УДК 629.7.05, 621.398

© Кукушкин С.С., Нестеровский И.С.
Kukushkin S., Nesterovskiy I.

ПРИМЕНЕНИЕ СТАТИСТИКО-ЛИНГВИСТИЧЕСКИХ МЕТОДОВ К РЕШЕНИЮ ЗАДАЧИ ИДЕНТИФИКАЦИИ ТЕЛЕМЕТРИЧЕСКОЙ ИНФОРМАЦИИ

APPLICATION OF STATISTICAL AND LINGUISTIC METHODS TO SOLVE THE PROBLEM OF IDENTIFICATION OF TELEMETRY DATA

Аннотация. *Статья посвящена прикладному применению статистико-лингвистических методов к решению задачи идентификации телеметрической информации. В ней изложены результаты исследования по применению метода ранговой статистики на основе классического закона распределения Цифпа и Ципфа-Мандельброта для решения задач идентификации с последующей классификацией параметров телеметрической информации. Показано, что для ТМИ распределение частоты встречаемости словоформ, как функции от ранга словоформ, существенно отличается как от классического закона Ципфа, так и от распределений, полученных при помощи расширения данного закона. При решении задач аппроксимации практических данных от телеметрических средств предложено использовать новый закон распределения случайных величин, являющийся расширением закона Ципфа.*

Annotation. *The article is devoted to applications of statistical and linguis-tic methods to solve the problem of identification of telemetry data. It sets out the results of a study on the application of the rank statistics based on the classical law of distribution of Zipf and Zipf-Modelbrota to address issues of identification and subsequent classification of the parameters of the telemetry information. It is shown that for TMI distribution of frequencies of word forms, as a function of the rank of word forms differs significantly from both the classical law of Zipf, and on the distributions received by the expansion of the law. When solving practical problems of approximation of data from the telemetry means is proposed to use the new law of random variables, which are an extension of the Zipf law.*

Ключевые слова. *Телеметрия, нетрадиционное представление данных образами-остатками, восстановление и обработка цифровых сигналов, искаженных помехами.*

Key words. *Telemetry, non-traditional presentation of data residuals, recovery and pro-cessing of digital signals, the distortion noise.*

Введение

В настоящее время чрезвычайно быстро развиваются цифровые методы преобразования данных, относящиеся к передаче сообщений.

От достигнутого при этом уровня решения существующих проблем значительно отстают другие технологии информационно-телеметрического обеспечения (ИТО) испытаний ракетно-космической техники (РКТ), связанные с приемом, обработкой и идентификацией измерительной информации.

Появляется потребность в совершенствовании всей системы ИТО испытаний РКТ, иначе положительный эффект от нововведений не достигнет ожидаемых величин. Об этом свидетельствует существующая практика модернизации программно-математического обеспечения (ПМО), предполагающая перенос существующих алгоритмов на новую аппаратно-программную платформу обработки и идентификации телеметрической информации (ТМИ) без внесения каких-либо корректив. Кроме того, ощущается острая необходимость в новых методах обработки, анализа и идентификации ТМИ, которые были ориентированы на символьный (знаковый) уровень представления цифровых данных. Прежде всего, это связано с решением ряда новых задач. К их числу относятся: 1) цифровая обработка данных, направленная на повышение качества принимаемых сигналов, пред-

Кукушкин Сергей Сергеевич – доктор технических наук, профессор. ведущий научный сотрудник, 4 ЦНИИ Минобороны России, тел.+7(495)515-19-82;
Нестеровский Игорь Сергеевич – научный сотрудник, 4 ЦНИИ Минобороны России.

Kukushkin Sergey – doctor of the technical sciences, professor, the main scientific employee, 4 Central Scientific Research Institute Ministry of Defence of Russia, tel. +7(495)515-19-82;
Nesterovskiy Igor – scientific employee, 4 Central Scientific Research Institute Ministry of Defence of Russia.

ставленных в виде двоичных и других видов кодов; 2) реализация оптимальных процедур декодирования ТМИ; 3) идентификация данных на знаковом уровне.

Применение предлагаемого специализированного математического аппарата необходимо, прежде всего, для автоматизации процесса анализа больших объемов ТМИ на знаковом уровне. Такой подход позволяет значительно сократить человеческие и временные ресурсы, так как требуется оценка нескольких тысяч разнообразных по динамике функционирования, диапазонам и точности измерения, формам представления информационных параметров. В данной статье предлагается использовать статистико-лингвистический алгоритм для решения проблемы оценки достоверности классификации ТМИ. Задачи идентификации телеметрической информации в условиях отсутствия априорных данных можно решить нетрадиционными методами, к которым можно отнести статистико–лингвистические методы оценивания, основанные на гиперболических законах распределения.

В качестве оценок достоверности используются качественные количественные оценки ТМИ. Количественные оценки, получаемые на основе частотно-ранговой модели, наиболее полно отражают ее структурные и статистические свойства. В основу статистико-лингвистического оценивания положены законы распределения словоформ.

