

УДК 519.2; 004.056.53

doi: 10.21685/2072-3059-2024-1-3

Варианты гиперболического закона Ципфа для фрактального описания вероятностно-рангового распределения букв в технических текстах на английском языке

А. И. Иванов¹, А. П. Иванов², А. П. Юнин³, Р. В. Еременко⁴^{1,3}Пензенский научно-исследовательский электротехнический институт, Пенза, Россия^{2,4}Пензенский государственный университет, Пенза, Россия¹ivan@pniei.penza.ru, ²ap_ivanov@pnzgu.ru, ³unin_ap@pniei.penza.ru, ⁴tsib@pnzgu.ru

Аннотация. *Актуальность и цели.* Рассматривается расширение области применения классического гиперболического закона Ципфа. Ранее этот закон применялся для статистического описания вероятности появления слов в текстах на одном из европейских языков. *Материалы и методы.* Выполнено упорядочивание букв по вероятности их использования в английских текстах длиной в 3799 знаков. *Результаты.* Показано, что статистики часто используемых малых букв оказываются легко отделимы от статистик больших букв. Даже на текстах малого объема упорядочивание кодировки букв по вероятности их применения дает еще одно гиперболическое распределение Ципфа для букв и знаков препинания. *Выводы.* Приводится распределение длин слов, выделенных с двух сторон пробелами по методу Мандельброта. Дополнительно приводится выделение слов между ASCII кодами «101», соответствующими английским буквам «e». Показано, что для всех иных часто употребляемых букв английского языка «b», «a», «i», «r» могут быть построены соответствующие функционалы, имеющие линейную вычислительную сложность. Как результат – мы получаем ряд новых статистических функционалов для более глубокого анализа текстов на европейских языках. Новые статистические функционалы могут быть использованы для оценки стойкости длинных парольных фраз.

Ключевые слова: закон распределения, гиперболический закон распределения слов Ципфа, гиперболический закон распределения букв

Для цитирования: Иванов А. И., Иванов А. П., Юнин А. П., Еременко Р. В. Варианты гиперболического закона Ципфа для фрактального описания вероятностно-рангового распределения букв в технических текстах на английском языке // Известия высших учебных заведений. Поволжский регион. Технические науки. 2024. № 1. С. 39–47. doi: 10.21685/2072-3059-2024-1-3

Variants of Zipf's hyperbolic law for fractal description of probability-rank distribution of letters in English technical texts

A.I. Ivanov¹, A.P. Ivanov², A.P. Yunin³, R.V. Eremenko⁴^{1,3}Penza Scientific Research Electrotechnical Institute, Penza, Russia^{2,4}Penza State University, Penza, Russia¹ivan@pniei.penza.ru, ²ap_ivanov@pnzgu.ru, ³unin_ap@pniei.penza.ru, ⁴tsib@pnzgu.ru

Abstract. *Background.* Extension of the scope of the classical hyperbolic Zipf's law. Previously, this law was used to statistically describe the probability of occurrence of words in

texts in one of the European languages. *Materials and methods.* The letters are sorted according to the probability of their use in English texts with a length of 3799 characters. *Results.* It is shown that the statistics of frequently used small letters are easily separable from the statistics of large letters. Even on short texts, ordering letter encodings by their probability of occurrence yields another hyperbolic Zipf distribution for letters and punctuation marks. *Conclusions.* The distribution of the lengths of words marked with spaces on both sides according to the Mandelbrodt method is given. Additionally, the selection of words between the ASCII codes “101” corresponding to the English letters “e” is given. It is shown that for all other frequently used letters of the English language “t”, “a”, “i”, “r” the corresponding functionals with linear computational complexity can be constructed. As a result, we get a number of new statistical functionals for a deeper analysis of texts in European languages. New statistical functionals can be used to evaluate the strength of long passphrases.

Keywords: distribution law, Zipf’s hyperbolic word distribution law, hyperbolic letter distribution law

For citation: Ivanov A.I., Ivanov A.P., Yunin A.P., Eremenko R.V. Variants of Zipf’s hyperbolic law for fractal description of probability-rank distribution of letters in English technical texts. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskoe nauki = University proceedings. Volga region. Engineering sciences.* 2024;(1):39–47. (In Russ.). doi: 10.21685/2072-3059-2024-1-3

Введение

Энтропия Шеннона, которая широко используется на практике специалистами по информационной безопасности для оценки стойкости парольных фраз, имеет экспоненциальную вычислительную сложность. По этой причине энтропию Шеннона невозможно оценить для длинных парольных фраз. Поэтому оценка информационной энтропии длинных парольных фраз с полиномиальной вычислительной сложностью представляет интерес для практики. Например, для практики оценки стойкости хорошо запоминаемых парольных фраз со смыслом. Для этой цели, например, вполне может быть использована энтропия Ципфа.

