители.

5.1.8 Надежность передачи информации

Если один бит передается за время τ, то полоса частот Δf для передачи сигнала равна $\Delta f \simeq 1/\tau$. При этом мощность тепловых шумов в согласованной линии составляет $W_T = \kappa T \Delta f$ [3]. Это означает, что передача каждого бита с помощью энергии импульса величины $\mathcal{E}$ сопровождается шумами, энергия которых равна κT. Таким образом, минимальная энергия, которую можно использовать для передачи одного бита должна превосходить κT.

Величина $\mathcal{E}/\kappa T$ постоянно уменьшается. Если средняя энергия, рассеиваемая процессором W, тактовая частота ν, а количество элементов N, то очевидно, что

$$\frac{\mathcal{E}}{\kappa T} = \frac{W}{\nu N \kappa T}.$$

Приведем оценки для различных процессоров:

Процессор	W	ν	N	$\frac{\mathcal{E}}{\kappa T}$
8086	1 Вт	5 МГц	$5 \cdot 10^4$	10^9
Pentium 4	100 Вт	2 ГГц	10^8	10^5
Core i7 Extreme edition(6 cores)	130 Вт	3.3 ГГц	$2 \cdot 10^9$	$2.5 \cdot 10^4$

В оптическом диапазоне $\kappa T \ll \hbar\omega$, поэтому основные шумы при передаче по оптическому волноводу являются квантовые шумы, связанные с потерями в волноводе. Если представить себе, что научились делать волноводы без потерь (или расстояние передачи много меньше характерной длины затухания) то основными шумами, сопровождающими передачу информации, будут квантовые шумы,

[3] Здесь мы имеем в виду передачу информации в СВЧ диапазоне, где $\hbar\omega_0 \ll \kappa T$ (в оптике наоборот).

точнее квантовая неопределенность энергии состояния, которое используется для передачи. Можно показать, что предельная величина для передачи одного бита за время τ составляет

$$\mathcal{E} \simeq \frac{\hbar}{\tau}.$$

5.1.9 Хранение информации

Для хранения оперативной информации (т.е. информации, доступ к которой должен быть быстрым, но которая стирается после выключения компьютера) в компьютере используются элементы, работающие по принципу триггера. Для хранения долговременной информации используются оптические и магнитные диски, а так же твердотельные (полупроводниковые) накопители. Современный оптический диск (Blu-ray) имеет емкость до 100 Гигабайт. Для сравнения Большая Советская Энциклопедия из 50 томов содержит $\times$ 500 стр. $\times$ 20 Kb $= 5 \times 10^8$ байт[4]. Плотность записи на Blu-ray по современным представлениям невелика, на один байт расходуется площадка размером $\simeq 4 \times 10^{-4}$ см. На жестком магнитном диске (винчестере) достигнута плотность записи до 200 Гбит на квадратный сантиметр (2014 год), есть перспективы дальнейшего повышения. Однако, все большее распространение получает память, основанная на изменении и регистрации электрического заряда в изолированной области полупроводниковой структуры. Так, изготавливаемый по технологии 40 нм чип Samsung имеет емкость 86 Гбит при площади 0.01 квадратный сантиметр. Такие чипы используются во флэш накопителях и SSD (Solid State Drive, дословно — “твердотельный накопитель”) “дисках”, заменяющих винчестеры.

5.2 Коды

Обсуждая в предыдущем разделе цифровые сигналы мы неявно предполагали, что каждый отсчет - это число, не задаваясь вопросом о том, в каком виде оно представлено. Привычное для нас десятичное представление (с плавающей запятой) не удобно для практической реализации. Основанием для представления информации и цифровых системах обработки и передачи данных обычно является двоичная система (0 и 1, "да” и "нет”). Однако, кодировать информацию с помощью нулей и единиц можно по-разному. На практике, часто каждое число кодируется определенной комбинацией. Соответственно, коды бывают:

- Неизбыточные - каждая комбинация нулей и единиц кодирует число.
- Избыточные – комбинаций больше, чем необходимо (лишние могут быть использованы для обнаружения ошибок).
- Равномерные (блоковые) - комбинации содержат постоянное число разрядов.
- Неравномерные (пример – азбука Морзе: буквы кодируются различным числом точек и тире).

[4] Это *без* картинок. Каждая картинка содержит информацию порядка $\sim 10^6$ бит

- Взвешенные - каждый разряд имеет вес, например n – степень 2 – натуральный двоичный код: $5_{10} = 1 \times 2^0 + 0 \times 2^1 + 1 \times 2^2 = 101_2$.

- Невзвешенные.

Можно привести примеры кодов, использующихся в цифровых системах хранения, обработки и передачи информации.

- Код 8421 : каждый десятичный знак заменяется на 4 двоичных: $N_{10} = 8a_3 + 4a_2 + 2a_1 + 1a_0$. (избыточный блоковый код)

- Натуральный двоичный код (неизбыточный, непрерывный.) Используются 8-и, 16-и, 32-ух... разрядные блоки.

- Код Грея – получается суммированием по модулю 2 соседних разрядов натурального двоичного. кода. Достоинство: коды соседних чисел отличаются только одним разрядом (циклический неизбыточный код).

- Код Джонсона – получается последовательным сдвигом блока единиц (избыточный) – легко формируется и дешифрируется.

- Код «1 из n» - только одна единица в кодовой комбинации. Очень простой, очень избыточный блоковый код.

Число	Натуральный двоичный код	Код Грея	Код Джонсона	Код «1 из 8»
0	0000	0000	0000	00000001
1	0001	0001	0001	00000010
2	0010	0011	0011	00000100
3	0011	0010	0111	00001000
4	0100	0110	1111	00010000
5	0101	0111	1110	00100000
6	0110	0101	1100	01000000
7	0111	0100	1000	10000000

Рис. 5.5: Пример различных двоичных кодов.

Использование специальных кодов позволяет бороться с ошибками, возникающими при передаче цифровых сигналов. Такие ошибки могут возникать, в том числе, под воздействием шумов. Простой способ - разбиение потока информации на блоки и вычисление некоторого контрольного числа для каждого блока (это может быть например, простая арифметическая сумма всех единиц в блоке, которая затем кодируется и передается в конце блока) позволяет обнаруживать искажение информации. Искаженные блоки можно затем передать повторно, но это не всегда удобно, например, при передаче речи и изображения в реальном времени. Более сложные коды позволяют не только обнаруживать, но и исправлять ошибки. На

рис. 5.6 показан пример такого кода. К трем информационным битам b1-b3 надо добавить три контрольных, так, что бы в каждой окружности сумма была четной, а затем – еще один, что бы четной была сумма всех. Такой код исправляет ошибки кратности 1 и обнаруживает двукратные. Платой за это является более чем двукратное увеличение объема передаваемых данных.

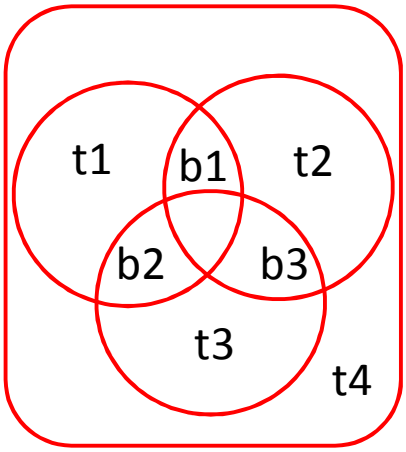


Рис. 5.6: Пример построения кода, исправляющего ошибки.

Математический анализ позволяет создавать оптимальные коды. Так, код Рида - Соломона, исправляющий t ошибок, требует 2t проверочных символов и с его помощью исправляются произвольные пакеты ошибок длиной t и меньше. Впервые он использовался при записи CD дисков (избыточность - 25%, корректирующая способность - 87 Мб из 700), алгоритмы на его основе применяются в штрих-кодах, мобильной связи.

5.3 Основы Булевой алгебры

Описание цифровых систем передачи и обработки информации удобно производить с помощью математического аппарата, оперирующего с *двоичными* переменными. Такая переменная (также называемая логической) может принимать всего два возможных значения: $x = 0$ ("ложь "false") и $x = 1$ ("истина "true"). Основными операциями над двоичными переменными являются логическое НЕ (NOT, отрицание, инверсия): если $x = 0$ то $\overline{x} = 1$, если $x = 1$ то $\overline{x} = 0$, а также операция ИЛИ (дизьюнкция, OR, $\bigvee$, $+$), операция И (коньюнкция, AND, $\bigwedge$, $\cdot$) а так же операция ИСКЛЮЧАЮЩЕЕ ИЛИ (XOR, $\forall$, $\bigotimes$). Комбинируя эти операции, можно осуществить любое преобразование двоичных переменных, или другими словами, определить функцию двоичных переменных. Очевидно, такая функция также может принимать только два значения, 0 и 1. Для того, что бы задать такую функцию, необходимо указать, при каких комбинациях переменных $F(A, B, C \ldots)$ равна 0, а при каких - 1. Функция, заданная для всех возможных комбинаций значений переменных, называется полностью определенной, заданная только для части комбинаций - недоопределенной (или факультативной). Задать функцию можно путем

словесного описания (перечисления), с помощью таблицы, называемой *таблицей истинности*, либо алгебраически. На рис. 5.7 приведены таблицы истинности для функций И и ИЛИ.

x	y	$x+y$
0	0	0
0	1	1
1	0	1
1	1	1

x	y	$x \cdot y$
0	0	0
0	1	0
1	0	0
1	1	1

Рис. 5.7: Таблицы истинности для операций: слева - логическое ИЛИ, справа - логическое И.