Как известно [1,2], изучение количественных характеристик текстов любой природы приводит к ранговым распределениям. Особое место среди ранговых распределений занимает закон Ципфа и его обобщение – закон Ципфа-Мандельброта. Зависимость между частотой слова в лингвистической выборке и его номером в последовательности, ранжированной по убыванию частоты, носит гиперболический характер и может быть описана этим законом. Распределение (1) является канонической формой закона Ципфа, распределение (2) – его расширением, (3) – распределение Ципфа-Мандельброта [1,2].

$$\Phi(r) = C_n\, r^{-1}; \qquad (1)$$

$$\Phi(r) = C_n\, r^{-\gamma}; \qquad (2)$$

$$\Phi(r) = C_n (B_м + r)^{-\gamma}, \qquad (3)$$

где r – ранг элемента словаря;

Φ – относительная частота слова, имеющего ранг r;

C_n – константа, определяемая объемом выборки;

γ – коэффициент частотно-рангового соотношения;

$B_м$ – поправочный коэффициент Мандельброта (ранговое искажение).

Представим поток телеметрической информации в виде целостного текста, где каждое из слов ТМИ состоит из постоянного числа символов двоичного кода. Различные словоформы образуют словарь (алфавит). На выбор словаря влияет наличие априорной информации о структуре анализируемой ТМИ: разрядность информационного слова, тип модуляционного кода, длина кадра и особенности организации вычислительного процесса. Если структура ТМИ неизвестна, то ее можно рассматривать в виде последовательности двоичных бит данных «0» и «1» и в качестве элементов словаря (алфавита) рассматривать блоки «0» и «1» равной длины. Если же структура анализируемой ТМИ априорно известна, то предпочтительнее в качестве алфавита выбирать все возможные элементы шкалы телеметрирования – значения измерений параметров (при восьмиразрядном информационном слове – всего 256 возможных значений шкалы). В соответствии с выбранным алфавитом представления данных $V_1 = \{0,1\}$ и $V_2 = \{0 – 255\}$ формируются частотно-ранговые распределения словоформ ТМИ. Такой подход впервые был предложен в работе [3].

Пусть все слова телеметрического сигнала образуют множество

$$X_г(t) = \{x_1, x_2, ..., x_{N_г}\}, \qquad (4)$$

где $N_г = Card(X_г)$, а различные словоформы образуют алфавит

$$A_x = \{a_1, a_2, ..., a_n\};\ a_i \neq a_j;\ i,j \in \{1,2,3, ..., n\}, \qquad (5)$$

где $n = Card(A_x)$ – мощность алфавита, причём $A_x \subset X_г(t)$.

В качестве словоформ алфавита (5) для телеметрического сигнала могут использоваться информационные слова сообщений в шкале телеметрирования, тогда

$$A_x = \{ a_1 = x_1, a_2 = x_2, a_3 = x_3, ... \}. \qquad (6)$$

В условиях неизвестной структуры ТМИ в качестве словоформ будем использовать длины блоков одноименных кортежей “0” и “1”

$$A_x = \{a_1 = \{0,1\}; a_2 = \{00,11\}; a_3 = \{000,111\}; ...; a_n = \{0...0,1...1\}\}. \qquad (7)$$

По выбранному словарю A_x (6) или (7) построим эмпирический ряд частот появления словоформ, гистограмму частот встречаемости словоформ (рис.1)

$$f(x) = \{ f(x)_i = Q_i^a / N_г,\ i = 1,2,3,...,n \}, \qquad (8)$$

где Q_i^a – число вхождений слов $x \in A_x$ в $X_г(t)$.

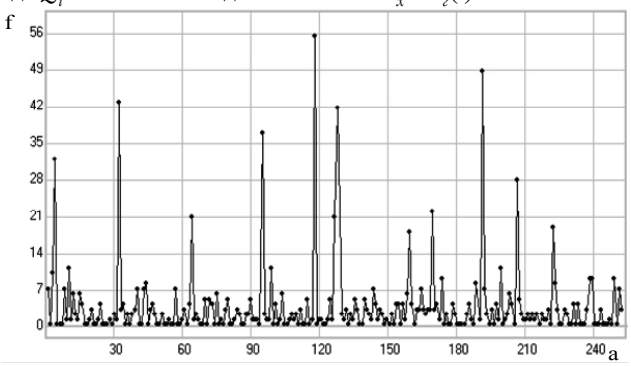

Рис.1. Эмпирический ряд частот появления словоформ, представленных в словаре V_2={0-255}

В результате значительного объема ранее проведенных статистических исследований было доказано, что распределение (8) для неискаженных телеметрических сообщений принадлежит к распределениям с ярко выраженными негауссовыми свойствами(степенные распределения), т.е. гистограммы на рис.1 и рис.2 характеризуются значительными отклонениями значений от средних величин, по сравнению с гауссовым распределением, и поэтому «хвосты» описываемых распределений («распределений с тяжелыми хвостами») убывают гораздо медленнее, чем у нормального закона.

