Вблизи границы Шеннона

В. Варгаузин

В статье обсуждаются задачи, идеи и практические аспекты помехоустойчивого кодирования с коррекцией ошибок, называемого в зарубежной литературе Forward Error Correction (FEC). Термин "Forward" означает, что передача информации производится в прямом направлении от источника информации к получателю. При этом возможность уточнения достоверности полученной информации у получателя через обратный канал отсутствует. Поэтому декодер всегда принимает окончательное решение о значении каждого переданного информационного бита. Этот режим характерен для вещательных систем. Качество декодирования в таком режиме оценивается вероятностью ошибки информационного бита P_b ***или, как говорят иначе, частотой следования ошибок (Bit Error Rate, BER) после декодирования. Типичным требованием в современной цифровой связи является величина*** $P_b \approx 10^{-5} - 10^{-7}$***. Для канала связи, который хорошо аппроксимируется каналом с аддитивным гауссовским шумом (что характерно, например, для спутникового канала связи), значение*** P_b ***непосредственно зависит от величины энергетической эффективностью кодирования. Эту величину оценивают отношением*** E_b/N_0***, где*** E_b ***– средняя энергия, приходящаяся на информационный бит,*** N_0 ***– спектральная плотность мощности шума. В то же время, после открытия в 1993 г. турбо-кодов [1], качество схем кодирования и алгоритмов декодирования часто оценивают и величиной близости энергетической эффективности к предельно-достижимой величине – границе Шеннона.***

Материал содержит три раздела: "Теория", "Идеи", "Практика".

В разделе "Теория" излагаются предельные возможности помехоустойчивого кодирования, эффект ухудшения эффективности кодирования при ограничении на конечную величину информационного пакета, а также понятие "хороший" код.

В разделе "Идеи" приведены современные "хорошие" коды, кодовые конструкции и алгоритмы декодирования, на основе которых удалось добиться достаточной для практики близости к границе Шеннона. Материал раздела не претендует на обзор "всех кодов и алгоритмов декодирования", излагаются лишь ключевые, по мнению автора, моменты. Значительное внимание уделено исторической ретроспективе.

В разделе "Практика" приведено несколько современных стандартов цифровой связи с использованием "хороших" кодов.

В статье рассматриваются только вопросы энергетической эффективности помехоустойчивого кодирования. Вопросы частотной эффективности в явном виде не затрагиваются.

Теория

Предельные возможности кодирования

Отправной точкой теоретического анализа эффективности кодирования является теорема помехоустойчивого кодирования Шеннона. Согласно этой теореме, помехоустойчивым кодированием и декодированием можно добиться сколь угодно малой величины P_b при условии $R_c < C$. Здесь R_c (бит/символ) – скорость кода; C (бит/символ) – удельная пропускная способность информационного канала между кодером и декодером.

Скорость кода можно представить в виде $R_c = T_c / T_b$, где T_b (с) – длительность информационного бита; $1/T_b$ (бит/с) – скорость поступления информационных бит в кодер; T_c (с) – длительность кодового символа; $1/T_c$ – (символ/с) – скорость поступления кодовых символов из кодера.

Значение C существенно зависит от ограничений, накладываемых модуляцией и демодуляцией. Рассмотрим три наиболее важные для практики случая.

1. Оптимальная двоичная модуляция с жёсткими решениями демодулятора.

При жёстких оценках кодовых символов с вероятностью ошибки P_c при демодуляции, удельная пропускная способность информационного канала зависит только от этой вероятности: $C = 1 - H_b(P_c)$, где $H_b(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ – энтропийная функция Шеннона двоичного (binary, отсюда индекс "b") информационного источника. Такой информационный канал называют двоично-симметричным (ДСК, Binary Symmetric Channel, BSC) или каналом с двоичным входом и двоичным выходом (Binary Input Binary Output, BIBO). При условии $R_c = C$ получаем предельно допустимую вероятности ошибки кодового символа P_c (рис. 1). Анализируя график, к примеру, видим, что при $R_c = 0.5$

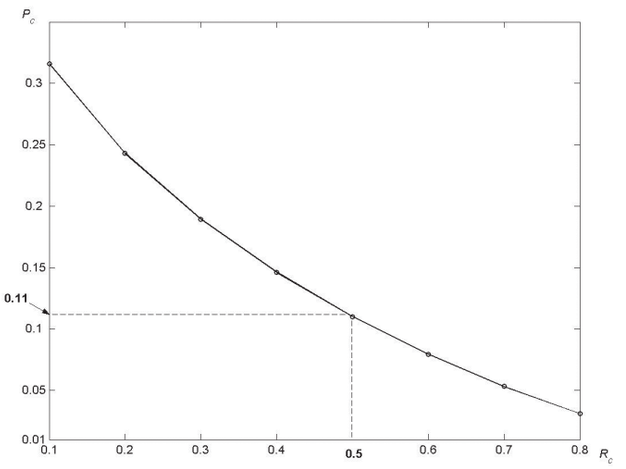


Рис. 1. Предельно допустимая вероятность P_c ошибки кодового символа для достижения сколь угодно малой вероятности ошибки P_b декодером жёстких решений.

допустимая вероятность ошибки кодового символа $P_c = 0.11$. При этом помехоустойчивым кодированием потенциально можно добиться сколь угодно малой вероятности ошибки информационного бита P_b.

Более того, теорема помехоустойчивого кодирования также утверждает, что при $R_c > C$ минимально достижимая величина P_b находится из уравнения $R_c = C/\left[1 - H_b\left(P_b\right)\right]$ (доказательство приведено в [5]). Анализ зависимости $P_b\left(R_c\right)$, полученной из этой формулы, показывает, что при скорости кода R_c, незначительно превышающей C, минимальная величина P_b может быть также весьма малой, например, $P_b = 10^{-12}$ при $R_c = 0.500084041856125$ и $C = 0.5$ (рис. 2).