На рисунке 5.8 приведена таблично заданная факультативная функция трех переменных F. Алгебраически можно задать ту же функцию либо в *нормальной дизъюнктивной* форме, выбрав все комбинации переменных, при которых эта функция

N	x	y	z	F
1	0	0	0	0
2	0	0	1	1
3	0	1	0	0
4	0	1	1	1
5	1	0	0	1
6	1	0	1	?
7	1	1	0	1
8	1	1	1	0

Рис. 5.8: Пример таблично заданной двоичной функции.

принимает значение 1 и объединив их операцией логическое ИЛИ:

$$F = \overline{x} \cdot \overline{y} \cdot z + \overline{x} \cdot y \cdot z + x \cdot \overline{y} \cdot \overline{z} + x \cdot y \cdot \overline{z}. \tag{5.28}$$

Альтернативным способом является задание в *нормальной конъюнктивной* форме: перебираем все комбинации переменных, при которых функция равна 0 (здесь надо использовать операцию ИЛИ) и объединяем их операцией И:

$$F = (x + y + z) \cdot (x + \overline{y} + z) \cdot (\overline{x} + \overline{y} + \overline{z}). \tag{5.29}$$

Порядок выполнения логических операций такой же, как в алгебре, только инверсия над одной переменной всегда выполняется первой а над алгебраическим выражением - последней.

Раздел математики, посвященный двоичным функциям, называется Булевой алгеброй. Приведем основные соотношения для случая одной переменной:

$$x + 0 = x$$
$$x + x = x$$
$$x + 1 = 1$$
$$x + \overline{x} = 1$$
$$x \cdot 0 = 0$$
$$x \cdot x = x$$
$$x \cdot 1 = x$$
$$x \cdot \overline{x} = 0$$
$$x + 0 = x$$
$$\overline{\overline{x}} = x.$$

Основными теоремами Булевой алгебры являются следующие соотношения (знак $\cdot$, также, как в обычной алгебре, опускается):

1. Переместительный закон:
$$x + y = y + x$$
$$xy = yx.$$

2. Сочетательный закон:
$$x + y + z = x + (y + z) = (x + y) + z$$
$$xyz = x(yz) = (xy)z.$$

3. Распределительный закон:
$$x(y + z) = xy + xz$$
$$xy + z = (x + z)(y + z).$$

4. Закон поглощения:
$$x + xy = x$$
$$x(x + y) = x.$$

5.
$$(x + \overline{y})y = xy$$
$$x\overline{y} + y = x + y.$$

6. Закон склеивания:
$$xy + \overline{x}y = y$$
$$(x + y) + (\overline{x} + y) = y.$$

Все они достаточно просто проверяются подстановкой.

Важным утверждением является теорема Де Моргана, которая позволяет переходить от одной логической операции к другой:

$$\overline{x + y + z + \ldots} = \overline{x} \cdot \overline{y} \cdot \overline{z} \cdot \ldots \tag{5.30}$$

$$\overline{xyz + \ldots} = \overline{x} + \overline{y} + \overline{z} + \ldots . \tag{5.31}$$

5.4 Основные логические элементы

В электронных устройствах для реализации двоичных функций используются *логические элементы*. Фактически, транзисторный усилитель, имеющий достаточно большой коэффициент усиления, можно условно считать логическим элементом НЕ, поскольку при положительном входном сигнале, превышающем некоторое значение, на выходе будет напряжение, близкое к нулю, а при нулевом - близкое к напряжению питания. На практике используются как простые микросхемы, содержащие отдельные логические элементы (обычно - по несколько элементов в одном корпусе), так и сложные, функции которых можно программировать - процессоры. На рис. 5.9 приведены обозначения основных логических элементов на электрических схемах.

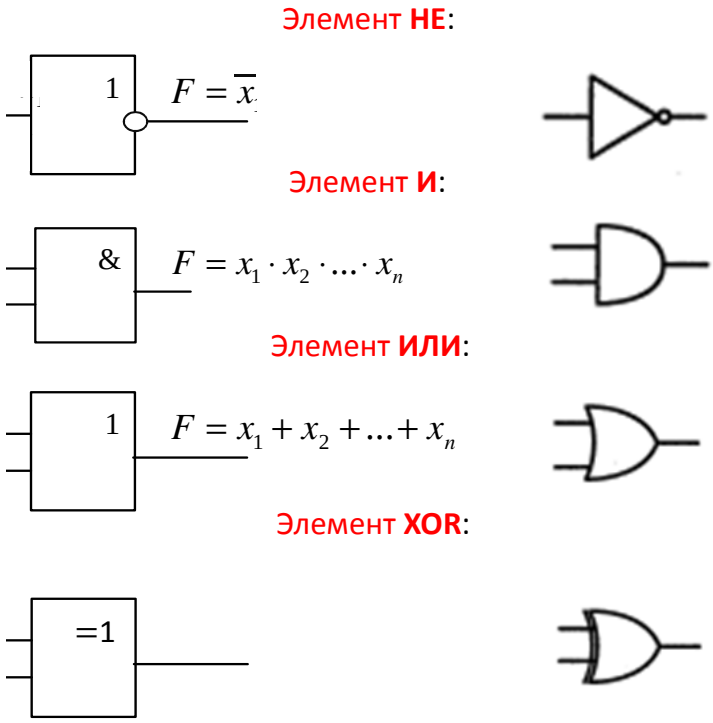


Рис. 5.9: Обозначения основных логических элементов на электрических схемах: слева - принятые в отечественной литературе, справа - в зарубежной.

Отметим, что не обязательно реализовывать все основные элементы: достаточно иметь в распоряжении *функционально-полную систему* - набор логических элементов, позволяющий реализовать любую Булеву функцию. Например, элементы ИЛИ и НЕ образуют функционально-полную систему. Возможный простейший вариант реализации таких элементов показан на рисунке 5.10. В реальных микросхемах цепи сложнее, но при оптимизации принято считать, что один выход логического элемента И или ИЛИ эквивалентен одному диоду, а операция НЕ эквивалентна одному транзистору.

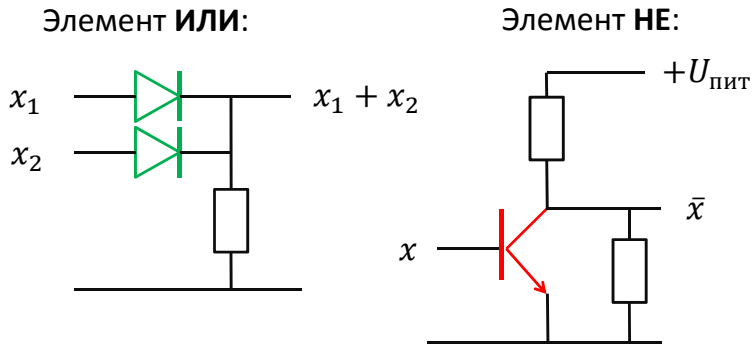


Рис. 5.10: Возможный вариант реализации логических элементов ИЛИ и НЕ.

Используя простейшие элементы можно конструировать более сложные. На рис.5.11 приведена возможная реализация RS триггера. У него два входа: установка S[et] и сброс R[eset]. Работает он следующим образом: При R=0 Подача 1 на вход S Устанавливает 1 на выходе (обозначается Q), а после перехода S в 0 выход остается Q = 1. После этого подача 1 на вход R устанавливает 0 на выходе Q. Одновременное S = 1 и R = 1 не допустимо (Q будет не определен). Таким образом, состояние триггера зависит не только от сигналов на его входах в данный момент времени, но и от предыстории (описать него работу с помощью одной только таблицы истинности не получится). Триггер является элементарной ячейкой памяти.

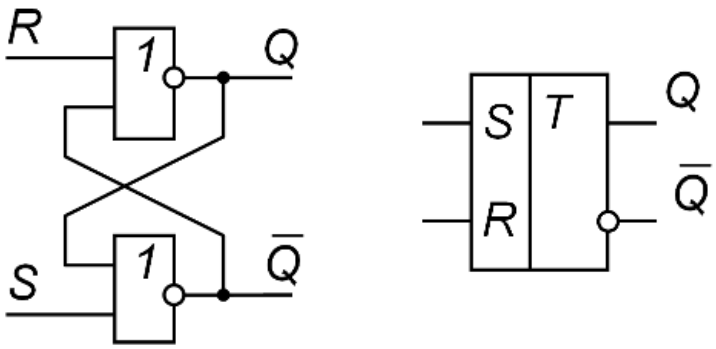


Рис. 5.11: RS триггер. Слева: реализация на элементах ИЛИ-НЕ, справа: обозначение на схемах.

