

V. I. Lutin¹, Senior Scientist, e-mail: lyutin_v_i@mail.ru,

F. A. Desyatirikov², Student, e-mail: felixdes@ieee.org,

A. S. Belyaev³, Developer, e-mail: belyaev.andrey@gmail.com,

E. V. Lutina⁴, Engineer, e-mail: lyutin@mail.ru,

E. N. Desyatirikova², Professor, e-mail: science2000@ya.ru

¹Military Educational and Scientific center of the Air Force “The Air Force Academy named after Professor Zhukovsky and Yuri Gagarin”, Voronezh, Russia

²Voronezh State University, Voronezh, Russia

³JetBrains, Belgrade, Serbia

⁴Voronezh State Technical University, Voronezh, Russia

Synthesis of an Algorithm for Searching a Given Segment in a Data Sequence of Finite Length

Under conditions of an unknown shift of the reference points of a continuous signal and its given fragment, an approach to the synthesis of a fragment search algorithm is proposed. The approach is based on the classical apparatus for testing hypotheses of mathematical statistics. A distinctive feature of the approach is taking into account the noise of the signal and fragment reading, and on this basis, the formalization of the problem of choosing the number of channels for monitoring the signal. Thus, in this work, using a probabilistic-statistical method, we synthesized an algorithm for a two-stage procedure for searching for a given segment in a long data sequence, which consists in determining the most probable position of the desired segment and checking whether the found segment matches the given one. Analytical relations are obtained to assess the quality of detection, which take into account the autocorrelation properties of sequences, and detection characteristics are calculated, which determine the sample size of a given segment and a sufficient number of detector channels to ensure the required quality of detection under given observation conditions. A block diagram of a device for making decisions about the most probable hypothesis and determining the average value and standard deviation of the

start time of the desired segment is proposed, which allows taking into account the least favorable conditions for the beginning of the signal countdown, namely, exactly in the middle of the specified fragment.

Keywords: *Deep Packet Inspection (DPI), Neyman-Pearson Criterion, Bayes Criterion, statistically optimal algorithm*

Введение

В задачах поиска заданного сегмента в длинной последовательности данных не всегда представляется возможным применить простое совпадение. Например, при навигации летательных аппаратов по рельефу земной поверхности невозможно получить последовательность отсчётов высот, строго совпадающую с известными высотами, имеющимися в геоинформационных системах. При автоматическом преобразовании речи человека в последовательность букв также невозможно получить строгое совпадение отсчётов наблюдаемого процесса с отсчётами эталонных последовательностей. Аналогичные задачи решаются и в целях обеспечения информационной безопасности. Так, для последующего анализа текущего по сети трафика необходимо сохранять всю полезную нагрузку, что неудобно. Поэтому необходим инструментарий, позволяющий решить основную задачу глубокой проверки трафика (DPI-Deep Packet Inspection), а именно, определение протокола прикладного уровня, на основе небольшого количества информации. В соответствующих задачах анализа протокола прикладного уровня объектом является поток трафика транспортного уровня, в котором различные приложения генерируют потоки с различными статистическими характеристиками. По ограниченному набору статистических метрик, таким образом, можно определить вклад конкретного приложения, а значит, и протокола. В связи с этим исследование принципа построения систем поиска и обнаружения заданного сегмента в длинной последовательности данных, разработка оптимального алгоритма и методики анализа качества поиска заданного сегмента данных являются актуальными.

Аналогом задаче поиска заданного сегмента в длинной последовательности данных является задача обнаружения сигналов, которая обычно решается с применением методов, инвариантных ко времени прихода в точку приёма сигнала с известным аналитическим описанием, основанных на принципах согласованной фильтрации [1, 2]. В рассматриваемой задаче сигнал задаётся последовательностью отсчётов, поэтому алгоритм синтезируется с применением теории различения статистических гипотез.

Цель работы – повышение эффективности и качества автоматических систем поиска заданного сегмента в длинной последовательности данных.

Для достижения поставленной цели проводится синтез и анализ оптимального алгоритма принятия решений с применением вероятностно-статистических методов, что указывает на привлечение элементов искусственного интеллекта при обработке последовательностей данных.