Минимальная вероятность ошибки P_c в канале связи с аддитивным гауссовским шумом (AWGN channel) достигается при использовании оптимальной двоичной модуляции с противоположными сигналами (бифазной модуляции, BPSK). При этом $P_c = Q\left(\sqrt{2R_cE_b/N_0}\right)$, где $Q(x) = \int_x^\infty \exp\left(-t^2/2\right)dt/\left(2\pi\right)$ [3]. Используя эту формулу и рис. 1, можно построить зависимость минимально требуемой величины E_b/N_0 от P_c. Результат представлен на рис. 3. Эта зависимость иллюстрирует предельные возможности кода со скоростью R_c.

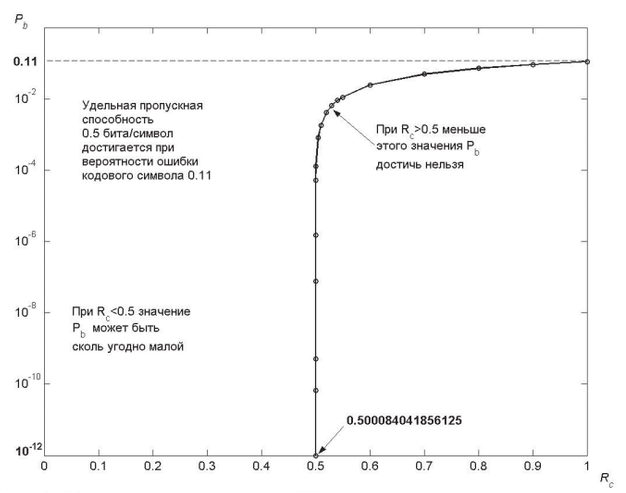


Рис. 2. Иллюстрация теоремы Шеннона.

2. Оптимальная двоичная модуляция с мягкими решениями демодулятора.

Использование мягких оценок символов (вероятностных величин правдоподобий значений "0" и "1" кодовых символов) после демодуляции повышает удельную пропускную способность информационного канала между кодером и декодером и ещё более повышает потенциальные возможности кодирования. Такой информационный канал называют каналом с двоичным входом и аддитивным гауссовским шумом (Binary-Input AWGN channel, BI-AWGN). С учётом пропускной способности такого канала [4] нетрудно построить зависимость минимально требуемой величины E_b/N_0 от P_c. Эти зависимости приведены на рис. 3. Из рисунка видно, что использование мягких решений при декодировании потенциально снижает энергетические затраты на 1.8-0.2=1.6 дБ ($R_c = 0.5$).

3. Канал без ограничения на вид модуляции.

В этом случае физический канал передачи сигналов совпадает с информационным каналом. Его удельная пропускная способность оказывается равной $C = 0.5\left(1 + 2E_bR_c/N_0\right)$. При этом кодер и модулятор представляют собой единое цифро-аналоговое устройство, на вход которого поступают информационные биты, а выходными кодовыми символами являются сигналы. Демодулятор и декодер также представляют собой единое аналогово-цифровое устройство, декодирующее сигналы в оценки переданных информационных бит. Это оптимальная схема кодирования и декодирования. Для этой схемы при $R_c = 0.5$ минимальная величина $E_b/N_0 = 0$ дБ (рис. 3). Эту величину можно назвать пределом кода со скоростью $R_c = 0.5$, или границей Шеннона.

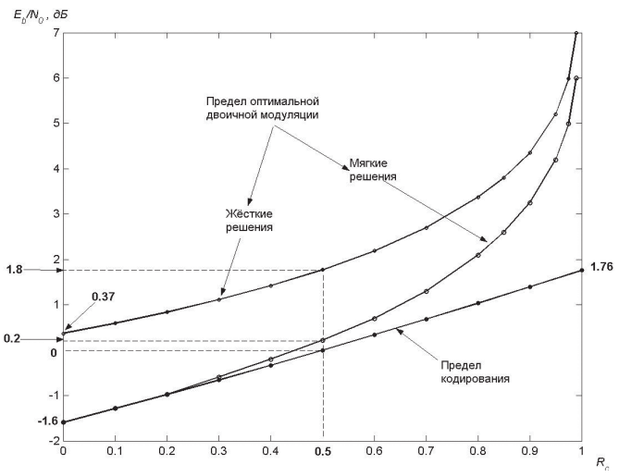


Рис. 3. Предельная энергетическая эффективность кода со скоростью R_c в гауссовском канале.

Из рисунка также следует, что энергетический проигрыш при ограничении на двоичную модуляцию составляет лишь 0.2-0=0.2 дБ ($R_c = 0.5$). Другими словами, величина поправки к границе Шеннона, связанная с ограничением на двоичную модуляцию, равна 0.2 дБ ($R_c = 0.5$).

На практике часто используют термин "энергетический выигрыш от кодирования" (ЭВК). Проанализируем предельный ЭВК кода со скоростью R_c. Для этого воспользуемся ранее рассмотренным урав-

нением $R_c = C/\left[1 - H_b(P_b)\right]$. Учитывая зависимость C от E_b/N_0, можно построить график от E_b/N_0 (рис. 4). График при отсутствии кодирования получается из этого же уравнения при $R_c = 1$ с учётом $C = 1 - H_b(P_c)$, откуда следует очевидный результат: $P_b = P_c = Q\left(\sqrt{2E_b/N_0}\right)$. При требовании $P_b = 10^{-5}$ без кодирования $E_b/N_0 = 9.6$ дБ. С кодированием со скоростью $R_c = 0.5$ это значение можно уменьшить примерно на 9.4 дБ для мягких решений и на 9.4-1.6=7.8 дБ для жёстких решений. Это и есть предельное значение

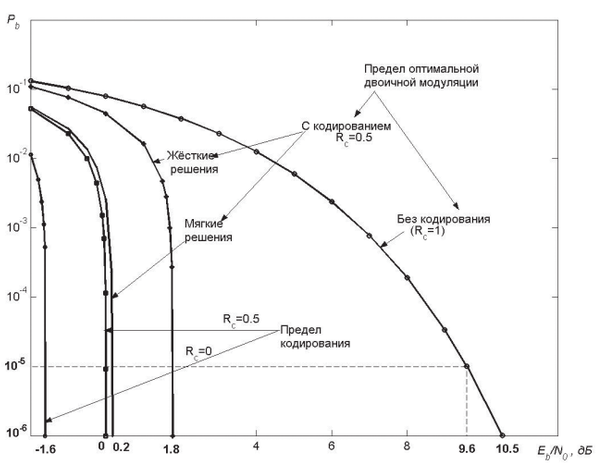


Рис. 4. Иллюстрация предельного энергетического выигрыша от кодирования.