Более удобным для многих задач является так называемый D триггер («защелка»). Он имеет два входа: вход данных D[ata] и счетный вход C[ount]. Работает он так: в момент прихода 1 на вход C запоминается значение D и устанавливается Q=D (рис.5.12).

Совсем коротко упомянем некоторые другие важные элементы цифровых логических схем. К ним относятся счетчики и регистры. Они имеют не один, а много выводов, состояние которых зависит от количества импульсов, пришедших на счетный вход с момента подачи единицы на вход установки. В частности, комбинация

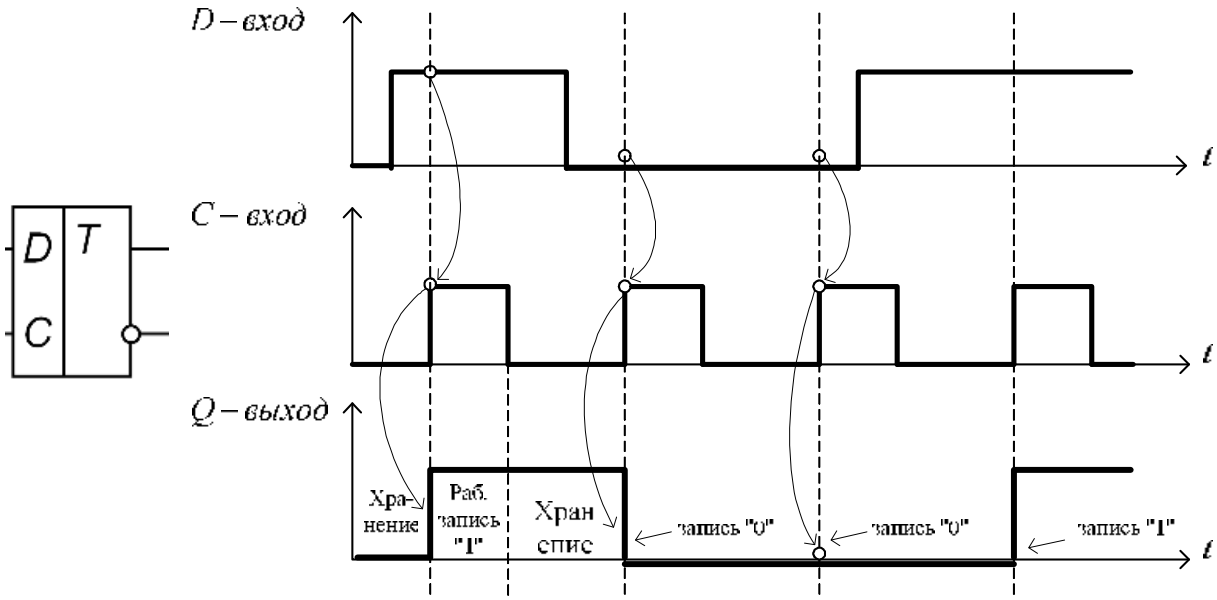


Рис. 5.12: D триггер. Слева: обозначение на схемах, справа: диаграмма, иллюстрирующая его работу.

нулей и единиц на выходах может представлять собой двоичный код числа импульсов. На рисунке 5.13 показан пример счетчика-сдвигового регистра, собранного из D - триггеров. Если на вход такого устройства поступает поток данных, состоящий из нулей и единиц, строго синхронизированный с тактовыми импульсами (синхроимпульсами - периодическим прямоугольным сигналом), то на выходах триггеров будет формироваться параллельный код, в котором каждый следующий разряд соответствует предшествующему значению во входном потоке данных. С каждым тактовым импульсом этот код "перемещается"слева на право.

5.5 ЦАП и АЦП

Для преобразования аналоговых сигналов в цифровые и обратно используются соответственно, аналого-цифровые и цифро-аналоговые преобразователи, сокращенно - АЦП и ЦАП. Начнем с ЦАП, они проще. Построить ЦАП можно, например, используя сумматор на операционном усилителе с весовыми резисторами в цепи обратной связи (см. рис. 5.14).

Работу сумматора мы уже рассматривали в разделе, посвященном операционным усилителям. Отличие данной схемы только в том, что входные сигналы усиливаются с различными коэффициентами усиления. Напряжение на выходе можно записать как

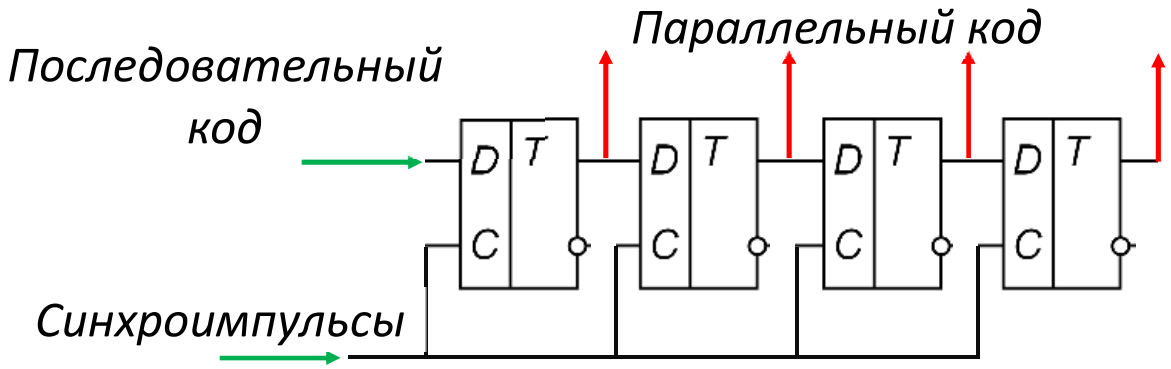


Рис. 5.13: Сдвиговый регистр на D триггерах.

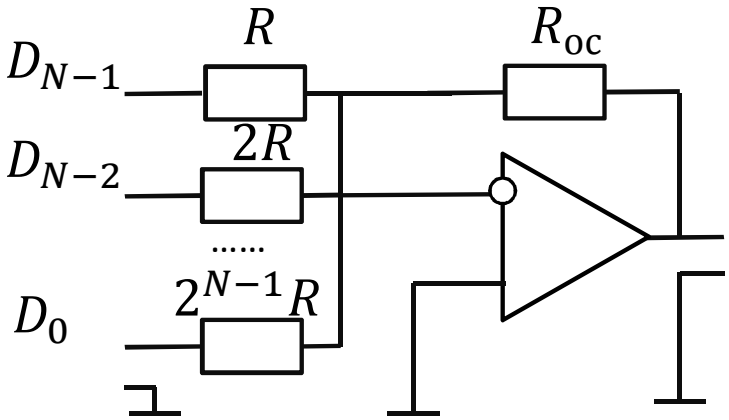


Рис. 5.14: Цифроаналоговый преобразователь на операционном усилителе с весовыми резисторами в цепи обратной связи.

$$U_{вых} = -U_1 R_{oc} \sum \frac{D_i}{2^{N-1-i}} R, \tag{5.32}$$

где $D_i = [0, 1]$.

На практике используют более сложные схемы, обычно оформленные в виде интегральных микросхем. Перечислим основные параметры, характеризующие ЦАП. *статические* характеристики это

- Разрядность – число двоичных разрядов входного кода (N). Распространенные ЦАП имеют от 8 до 14 разрядов, специализированные - 16, 24, возможны другие значения.

- Диапазон выходной величины – интервал значений выходного напряжения $U_{min} - U_{max}$.

- Относительная разрешающая способность - величина, обратная числу уровней квантования $d_r = \frac{1}{2^{N-1}}$ а абсолютная разрешающая способность численно равна шагу квантования в Вольтах: $d_A = \frac{U_{max}}{2^{N-1}} = \triangle U$.
- Абсолютная погрешность преобразования δd - максимальное отклонение выходного напряжения в конечной точке реальной характеристики преобразования от идеальной.
- Интегральная нелинейность преобразования ЦАП d_{int} - определяет максимальное отклонение реальной характеристики от идеальной.
- Дифференциальная нелинейность преобразования ЦАП d_{dif} - численно равна максимальной разности двух соседних шагов квантования.

Иллюстрация этих параметров приведена на рис. 5.15.

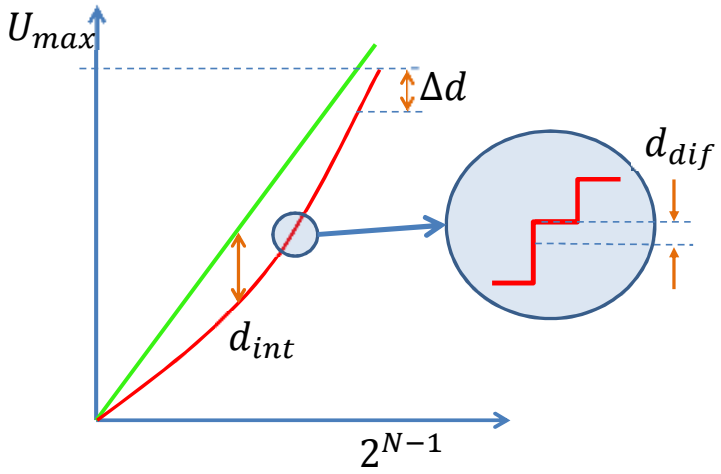


Рис. 5.15: Иллюстрация погрешностей, возникающих при работе ЦАП.