Постановка задачи

Искомый сегмент находится в последовательности чисел, полученных в результате взятия отсчётов мгновенных значений непрерывного процесса $x(t)$:

$$\vec{X} = [x_1, x_2, \dots, x_I]^T. \quad (1)$$

Пусть среднее значение последовательности чисел (математическое ожидание) m равно нулю, дисперсия равна σ_0^2 .

Непрерывный процесс $x(t)$ – это может быть сигнал, получаемый от видеокамеры, формирующей изображение, или от микрофона, преобразующего речь человека в электрическое колебание. Все сигналы, используемые для передачи данных, являются непрерывными, когда в каждый момент времени есть только одно-единственное значение сигнала, отображаемого физическим процессом, пригодным для прохождения в выбранной человеком физической среде: в металлических проводах – это электрический ток или напряжение, в стекловолокне – это свет, в свободном пространстве – это радиоволны.

Цифровыми данными непрерывный сигнал становится после дискретизации по времени и квантования по уровню, когда появляется возможность записать его в виде чисел. Для передачи числовых данных по каналам связи их всегда преобразовывают в непрерывный сигнал [3].

На рисунке 1а изображена реализация непрерывного сигнала $x(t)$ с показанными отсчётами мгновенных значений $x_i, i=\overline{1, I}$ в равноотстоящие моменты времени t_i с интервалом дискретизации Δt . На реализации выделен сегмент в виде показанного утолщённой линией участка кривой с границами, отмеченными прямоугольниками с моментами времени начала t_n и окончания t_k .

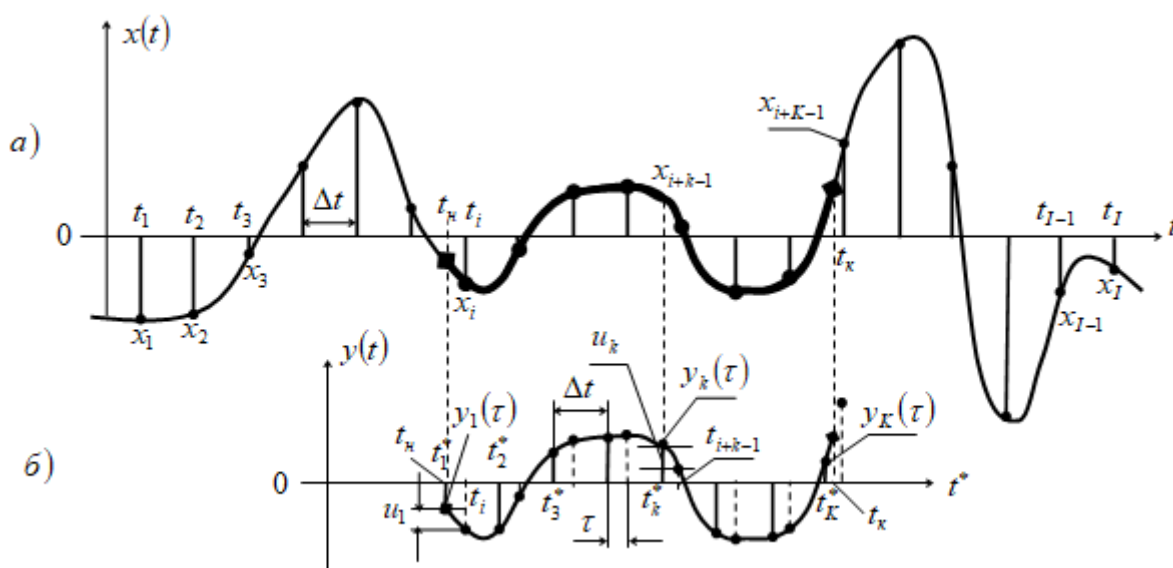


Рис.1. Выделение сегмента в длинной последовательности данных (1а – дискретизация сигнала по отсчётам, 1б – выделенный сегмент сигнала с несовпадающими с отсчётами границами)

На рисунке 1б изображён выделенный сегмент или фрагмент в виде реализации $y(t)$ с отсчётами $y_k(\tau), k=\overline{1, K}, K < I$, взятыми в моменты времени $t_k^{\square}$, не совпадающими с моментами времени t_i , но с таким же интервалом дискретизации Δt . Сдвиг между отсчётами обеих последовательностей равен τ .