Закон, сформулированный Джорджем Кингсли Ципфом, показывает линейную связь вероятности использования того или иного слова в тексте с его рангом по словарю. При этом Ципф использовал статистику слов своей собственной книги [1]. Распределение для английских текстов, взятое из первоисточников [2–4], приведено на рис. 1. Размер текста может быть любой, при увеличении размера текста число точек увеличивается, однако общая тенденция их расположения рядом с прямой сохраняется.

Позднее это свойство распределения слов было использовано Шенноном [5] при оценке энтропии английского языка. Ципф выполнял свои расчеты, не обосновывая объем используемого текста и словаря [1]. Шеннон вынужден был поступать иначе, он дал обоснование достаточного объема текста в 8727 слов. Размер текста и число букв в каждом слове для Шеннона играли важную роль, так как вычисление энтропии букв и их сочетаний его методом является задачей экспоненциальной вычислительной сложности, которая и сейчас недоступна для обычных вычислительных машин в случае длинных текстов.

Анализ текста из 8727 слов не является сегодня сложной задачей. Ее решение вполне возможно даже силами одного человека. Тем не менее всегда

остается желание снизить вычислительные затраты оценок. Например, можно попытаться корректно снизить размер анализируемого текста. Кроме того, можно отказаться от оценки вероятностей обнаружения длинных кодов в длинных текстах и перейти к анализу расстояний Хэмминга. В этом случае удастся снизить вычислительную сложность с экспоненциальной до линейной¹ [6, 7], однако шкала энтропии Шеннона и шкала энтропии Хэмминга не совпадают. Кроме того, существуют еще несколько методов оценки энтропии. В связи с этим возникает двуединая задача синтеза новых экономных в вычислительном плане методов оценки энтропии и одновременного создания таблиц преобразования их шкал к наиболее часто применяемой (классической) шкале энтропии Шеннона.

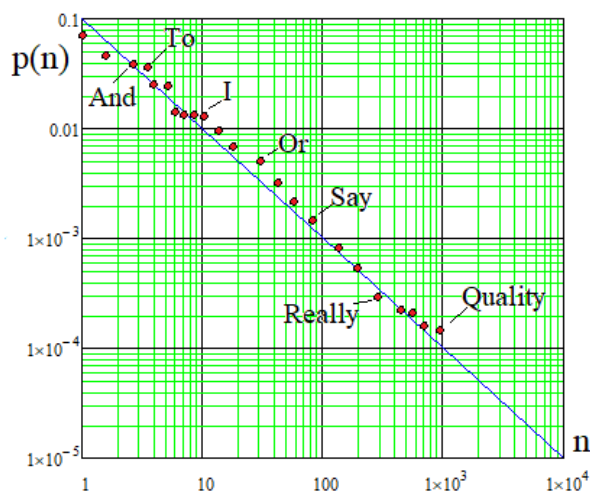


Рис. 1. Гиперболическое распределение вероятности появления слов в английских текстах

Переход от слов текста по Ципфу к буквам и знакам текста

Одним из методов потенциального снижения вычислительных затрат оценки энтропии по Ципфу является отказ от анализа крупных элементов текста – слов, с переходом к более мелким элементам – знакам текста. Оценка энтропии по Ципфу не требует находить на огромных текстах вероятности сочетаний разных знаков. Уже имеется заранее созданный словарь слов, где указаны вероятности их появления в длинных текстах того или иного языка. В этом случае оценка энтропии Ципфа имеет линейную вычислительную сложность: увеличение длины анализируемого текста в 2 раза приведет примерно в двукратному росту вычислений.

Пробел (его ASCII-код), использованный Ципфом как разделитель кусков текста, является наиболее часто используемым знаком. Пользователи, которым недостаточно приближенной оценки энтропии по Ципфу, могут создать свои словари, добавив как дополнительный разделитель ASCII-код сле-

¹ ГОСТ Р 52633.3–2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора. М. : Стандартинформ, 2011.