Кроме этого, преобразование не может выполняться мгновенно. Важными *динамическими* параметрами являются время установления $t_{уст}$ выходного напряжения или тока – интервал времени от начала изменения входного двоичного кода от минимального до максимального значения до момента, когда выходной аналоговый сигнал достигнет заданной величины. ЦАП характеризуют также максимальной частотой преобразования $f_{пр}$ – наибольшей допустимой частотой смены входного кода.

Обратную задачу решают с помощью *аналого-цифровых* преобразователей (АЦП). Пример возможной реализации такого преобразователя с помощью операционных усилителей, работающих в режиме компараторов, показан на рис 5.16.

Цепочка резисторов образует делитель, таким образом, на неинвертирующий вход операционного усилителя с номером n (считая снизу) подается напряжение $U_n = U_{ВХ}\frac{N-1}{n}$, где N - полное число резисторов в делителе. На все инвертирующие входы подается опорное напряжение, равное $U_{ВХмакс}\frac{N-1}{N}$. Чем выше входное напряжение, тем для большего числа операционных усилителей (считая сверху) будет выполняться условие $U_{инв} > U_{неинв}$, соответственно, на их выходах будет

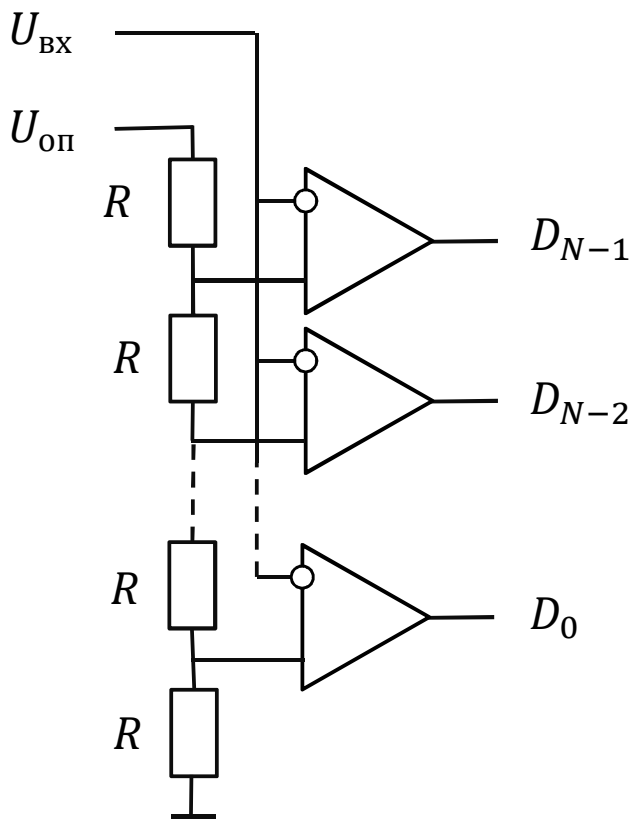


Рис. 5.16: Аналого-цифровой преобразователь параллельного преобразования.

напряжение, близкое к $U_{пит}$ (логическая единица), а на выходах остальных ОУ - близкое к $-U_{пит}$ (логический ноль). Таким образом, на выходах ОУ формируется значение входного напряжения в коде Джонсона, который с помощью цифрового вычислительного устройства можно преобразовать в любой другой.

Статические параметры АЦП аналогичны ЦАП, в качестве динамических обычно указывают

- максимальную частоту преобразования – частоту дискретизации входного сигнала;
- апертурное время – время, в течение которого сохраняется неопределенность между значением выборки и временем, к которому оно относится;
- апертурную неопределенность – случайное изменение апертурного времени в конкретной точке характеристики преобразования;
- время кодирования – время, в течение которого осуществляется непосредственное преобразование установившегося значения входного сигнала.

Так же, как и в случае с ЦАП приведенная схема не является единственно возможной. Выпускается множество различных интегральных микросхем ЦАП.

Кроме того, ЦАП и АЦП часто содержатся внутри микропроцессоров, предназначенных для обработки аналоговых сигналов (иногда называемых сигнальными процессорами).

5.6 Основы z-преобразования

При анализе и синтезе дискретных и цифровых устройств широко используют так называемое z-преобразование, играющее по отношению к дискретным сигналам такую же роль, как интегральные преобразования Фурье и Лапласа по отношению к непрерывным сигналам.

Определение z-преобразования: пусть $x_k = (x_0, x_1, x_2...)$ — числовая последовательность, конечная или бесконечная, представляющая собой отсчеты некоторого сигнала. Поставим ей в однозначное соответствие сумму ряда по отрицательным степеням комплексной переменной z:

$$X(z) = x_0 + \frac{x_1}{z} + \frac{x_2}{z^2} + \cdots = \sum_{k=1}^{\infty} x_k z^{-k}. \tag{5.33}$$

Назовем эту сумму, если она существует, z-преобразованием последовательности $\{x_k\}$. Целесообразность введения такого математического объекта связана с тем, что вместо дискретных последовательностей чисел мы получаем функции, свойства которых можно исследовать методами математического анализа.

По формуле (5.33) можно непосредственно найти z-преобразование дискретных сигналов с конечным числом отсчетов. Так, простейшему дискретному сигналу с единственным отсчетом $\{x_k\} = (1, 0, 0, ...)$ соответствует z-образ $X(z) = 1$. Другой пример:

$$\{x_k\} = (1, 1, 1, 0, 0, \dots) \quad \Rightarrow \quad X(z) = \frac{z^2 + z + 1}{z^2}. \tag{5.34}$$

Важно, что, для широкого класса бесконечных последовательностей получаемый ряд является сходящимся. Так, в теории функций комплексного переменного доказывается, что, если коэффициенты ряда 5.33 удовлетворяют условию

$$|x_k| < mR^k, \tag{5.35}$$

для всех $k \geq 0$ ($m > 0$, $R > 0$ - вещественные числа) то он сходится при всех z таких, что $|z| > R$.

z-преобразование можно определить и для аналоговых сигналов. Для этого берем отсчеты непрерывной функции $x(t)$ в точках $t = k\Delta$ так, что бы шаг дискретизации Δ удовлетворял критерию Найквиста: $\Delta < 2/\nu$, где ν — верхняя частота (в герцах) в спектре $x(t)$. Тогда z-образ для $x(t)$:

$$X(z) = \sum_{k=0}^{\infty} u(k\Delta) z^{-k}.$$

Например, если $x(t) = \exp(at)$, то ее z-образ:

$$X(z) = \sum_{k=0}^{\infty} \exp(ak\Delta)\, z^{-k} = \frac{z}{z - \exp(a\Delta)}.$$

Обратное z-преобразование

Замечательное свойство z-преобразования состоит в том, что функция $X(z)$ полностью определяет *всю, вообще говоря, бесконечную,* последовательность отсчетов $(x_0, x_1, x_2, \ldots)$. Умножим обе части (5.33) на множитель z^{m-1}:

$$z^{m-1}X(z) = x_0 z^{m-1} + x_1 z^{m-2} + \ldots x_m z^{-1}, \tag{5.36}$$

а затем вычислим интегралы по контуру от обеих частей полученного равенства, взяв в качестве контура интегрирования произвольную замкнутую кривую, лежащую целиком в области аналитичности и охватывающую все полюсы функции $X(z)$. По теореме Коши:

$$\oint z^n dz = \begin{cases} 2\pi i & \text{если } n = -1\,, \\ 0 & \text{если } n \neq -1\,. \end{cases}$$

Очевидно, интегралы от всех слагаемых правой части обратятся в нуль, за исключением слагаемого с номером m, поэтому:

$$x_m = \frac{1}{2\pi i}\oint z^{m-1}X(z)\,dz. \tag{5.37}$$

Формула (5.37) называется обратным z-преобразованием.

Свойства z-преобразования

1. Линейность.

   Если x_k и y_k некоторые дискретные сигналы с z-образами $X(z)$ и $Y(z)$, то сигналу $u_k = ax_k + by_k$ соответствует z-образ $U(z) = aX(z)+bY(z)$ (очевидно, проверяется подстановкой).

2. z-преобразование смещенного сигнала.