Из-за временного сдвига τ появляется отличие отсчётов (невязка) обеих последовательностей u_k

$$u_k = y_k(\tau) - x_{i+k-1}. \quad (2)$$

Несовпадение расположения отсчётов и появление последовательности случайных величин u_k обусловлено тем, что изначально были заданы непрерывный процесс $x(t)$ и выделенный из него сегмент (фрагмент) $y(t)$, а после этого выполнена дискретизация обоих непрерывных процессов.

Постановка задачи: по заданному сегменту $y(t)$ требуется найти его на дискретном представлении по отсчётам непрерывного процесса $x(t)$ и определить время его начала t_n .

Синтез алгоритма

Для решения этой задачи применим теорию различения статистических гипотез по критерию максимума апостериорной вероятности. По каждой из гипотез H_i выдвигается предположение, что искомый сегмент начинается в момент времени t_i . Это соответствует поочерёдному прикладыванию отсчётов искомого сегмента к отсчётам заданного процесса и вычислению вероятностей совпадения отсчётов. За истинное положение сегмента принимается то, вероятность которого наибольшая, и функция правдоподобия превышает порог сравнения по критерию Неймана-Пирсона.

Для каждой из гипотез выразим отсчёты искомого сегмента через отсчёты заданной последовательности

$$H_i: y_k(\tau) = x_{i+k-1} + u_k \quad k = \overline{1, K} \quad i = \overline{1, L} \quad H_i. \quad (3)$$

Так как при последовательном совмещении искомого сегмента с заданной последовательностью данных конечное положение определяется при совмещении крайнего справа отсчёта $y_K(\tau)$ искомого сегмента с крайним справа отсчётом заданной последовательности x_L , то число различаемых гипотез равно

$$L = I - K + 1. \quad (4)$$

Для сокращения записей введём обозначение для вектора отсчётов искомого сегмента

$$\vec{Y} = [y_1(\tau), y_2(\tau), \dots, y_K(\tau)]^T. \quad (5)$$

Ввиду многочисленности причин, приводящих к несовпадению взятых отсчётов от заданных в памяти, и ошибок измерения, (отсчёты) невязки u_k считаются помехой и принимаются нормально распределёнными с нулевым матожиданием и дисперсией σ^2

$$w(u_k) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{-\frac{u_k^2}{2\sigma^2}\right\}. \quad (6)$$

Условные (обратные) вероятности гипотез при получении векторного наблюдения $\vec{Y}$ определяются по формуле Байеса [4]

$$P(H_i|\vec{Y}) = P(H_i) \cdot P(\vec{Y}|H_i) / \sum_{r=1}^L P(H_r) \cdot P(\vec{Y}|H_r), \quad (7)$$

где $P(H_i)$ – априорные вероятности гипотез, $P(\vec{Y}|H_i)$ – апостериорные вероятности наблюдений $\vec{Y}$ при условии, что они соответствуют гипотезе H_i , $i = \overline{1, L}$.

При равенстве априорных вероятностей гипотез и переходе к плотностям распределения вероятностей, с учётом того, что знаменатель формулы Байеса для всех гипотез одинаков, получим

$$P(H_i|\vec{Y}) = C \cdot w(\vec{Y}|H_i), \quad (8)$$

где $C^{-1} = \sum_{r=1}^L w(\vec{Y} | H_r)$ – постоянная нормировки

Выбор истинной гипотезы осуществляется по критерию максимума обратной вероятности

$$H_i: i = \max_r \left\{ P(H_r | \vec{Y}) \right\} \quad r = \overline{1, L}. \quad (9)$$

При использовании плотности распределения вероятностей

$$H_i: i = \max_r \left\{ w(\vec{Y} | H_r) \right\} \quad r = \overline{1, L}. \quad (10)$$

при статистической независимости отсчётов шума (ошибок измерения) u_k $k = \overline{1, K}$ условная плотность распределения вероятностей равна

$$w(\vec{Y} | H_i) = \frac{1}{(\sigma \sqrt{2\pi})^K} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^K (y_k(\tau) - x_{i+k-1})^2 \right\}. \quad (11)$$