ЭВК.

Ограничение на конечный размер информационного пакета

Метод передачи без кодирования можно рассматривать как оптимальный для передачи информации при ограничении числа бит $K = 1$ в сеансе связи. Если, к примеру, нужно передать только одну команду со значением "да" (значение бита равно 1) или "нет" (значение бита равно 0) за время T_b, то оптимальным является метод передачи противоположными сигналами. Это оптимальный метод передачи в смысле минимума вероятности ошибки P_b, и кодированием её нельзя уменьшить. Схожая постановка вопроса возможна и при ограничении на время доставки $K > 1$ бит получателю информации. При этом эффективность кодирования уменьшается в сравнении с рассмотренными предельными возможностями.

Действительно, предельные возможности кодирования предполагают, что информационные биты могут быть переданы пакетами любого размера K. Более того, доказательство теоремы Шеннона основано на том, что именно за счёт увеличения K можно добиться сколь угодно малого значения P_b при некоторой величине E_b/N_0. Однако если такой возможности на практике нет и величина K ограничена максимально-допустимой задержкой приёма пакета, то значение E_b/N_0 увеличивается. Иллюстрация этого положения приведена на рис. 5. Видно, что при $K \simeq 10^6$ (пакет из миллиона бит) энергетическая эффективность ухудшается несущественно. Однако при $K = 10^2 - 10^3 - 10^4$ ухудшение эффективности составляет 2.5-1-0.4 дБ ($R_c = 0.5$). Это существенная поправка к границе Шеннона, и её следует иметь в виду для оценки эффективности реальных кодов.

Из этого анализа также следует весьма известный вывод: до-

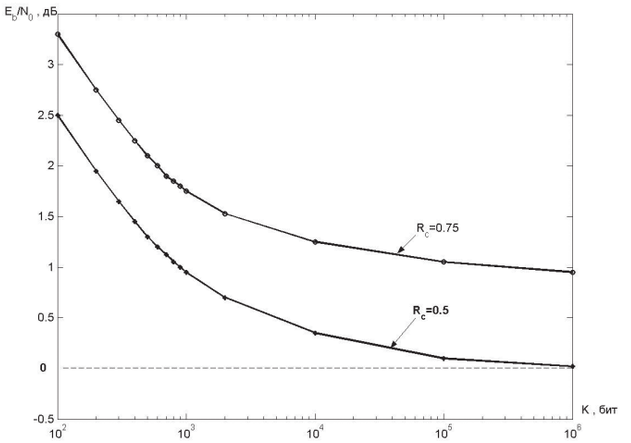


Рис. 5. Предельная энергетическая эффективность в зависимости от размера информационного пакета .

биться заданной близости к границе Шеннона можно лишь кодами с весьма большими информационными блоками (пакетами). Для этого, в свою очередь, первостепенное значение имеют эффективные алгоритмы декодирования. Отсутствие таких алгоритмов даже при наличии "хороших" кодов делает бесполезными для практики такие коды.

Хорошие коды

Для приближения к границе Шеннона реальными кодами недостаточно их строить лишь исходя из классического критерия "хорошего" кодового расстояния. Эта мысль поясняется в этом разделе на примере ДСК и наиболее простых и распространённых на практике двоичных линейных кодов. Однако основной вывод раздела применим и к рассмотренным выше информационным каналам, и к произвольным линейным кодам. Материал раздела в основном базируется на [5].

На рис. 6 приведена классическая иллюстрация корректирующей способности двоичного линейного (N, K) кода, где $N = K/R_c$. Точками изображаются два кодовых слова. Расстояние Хэмминга между ними равно минимальному кодовому расстоянию $d_{\min}$. Вокруг этих точек – сферы радиуса $t \simeq d_{\min}/2$. Величина t представляет собой корректирующую способность кода, а сфера является сферой декодирования (иногда её называют сферой Хэмминга). Это означает следующее. Если число ошибок e в кодовом слове размера N не превосходит значения t, декодирование по алгоритму максимального правдоподобия производится без ошибок. Декодер, гарантированно исправляющий t ошибок, называется декодером с ограниченным расстоянием (bounded – distance decoder). Такого декодера достаточно для получения произвольно малой вероятности P_b путём увеличения $d_{\min}$. По этой причине в классической теории помехоустойчивого кодирования значительное внимание уделяется построению кодов с возможно

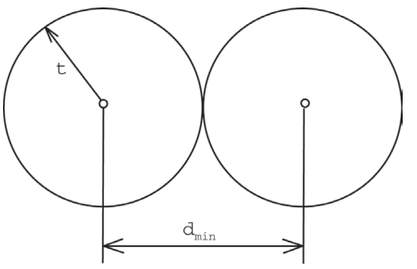


Рис. 6. Корректирующая способность и минимальное расстояние кода .

большим значением $d_{\min}$. Считается, что код имеет "хорошее" расстояние, если отношение $d_{\min}/N$ стремится к положительной величине при неограниченном увеличении N. К таким кодам относятся случайные линейные коды. Согласно теореме Варшамова–Гильберта, при достаточно большом N не существует двоичного кода со скоростью R_c, имеющего кодовое расстояние больше, чем расстояние, которое находится из уравнения $R_c = 1 - H_b(d_{\min}/N)$. Случайные линейные коды удовлетворяют этой теореме. Коды, для которых отношение $d_{\min}/N$ стремится к нулю с ростом N, следует отнести к кодам с "плохим" кодовым расстоянием. Коды с постоянной величиной $d_{\min}$, независимо от N, – к кодам с "очень плохим" кодовым расстоянием. Примером такого кода является классический код Хэмминга, для которого $d_{\min} = 3$.