   Рассмотрим дискретный сигнал $\{y_k\}$, получающийся из дискретного сигнала $\{x_k\}$ путем сдвига на одну позицию в сторону запаздывания: $y_k = x_{k-1}$. Непосредственно вычисляя z-преобразование, получаем:

$$Y(z) = \sum_{k=0}^{\infty} x_{k-1}z^{-k} = z^{-1}\sum_{n=0}^{\infty} x_n z^{-n} = z^{-1}X(z).$$

3. z-преобразование свертки. Пусть x_k и y_k -дискретные сигналы, для которых определена свертка:

$$f_m \equiv \sum_{k=0}^{\infty} x_k y_{m-k}.$$

   Тогда z-образ свертки равен произведению образов. Доказываем:

$$F(z) = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty} x_k y_{m-k} z^{-m} = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty} x_k z^{-k} y_{m-k} z^{-(m-k)} =$$
$$= \sum_{k=0}^{\infty} x_k z^{-k} \sum_{n=0}^{\infty} y_n z^{-n} = X(z)Y(z). \tag{5.38}$$

Таблица 5.2: Таблица некоторых z-преобразований

Сигнал	z-преобразование	Область сходимости
$\delta_n = \{1, 0, 0, \ldots\}$	1	$\forall$
δ_{n-n_0}	z^{-n_0}	$z \neq 0$
$\mathcal{H}_n = \{1, 1, 1, \ldots\}$	$\frac{z}{z-1}$	$\lvert z\rvert > 1$
$a^n\mathcal{H}_n$	$\frac{z}{z-a}$	$\lvert z\rvert > \lvert a\rvert$
$na^n\mathcal{H}_n$	$\frac{za}{(z-a)^2}$	$\lvert z\rvert > \lvert a\rvert$
$\mathcal{H}_n \cos n\omega_0$	$\frac{z\left(z-\cos\omega_0\right)}{1-2z\cos\omega_0+z^2}$	$\lvert z\rvert > 1$
$\mathcal{H}_n a^n \cos n\omega_0$	$\frac{z\left(z-a\cos\omega_0\right)}{a^2-2za\cos\omega_0+z^2}$	$\lvert z\rvert > \lvert a\rvert$
$\mathcal{H}_n \sin n\omega_0$	$\frac{z\sin\omega_0}{1-2z\cos\omega_0+z^2}$	$\lvert z\rvert > 1$
$\mathcal{H}_n a^n \sin n\omega_0$	$\frac{za\sin\omega_0}{a^2-2za\cos\omega_0+z^2}$	$\lvert z\rvert > \lvert a\rvert$

5.7 Цифровая фильтрация

Теория цифровых фильтров переносит на случай дискретных сигналов все основные положения теории линейных систем, преобразующих непрерывные сигналы. Пусть известна импульсная характеристика $g(t)$ линейной стационарной системы. Тогда если на ее вход подается *аналоговый* сигнал $x(t)$, то сигнал на ее выходе равен:

$$y(t) = \int_{-\infty}^{t} x(\tau)g(t-\tau)d\tau. \tag{5.39}$$

Линейный цифровой фильтр — это дискретная система (физическое устройство или программа, не обязательно предназначенная именно для фильтрации в узком смысле этого слова), которая преобразует последовательность x_k в y_k.

Для того чтобы обобщить формулу (5.39) на случай дискретных сигналов, вводят понятие *импульсной характеристики* цифрового фильтра. По определению, она представляет собой дискретный сигнал g_k, который является реакцией цифрового фильтра на входную последовательность, которую можно назвать единичным импульсом: $\delta_k = \{1, 0, 0, 0, \ldots\}$ (аналог дельта-функции для аналогового сигнала):

$$\{1, 0, 0, 0, \ldots\} \Rightarrow \{g_0, g_1, g_2, \ldots\}.$$

Тогда для любого дискретного входного сигнала x_k можно записать любой отсчет выходного как:

$$y_m = x_0 g_m + x_1 g_{m-1} + \ldots + x_m g_0 = \sum_{k=0}^{m} x_k g_{m-k}.$$

— выходная последовательность есть дискретная свертка входного сигнала и импульсной характеристики фильтра. Смысл этой формулы очевиден: в момент каждого отсчета цифровой фильтр производит взвешенное суммирование всех предыдущих значений входного сигнала, причем роль последовательности весовых коэффициентов играют отсчеты импульсной характеристики. Можно сказать, что циф-

ровой фильтр обладает "памятью" по отношению к прошлым входным воздействиям. Практический интерес представляют лишь физически реализуемые фильтры, импульсная характеристика которых не может стать отличной от нуля в отсчетных точках, предшествующих моменту подачи входного импульса.

Системной функцией стационарного линейного цифрового фильтра называют отношение z-преобразования выходного сигнала $\{y_k\}$ к z-преобразованию сигнала на входе $\{x_k\}$. Легко убедиться, что системная функция фильтра — это z-образ его импульсной характеристики:

$$G(z) = Y(z)/X(z) = \sum_{k=0}^{\infty} g_k z^{(-k)}. \tag{5.40}$$

Для анализа аналоговых линейных цепей мы использовали гармонические сигналы. Для исследования свойств цифровых фильтров используем дискретные гармонические последовательности, получаются путем дискретизации гармонических сигналов:

$$\{x_k\} = \{A\exp\left[i\left(\omega k\triangle + \varphi\right)\right]\}, \tag{5.41}$$

$$Re\{x_k\} = \{A\cos\left[\left(\omega k\triangle + \varphi\right)\right]\}.$$

Заметим, что представление неоднозначно: $\{x_k\}$ не меняется при замене

$\omega \to \omega + 2\pi n/\triangle$.

Пусть такая бесконечная последовательность подается на вход цифрового фильтра.Тогда отсчеты на его выходе:

$$y_m = \sum_{k=-\infty}^{m} x_k h_{m-k} = Ae^{i\varphi} \sum_{k=-\infty}^{m} e^{i\omega k\triangle} h_{m-k}.$$

Преобразуем:

$$y_m = Ae^{i\varphi} \sum_{k=-\infty}^{m} e^{i\omega k\triangle} h_{m-k} = Ae^{i(\omega m\triangle+\varphi)} \sum_{k=-\infty}^{m} e^{i\omega(k-m)\triangle} h_{m-k},$$

заменим: $n = m - k$:

$$y_m = Ae^{i(\omega m\triangle+\varphi)} \sum_{n=0}^{\infty} e^{-i\omega n\triangle} h_n$$

- дискретная гармоническая последовательность!

$$K\left(i\omega\right) = \sum_{n=0}^{\infty} e^{-i\omega n\triangle} h_n$$

-её частотный коэффициент передачи.

Трансверсальные цифровые фильтры

Трансверсальным фильтром называется устройство, которое преобразует цифровую последовательность по следующему алгоритму (см. рис.5.17):

$$y_i = a_0 x_i + a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_m x_{i-m}, \tag{5.42}$$

здесь $a_0, a_1, a_2 \ldots$ - коэффициенты, m – порядок фильтра.

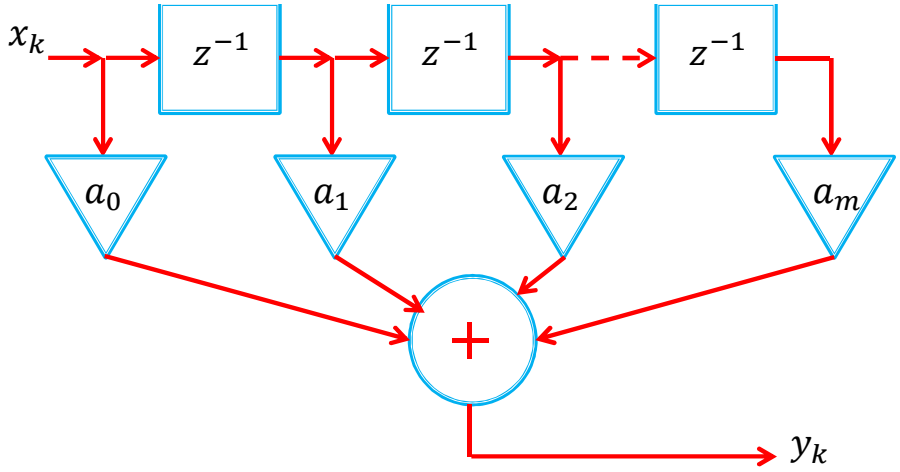


Рис. 5.17: Трансверсальный цифровой фильтр.

Применим z - преобразование к обоим частям:

$$Y(z) = (a_0 + a_1 z^{-1} + a_2 z^{m-2} + \ldots a_m z^{-m}) X(z).$$

Системная функция:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{a_0 z^m + a_1 z^{m-1} + \ldots a_m}{z^m}$$

имеет m - кратный полюс при $m = 0$ и m нулей.

Импульсная характеристика трансверсального фильтра содержит конечное число членов: Finite Impulse Response filter (FIR).

Доказательство: каждое слагаемое функции $H(z)$ дает вклад, равный соответствующему коэффициенту a_m, смещенному на n позиций в сторону запаздывания. Подадим на вход единичный импульс $(1,0,0,\ldots)$, получим: $(a_0, a_1, a_2 \ldots)$.

Частотная характеристика: сделаем замену: $z = exp(i\omega\triangle)$

частотный коэффициент передачи:

$$K(i\omega) = a_0 + a_1 e^{-i\omega\triangle} + a_2 e^{-i2\omega\triangle} \ldots a_m e^{-im\omega\triangle}.$$

При заданном шаге дискретизации $\triangle$ можно реализовать самые разнообразные формы АЧХ, подбирая должным образом весовые коэффициенты a_n.