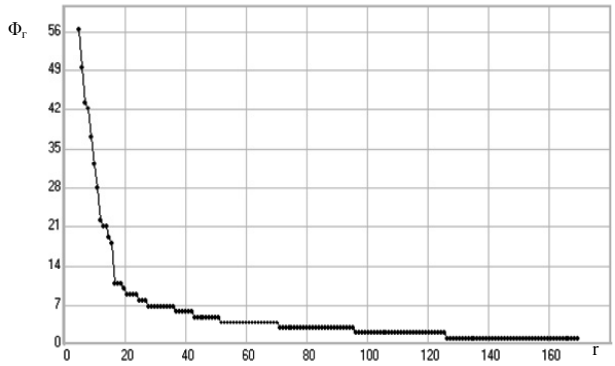

Рис.2. Частотно-ранговое распределение словоформ ГТС в словаре значений телеметрируемого параметра V_2={0-255}

Предложено для процесса аппроксимации экспериментальных данных использовать новый закон распределения частотно-рангового характера представления данных на основе ранее описанного закона распределения Ципфа-Мандельброта, т.е функционала $\Phi(r)$, который ставит в соответствие номеру (рангу) словоформы частоту его появления. При таком представлении максимальный ранг r, для которого выполняется условие $f_r \geq 0$, характеризует мощность алфавита и является количественным параметром оценки достоверности (информативности) оцениваемой выборки.

Практические исследования [3] распределений (1, 2, 3) показали, что они недостаточно точно отражают статистические закономерности ТМИ, то есть хорошо описывают только начальные ранги и, как следствие, делают более или менее бессмысленным минимизацию суммы квадратов отклонений теоретической кривой от измеренных точек. Поэтому для получения количественных оценок достоверности анализируемой информации логарифмируют частоты f_r и ранги r эмпирической зависимости (2). Для этого в частотно-ранговом распределении (2) с константами C_n и γ, представляют γ, зависящим от r. Тогда

$$\Phi(r) = C_n\, r^{-\gamma(r)}, \qquad (9)$$

где $\gamma(r)=\gamma_0+\gamma_0(r)$.

Если $\gamma(r)$=const, то (9) имеет вид прямой (0, Φ_{max}; r_{max}, 0) (рис.3), а γ выражается через соотношение логарифмов Φ_{max} и r_{max}. Функция рангового распределения $\Phi(r)$ на рис. 3 представлена кусочно-линейным приближением. Для характеристики увеличения начального значения γ_0 введём коэффициент его прироста d, тогда для r-го ранга: $\gamma(r)=\gamma_0+\gamma_0 dr$. Используя первые члены разложения функции exp в ряд, получим $\gamma(r)=\gamma_0 exp(dr)$. В результате данных преобразований выражение (9) примет следующий вид:

$$\Phi(r) = C_n\, r^{-\gamma_0 exp(dr)}, \qquad (10)$$

где d – параметр, характеризующий изменение коэффициента частотно-рангового соотношения статистической выборки;

γ_0 – начальный коэффициент частотно-рангового соотношения.

Полученные ранее логарифмические зависимости частота-ранг апроксимируются функциональной зависимостью (10).

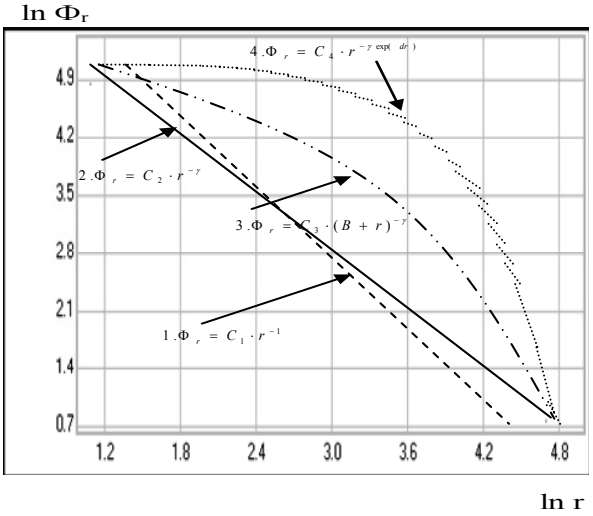

Рис.3. Частотно-ранговые законы распределения в логарифмической форме

Форма графиков зависимости частота встречаемости – словоформа и частота встречаемости – ранг используется в практике для получения параметров качественной оценки анализируемой информации. Сравнив полученные графики с эталонным (ранее определенными и достоверным), можно качественно (визуально) оценить достоверность оцениваемой информации.

Коэффициенты C, γ и d в формуле (10) определяются в два приема. Вначале определяются начальные значения коэффициентов C_o, γ, d_o по следующим соотношениям:

$$C_0 = \max_{r=1}^{N}(f_r) = f_1; \qquad (11)$$

$$\gamma(r) = \frac{\left(\log 10\left(\max_{r=1}^{N}(f_r)\right) - \log 10(f_r)\right)}{\log 10(r)},$$

$$d_0 = -\frac{\ln(\gamma(1)) - \ln(\gamma(r))}{1-r};$$

$$\gamma_0 = \frac{\gamma(r)}{\exp(d_0 r)}.$$

Вторым этапом производят их коррекцию с целью минимизации квадрата разности эмпирического ряда частот и аппроксимирующей зависимости

$$min\left\{\sum_{r=1}^{N}\left(f_r - \Phi_r\right)^2\right\}. \qquad (12)$$

Проведенные исследования [3] показали, что модель (10) является более точной по сравнению с аналогами (1–3). Она достаточно корректно описывает закономерности эмпирического ряда в выборках различного объёма. Вычисление оценок разностей и суммы квадратов разностей коэффициентов *C, d, γ* соотношения (10), при этом дает количественное оценивание расхождений, которое несет информацию о достоверности анализируемой информации. Результаты данных операций выступают в качестве оценочных показателей достоверности, шумовых и других характеристик анализируемых данных. Качественная оценка достоверности структуры ТМИ определяется с помощью частотно-ранговых распределений – по форме графиков кривых аппроксимации при их сравнении с эталонами, количественная оценка по значениям коэффициентов *C, d, γ* и минимизации суммы квадратов разностей. Статистико-лингвистический подход к идентификации данных телеизмерений позволяет получать оценки достоверности ТМИ при анализе как перекрывающихся, так и неперекрывающихся выборок из потока ТМИ.