Построение кодов с "хорошим" кодовым расстоянием целесообразно при условии $d_{\min}/2 > e$, где e – некое среднее число ошибок в кодовом слове размера N. Однако вблизи границы Шеннона, когда $R_c \approx 1 - H_b(P_c)$, это условие не выполняется. Для случайных кодов $d_{\min} \approx P_c N$ и $e \approx P_c N$! В результате сферы радиуса e вокруг кодовых слов перекрываются (рис. 7). Отсутствие перекрытия этих сфер наблюдается лишь до вероятности ошибки $P_c/2$, где P_c – предельно допустимая вероятность ошибки согласно теореме Шеннона. Иначе говоря, декодер с ограниченным расстоянием допускает в два раза большую вероятность ошибки, чем допускает граница Шеннона.

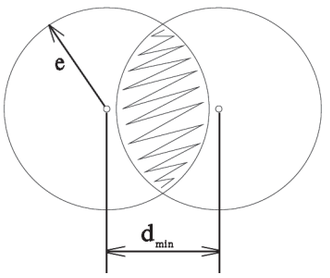


Рис. 7. Эффект перекрытия e – сфер вблизи границы Шеннона.

Итак, с точки зрения классической теории помехоустойчивого кодирования, декодирование вблизи границы Шеннона сводится к декодированию за пределом корректирующей способности кода. При этом минимальное кодовое расстояние не является столь фундаментальным понятием, каким оно является при относительно малом уровне шума. Для достижения произвольно малой вероятности ошибки P_b первостепенное значение имеет распределение числа кодовых слов заданного веса. Среди этого распределения главную роль играет число кодовых слов малого веса.

В результате обретает чёткость понятие "хороший" код. С помощью такого кода на некоторой ненулевой скорости R_c можно добиться произвольно малой вероятности ошибки P_b.

Для практических целей коды часто строятся в виде кодовых конструкций из нескольких простых кодов. Такой подход позволяет использовать эффективные в вычислительном отношении алгоритмы декодирования. В этой связи рассмотрим пример кодовой конструкции, иллюстрирующий возможность построения "хорошего" кода и обладающий "плохим" кодовым расстоянием [5]. Конструкция представляет собой код – произведение из последовательности простых кодов Хэмминга (N_i, K_i), где $N_i = 2^{M_i} - 1$, $K_i = N_i - M_i$, $M_i = 2,3,4,5...$. С ростом числа кодов Хэмминга в этой последовательности длина составного кода оказывается равной $N = 3,21,315,6125...$ При этом скорость составного кода R_c стремится к величине 0.093. В целом конструкция имеет "плохое" кодовое расстояние, поскольку $d_{\min}/N$ стремится к нулю. Однако анализ зависимости P_b от скорости R_c свидетельствует, что вероятность ошибки P_b стремится к нулю (рис. 8, где $P_c = 0.0588$). Поэтому кодовая конструкция представляет собой "хороший" код, хотя и "далёкий" от границы Шеннона приведенной на рисунке справа (график построен аналогично рис. 2.).

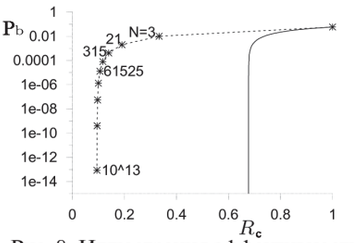


Рис. 8. Иллюстрация эффективности "хорошего" кода, представляющего кодовую конструкцию из кодов Хэмминга.

Примером "плохого" кода является тривиальный код с повторением. Этим кодом можно добиться произвольно малой вероятности ошибки P_b лишь ценой уменьшения скорости кода R_c до нуля, что эквивалентно нулевой скорости передачи информации.

В следующем разделе рассмотрены идеи построения линейных кодов, с помощью которых удалось значительно приблизиться к границе Шеннона для гауссовского канала. Их характеристики приближаются к характеристикам "очень хороших" кодов.

Идеи

Турбо-коды

Схема турбо-кодирования позволила получить значение $P_b = 10^{-5}$ при $E_b/N_0 = 0.7$ дБ для скорости кода $R_c = 0.5$ для информационного блока из $K = 6.5 \cdot 10^4$ бит. Как следует из предыдущего раздела, это значение лишь на 0.7-0.2=0.5 дБ уступает оптимальной двоичной схемой кодирования. Результат получен с использованием алгоритма декодирования мягких решений. Тем не менее, для наглядности его полезно сформулировать следующим образом: при вероятности ошибки кодового символа на входе декодера $P_c = Q\left(\sqrt{2R_c E_b/N_0}\right) \approx 0.14$ вероятность ошибки информационного бита на выходе декодера $P_b = 10^{-5}$. Принципиален алгоритм декодирования, с помощью которого получен такой уникальный результат. Поучительна и его история.

Турбо-код формируется кодовой конструкцией из двух параллельно включённых простых свёрточных кодеров и псевдослучайного перемежителя [1]. Кодовое слово можно разделить на два систематических кодовых слова, информационные части которых идентичны с точностью до перемежения. Поэтому для декодирования могут быть использованы два последовательно включённых свёрточных декодера, разделённые перемежителем. При этом второй декодер использует результат оценки правдоподобий информационных символов первого декодера в качестве априорной информации для декодирования. Затем первый декодер использует результат оценки правдоподобий информационных символов второго декодера в качестве априорной информации и т.д. В результате приходим к итеративной процедуре уточнения правдоподобий информационных бит, в которой декодеры обмениваются между собой сообщениями. Подобные обменные вероятностные алгорит-

мы декодирования в зарубежной литературе именуются термином Message Passing Algorithms (MPA). Процедура обмена прекращается принудительно, результатом декодирования являются жёсткие решения на выходе второго декодера.

Возможна эффективная реализация декодеров, участвующих в обменном алгоритме. Действительно, турбо-код представляет собой блоковый код, формирование которого производится с помощью свёрточных кодеров с небольшим, в сравнении с длиной кодового слова N, числом состояний S кодовой решётки. В начале и в конце информационного пакета производится сброс памяти декодеров, тем самым начальный и конечный узлы решётки оказываются заданными. Использование свёрточных кодеров в таком режиме называют режимом терминирования. Поэтому для декодирования в принципе можно применить эффективный алгоритм декодирования Витерби нахождения максимально-правдоподобного пути на кодовой решётке между двумя заданными узлами. Для его использования в итеративной обменной вероятностной процедуре требуется модификация, заключающаяся в замене жёстких решений правдоподобиями (мягкими решениями) информационных бит. Такой модифицированный алгоритм Витерби (Soft Output Viterbi Algorithm, SOVA) был разработан за несколько лет до изобретения турбо-кодов.