Пример: фильтр 2 порядка. Пусть $y_i = \frac{1}{3}(x_i + x_{i-1} + x_{i-2})$ тогда его системная функция:

$$H(z) = \frac{1}{3}\left(1 + z^{-1} + z^{-2}\right).$$

Частотный коэффициент передачи:

$$K(i\omega) = \frac{1}{3}\left(1 + e^{-i\omega\triangle} + e^{-i2\omega\triangle}\right) = \frac{1}{3}[(1 + \cos(\omega\triangle) + \cos(2\omega\triangle)) - i\,(1 + \sin(\omega\triangle) + \sin(2\omega\triangle)].$$

Амплитудная характеристика этого фильтра (см.рис.5.18):

$$|K(i\omega)| = \frac{1}{3}\sqrt{3 + 4\cos\omega\triangle + 2\cos 2\omega\triangle}.$$

Таким образом, трансверсальный фильтр 2 порядка $y_i = \frac{1}{3}(x_i + x_{i-1} + x_{i-2})$ при $\omega\triangle < 2$ может играть роль ФНЧ. Однако, при $\omega\triangle > \pi$ снова появляются полосы пропускания (характеристика периодична). Это не удивительно: на этих частотах теорема Котельникова не выполняется. По этой причине при оцифровке аналоговых сигналов перед АЦП обычно ставят аналоговый ФНЧ (так называемый anti-aliasing фильтр).

Фазовая характеристика:

$$\varphi_K(\omega) = -\omega\triangle$$

представляет собой линейную зависимость.

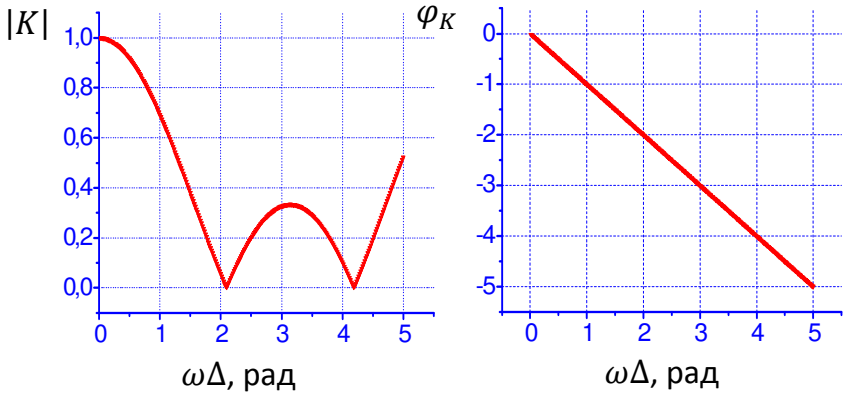


Рис. 5.18: Амплитудная и фазовая характеристики трансверсального цифрового фильтра.

Рекурсивные цифровые фильтры.

При обработке цифровой последовательности могут используются предыдущие значения не только входного, но и выходного сигнала:

$$y_i = a_0x_i + a_1x_{i-1} + a_2x_{i-2} + \cdots + a_mx_{i-m} + b_1y_{i-1} + b_2y_{i-2} + \cdots + b_my_{i-m}.$$

Если $b_1 = b_2 = \cdots = b_m = 0$ – получаем трансверсальный (нерекурсивный) фильтр (см. рис. 5.19).

Получим системную функцию рекурсивного фильтра. Перенесем члены с b_n в левую часть, применим Z – преобразование:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{a_0z^m + a_1z^{n-1} + \ldots + a_mz^{n-m}}{z^n - b_1z^{n-1} - \cdots - b_n}.$$

В рекурсивном фильтре возможны свободные колебания – генерация ненулевой последовательности на выходе при нулевой входной. Цифровой фильтр называется устойчивым, если возникающий в нем свободный процесс есть невозрастающая

последовательность, т. е. значения $|y_n|$ не превышают некоторого положительного числа независимо от выбора начальных условий.

Рекурсивный фильтр устойчив, если все полюса системной функции $H(z)$ по модулю не превосходят единицы.

Импульсная характеристика имеет вид неограниченно-протяженной последовательности: Infinite Impulse Response filter (IIR).

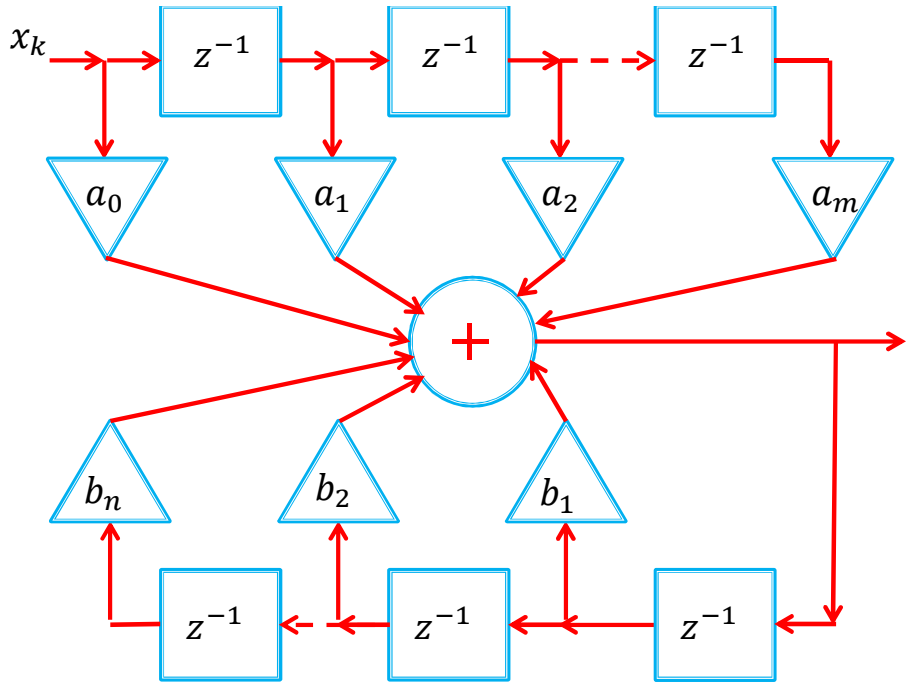


Рис. 5.19: Рекурсивный цифровой фильтр.

Основным преимуществом рекурсивных фильтров является их эффективность (требуется гораздо меньше вычислительных ресурсов), особенно когда требуется задать сложную характеристику фильтра. Отметим, что современные сигнальные процессоры содержат аппаратные средства, позволяющие реализовать трансверсальные фильтры почти так же эффективно, как рекурсивные. Рекурсивный фильтр можно сделать гораздо ближе к аналоговому прототипу. Основным недостатком является ограниченная устойчивость.

5.8 Понятие об оптимальной фильтрации

Как следует из названия, оптимальный фильтр должен выделять желаемый сигнал из шумов наилучшим образом. Заметим, что построение такого фильтра существенно зависит от априорной информации об этом сигнале. Задачами оптимальной фильтрации могут быть:

1. Измерение параметров стационарного сигнала известной формы на фоне шумов («задача радиосвязи»)

2. Обнаружение сигнала известной формы, время прихода которого не определено («задача радиолокации»)

3. Поиск сигналов, о форме которого имеются лишь предположения («задача радиоастрономии»)

Во всех случаях на входе системы, кроме сигнала, присутствует шум.

Предположим, что нам необходимо установить факт присутствия сигнала, форма которого известна *a priori*. Применим линейный фильтр $h(t)$. На его выходе получим:

$$s_{вых}(t) = \int_{-\infty}^{\infty} s_{вх}(\tau)\, h(t-\tau)\, d\tau.$$

Зафиксируем t_0 и будем искать $h(t)$ дающую максимум $|s(t_0)|$ и используем неравенство Коши — Буняковского:

$$\left|\int_{-\infty}^{\infty} s_{вх}(\tau)\, h(t_0-\tau)\, d\tau\right| \le \sqrt{\int_{-\infty}^{\infty} s_{вх}^2(\tau)\, d\tau \int_{-\infty}^{\infty} h^2(t_0-\tau)\, d\tau}.$$

Равенство здесь достигается только если $h(t_0-\tau) = ks(\tau)$. После замены $t = t_0-\tau$ получаем:

$$h_{согл}(t) = k s_{вх}(t_0-t) \tag{5.43}$$

- согласованный фильтр.

Импульсная характеристика согласованного фильтра - это масштабная копия входного сигнала, зеркально отраженная во времени (см. рис 5.20).

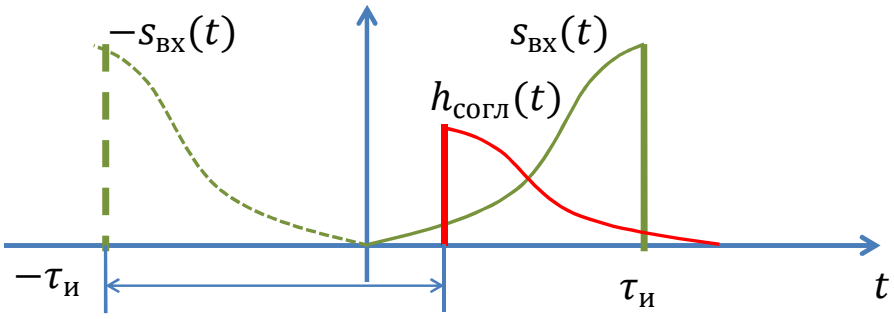


Рис. 5.20: Иллюстрация построения согласованного цифрового фильтра.