дующей по частоте появления буквы. Тогда словарь новых фрагментов текста должен увеличить свой объем. Точность оценки с использованием нового словаря увеличится (приблизится к оценкам по шкале Шеннона). Такой прием впервые применил Бенуа Мандельброт [8–10], в том числе для текстов из случайных знаков. В этом случае можно выбирать как разделители любые буквы, поскольку они равновероятны.

Далее даются примеры применения этого простого правила для анализа текстов на английском языке. На некоторых рисунках видны характерные изломы, обусловленные скачком от часто используемых фрагментов к группе редко используемых фрагментов. При этом объем малых выборок для корректного анализа остается приемлемым для быстрых вычислений.

Если воспользоваться сравнительно небольшим текстом объемом в 3799 знаков, то можно достаточно точно оценить вероятности появления в тексте того или иного знака. При этом мы знаем, что Ципф при выводе своего закона использовал данные о вероятности появления слова и ранге слова по словарю. Для того чтобы получить ранг букв, упорядочим их в соответствии с их вероятностью появления в тексте. Перейдя к логарифмическим координатам вероятности обнаружения букв и их рангов, мы получим кривую, отображенную на рис. 2.

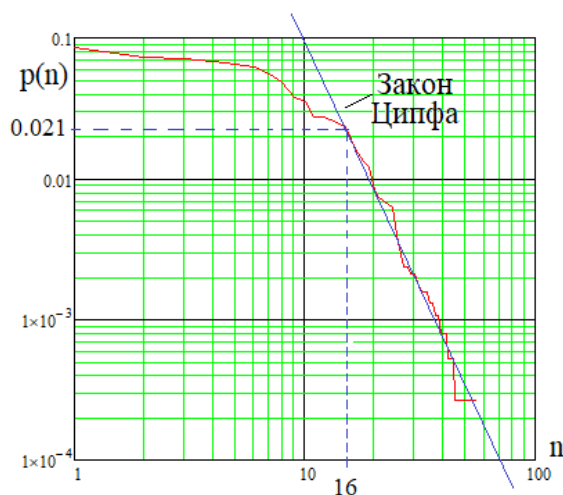


Рис. 2. Связь вероятности появления букв и их рангов с прямой закона Ципфа (технический текст на английском языке в 3799 знаков)

Из рис. 2 видно, что данные, представленные в логарифмических координатах начиная со знака (буквы) 16-го ранга, хорошо ложатся вокруг прямой линии, т.е. закон Ципфа соблюдается не только для слов текста, но и для знаков (букв) текста. Всего в анализируемом английском тексте объемом 3799 знаков было обнаружено 59 разных ASCII-кодов. При этом английский алфавит состоит из 26 малых букв и 26 заглавных букв. Знак пробела (ASCII-код – 32) появляется в тексте с наибольшей вероятностью 0,141. Следующий по частоте появления является ASCII-код – 101, кодирующий английскую букву «e» (вероятность появления 0,085).

Следует также отметить, что Ципф при выводе своего закона (см. рис. 1) пользовался для ранжирования слов словарем. Естественно, что в словаре нет

«пробела», цифр и знаков препинания, тем не менее все эти элементы активно используются в текстах. Если уйти от слов и словарей, то наряду с кодами букв приходится использовать коды пробелов и коды знаков препинания. Кроме того, ASCII-кодов мало – 256 штук, что намного меньше объемов словарей для естественных языков. В этом контексте целесообразно отказаться от логарифмирования рангов. В итоге мы получаем вариант закона Ципфа для букв и знаков с логарифмической шкалой только для вероятностей (рис. 3).

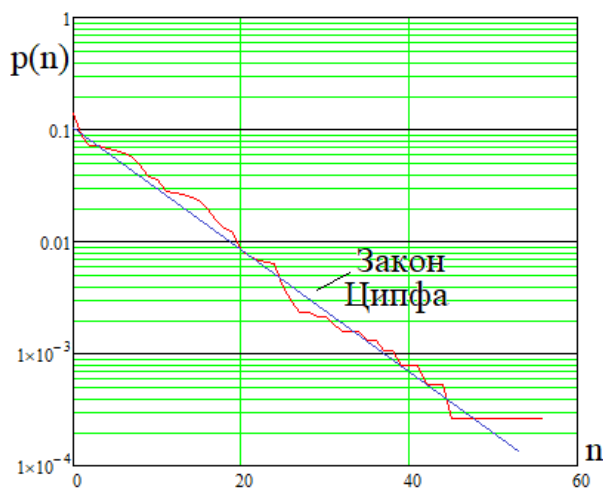


Рис. 3. Второй вариант практически линейного закона Ципфа для букв и знаков текста

Еще одним полезным вариантом трактовки закона Ципфа для букв и знаков является инверсия следования словарных рангов обнаруженных ASCII-кодов при сохранении логарифмических масштабов по координатам. Эта ситуация отображена на рис. 4.