Анализ статистических данных позволил сделать вывод о том, что области возможных изменений коэффициентов γ_0 и d в (10) представляют собой своеобразные статистические образы различных информационных процессов и могут использоваться для их идентификации. Искажение информационной структуры телеметрического сигнала в целом или какой-либо его части при помощи помех или кодирующих, шифрующих преобразований приводит к изменению частотно-рангового распределения, т.е к увеличению размеров областей коэффициентов γ_0 и d, что подтверждается результатами проведенного моделирования [3]. При таком подходе можно выделить области $U_{\gamma d} = \{\gamma_{min}<\gamma_0<\gamma_{max};\ d_{min}<d<d_{max}\}$ для информации с нарушенной структурой сообщений $U_{\gamma d}^{ш}$ и области $U_{\gamma d}^{э}$, характерные для смысловой информации (табл.1 и табл. 2). Для анализа полученных эмпирических частотных и частотно-ранговых соотношений для информационного процесса $П_i$ по определенным поисковым признакам (тип информации, тип словаря) на основе имеющейся информации по характерным параметрам строятся полигоны допусков по каждому из параметров. Определяются оценки мер расхождения в виде отношения площадей выхода значений за область полигонов допусков.

Используя характеристики распределений и зная $U_{\gamma d,\,i}=\{U_{\gamma d,i}^{ш};\ U_{\gamma d,i}^{э}\}$ словарей (6), (7) для каждого *i*-го типа данных, можно строить алгоритмы оценки качества, отбраковки недостоверных участков, определения отношения сигнал/шум и др.

Таблица 1

Оценки γ_0 и d для словаря V_2, основу которого составляют значения выделенных информационных слов: $P_{ис}$=8, n=256, $N_г$=20000, $N_{оп}$=120, $P_д$=0,95, $S_ф^{н}$=0,05

Тип ТМИ	Условное обозначение	Математическое ожидание	СКО	Асимметрия	Эксцесс	Мода
РТС-9	γ_o	0,94435	0,18915	-0,01635	-0,1724	0,94323
	d	0,01134	0.09834	0,01594	-0,1203	0,01285
БРС4-М	γ_o	1,11594	0,20156	-0,10394	-0,0237	1,09456
	d	0,000024	0,000016	-0,09841	-0,0094	0,000022
Trident-1	γ_o	0,50025	0,01835	-0,02383	0,0167	0м50125
	d	0,00286	0,00084	-0,03781	0,0836	0,00281
Trident-2	γ_o	1,49147	0,10891	-0,00392	-0,1648	1,48352
	d	0,00235	0,00153	0,01974	-0,1973	0,00231
Menuteman-3	γ_o	0,33715	0,09327	0,00093	-0,0831	0,32762
	d	0,04729	0,00103	0,00264	-0,0975	0,04863
Имитационный	γ_o	0,26143	0,05248	-0,01674	-0,0794	0,26299
	d	0,00492	0,00089	-0,03973	-0,0886	0,00458
Имитационный +25% шум	γ_o	0,29051	0,08012	-0,01218	-0,0591	0,28310
	d	0,00616	0,00127	-0,02364	-0,0475	0,00629
Имитационный +50% шум	γ_o	0,30025	0,09095	-0,02050	-0,0126	0,29093
	d	0,00833	0,00383	0,01744	-0,0202	0,00840
Имитационный +75% шум	γ_o	0,39467	0,14083	-0,00921	0,0094	0,40052
	d	0,01052	0,00420	0,00283	-0,0038	0,09693

Таблица 2

Оценки γ_0 и d для словаря V_1, основой которого являются словоформы, составленные из бит ГТС: n=64, $N_г$=20000, $N_{оп}$=120, $P_д$=0,95, $S_ф^n$=0,05

Тип ТМИ	Условное обозначение	Математическое ожидание	СКО	Асимметрия	Эксцесс	Мода
РТС-9	γ_0	0,79825	0,08123	0,01291	-0,00228	0,78741
	d	0,16451	0,09285	0,02274	-0,01516	0,16217
БРС-4ЛР	γ_0	1,35372	0,08791	-0,00241	-0,05926	1,34459
	d	0,07861	0,00183	-0,01159	-0,08820	0,07984
Trident-1	γ_0	0,95175	0,07329	-0,02673	-0,18736	0,94853
	d	0,05612	0,00381	-0,08861	-0,12752	0,00558
Trident-2	γ_0	1,02592	0,08491	-0,48813	-0,03450	1,00481
	d	0,12034	0,07601	-0,05519	-0,02452	0,14006
Menuteman-3	γ_0	1,12179	0,09596	-0,00830	-0,19424	1,13494
	d	0,04653	0,00570	-0,00955	-0,01035	0,04820
Имитационный	γ_0	1,98360	0,29705	-0,07302	-0,11790	2,00328
	d	0,01803	0,00238	-0,06477	-0,10353	0,01774
Имитационный +25% шум	γ_0	1,84939	0,40161	0,07021	-0,10457	1,90072
	d	0,01892	0,00449	-0,00597	0,09063	0,01847
Имитационный +50% шум	γ_0	1,94922	0,55210	-0,00901	0,06842	1,90024
	d	0,01907	0,00883	-0,01627	0,01721	0,01947
Имитационный +75% шум	γ_0	1,99042	0,66027	0,00766	-0,03944	1,89240
	d	0,01935	0,00890	-0,09938	-0,01061	0,02066