Стоящая в показателе экспоненты сумма квадратов разностей поступивших данных со сравниваемыми значениями в теории различения гипотез называется достаточной статистикой, представляющей собой случайное число и содержащей всю информацию о поступивших данных [1]

$$\xi_i(\tau) = \sum_{k=1}^K (y_k(\tau) - x_{i+k-1})^2. \quad (12)$$

Максимизация вероятности соответствует минимизации достаточной статистики

$$H_i: i = \min_r \left\{ \xi_i(\tau) \right\} \quad r = \overline{1, L}. \quad (13)$$

Достаточная статистика в рассматриваемом случае есть дисперсия отклонения поступивших данных от ожидаемых значений по различным гипотезам.

Алгоритм определения истинной гипотезы основывается на определении вероятностей гипотез и принятии решения о гипотезе, имеющей наибольшую вероятность.

Обратные вероятности гипотез после сокращения одинаковых множителей в числителе и знаменателе равны

$$P(H_i|\vec{Y}) = \exp\left\{-\frac{1}{2\sigma^2}\xi_i(\tau)\right\} / \sum_{r=1}^L \exp\left\{-\frac{1}{2\sigma^2}\xi_r(\tau)\right\}. \quad (14)$$

Определим функцию правдоподобия в виде

$$\Lambda_i(\tau) = \exp\left\{-\frac{1}{2\sigma^2} \cdot \xi_i(\tau)\right\}. \quad (15)$$

Тогда

$$P(H_i|\vec{Y}) = \Lambda_i(\tau) / \sum_{r=1}^L \Lambda_r(\tau). \quad (16)$$

Таким образом, получены аналитические соотношения для расчёта апостериорных вероятностей при фиксированном сдвиге τ отсчётов искомого сегмента и наблюдаемой последовательности данных.

Если раскрыть скобки в выражении для достаточной статистики, то при нулевых средних и равенстве дисперсий отчётов заданной и искомой последовательностей нетрудно получить

$$P(H_i|\vec{Y}) = \exp\left\{\frac{1}{\sigma^2} \sum_{k=1}^K y_k(\tau) x_{i+k-1}\right\} / \sum_{r=1}^L \exp\left\{\frac{1}{\sigma^2} \sum_{k=1}^K y_k(\tau) x_{r+k-1}\right\}. \quad (17)$$

Сумма в показателе экспоненты представляет собой ковариационную функцию последовательностей $y_k(\tau)$ и x_{i+k-1}

$$B_i(\tau) = \sum_{k=1}^K y_k(\tau) x_{i+k-1}. \quad (18)$$

Выразим функцию правдоподобия через ковариационную функцию

$$L_i^{\square}(\tau) = \exp \left\{ \frac{1}{\sigma^2} \cdot B_i(\tau) \right\}. \quad (19)$$

Для исключения влияния задержки τ необходимо усреднить функцию правдоподобия по распределению задержки.

Пусть длительность интервала дискретизации Δt известна точно, а задержка τ имеет равномерное распределение вида

$$p(\tau) = 1/\Delta t, \quad \tau \in [-\Delta t/2, \Delta t/2]. \quad (20)$$

Усреднение функций правдоподобия (15) и (19) по распределению (20) выполнить затруднительно, поэтому аппроксимируем распределение (20) суммой M дискретных вероятностей, равных $1/M$, соответствующих фиксированным значениям задержки, равномерно размещённым на интервале от $-\Delta t/2$ до $+\Delta t/2$

$$\tilde{p}(\tau) = \frac{1}{M} \sum_{m=1}^M \delta(\tau_m - \tau) \quad (21)$$

где $\delta(\cdot)$ – функция Дирака, $\tau_m = \Delta t \cdot (2m - M - 1)/(2M)$.