Условие реализуемости такого фильтра: $t_0 \ge \tau_и$

Покажем, что согласованный фильтр является *коррелятором*:

Пусть на входе $u(t) \ne s(t)$:

$$u_{вых}(t) = \int_{-\infty}^{\infty} u_{вх}(\tau)\, h_{согл}(t-\tau)\, d\tau =$$

$$= k\int_{-\infty}^{\infty} u_{вх}(\tau)\, s_{вх}[t_0-(t-\tau)]\, d\tau = k\int_{-\infty}^{\infty} u_{вх}(\tau)\, s[\tau-(t-t_0)]\, d\tau = kB(t-t_0)$$

- кросс-корреляционная функция входного сигнала $u_{вх}(t)$ и сигнала, для которого этот фильтр является согласованным $s_{вх}(t)$.

В момент времени t_0 $u_{вых}(t)$ - их скалярное произведение:

$$u_{вых}(t_0) = k\int_{-\infty}^{\infty} u_{вх}(\tau)\, s_{вх}(\tau)\, d\tau.$$

Получим частотный коэффициент передачи для согласованного фильтра. Из общих соображений можно предположить, что модуль частотного коэффициента передачи должен быть пропорционален модулю спектральной плотности сигнала. Простой пример – это гребенчатый фильтр для сигнала, содержащего три гармонические составляющие. На рис 5.21 приведена спектральная плотность мощности для такого сигнала и качественно показана возможная зависимость модуля коэф-

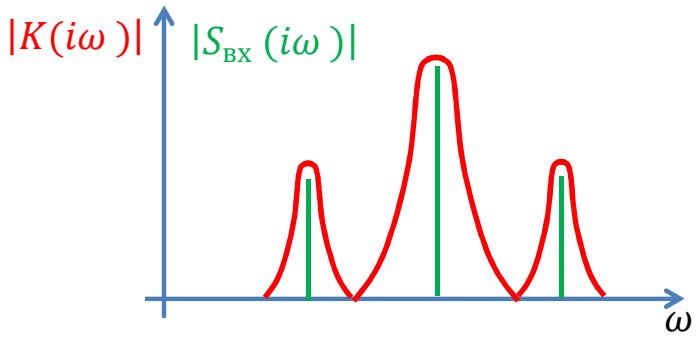


Рис. 5.21: Гребенчатый фильтр.

фициента передачи фильтра от частоты. Однако, такое рассмотрение не принимает во внимание фазовые соотношения между сигналами. Оптимальный фильтр, построенный в соответствии с формулой, приведенной выше, обладает замечательной особенностью, заключающейся в том, что возможность обнаружения сигнала оказывается зависящей от его энергии, а не от формы. Такие фильтры называют *согласованными* с сигналом.

Пример: Согласованный фильтр для прямоугольного импульса. Пусть нашим искомым сигналом является прямоугольный импульс (см. рис. 5.22).

Его спектр:

$$S_{вх}(\omega) = \int_{-\infty}^{\infty} s_{вх}(t)\, e^{-i\omega t} dt = U_0 \int_0^{\tau_и} e^{-i\omega t} dt = \frac{U_0}{i\omega}\left(1 - e^{-i\omega\tau_и}\right).$$

Тогда коэффициент передачи согласованного фильтра, отклик которого максимален в момент окончания импульса:

$$K_{согл}(i\omega) = k\frac{1 - e^{i\omega\tau_и}}{-i\omega} e^{-i\omega\tau_и} = \frac{k}{i\omega}\left(1 - e^{i\omega\tau_и}\right)$$

Реализовать такой коэффициент передачи с помощью простой аналоговой цепи не получится. Составим из ее из блоков (см. рис. 5.23). В соответствии с формулой,

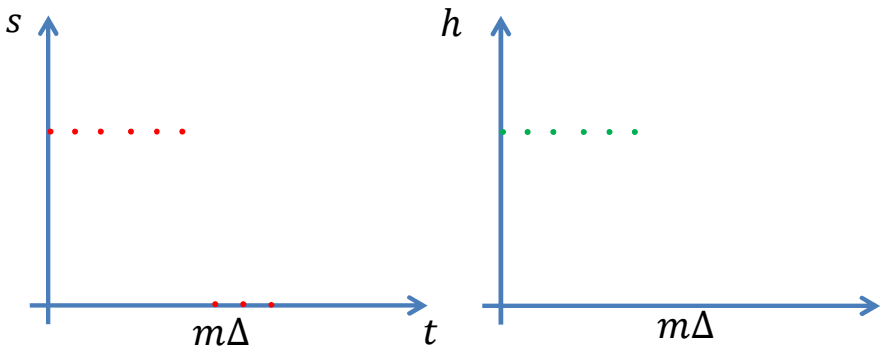


Рис. 5.22: Дискретный прямоугольный импульс.

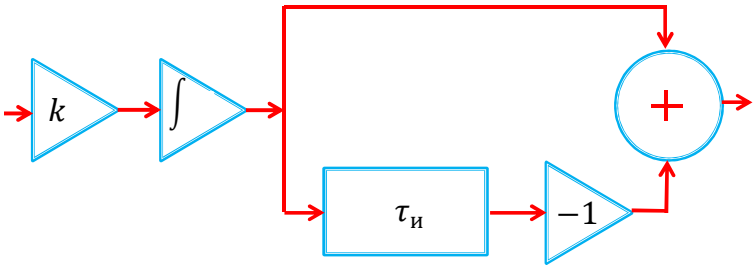


Рис. 5.23: Оптимальная аналоговая фильтрация прямоугольного импульса.

сигнал сначала умножается на коэффициент k, затем поступает на интегрирующую цепочку после чего разделяется на две части одна из которых задерживается на время $\tau_{И}$, инвертируется и складывается со второй. А если оцифровать такой импульс и сделать трансверсальный фильтр? На рис.5.24 представлена получаемая цифровая последовательность и импульсная характеристика оптимального фильтра:

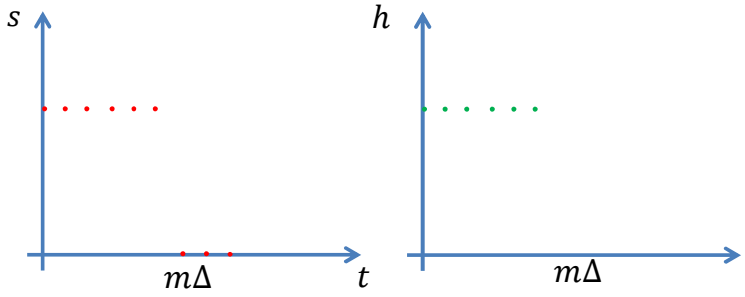


Рис. 5.24: Оптимальная цифровая фильтрация прямоугольного импульса.

$$y_i = kx_i + kx_{i-1} + kx_{i-2} + \cdots + kx_{i-m}.$$

фильтр просто складывает отсчеты в течение времени, равного его длительности

- выглядит значительно проще!

Глава 6

Приложение

6.1 Обобщенные функции

Рассмотрим некоторые математические тонкости преобразования Фурье для дельта-функции и функции Хевисайда, а также связь между ними.

6.1.1 Дельта-функция и Фурье образ

Дельта-функция $\delta(t)$ является обобщенной функцией и математически она определяется так:

$$\int_{-\infty}^{\infty} f(\tau)\,\delta(x-\tau)\,d\tau = \frac{1}{2}\big(f(x-0)+f(x+0)\big),$$

где $f(x)$ — произвольная кусочно-непрерывная функция.

Для физика полезно представить дельта-функцию в виде предела некоторой обычной функции (это называют представлением дельта-функции, которых существует множество), например, такого:

$$\delta(x) = \lim_{\alpha\to 0} D(x,\alpha), \qquad D(x,\alpha) = \frac{1}{\sqrt{2\pi\alpha^2}}\,\exp\left(\frac{-x^2}{2\,\alpha^2}\right), \quad \int_{-\infty}^{\infty} D(x,\alpha)\,dx = 1. \tag{6.1}$$

Из этого определения видно, что дельта-функцию можно понимать как "колокол" с центром в начале координат, ширина которого стремится к нулю при постоянной площади под "колоколом". Функцию $D(x,\alpha)$ часто называют "размазанной" дельта-функцией[1]. Ей полезно пользоваться для понимания свойств дельта-функции.