Использовалась следующая концепция классификации информационных процессов: отыскивается такое разбиение на изолированные классы, при котором каждому из них свойственен свой закон распределения вероятностей. Для идентификации процессов предложено использовать меры, которые используются в теории распознавания образов, рассматривая в качестве признаков коэффициенты γ_0 и d,.

Полученные количественные оценки ТМИ, в том числе коэффициенты γ_0 и d,, вместе с сопроводительной информацией заносятся в базу данных (БД), которая служит для накопления статистической информации о достоверности ТМИ и получения обобщенных оценок, а также для самообучения программной оболочки экспертной системы, построенной на основе рассмотренного подхода. Положительные результаты быстрой и надёжной классификации при последующем гарантированном распознавании информации и наполнении этими данными эталонной базы позволяют в качестве меры качества использовать расстояние между полученными координатами (γ_0,d) и центрами $Z_{u,i}$ кластеров $U_{\gamma d,\, i}$ в признаковом пространстве γ_0, d.

Предложенный метод доказал свою эффективность при идентификации ТМИ с неизвестными структурными характеристиками при наличии искажающих факторов в виде шумов, шифрования и кодирования.

Для гарантированной идентификации информационных телеметрических процессов при представленном ранее общем подходе, применяющем математический аппарат ранговой статистики и распознавания образов, можно также включить процедуры анализа законов частотно-ранговых распределений, основанных на получении оценок тяжести “хвостов” распределений.

Как раннее было установлено [3], телеметрическая информация подчиняется закону распределения (10) с явно выраженными негауссовыми свойствами. Для данного типа распределений характерно наличие длинной “хвостовой” части, описывающей статистику данных больших рангов при малых частотах появления и резкой асимметрии, что приводит к использованию моментов распределения в качестве дополнительных идентифицирующих показателей.

При этом для каждого закона распределения по выборке (4) будем определять величину показателя тяжести “хвостов” распределений [3]

$$Q_n = (U_n(0{,}2) - L_n(0{,}2))/(U_n(0{,}5) - L_n(0{,}5)), \qquad (13)$$

где $L_n(0{,}2)=1/N_1\sum_{j=1}^{N_1} x_j$; $L_n(0{,}5)=1/N_2\sum_{j=1}^{N_2} x_j$;

$$U_n(0{,}2)=1/N_1\sum_{j=N_г-N_1+1}^{N_1} x_j;\quad U_n(0{,}5)=1/N_2\sum_{j=N_г-N_2+1}^{N_2} x_j.$$

По результатам применения алгоритма кластерного анализа для эмпирического распределения установлены объёмы выборки N_1 и N_2 , определяемые ниж-

ним и верхним квантилями (20% и 50%) распределения, которые принимались за границы критической области при тестировании соответствующей статистической гипотезы. При $N_г \to \infty$ (13) эмпирический показатель тяжести хвостов Q_n для нормального закона Q_n=1,757, для законов Стьюдента Q_n=1,895, Коши Q_n=2,5. Из данных, приведенных в табл. 3, следует, что для различных распределений оценки Q_n, полученные методом статистических исследований, что негауссовые распределения имеют большие значения Q_n по сравнению с нормальными законами распределения. Они были получены при следующих условиях: словарь (6), байтовые информационные слова-измерения ($P_{ис}$=8), объем выборки (N=70). Проведенные исследования также показали, что значения "показателя тяжести хвостов", в свою очередь, имеют нормальное распределение. Проверка гипотезы о нормальном распределении Q_n осуществлялась с использованием критерия Колмогорова-Смирнова

$$H_o(\{f(x)\} \in (Q_n, \sigma_{Qn})): D_{max} < D_{кр} = 0{,}163; \qquad (14)$$

$$D_{max} = \max_x(|f(x) - f^m(x)|); P_д = 0{,}95.$$

эффект даёт исключение из анализа элементов $\Phi(r)$=0, а также более точный подбор коэффициентов при аппроксимации зависимости γ_o и d.

В качестве критерия наилучшего приближения (10) к (9) выбран минимум суммы квадратов разностей эмпирического $\Phi_э(r)$ и теоретического $\Phi_m(r)$ законов распределения:

$$S_ф = 1/n \sum_{i=1}^{n} (\Phi^э(r)_j - \Phi^m(r)_j)^2. \qquad (15)$$

При достижении $S_ф$ заданного порогового уровня $S_ф^n$ по критерию

$$\lim_{n\to\infty} S_ф \to S_ф^П$$

фиксируются коэффициенты статистической модели (10) и формируется точка (C_n, γ_o, d) в признаковом пространстве. Далее, применяя критерий «минимаксного» расстояния, описанный в работе [3], проводим проверку гипотезы о принадлежности анализируемой информации к одной из эталонных $U_{gd,i}$. При получении положительного результата происходит дополнение эталонной базы, при отрицательном результате выдвинутая гипотеза отвергается.