Однако при декодировании турбо-кодов такой алгоритм не позволил приблизиться к границе Шеннона столь близко, как указано выше. Причина заключается в том, что алгоритм Витерби основан на критерии максимального правдоподобия кодового слова и минимизирует вероятность ошибки кодового слова P_w. Добиться указанного выше результата позволил алгоритм на основе критерия максимального правдоподобия отдельного информационного бита. Этот алгоритм минимизирует вероятность информационного бита P_b. На научном языке оба алгоритма декодирования являются алгоритмами на основе критерия максимума апостериорной вероятности (Maximum A Posteriori, MAP): первый – кодового слова, второй – информационного бита (Bit-By-Bit Decoding) [6]. В зарубежной литературе последний алгоритм связывают с Балом и др. (Bahl, Cocke, Jelinek, Raviv, 1974 г) и часто называют "the BCJR algorithm" по первым буквам фамилий авторов.

Из-за большей вычислительной сложности, в сравнении с классическим алгоритмом Витерби он был мало востребован. И вот почти через 20 лет создатели турбо-кодов (Berrou, Glavieux, Thitimajshima) модифицировали алгоритм и использовали его для декодирования турбо-кодов. Модификация алгоритма Бала для обменной вероятностной процедуры заключалась в замене жёстких решений правдоподобиями информационных бит, как и для алгоритма SOVA. Такой алгоритм иногда называют просто алгоритмом вычисления апостериорных вероятностей бит (A Posteriori Probability algorithm, APP algorithm)[1]. Также этот алгоритм называют "Belief Propagation Algorithm (BPA, алгоритм с распространением доверия)", или "the Sum-Product Algorithm (SPA, алгоритм сумма произведений)".

Вычисление апостериорных вероятностей в алгоритме производится с использованием двойного прохода кодовой решётки в прямом и обратном направлении ("the Forward-Backward Algorithm"). Поэтому он может быть использован только для декодирования кодовых последовательностей конечной длины. Для этого и производится терминирование свёрточных кодеров.

Существует несколько версий этого алгоритма. На практике удобнее максимизировать не апостериорную вероятность, а её логарифм. При этом используется формула $\log\left(e^a + e^b\right) = \max\left(a,b\right) + \log\left(1 + e^{|a-b|}\right)$. Эту версию алгоритма называют Log-MAP (или, к примеру, log-SPA). Ещё более простую, но подоптимальную версию алгоритма с аппроксимацией $\log\left(e^a + e^b\right) \approx \max\left(a,b\right)$ называют Max-Log-MAP алгоритмом.

Однако не только алгоритм декодирования позволил приблизиться к границе Шеннона. Использование систематического свёрточного кодера также является принципиальным моментом в схеме турбо-кодирования. В этом существенное отличие от оптимальных свёрточных кодов, декодируемых по алгоритму Витерби. Этими оптимальными кодами (в смысле максимума кодового расстояния) являются несистематические нерекурсивные коды. К примеру, при числе состояний кодера $S = 4,8,16$ оптимальные полиномы кодера в восьмеричном представлении имеют вид $\left(5,7\right)$, $\left(15,17\right)$ и $\left(23,35\right)$ соответственно. Несмотря на то, что несистематический нерекурсивный кодер с полиномами вида $\left(a,b\right)$ и систематический рекурсивный кодер с полиномами вида $\left(1,a/b\right)$ формируют одно и то же множество кодовых слов [5], отображение информационного слова в кодовое слово для них различно. Это различие как раз и было использовано при оптимизации кода по критерию минимума вероятности ошибки P_b путём перераспределения числа кодовых слов заданного веса и минимизации числа кодовых слов с минимальным весом, на необходимость которого было указано в разделе "Теория/Хорошие коды". В результате были найдены оптимальные свёрточные коды для турбо-схемы. К примеру, при числе состояний кодера $S = 4,8,16$ оптимальные полиномы кодера имеют вид $\left(1,5/7\right)$, $\left(1,15/17\right)$ и $\left(1,21/37\right)$ соответственно. Турбо-кодирование с полиномами свёрточного кодера $\left(1,21/37\right)$ было использовано при $Q = 10$ итерациях декодирования для получения результата, приведённого в начале этого раздела.

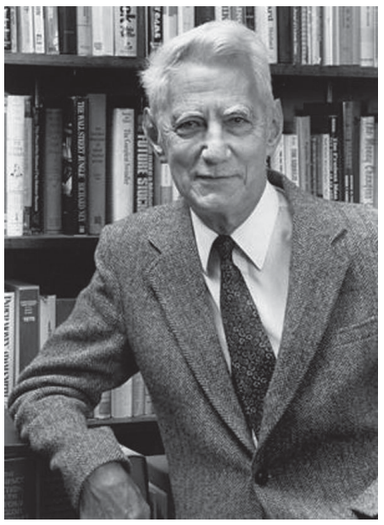
Коды с мажоритарным декодированием

Теоретически сформулировать алгоритм MAP для информационного символа не представляет труда для любого кода. Алгоритм, однако, в общем случае оказывается более сложным в реализации, чем алгоритм максимального правдоподобия кодового слова[2]. Однако для некоторых кодов возможна подоптимальная реализация алгоритма, а и иногда итеративная процедура, сходящаяся к оптимальному решению.

Таким алгоритмом, к примеру, является алгоритм голосования по большинству или, как его ещё называют, мажоритарный алгоритм. Идея весьма проста. Предположим, что каждый информационный бит

1. Различие по энергетической эффективности алгоритмов APP и SOVA незначительно лишь при относительно небольших размерах $K \approx 10^3$ информационных пакетов, при которых невозможно приблизиться к границе Шеннона (см. результаты моделирования на www.ee.vt.edu/%7Eyufei/turbo.html). При больших значениях оно может достигать 1 дБ.