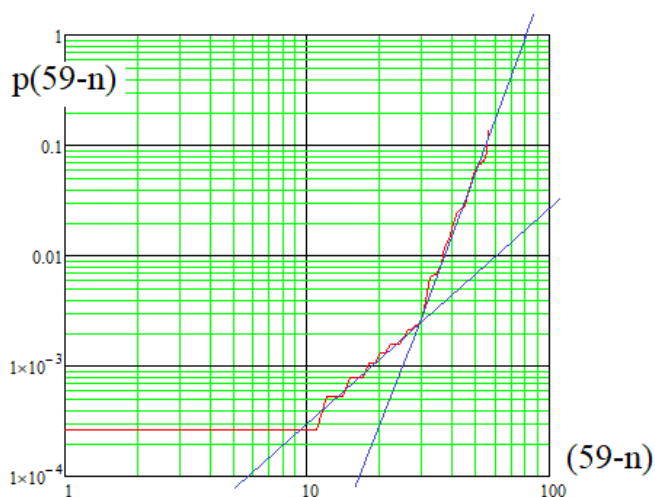


Рис. 4. Три фрагмента функции Ципфа, хорошо приближаемых линейно

Первый горизонтальный участок соответствует ситуации, когда 11-ков в анализируемой вектор-строке текста обнаружены однократно. Второй участок соответствует кодам редко встречающихся заглавных букв текста. Третий участок соответствует часто встречающимся кодам малых букв, пробелов и знаков препинания. Эта форма представления данных позволяет утверждать, что очень редкие (однократные) события, а также события с низкой и высокой вероятностью появления в выборке реальных данных могут быть легко линейно разделимы.

Рост объема извлекаемой информации из относительно небольших текстов

Одним из приложений более глубокого статистического анализа Ципфа являются более поздние работы Шеннона и Мандельброта. Шеннон [5] показал, что, пользуясь законом Ципфа, удастся смягчить проблему экспоненциальной сложности оценки энтропии естественных языков. Мандельброт в своих ранних работах [9–11] показал, что обезьяна, случайно нажимающая на кнопки клавиатуры, порождает случайный язык с распределением Ципфа. При этом Мандельброт предложил в непонятном обезьяньем языке как слова рассматривать интервалы между ASCII-кодами пробелов.

Если перенести этот статистический прием на статистический анализ коротких текстов, то мы получим распределение, представленное на рис. 5,а.

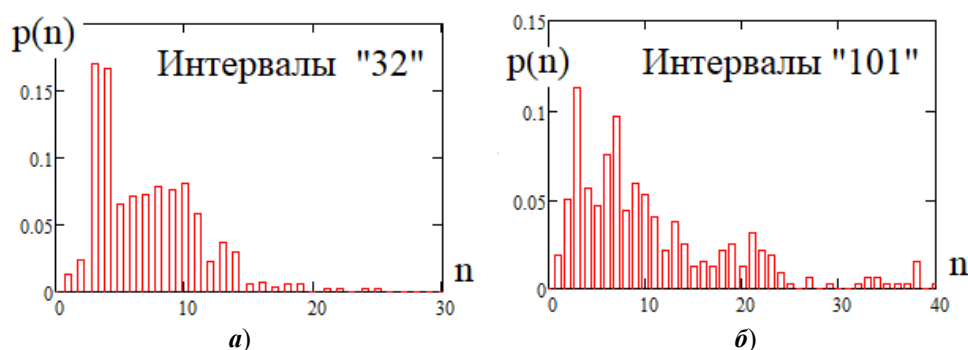


Рис. 5. Интервалы между пробелами (или длина слов по Мандельброту) (а) и распределение интервалов между английскими буквами «е» (ASCII-код «101») (б)

Следующей по интенсивности появления в англоязычных текстах является буква «е». Если интервалы между буквами «е» рассматривать как слова в тексте, то распределение длин таких слов будет существенно отличаться от распределения длин обычных слов. Сравнить два разных распределения можно по данным рис. 5.