Таблица 3

Оценки Q_n «тяжести хвостов» распределений

Вид закона распределения	Форма закона распределения					
	Частотная			Ранговая		
	Q_n	σ_{Qn}	D_{max}	Q_n	σ_{Qn}	D_{max}
Нормальный	1,7320	0,0258	0,0729	1,7293	0,0234	0,0902
Пуассона	1,7832	0,0472	0,0601	1,7809	0,0291	0,0734
Коши	2,1939	0,0745	0,0653	2,2137	0,0914	0,0721
Ципфа α=1,3	2,1817	0,0971	0,0548	2,1773	0,0881	0,0551
Ципфа α=0,7	2,2823	0,085	0,0365	2,2791	0,0861	0,0822
Двумерное экспоненциальное	1,8794	0,0456	0,0545	1,7983	0,0320	0,0631
Стьюдента	1,8636	0,0431	0,0968	1,7882	0,0392	0,0935

В результате на основе статистических распределений (10)–(11) предлагается построить статистико-лингвистический алгоритм оценки качества ТМИ. Вначале на основе выбранного словаря (6) или (7) строится эмпирический ряд частот (8) и рангов (9). Для выполнения процедуры надёжной оценки необходимо выбирать одинаковые типы словарей, $N_г$, $R_{i,зад}$, $S_ф^n$<1% от $P_д$. Затем выполняется операция аппроксимации закона распределения (9) зависимостью (10) с последовательно-итерационным подбором коэффициентов C_n, γ_o и d согласно (11).

При проведении процесса аппроксимации частотно ранговых распределений некоторых информационно-выраженных процессов зависимостью (10) был выявлен эффект получения отрицательного результата $R \to max$ при $\Phi(r \to max)$ = 0, $n_i < n_{max}$. Положительный

Проведенные исследования показали, что оценка достоверности ТМИ, ориентируясь на показатель «сигнал/шум» для цифровых данных, может быть выполнена по критерию минимума расстояния между полученными координатами (γ_o, d) и центром $Z_{u,i}$ наиболее близкого кластера $U_{gd,i}$

$$\min_i(R_i = |(\gamma_o, d) - Z_{u,i}|). \qquad (16)$$

Одно из направлений полезного использования предлагаемого статистико-лингвистического анализа ТМИ, которое особенно востребовано современной практикой ИТО испытаний РКТ, связано с исследованием различных алгоритмов преобразования данных телеизмерений. Предположим, что в результате преобразований, исходный телеметрируемый параметр (ТМП), графическое представление которого приведено на рис. 5