2. В терминологии Л.М. Финка [4] алгоритм декодирования на основе критерия максимального правдоподобия кодового слова с мягкими решениями называется "приёмом в целом", в отличие от "поэлементного приёма", под которым в статье понимается алгоритм декодирования на основе критерия максимального правдоподобия жёстких решений демодулятора.



Клод Шеннон

может быть представлен несколькими уравнениями, в которые входят символы кодового слова (в общем случае как информационные, так и проверочные). Тогда решение о том, что бит равен единице (нулю) принимается в случае предпочтения единицы (нуля) в большинстве уравнений для этого бита.

Мажоритарный алгоритм декодирования может быть использован далеко не для всех кодов. Тем не менее, классические блоковые коды Хэмминга, расширенные одной проверкой на чётность, могут быть декодированы мажоритарным алгоритмом. Для такого кода каждый информационный бит может быть представлен четырьмя проверочными уравнениями, поэтому решение о равенстве единицы (нуля) значения бита принимается при наличии не менее чем трех голосов в пользу единицы (нуля). В терминологии мажоритарного декодирования, три голоса – это порог. Поэтому такое декодирование также называют пороговым. Расширенные проверкой на чётность коды Хэмминга принадлежат более широкому классу блоковых кодов Рида–Маллера. Алгоритм декодирования Рида для этих кодов представляет собой итеративный алгоритм мажоритарного декодирования.

Расширенный проверкой на чётность код Хэмминга $(N,K)=(8,4)$ также является примером систематического самоортогонального кода (СОК, Self-Dual Code, SDC). Это означает чёткое разделение информационных и проверочных бит при мажоритарном декодировании. При этом информационный бит представлен проверочными уравнениями, в которые входят только проверочные символы (информационные в эти уравнения не входят). Или иначе: в каждое из $M=K$ уравнений проверок на чётность входит лишь один информационный бит. На языке теории кодирования это означает, что порождающая матрица кода G (размерности $K\times N$) совпадает с проверочной матрицей H (размерности $M\times N$, где $M=N-K$ – число проверок на чётность). Скорость такого кода имеет одно значение $R_c=0.5$.

Теория порогового декодирования была разработана Дж. Месси в 70-х годах прошлого века. Значителен вклад в теорию и практику отечественной школы [2, 4, www.mtdbest.iki.rssi.ru]. Разработанный В.В. Золотарёвым и др. мажоритарный алгоритм (алгоритму более 20 лет) многопорогового декодирования (МПД) свёрточных СОК является итеративным алгоритмом MAP информационных бит [6]. Критерием окончания итеративной процедуры является равенство всех проверок на чётность, что эквивалентно равенству нулю вектора синдрома $s=Hy$, вычисляемого на каждой итерации, где y – обработанный на итерации кодовый вектор. Поэтому процедуру декодирования можно рассматривать как итеративное решение синдромного уравнения относительно информационных бит. Принципиально возможно декодирование как жёстких, так и мягких решений демодулятора. В то же время алгоритм МПД особенно прост при декодировании жёстких решений.

Судя по [6], эффективность мажоритарного декодирования не уступает декодированию по классическому алгоритму Витерби. Кодовая конструкция из внутреннего мажоритарно декодируемого свёрточного СОК при жёстких решениях демодулятора и простейшего внешнего кода проверки на чётность обеспечивает значение $P_b=10^{-5}$ при $E_b/N_0=2.2$ дБ при задержке в $K=4\cdot10^4$ бит. Анализируя рис. 4,1 видим, что энергетический проигрыш такой кодовой конструкции в сравнении с оптимальной схемой декодирования жёстких решений составляет 2.2-1.8=0.4 дБ. Думается, что этот результат заслуживает внимания.

Коды с низкой плотностью проверок на чётность

Существует мнение, что на сегодняшний день коды с низкой плотностью проверок на чётность (Low Density Parity Code, LDPC) являются лучшими блоковыми кодами [5]. На это есть основания.

1. Коды являются "хорошими" (см. раздел Теория/Хорошие коды).

2. Коды имеют "хорошее" кодовое расстояние (см. раздел Теория/Хорошие коды).

3. Существует эффективный обменный вероятностный алгоритм декодирования.

С помощью одного из таких кодов удалось получить значение $P_b=10^{-5}$ при $E_b/N_0=0.0045$ дБ и скорости кода $R_c=0.5 (N=10^7)$.

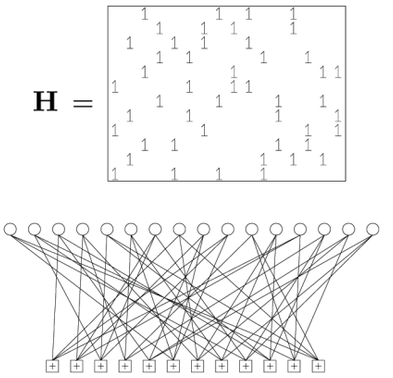


Рис. 9. Проверочная матрица с низкой плотностью проверок на чётность и соответствующий ей проверочный граф.

На рис. 9 приведён пример проверочной матрицы с низкой плотностью проверок на чётность (единиц в проверочной матрице) для кода $(N,K)=(16,4)$. Матрица имеет размерность $M\times N=12\times16$, в каждом столбце матрицы – $j=3$ единицы, в каждой строке – $k=4$ единицы. Это означает, что каждый принятый символ кодового слова участвует в $j=3$ проверках на чётность, а в каждой проверке на чётность участвуют $k=4$ принятых символов.

Для практики важны коды с очень низкой плотностью проверок на чётность, для которых $j\ll M$ и $k\ll N$. Очень низкая плотность значительно снижает вычислительные затраты на реализацию алгоритма декодирования при больших размерах матриц. Таким примером является код $(N,K)=(2\cdot10^4,10^4)$, имеющий проверочную матрицу с параметрами $j=3$, $k=6$.

Заметим, что для подобных кодов

$$R_c=K/N=(N-M)/N=1-j/k.$$

История этих кодов, как и турбо-кодов, весьма поучительна. Этим кодам более 40 лет, их изобретателем является Р. Галлагер, и приведённая на рис. 9 матрица построена в соответствии с псевдослучайным алгоритмом Галлагера. "Вспомнили" об этих кодах лишь в середине 90-х годов после открытия турбо-кодов. Почему? Потому что для этих кодов Галлагером была предложена итеративная обменная вероятностная процедура декодирования.