После выполнения усреднения функций правдоподобия (15) и (19) получим

$$\tilde{\Lambda}_i \approx \frac{1}{M} \sum_{m=1}^M \exp \left\{ -\frac{1}{2\sigma^2} \cdot \xi_i(\tau_m) \right\}. \quad (22)$$

$$\tilde{\Lambda}_i^{\square} \approx \frac{1}{M} \sum_{m=1}^M \exp \left\{ \frac{1}{\sigma^2} \cdot R_i(\tau_m) \right\}. \quad (23)$$

Тогда вероятности гипотез равны

$$P(H_i | \vec{Y}) = \begin{cases} \tilde{\Lambda}_i / \sum_{r=1}^L \tilde{\Lambda}_r, & \text{при усреднении достаточных статистик,} \\ \tilde{\Lambda}_i^{\square} / \sum_{r=1}^L \tilde{\Lambda}_r^{\square}, & \text{при усреднении ковариационных функций.} \end{cases} \quad (24)$$

Таким образом, алгоритм выбора истинной гипотезы реализуется как выбор по критерию минимума достаточной статистики (13) или максимума ковариационной функции (18), или по критерию максимума функций правдоподобия (22) или (23).

Среднее значение времени начала искомого сегмента равно

$$m_{t_n} = \Delta t \sum_{i=1}^L P(H_i | \vec{Y}) \cdot i. \quad (25)$$

Среднеквадратическое отклонение определения времени начала искомого сегмента равно

$$\sigma_{T_n} = \sqrt{\sum_{i=1}^L P(H_i | \vec{Y}) \cdot (i \cdot \Delta t - m_{T_n})^2}. \quad (26)$$

Структурная схема алгоритма

Структурная схема устройства вычисления достаточных статистик ξ_j приведена на рисунке 2. Выборка отсчётов заданной последовательности

$x_i, \dots, x_{i+K}$ при проверке каждой гипотезы H_i $i=\overline{1, L}$ поступает на вычитающие устройства ВУ, на другие входы которых подаются отсчёты искомого сегмента $y_1, \dots, y_K$. В квадраторах КВ вычисляются квадраты разностей, которые суммируются в сумматоре Σ .

Структурная схема устройства принятия решений о наиболее вероятной гипотезе и определения среднего значения и среднеквадратического отклонения времени начала искомого сегмента приведена на рисунке 3.