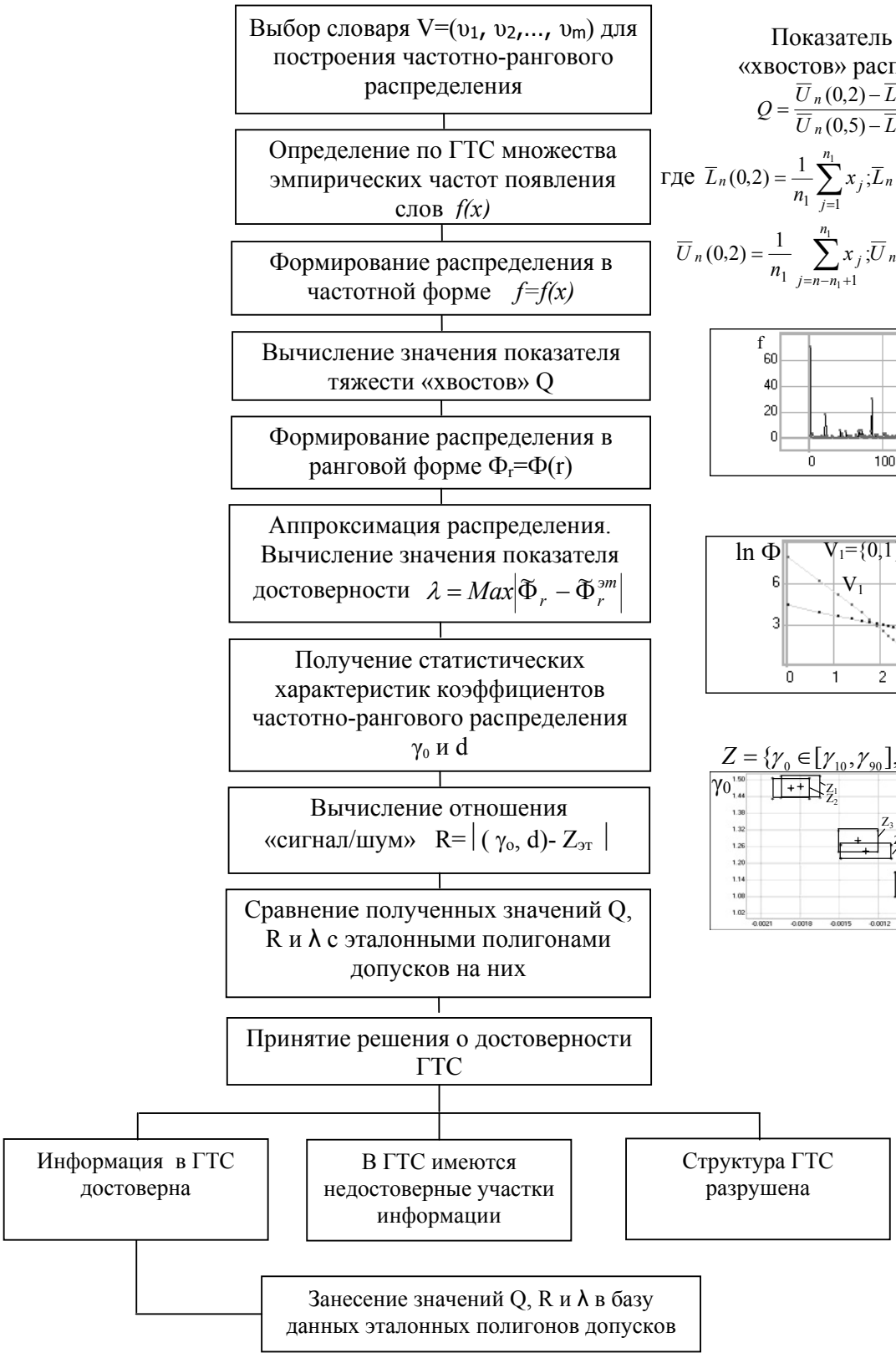

Рис.4. Методика статистико-лингвистической идентификации структур цифровых групповых телеметрических сигналов

вверху, приведен к виду, изображенному на рис. 5 внизу. Необходимо определить, как при этом изменятся кластеры, составленные из коэффициентов γ_0 и d ?

Проведем статистический анализ слов исходного кода. Длина переданных данных в битах равна N_b=5864. Найдем количество одинарных единиц и нулей («1» и «0»), двойных одноименных блоков «11» и «00», тройных блоков, состоящих из одних и тех же символов и т.д. Под-

считаем частоту повторений выделенных блоков относительно всего сигнала и упорядочим полученные значения в порядке убывания. Полученный результат отображен в табл. 4, а ранжированное эмпирическое распределение представлено на рис 6.

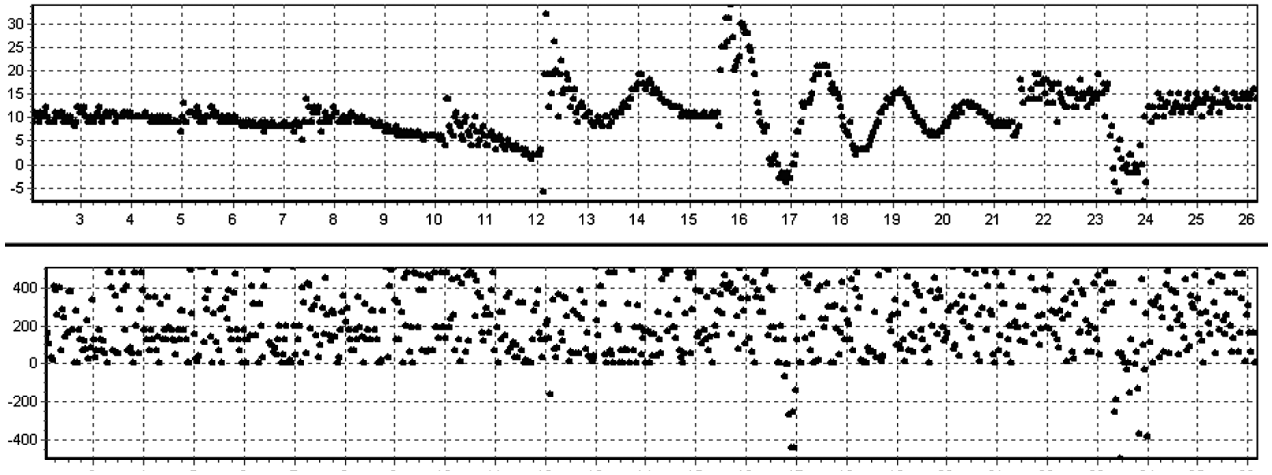

Рис.5. Графическое представление исходного и преобразованного телеметрируемого параметра

Таблица 4

Результаты расчета

Name	Kol	Chast
1	987	0,1683151432469 3
2	376	0,0641200545702592
4	306	0,0521828103683492
5	179	0,030525238744884
3	131	0,0223396998635744
6	124	0,0211459754433834
7	73	0,0124488403819918
8	31	0,00528649386084584
10	2	0,00034106412005457

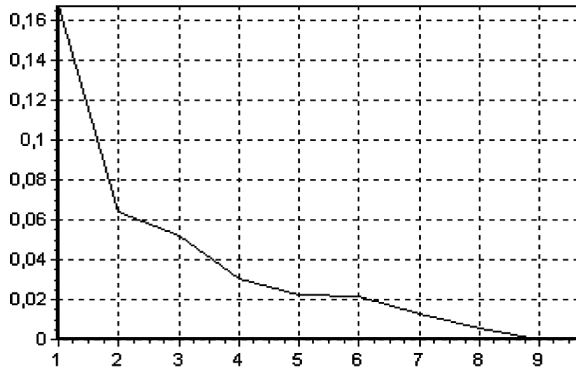

Рис. 6. Ранжированное эмпирическое распределение частости встречаемости блоков, составленных из одноименных двоичных символов

Далее найдем коэффициенты частотно-рангового распределения $\Phi(r)$.