Применительно к обработке жёстких решений сущность этой процедуры сводится к следующему. На каждой итерации проверяются соотношения на чётность в соответствии с проверочной матрицей кода H. После первой проверки исправляется символ (ноль заменятся на единицу и наоборот), входящий в наибольшее число невыполненных проверочных соотношений на чётность, после чего проверка производится повторно, пока все соотношения не будут удовлетворены. Это "наибольшее число невыполненных проверочных соотношений" может задаваться порогом. Не правда ли, в этом есть что-то от многопорогового декодирования?

Интерпретация обменного алгоритма декодирования особенно наглядна на проверочном графе кода, который представлен на этом же рис. 9. Этот граф также называют графом Таннера. Он имеет две группы проверочных узлов (Bipartite Graph). Первый набор представляет правдоподобия принятых N кодовых символов. Второй представляет правдоподобия M проверок на чётность. Итеративная обменная процедура декодирования строится на основе обмена правдоподобий между двумя группами узлов. Каждый принятый узел обменивается информацией с проверочным узлом. В качестве начальной информации для алгоритма используются мягкие решения принятых символов. Декодирование заканчивается при условии выполнения всех уравнений проверок на чётность, что эквивалентно нулевому вектору синдрома.

Отметим, что для рассмотренных кодов порождающая матрица G в корне отличается от проверочной матрицы (в противоположность СОК) и имеет высокую плотность единиц (High Density). По этой причине для LDPC существует проблема эффективного кодирования.

В настоящее время, помимо алгоритма Галлагера, предложено большое число алгоритмов построения матриц с рассмотренным выше свойством. Отметим алгоритмы, разработанные Маккеем (J. MacKay). На его web-странице (www.inference.phy.cam.ac.uk/mackay) демонстрируется итеративное декодирование упомянутого в начале этого раздела кода с очень низкой плотностью проверок на чётность.

В заключение раздела заметим, что для любого блокового кода существует проверочная матрица. В этой связи вопрос: что представляет собой матрица турбо-кода? Она имеет явно низкую плотностью проверок на чётность (единиц), хотя структура её, естественно, отличается от матриц Галлагера.

Прочие кодовые конструкции

После создания классической конструкции турбо-кодов из параллельных свёрточных кодеров была исследована и последовательная конструкция. По эффективности в целом она несколько уступает классической.

К турбо-подобным кодам сегодня относят и коды с повторением и накоплением (Repeat-Accumulate Codes, RAC). По сути эти коды сочетают в себе идею тривиальных кодов повторения с рекурсивной процедурой вычисления проверочных символов путём накопления информационных бит. Обменный алгоритм декодирования может быть проинтерпретирован на проверочном графе. Проверочная матрица кода имеет низкую плотность проверок на чётность.

Появилась и конструкция из простых блоковых кодов, которую стали именовать "блоковыми турбо-кодами" (Block Turbo Codes, BTC). Эти конструкции также получили название матричных, или двумерных турбо-кодов произведений (2-D Turbo Product Coduct, 2D-TPC).

Конструкция кодирования в виде произведения кодов давно известна. Информационное слово размещается в матрице. Далее кодируется каждая строка, а затем каждый столбец. Кодовое слово длины N формируется последовательным считыванием строк матрицы. При декодировании каждый информационный бит участвует в проверках на четность, как по строкам, так и по столбцам. Этот факт и используется в итеративной вероятностной обменной процедуре.

В этой связи отметим, что идея обменной процедуры для подобных кодов известна, по меньшей мере, 50 лет. Принадлежит она П. Элайесу, который ввел итерированные коды. Им, кстати, доказано, что эти коды являются "хорошими".

Практика

Турбо-коды

Немаловажным значением для распространения этих кодов является тот факт, что компаниями France Telecom и T'el'ediffusion de France в США запатентован широкий класс турбо-кодов (US Patent 5,446,747).

Ставшая классической схема турбо-кодирования с числом состояний свёрточного кодера $S = 16$ утверждена американским комитетом CCSDS (Consultative Committee for Space Data Systems) в стандарте передачи телеметрической информации с космических аппаратов. Ранее в CCSDS использовалась традиционная каскадная схема помехоустойчивого кодирования с внутренним свёрточным кодированием и внешним кодом Рида–Соломона. В сравнении со старой схемой кодирования новая схема позволила добиться энергетического выигрыша на 1.5-2.8 дБ.

Турбо-код с $S = 8$ состояниями утверждён несколькими стандартами связи. О нём заявлено в американском стандарте 3G CDMA2000 для высокоскоростного режима передачи информации (больше 14.4 Кбит/с), как к абоненту (forward link), так и от абонента (reverse link). Этот же код рассматривается европейским стандартом 3G UMTS для высокоскоростного режима передачи информации (больше 32 Кбит/с) и приёма с высоким качеством ($P_b \approx 10^{-6}$).

Консорциум DVB утвердил турбо-коды в стандарте DVB-RCS (ETSI EN 301 790) для передачи информации по обратному спутниковому каналу (Return Channel for Satellite-RCS) в направлении от спутника к абоненту. Код формируются на основе циклического рекурсивного систематического свёрточного кодера. Основным требованием для спутниковых интерактивных систем является маленькая задержка. Поэтому при кодировании используется относительно небольшой размер информационного пакета, равный примерно $K \approx 10^2 - 10^3$ бит (16-188 байт).

Компанией TurboConcept, в партнёрстве с европейским спутниковым оператором Eutelsat, разработан турбо-декодер TC1000 в соответствии со стандартом DVB-RCS, использующий подоптимальный алгоритм Max-Log-MAP. Декодер обеспечивает величину $P_b = 10^{-8}$ при $E_b/N_0 = 5.1 - 2.5$ дБ при $Q = 5$ итерациях ($R_c = 0.5$).

Использование турбо-кодов принято также в стандарте спутниковой системы связи Inmarsat.