Очевидно, что подобные распределения длин слов и соответствующие словари с их разными рангами могут быть построены и для других часто встречающихся в текстах букв английского алфавита: «t», «a», «i», «r». Мы наблюдаем многократный рост информации, извлекаемой из коротких английских текстов. Для каждой из часто встречающихся в тексте букв возникает свой гиперболический закон Ципфа. При этом слова в новых словарях будут иметь собственные независимые статистические характеристики.

Заключение

Все вышеизложенное позволяет ожидать новых результатов по оценке стойкости к атакам подбора длинных осмысленных и потому легко запоминаемых парольных фраз [11, 12]. Принципиально важным в новом подходе к решению задачи является то, что энтропия Ципфа и энтропия Шеннона взаимно дополняют друг друга. Энтропия Шеннона дает конструктивные технически реализуемые алгоритмы оценок для коротких парольных фраз. Энтропия Ципфа, напротив, дает достаточно достоверные оценки для длинных осмысленных паролей (осмысленных текстов). Предположительно, предложенный в данной работе подход может позволить создать таблицы связи энтропии Шеннона и энтропии Ципфа. Если это удастся, методика синтеза таблиц связи двух типов энтропии может оказаться полезной для выполнения подобной работы для других известных на сегодня типов оценки энтропии.

Список литературы

1. Zipf G. K. Human behavior and the principle of least effort. Cambridge, (Mass.) : Addison-Wesley, 1949. P. 573.
2. Brillouin L. Science and information theory. New York : Academic Press, 1956.
3. Шредер М. Фракталы, хаос, степенные законы. Миниатюры из бесконечного рая. Ижевск : НИЦ «Регулярная и хаотическая динамика», 2001. 528 с.
4. Мандельброт Б. Фрактальная геометрия природы. М. : Институт компьютерных исследований, 2002. 656 с.
5. Shannon C. Prediction and entropy of printed English // Bell System Technical Journal (BSTJ). 1951. Vol. 30, № 1. P. 50–64. doi: 10.1002/j.1538-7305.1951.tb01366.x
6. Иванов А. И., Юнин А. П., Бояршинов М. А. Оценка энтропии длинных кодовых слов на выходе нейросетевого преобразователя биометрии в пространствах множества сверток Хэмминга // Интеллектуальные системы в производстве. 2019. Т. 17, № 2. С. 30–36. doi: 10.22213/2410-9304-2019-2-30-36
7. Иванов А. И., Юнин А. П. Эмбрион искусственного интеллекта: компактная нейросетевая проверка качества случайных последовательностей, полученных из биометрических данных : препринт. Пенза : Изд-во ПГУ, 2021. 68 с.
8. Mandelbrot B. B. On the theory of word frequencies and on related Markovian models of discourse. In Structure of Language and Its Mathematical Aspects: Proceedings of Symposia on Applied Mathematics Volume 3 / Jakobson R., Ed.; Amer. Math. Soc.: Providence, RI, USA, 1961. P. 190–219.
9. Mandelbrot B. B. The Variation of Certain Speculative Prices // The Journal of Business. 1963. Vol. 36, № 4. P. 394–419.
10. Mandelbrot B. B. New Methods in Statistical Economics // Journal of Political Economy. 1963. Vol. 71. P. 421–440.
11. Волчихин В. И., Иванов А. И., Безяев А. В., Елфимов А. В., Юнин А. П. Оценка эффекта ускорения вычислений, обусловленного поддержкой квантовой суперпозиции при корректировке выходных состояний нейросетевого преобразователя биометрии в код // Известия высших учебных заведений. Поволжский регион. Технические науки. 2017. № 1. С. 43–55. doi: 10.21685/2072-3059-2017-1-4
12. Юнин А. П., Иванов А. И., Ратников К. А., Кольчугина Е. А. Оценка качества «белого» шума: реализация теста «стаи обезьян» через множество сверток Хэмминга, построенных для разных систем счисления // Известия высших учебных заведений. Поволжский регион. Технические науки. 2018. № 4. С. 54–64. doi: 10.21685/2072-3059-2018-4-5