Для этого, разбивая всю последовательность данных на несколько перекрывающих интервалов, соответствующих установленным статистическим выборкам $N_в$ (рис. 7), находим значения γ_0 и d для каждого из выделенных интервалов. После чего определяется кластер статистико-лингвистических распределений К(γ_0, d) (рис. 8).

В результате получили графическое отображение

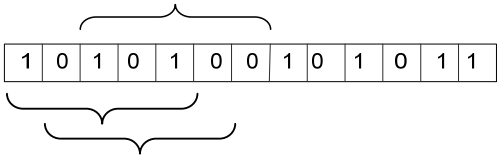

Рис. 7. Иллюстрация предлагаемого способа формирования статистических выборок

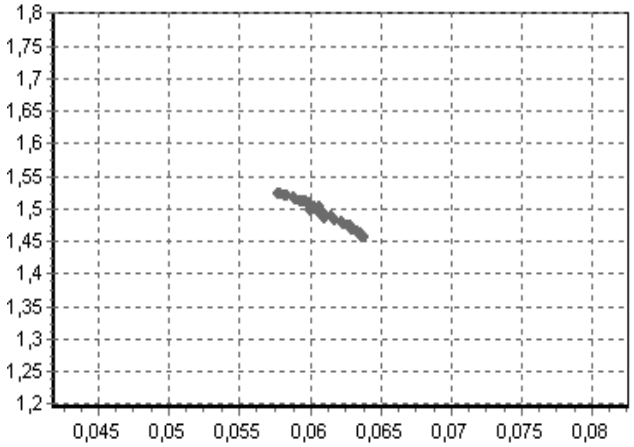

Рис. 8. Графическое отображение кластера $K_{mp}(\gamma_0, d)$

кластера $K_{mp}(\gamma_0, d)$ для случая традиционного представления анализируемого ТМП $x(t)$.

Выполним те же операции, но по отношению к преобразованному ТМП $x_n(t)$. Полученные конечные результаты в виде кластера $K_n(\gamma_0, d)$ представлены на рис. 9.

В результате преобразований получили более компактное представление кластера $K_n(\gamma_0$, d) по отношению к кластеру $K_{mp}(\gamma_0$, d). Это свидетельствует о том, что в результате преобразований устраняются демаскирую-

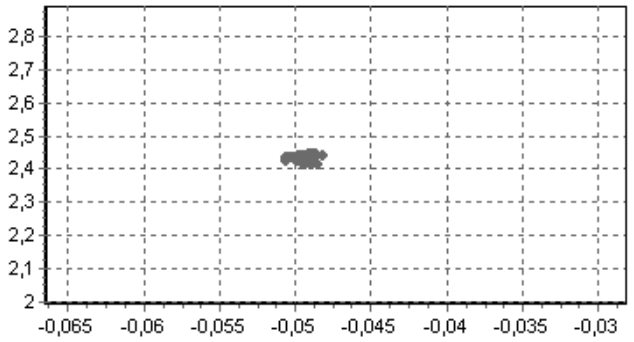

Рис. 9. Графическое отображение кластера $K_n(\gamma_0, d)$ преобразованного ТМП $x_n(t)$

щие признаки, позволяющие идентифицировать ТМП в его исходном виде $x(t)$.

При проведении исследований установлено, что семантически нагруженные последовательности данных цифровых ГТС характеризуются частотно-ранговым распределением с ярко выраженными свойствами асимметрии, которая свойственна гиперболическим законам. Рассмотренные частотно-ранговые распределения и отмеченные закономерности их параметров являются важной характеристикой, отражающей статистические и структурные свойства ГТС. Совместное их использование с научно-методическим аппаратом теории распознавания образов даёт гибкий и мощный инструмент идентификации разнообразных информационных процессов.

Литература

1. Мандельброт Б. Фракталы, случай и финансы: Пер. с англ. - Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2003. – 255 с.

2. Арапов М.В. ,Ефимов Е.Н., Шредер Ю.А. О смысле ранговых распределений.- http://www.kudrinbi.ru/public/442/index.htm.

3. Кукушкин С.С. Статистико-лингвистические методы идентификации телеизмерений и оценивания достоверности телеметрической информации – М.: Измерительная техника, изд. «Стандарты», №8, 2000. – с.13-19.

4. Кукушкин С.С. Теория конечных полей и информатика: том.1 Методы и алгоритмы, классические и нетрадиционные, основанные на использовании конструктивной теоремы об остатках – М: МО РФ, 2003. – 281с.

5. Кукушкин С.С. Повышение эффективности передачи телеметрической информации на основе представления результатов телеизмерений в системе остаточных классов // Измерительная техника. М.: "Стандарты", №7, 1995.- с.11.

6. Кукушкин С.С. Теоретические основы построения перспективных систем телеизмерений, сбора и обработки телеметрической информации // Измерительная техника. М.: "Стандарты", №12, 1995.- с34.

Материал поступил в редакцию 28. 05. 2012 г.