Коды с низкой плотностью проверок на чётность

По-видимому, стандарт DVB-S2 (ETSI EN 302 307 v.1.1.1 2005-03) стал первым международным стандартом, взявшим на вооружение коды с низкой плотностью проверок на чётность. Были попытки внедрить в стандарт классические турбо-коды. Однако благодаря настойчивости компании Hughes LDPC были утверждены в стандарте.

Стандартом требования к помехоустойчивости определены величиной $P_b = 10^{-7}$. Это значение достигается каскадной конструкцией блокового LDPC и кода БЧХ. Код LDPC может иметь несколько кодовых скоростей R_c и только две длины кодового слова $N = 64800$ и $N = 16200$.

Поясним идеологию этой кодовой конструкции. Для декодирования LDPC стандартом рекомендовано использование итеративной обменной вероятностной процедуры на проверочном графе (см. раздел Идеи/Коды с низкой плотностью проверок на чётность). Учитывается, что при достаточно низком отношении сигнал-шум не все уравнения системы проверок на чётность могут быть выполнены. Поэтому процесс декодирования может быть принудительно прерван через приемлемое число итераций. Анализ показывает, что при этом ошибки в декодированном слове имеют тенденцию группирования в одной ограниченной области кодового слова. Такую группу называют пакетом ошибок. Для обнаружения пакетов ошибок и исправлений могут быть использованы простые циклические коды, поскольку любой циклический (n,k) код способен обнаруживать все пакеты ошибок из $r = n - k$ символов и меньше. Более того, циклические коды способны обнаруживать большую часть пакетов ошибок, длина которых превосходит r. Именно поэтому "вычищение" пакетов ошибок на выходе кода LDPC производится циклическим кодом БЧХ, позволяющим не только обнаруживать, но и гарантированно исправлять пакеты из $t = 8 - 12$ ошибок.

В стандарте приведены результаты имитационного компьютерного моделирования для гауссовского канала: требуемое значение $P_b = 10^{-7}$ достигается при $E_b/N_0 = 1.05$ дБ для большой длины блока $N = 64800$ ($R_c = 0.5$)[3]. Для малой длины блока $N = 16200$ дополнительные энергетические затраты не превышают 0.3 дБ, в результате чего $E_b/N_0 = 1.3$ дБ. В обоих случаях для декодирования LDPC использовался вероятностный обменный алгоритм с числом итераций $Q = 50$.

В настоящее время уже известно несколько декодеров и ресиверов стандарта DVB-S2. Компанией TurboConcept разработана серия декодеров TC4xxx кодов LDPC с числом итераций $Q \le 60$. Компанией AHA разработан декодер AHA4702. Компанией Efficient Channel Coding (ECC) разработана полнофункциональная платформа тестирования приёма сигналов стандарта DVB-S2. Для своего приёмного оборудования компанией приводится цифра энергетической эффективности: требуемое значение $P_b = 10^{-7}$ достигается при $E_b/N_0 = 2.1$ дБ ($R_c = 0.5$, QPSK). Отметим, что она отличается примерно на 1 дБ от результатов моделирования, приведённых в стандарте.

В марте 2005 г. компания ECC, совместно с оператором спутниковой связи компанией Telesat, впервые в мире продемонстрировали приём сигнала в стандарте DVB-S2.

Прочие коды и кодовые конструкции

Для свёрточных кодов с мажоритарным декодированием в Научно Исследовательском Институте Радио (НИИР) разрабатывается уже шестое поколение декодеров. О внедрении этих кодов в международных стандартах информация отсутствует.

Блоковые турбо-коды утверждены в стандарте IEEE 802.16a. Для кодирования строк и столбцов используются простые коды с двойной проверкой на чётность и расширенные проверкой на чётность коды Хэмминга длины $n = 64, 128, 256$. Для декодирования кода использует итеративный алгоритм. На каждой итерации производится обмен величин правдоподобий между декодером двойной проверки на чётность и расширенным кодом Хэмминга.

Стандартом предусмотрены несколько вариантов кодирования для низкой и высокой скорости в сочетании с низкой и высокой стоимостью декодеров. Это достигается сочетанием различных значений R_c и длины кодового слова $N = 4048\text{-}65536$. При этом, естественно, различна и энергетическая эффективность кода. К примеру, для высокоскоростного режима $R_c = 0.793 - 0.893$ и $E_b/N_0 = 3.8 - 5$ дБ, для низкоскоростного $R_c = 0.653 - 0.872$ и $E_b/N_0 = 2.9 - 4.1$ дБ.

Компанией TurboConcept разработано семейство декодеров TC3xxx с числом итераций $Q = 2 - 16$. Семейство состоит из низкоскоростных декодеров TC30xx и высокоскоростных TC34xx. Декодеры дополнительно к требованиям стандарта позволяют использовать в кодовой конструкции вместо расширенных проверкой на чётность кодов Хэмминга коды БЧХ, исправляющие две ошибки ($t = 2$). Компанией Comtech AHA Corporation для стандарта IEEE 802.16a также выпущено семейство кодеков AHA4525. Для декодирования используются $Q = 3$ итерации.

Литература

1. В.А. Варгаузин, Л.Н. Протопопов. Турбо-коды и итеративное декодирование: принципы, свойства, применение // ТелеМультиМедиа. 2000. №34. С.33-38.

2. Зубарев Ю. Б., Золотарёв В.В., Жуков С. Е., Строков В.В., Овечкин Г.В. Многопороговые декодеры для высокоскоростных спутниковых каналов связи: новые перспективы // Электросвязь. 2005. №2. С.10-12.

3. Дж. Проакис. Цифровая связь. М.: Радио и связь, 2000.

4. Л.Н. Финк. Теория передачи дискретных сообщений. М.: Издво "Советское радио". 1970.

5. David J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press. 2003.

6. A.J. Viterbi, An Intuitive Justification and a Simplifed Implementation of the MAP Decoder for Convolution Codes, IEEE Journal On Selected Areas In Communications. Vol.16. No.2. February 1998.

3. На протяжении всей статьи подразумевается модуляция BPSK. Этот результат получен при модуляции QPSK, которая обладает такой же энергетической эффективностью, что и BPSK, но в два раза лучшей частотной эффективностью. Поэтому такой результат может быть использован для сравнения с другими приведёнными в статье цифрами энергетической эффективности.