References

1. Zipf G.K. *Human behavior and the principle of least effort*. Cambridge, (Mass.): Addison-Wesley, 1949:573.
2. Brillouin L. *Science and information theory*. New York: Academic Press, 1956.
3. Shreder M. *Fraktaly, khaos, stepennye zakony. Miniatury iz beskonechnogo raya = Fractals, chaos, power laws. Miniatures from endless paradise*. Izhevsk: NITs «Regulyarnaya i khaoticheskaya dinamika», 2001:528. (In Russ.)
4. Mandel'brot B. *Fraktal'naya geometriya prirody = Fractal geometry of nature*. Moscow: Institut komp'yuternykh issledovaniy, 2002:656. (In Russ.)
5. Shannon C. Prediction and entropy of printed English. *Bell System Technical Journal (BSTJ)*. 1951;30(1):50–64. doi: 10.1002/j.1538-7305.1951.tb01366.x
6. Ivanov A.I., Yunin A.P., Boyarshinov M.A. Estimation of the entropy of long code-words at the output of a neural network biometric converter in the spaces of a set of Hamming convolutions. *Intellektual'nye sistemy v proizvodstve = Intelligent systems in production*. 2019;17(2):30–36. (In Russ.). doi: 10.22213/2410-9304-2019-2-30-36
7. Ivanov A.I., Yunin A.P. *Embrion iskusstvennogo intellekta: kompaktная neyrosetevaya proverka kachestva sluchaynykh posledovatel'nostey, poluchennykh iz biometricheskikh dannykh: preprint = Embryo of artificial intelligence: compact neural network quality check of random sequences obtained from biometric data: preprint*. Penza: Izd-vo PGU, 2021:68. (In Russ.)
8. Mandelbrot B.B. *On the theory of word frequencies and on related Markovian models of discourse. In Structure of Language and Its Mathematical Aspects: Proceedings of Symposia on Applied Mathematics Volume 3 / Jakobson R., Ed.; Amer. Math. Soc.: Providence, RI, USA, 1961:190–219.*
9. Mandelbrot B.B. The Variation of Certain Speculative Prices. *The Journal of Business*. 1963;36(4):394–419.
10. Mandelbrot B.B. New Methods in Statistical Economics. *Journal of Political Economy*. 1963;71:421–440.
11. Volchikhin V.I., Ivanov A.I., Bezyaev A.V., Elfimov A.V., Yunin A.P. Evaluation of the effect of accelerating calculations due to the support of quantum superposition when adjusting the output states of a neural network converter of biometrics into code. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2017;(1):43–55. (In Russ.). doi: 10.21685/2072-3059-2017-1-4
12. Yunin A.P., Ivanov A.I., Ratnikov K.A., Kol'chugina E.A. Assessing the quality of “white” noise: implementation of the “flock of monkeys” test through a set of Hamming convolutions constructed for different number systems. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2018;(4):54–64. (In Russ.). doi: 10.21685/2072-3059-2018-4-5

Информация об авторах / Information about the authors

Александр Иванович Иванов

доктор технических наук, профессор,
научный консультант, Пензенский
научно-исследовательский
электротехнический институт (Россия,
г. Пенза, ул. Советская, 9)

E-mail: ivan@pnici.penza.ru

Aleksandr I. Ivanov

Doctor of engineering sciences, professor,
scientific adviser, Penza Scientific
Research Electrotechnical Institute
(9 Sovetskaya street, Penza, Russia)

Алексей Петрович Иванов

кандидат технических наук, доцент,
заведующий кафедрой технических
средств информационной безопасности,
Пензенский государственный
университет (Россия, г. Пенза,
ул. Красная, 40)

E-mail: ap_ivanov@pnzgu.ru

Aleksey P. Ivanov

Candidate of engineering sciences, associate
professor, head of the sub-department
of technical means of information security,
Penza State University (40 Krasnaya
street, Penza, Russia)

Алексей Петрович Юнин

ведущий специалист, Пензенский
научно-исследовательский
электротехнический институт (Россия,
г. Пенза, ул. Советская, 9)

E-mail: unin_ap@pniei.penza.ru

Aleksey P. Yunin

Lead expert, Penza Scientific
Research Electrotechnical Institute
(9 Sovetskaya street, Penza, Russia)

Роман Викторович Еременко

старший преподаватель кафедры
радио- и спутниковой связи Военного
учебного центра, Пензенский
государственный университет
(Россия, г. Пенза, ул. Красная, 40)

E-mail: tsib@pnzgu.ru

Roman V. Eremenko

Senior lecturer of the sub-department
radio and satellite communications,
Military Training Center, Penza State
University (40 Krasnaya
street, Penza, Russia)

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflicts of interests.

Поступила в редакцию / Received 29.06.2023

Поступила после рецензирования и доработки / Revised 10.10.2023

Принята к публикации / Accepted 27.12.2023