УДК004.056.5:518:512.624.3 **Ю.П. Буценко, Н.С. Греков, Ю.Г. Савченко**

СТАТИСТИЧЕСКИЙ ПОДХОД К ВЫЯВЛЕНИЮ СКРЫТОГО СООБЩЕНИЯ В СТЕГОКОНТЕЙНЕРАХ ГРАФИЧЕСКОГО И АУДИО ФОРМАТА

Рассмотрена задача обнаружения скрытых сообщений в файлах аудио и графического форматов на основе статистического анализа соответствующих битовых последовательностей. Предлагаемый подход базируется на выявлении корреляционных связей между соседними отсчетами аудио сигнала или пикселями изображения. Приведены примеры анализа реальных файлов.

Ключевые слова: стеганография, статистический анализ, обнаружение скрытого вложения.

Рассмотрим задачу обнаружения скрытых стегосообщений, развивая подход [1], предложенный ранее для тестирования псевдослучайных битовых последовательностей. Ограничимся случаем использования метода наименее значащего бита (НЗБ). Эта задача на содержательном уровне может быть описана таким образом.

По определению метод предполагает модификацию (изменение в соответствие с битами скрываемого сообщения) младших разрядов отсчетов оцифрованного аудио сигнала либо пикселей изображения, представленного в цифровом формате [2]. Область **Q**, (подмножество битов которое потенциально может быть использовано для загрузки стего) в каждом конкретном случае определяется файлом контейнера и имеет вполне определенную статистическую структуру в смысле распределения частот (вероятностей) появления фрагментов определенной длины в файле контейнера (так, как было предложено для вычисления частных энтропий при тестировании «на случайность»). После загрузки стегосообщения структура контейнера, как можно ожидать, изменится. Выявить наличие (подозрение в наличии) таких изменений можно различными способами, в том числе и с помощью энтропийных соотношений.

Однако здесь возникает принципиальное затруднение. Структура незагруженного (пустого) контейнера стеганоаналитику *неизвестна*, поэтому для выявления изменений нет объекта для сравнения (эталона).В этой ситуации, по-видимому, возможен практически единственный выход – использовать в качестве такого эталона данные анализа структур файлов *реальных* и достаточно *типичных* аудио сигналов и изображений (контейнеров). Речь идет, условно говоря, только об «содержательных» файлах (можно использовать и другой какой-либо термин). Но смысл, который вкладывается, состоит в том, что файл не является произвольной или случайной комбинацией битов, а несет некоторое содержание – текст, звуки, изображение, воспринимаемое человеком как осмысленное. При компьютерной обработке и анализе таких объектов это битовые массивы, заданные в виде последовательности (таблицы) 0 и 1 или соответствующих байтов.

Может возникнуть вопрос, можно ли на основе *формальных* признаков отличить (идентифицировать) тип файла – графический, аудио или текстовый. На первый взгляд, вопрос выглядит лишенным практического смысла. Ведь структура перечисленных типов файлов различна: у графических файлов для растровых изображений это совокупность пикселей, каждый из которых задает цвет и яркость соответствующей точки изображения, для аудио файлов - это совокупность отсчетов, полученных в результате аналогово-цифрового преобразования электрического сигнала от микрофона, текстовый – совокупность байтов, каждый из которых соответствует некоторому символу текста.

Предположим, что структура файла (разбиение на пиксели, байты отсчетов аудио сигнала или символы текста) аналитику известна. Требуется определить, где, например, текстовый файл, а где аудио. Для определенности будем считать, что исходный материал – это достаточно длинная (в статистическом смысле) последовательность битов: в первом случае, символов осмысленного текста на украинском, русском или английском языке, во втором – фрагмент речи, записанной с микрофона и оцифрованной. Чем отличаются эти файлы? Очевидно, ограничиваясь анализом лишь отдельных байтов, различие вряд ли

удастся обнаружить – любое из 256 возможных значений каждого байта является допустимым. Но если анализировать относительные частоты появления отдельных значений, то окажется, что в текстовых файлах для конкретного языка эти частоты имеют вполне определенное и, главное, устойчивые значения. Соответствующее распределение частот (гистограмма) является надежным критерием, позволяющим отличить текст на одном языке от текста на другом [3]. Более того, практически со 100%-ой достоверностью можно идентифицировать язык, на котором написан текст (разумеется, если известно частотное распределение символов языка).