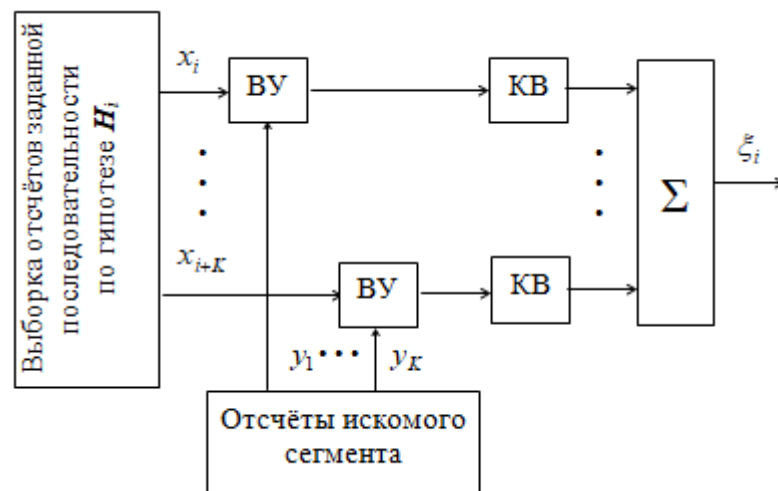


Рис. 2. Устройство вычисления достаточных статистик

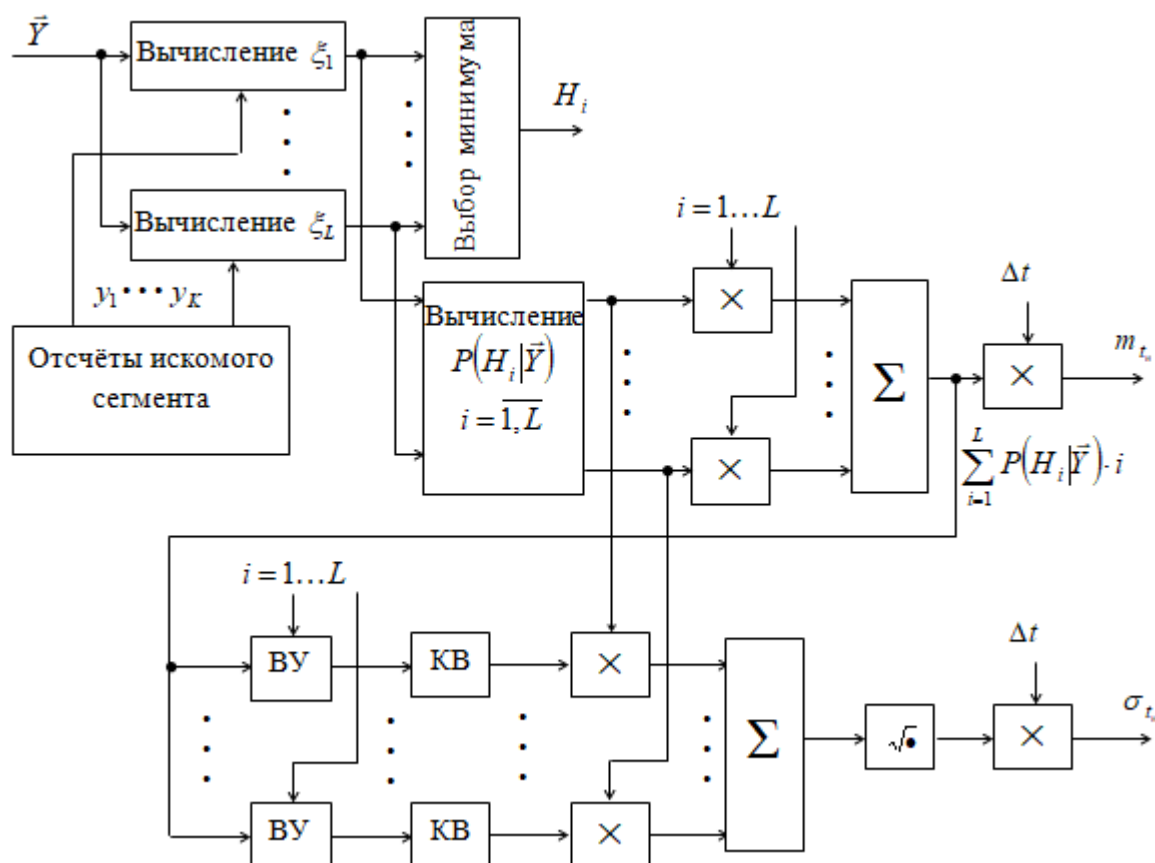


Рис. 3. Устройство принятия решений и вычисления параметров времени начала искомого сегмента

Анализ качества поиска заданного сегмента

Наименьшее значение усреднённой в соответствии с выражениями (22, 23) функции правдоподобия наблюдается при задержках $\tau_0 = \pm \Delta t / 2$, что соответствует взятию отсчётов входного процесса точно посередине между заданными отсчетами искомого сегмента. Эти значения задержки являются самыми неблагоприятными при обнаружении, и поэтому характеристики обнаружения будут рассчитаны для этого случая. Для многоканальной системы усреднения функции правдоподобия преобладающим в (22, 23) при $\tau = \pm \Delta t / 2$ будет слагаемое, получаемое либо в 1-ом, либо в M -ом канале, так как в одном из этих каналов отсчёты будут браться в точках совпадающих с заданными значениями. Характеристики обнаружения при этом допущении будут определяться распределением вероятностей либо $B_i(\tau_1)$, либо $B_i(\tau_M)$ в зависимости от задержки.

Нормированная ковариационная функция последовательности отсчётов искомого сегмента данных (4) при $k_1=0, \tau=\Delta t/2, \tau_m=\tau_1$ и при $L \gg 1$, равна

$$R(\tau_1 - \tau) = R \approx 1 - 1/2M \quad (27)$$

В отсутствие помех и при точном совпадении отсчётов искомого сегмента и последовательности данных ковариационная функция равна

$$B_i(0) = \sum_{k=1}^K (x_{i+k-1})^2 = K\sigma_0^2, \quad i=\overline{0, I} \quad (28)$$

Вероятность правильного определения гипотезы о положении искомого сегмента определяется в соответствии с методикой [2] и равна

$$P_{\text{прав}} = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-\infty}^{\infty} \left[\frac{1}{2} + \Phi_0(\sqrt{q} + y) \right]^I \cdot \exp[-y^2/2] \cdot dy \quad (29)$$

где $\Phi_0(z) = \frac{1}{\sqrt{2 \cdot \pi}} \int_0^z \exp[-t^2/2] \cdot dt$ – нормированная функция ошибок;

$q = q_i = B_i(0) / \sigma^2$ – параметр различения проверяемой гипотезы.