Нетрудно найти Фурье преобразование функции $D(x,\alpha)$

$$D(\omega,\alpha) = \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\alpha^2}}\,\exp\left(\frac{-x^2}{2\alpha^2} - i\omega x\right)\,dx = \exp\left(\frac{-\alpha^2\omega^2}{2}\right), \tag{6.2}$$

[1] Заметим, что дельта-функции соответствует множество таких "размазанных" функций (т.е. представлений дельта-функции), например

$$\delta(x) = \lim_{\alpha\to 0}\frac{\sin(x/\alpha)}{\pi x}, \quad \text{или} \quad \delta(x) = \lim_{\alpha\to 0}\frac{\alpha}{\pi\big(x^2+\alpha^2\big)}$$

Ниже для определенности и удобства мы будем пользоваться функцией $D(x,\alpha)$

Справка: $\dfrac{-x^2}{2\alpha^2} - i\omega x = -\left(\dfrac{x}{\sqrt{2}\,\alpha} + \dfrac{i\omega\alpha}{\sqrt{2}}\right)^2 - \dfrac{\alpha^2\omega^2}{2}.$

Теперь мы можем формально найти Фурье-образ дельта-функции, переходя к пределу:

$$\delta(\omega) = \lim_{\alpha\to 0} D(\omega,\alpha) = \lim_{\alpha\to 0} \exp\left(\frac{-\alpha^2\omega^2}{2}\right) = 1, \tag{6.3}$$

$$\delta(x) = \lim_{\alpha\to 0} D(x,\alpha) = \lim_{\alpha\to 0} \int_{-\infty}^{\infty} D(\omega,\alpha)\, e^{i\omega x}\,\frac{d\omega}{2\pi} = \int_{-\infty}^{\infty} e^{i\omega x}\,\frac{d\omega}{2\pi}. \tag{6.4}$$

Заметим, что иногда равенство (6.4) используют как еще одно определение дельта-функции.

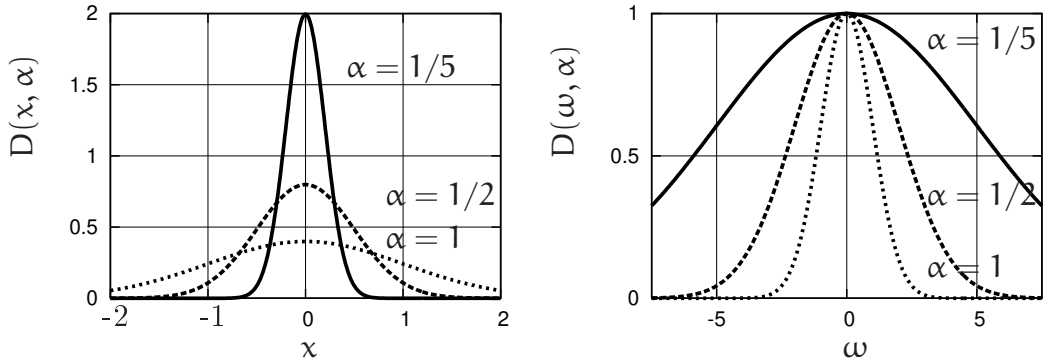


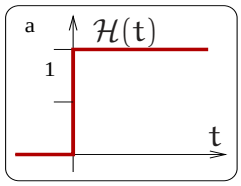
Рис. 6.1: Слева: графики "размазанной" дельта-функции $D(x,\alpha)$ при различных значениях параметра α. Справа: графики Фурье образа "размазанной" дельта-функции $D(\omega,\alpha)$ при различных значениях параметра α. Видно, что при уменьшении α функция $D(x,\alpha)$ становиться все более узкой, тогда как ее Фурье образ $D(\omega,\alpha)$ стремиться к постоянной величине ($\to 1$).

Переход к пределу иллюстрирует рис. 6.1. Мы видим, что чем "уже" функция $D(x,\alpha)$ тем "шире" ее Фурье образ.

6.1.2 Функция Хевисайда ("ступенька") и ее Фурье образ

Напомним определение функции Хевисайда:

$$\mathcal{H}(t) = \begin{cases} 1 & \text{если } t > 0, \\ 1/2 & \text{если } t = 0, \\ 0 & \text{если } t < 0 \end{cases} .$$



При вычислении Фурье образа от функции Хевисайда сразу возникает затруднение — интеграл не сходится. Действительно:

$$\mathcal{H}(\omega) = \int_{-\infty}^{\infty} \mathcal{H}(t)\, e^{-i\omega t}\, dt = \lim_{t_0\infty} \int_0^{t_0} e^{-i\omega t}\, dt = \lim_{t_0\to 0} \frac{1 - e^{-i\omega t_0}}{i\omega} = ?,$$

поэтому используют представление функции Хевисайда через введение "заваленной на ∞" ступеньки $H(t,\epsilon)$.

$$\mathcal{H}(t) = \lim_{\epsilon\to 0} H(t,\epsilon) = \begin{cases} e^{-\epsilon t} & \text{если } t > 0, \\ 1/2 & \text{если } t = 0, \\ 0 & \text{если } t < 0. \end{cases} \tag{6.5}$$

Теперь нетрудно найти Фурье образ "заваленной" функции $H(t,\epsilon)$

$$H(\omega,\epsilon) = \int_0^\infty e^{-\epsilon t - i\omega t}\, dt = \frac{1}{i\omega+\epsilon}.$$

Введенная бесконечно малая ϵ нужна и при вычислении обратного преобразования Фурье: наличие $\epsilon > 0$ дает возможность применить теорему о вычетах, полагая, что полюс функции $H(\omega,\epsilon)$ лежит в верхней полуплоскости. Для справки приведем выкладки для вычисления обратного преобразования Фурье с применением вычетов:

$$H(t,\epsilon) = \int_{-\infty}^{\infty} \frac{e^{i\omega t}}{i\omega+\epsilon}\,\frac{d\omega}{2\pi}, \quad \text{Полюс: } \omega = i\epsilon \quad \text{в верхней полуплоскости}$$

$t > 0$: интегрируем по верхней полуплоскости $\quad \mathbf{H(t>0,\epsilon) = e^{-\epsilon t}},$

$t < 0$: интегрируем по нижней полуплоскости $\quad \mathbf{H(t<0,\epsilon) = 0},$

$$t = 0 : \mathbf{H(t=0,\epsilon)} = \int_{-\infty}^{\infty} \frac{\epsilon - i\omega}{\omega^2+\epsilon^2}\,\frac{d\omega}{2\pi} = \frac{1}{2}.$$

Рассуждения с бесконечно малой ϵ кажутся спекулятивными, но они приняты и широко используются в физической литературе. С физической точки зрения очевидно, что сигналов в виде бесконечно-длинных ступенек нет, поэтому можно считать ϵ не бесконечно малой, а просто малой величиной, определяемой из конкретной физической модели.

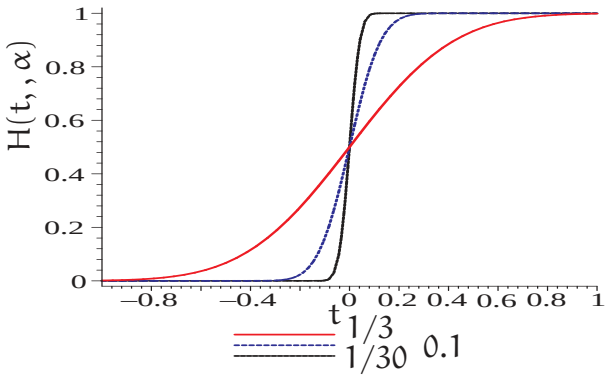


Рис. 6.2: Графики функции $H(t,\alpha)$ для $\alpha = 1/3$ (самая пологая функция), $\alpha = 0.1$ (более крутая) и $\alpha = 1/30$ (самая крутая функция).

6.1.3 Связь между дельта-функцией и функцией Хевисайда

Связь между этими функциями выражается следующим образом

$$\mathcal{H}(t) = \int_{-\infty}^{t} \delta(\tau)\, d\tau, \quad \Rightarrow \quad \frac{d\mathcal{H}(t)}{dt} = \delta(t). \tag{6.6}$$

Для иллюстрации опять рассмотрим "размазанную" дельта-функцию (6.1) и подставим ее в (6.6):

$$H(t,\alpha) = \int_{-\infty}^{t} D(\tau,\alpha)\, d\tau = \frac{1}{2}\left\{1 + \operatorname{erf}\left(\frac{t}{\sqrt{2}\,\alpha}\right)\right\}, \tag{6.7}$$

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-u^2}\, du\,. \tag{6.8}$$

Этот интеграл выражается через интеграл ошибок $\operatorname{erf}(x)$ — на рис. 6.2 приведены графики "размазанной" ступеньки $H(t,\alpha)$ для разных значений параметра α: чем меньше α, тем круче идет "размазанная" ступенька и в пределе $\alpha \to 0$ получаем функцию Хевисайда.

Литература

[1] А.А. Харкевич, Спектры и анализ. Москва, 1962.

[2] И.С. Гоноровский, Радиотехнические цепи и сигналы. Москва, 1986.

[3] С.И. Баскаков, Радиотехнические цепи и сигналы. Москва, 2000.

[4] А.Б. Сергиенко, Цифровая обработка сигналов, СПб, 2003.

[5] Л.И. Глюкман, Пьезоэлектрические кварцевые резонаторы, Москва, Радио и связь, 1981.

[6] М. Абрамовиц, И. Стиган Справочник по специальным функциям. Москва, 1979.

Учебное издание

Биленко Игорь Антонович,
Воронцов Юрий Иванович,
Вятчанин Сергей Петрович.

Введение в радиофизику.

Подписано в печать 17.10.2016.

Объем 12.75 п.л. Тираж 100 экз. Заказ №

Физический факультет им. М.В. Ломоносова
119991 Москва, ГСП-1, Ленинские горы, д.1, стр.2.

Отпечатано в отделе оперативной печати физического факультета МГУ.