В результате поиска сегмента в длинной последовательности минимум достаточной статистики (13) или максимум апостериорной вероятности (9) будут найдены всегда, даже при отсутствии искомого сегмента в просматриваемой последовательности. Поэтому необходимо вычислить вероятность правильного обнаружения D искомого сегмента при заданной вероятности ложной тревоги F в соответствии с выражением [5]

$$D = (\sqrt{2\pi})^{-1} \int_{y_0}^{\infty} \exp[-y^2/2] dy \quad (30)$$

$$\text{где } y_0 = v_0 - R\sqrt{q} - \eta, \quad v_0 = \sqrt{-2 \ln F}, \quad \eta = \frac{1}{2R\sqrt{q}} - \frac{v_0 + R\sqrt{q}}{(2R\sqrt{q})^2}, \quad q = q_M = B_i(\tau_M) / \sigma^2 -$$

отношение сигнал/шум при сдвиге по времени τ_M .

По формулам (29) и (30) рассчитаны и построены графики характеристик обнаружения искомого сегмента для различного числа элементов анализируемой последовательности при $\tau=0$, рисунок 4а, и для различного числа каналов обнаружителя при $|\tau|=\Delta t/2$, рисунок 4б.

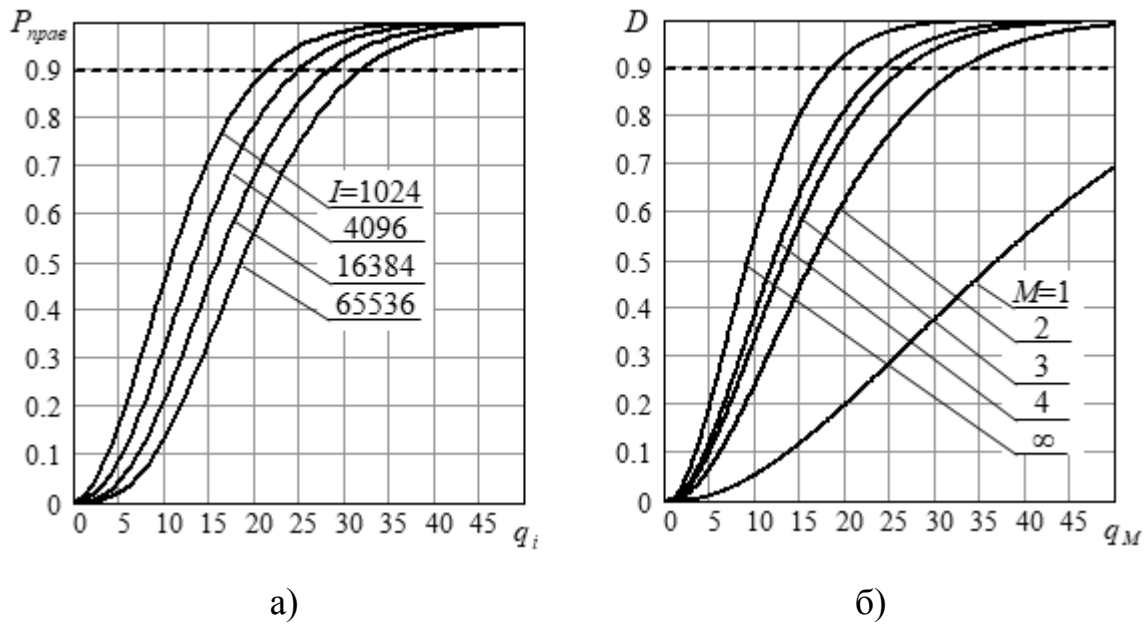


Рис. 4. Характеристики обнаружения заданного сегмента

(4а – при различном числе элементов анализируемой последовательности и при нулевом сдвиге точек отсчётов фрагмента и сигнала, 4б – при различном числе каналов обнаружителя)

Критерий обнаружения – превышение вероятностью обнаружения заданного критериального уровня 0,9, показанного пунктиром на рисунках. Из графиков на рисунке 4а определяется отношение сигнал/шум, при котором вероятность $P_{\text{прав}}$ достигает критериального уровня. При объёме данных $I=2^{16}$ $q=32,0$, а при $I=10^8$ $q=49,0$, как показывают отдельные расчёты по формуле (29).

С учётом (28) отношение сигнал/шум равно

$$q = K \cdot (\sigma_0 / \sigma)^2 \quad (31)$$

Данное соотношение определяет связь между ошибкой измерения отсчётов σ , среднеквадратическим отклонением наблюдаемых отсчётов σ_0 , числом отсчётов K в искомом сегменте и отношением сигнал/шум q , при котором обеспечивается заданная вероятность обнаружения заданного сегмента.

Из графиков на рисунке 4б следует, что область расположения характеристик при $|\tau_0| \leq T_0/2$ ограничена асимптотической кривой, соответствующей либо бесконечному числу каналов, либо любому числу каналов, но при $|\tau_0|=0$, и кривой, соответствующей выбранному числу каналов при $|\tau_0|=T_0/2$. Обнаружитель может быть одноканальным при очень высоких отношениях сигнал/шум $q > 100$. Увеличение числа каналов более двух даёт небольшой выигрыш по мощности. Существенное улучшение обнаружения наблюдается при переходе от одноканального обнаружителя к двухканальному. При числе каналов более 2-х улучшение незначительно. При этом ошибка измерения отсчётов будет не более среднеквадратического отклонения наблюдаемой входной последовательности данных.

Следовательно, величина параметра q определяет минимальный объём выборки данных для обеспечения требуемой вероятности обнаружения заданного сегмента.

Заключение

В условиях неизвестного сдвига точек отсчёта непрерывного сигнала и его заданного фрагмента предложен подход к синтезу алгоритма поиска фрагмента. Подход основан на классическом аппарате проверки гипотез математической статистики. Отличительной особенностью подхода является учет шумов отсчёта сигнала и фрагмента, и на этой основе формализация задачи выбора числа каналов наблюдения за сигналом.

Таким образом, в настоящей работе с применением вероятностно-статистического метода синтезирован алгоритм двухэтапной процедуры поиска заданного сегмента в длинной последовательности данных, заключающийся в

определении наиболее вероятного положения искомого сегмента и проверки соответствия найденного сегмента заданному.

Получены аналитические соотношения для оценки качества обнаружения, которые учитывают автокорреляционные свойства последовательностей, и рассчитаны характеристики обнаружения, с помощью которых определяются объём выборки заданного сегмента и достаточное число каналов обнаружителя для обеспечения требуемого качества обнаружения при заданных условиях наблюдения.

Предложена структурная схема устройства принятия решений о наиболее вероятной гипотезе и определения среднего значения и среднеквадратического отклонения времени начала искомого сегмента, которая позволяет учесть наименее благоприятные условия начала отсчета сигнала, а именно, точно посередине заданного фрагмента.

References

1. **Van Trees Harry L., Bell Kristine L.** Detection, Estimation, and Modulation Theory. P. I, 2nd Ed. Wiley, John Wiley & Sons Ltd., 2013, 1176 p.
2. **Gutkin L. S.** The Theory of Optimal Methods of Radio Reception with Fluctuation Interference. Moscow: Sovetskoe Radio, 1972, 448 p. (in Russian)
3. **Bocker P.** Data Transmission: Communications Technology in Remote Data Processing Systems. Springer Verlag, Berlin – Heidelberg – New York, 1976, 264 p. (in German)
4. **Wentzel E. S.** Probability Theory, 11-th Ed. Moscow: KNORUS, 2010, 664 p. (in Russian)
5. **Vainshtein L. A., Zubakov V. D.** Isolation of Signals Against the Background of Random Noise. Moscow: Sovetskoe Radio, 1960, 449 p. (in Russian)