Для графических и аудио файлов аналогичных частотных распределений, которые могли бы служить эталоном, не существует. При попытках построить такие распределения оказывается, что относительные частоты появления отдельных значений отсчетов или пикселей зависят от множества трудно формализуемых факторов, таких как характер первичного контента (речь, музыка, вокал, технические шумы). Для графических файлов – живопись, графика, цифровые фотографии, файлы, полученные в результате сканирования и т.п. В силу перечисленных причин сформулировать некоторые общие свойства файлов названных классов весьма затруднительно.

Однако, более глубокий анализ показывает, что такие свойства можно обнаружить – это существенная корреляционная связь между соседними (во времени) отсчетами первичного аналогового аудио сигнала или соседними (но в пространстве) пикселями изображения. То же можно сказать о соседних битах по уроню. В физических терминах эта связь обусловлена малыми изменениями акустического сигнала за период его дискретизации. Например, даже при оцифровке телефонного разговорного сигнала в многоканальных системах связи при частоте дискретизации 8 кГц период дискретизации составляет 1/8000=125мкс. Ясно, что за такое время реальный сигнал изменится мало, либо вообще не изменится на большинстве временных интервалов. А если взять аудио файлы DVD - качества, то этот интервал составляет 1/192000≈5мкс. Это означает, что за время звучания одной форманты (≈20мс) будет проведено более 1000 отсчетов аналогового сигнала, полученного, например, с микрофона. Ясно, что эти отсчеты будут мало отличаться друг от друга, а большинство из них, можно ожидать, будут одинаковыми.

Аналогичная ситуация, судя по всему, будет иметь место и для графических растровых файлов. Как показывают результаты статистической обработки значительного массива реальных аудио файлов (классическая музыка, джаз, речь) и оцифрованных изображений (архитектура, пейзажи, портреты и др.), в соответствующих битовых последовательностях практически всегда присутствует ожидаемая существенная корреляция между ***соседними*** (во времени) отсчетами аудиосигнала и соседними (на плоскости) пикселями изображения. То же самое можно утверждать относительно соседних (по уровню) битов сигнала яркости или громкости, т.е. между младшим битом и соседним с ним старшим.

Для иллюстрации изменения статистики, получаемой с аудио файлов при внесении в них скрытых сообщений было выбрано три файла: queen_dear_friends.wav, vivaldi_sinfonia_in_c_major_presto.wav и hs-noclue.wav. Основные их характеристики приведены в табл. 1.

Таблица 1

Основные характеристики аудио файлов

Имя файла	queen_dear_ friends.wav	vivaldi_sinfonia_in_c_ major_presto.wav	hs-noclue.wav
Вид сигнала	Популярная музыка	Симфоническая музыка	Речевой
Формат	WAV	WAV	WAV

Сжатие	-	-	-
Число каналов	1	1	1
Число бит в выборке	8	8	8
Частота дискретизации	44100 Гц	44100 Гц	22050 Гц
Число выборок	2929536	2880000	351360

Для получения статистической информации, была написана программа на языке программирования python, которая поочередно сравнивает последние биты и два последних с двумя последними битами двух соседних выборок. С помощью аналогичной программы на языке программирования python, которая позволяет записывать один аудио файл в другой аудио файл на основе метода, НЗБ был «замешан» речевой сигнал hs-noclue.wav в файл queen_dear_friends.wav и vivaldi_sinfonia_in_c_major_presto.wav.

Статистическая информация аудио файла vivaldi_sinfonia_in_c_major_presto.wav до и после внесения hs-noclue.wav изображена в табл.2.

Таблица 2

Статистическая информация аудио файла vivaldi_sinfonia_in_c_major_presto.wav до и после внесения речевого сигнала

Комбинация битов	До внесения речевого сигнала	После внесения речевого сигнала
00	0.308796875	0.298341667
01	0.201495486	0.199090625
10	0.201495486	0.199090972
11	0.288211806	0.303476389
0000	0.146387847	0.121317708
0001	0.049480556	0.045243403
0010	0.024662847	0.04120625
0100	0.049463194	0.045363542
1000	0.024648611	0.041062153
1100	0.053200694	0.057763889
1010	0.113097569	0.094755556
0110	0.049307986	0.044773611
0011	0.053169097	0.057739583
1110	0.049523611	0.051189931
1011	0.049556944	0.051165625
0111	0.024670486	0.036832292
1101	0.024672222	0.036784028
0101	0.113343056	0.074098264
1001	0.049288889	0.044942014
1111	0.125526042	0.155761806

Полученные результаты представлены в виде графиков на рис.1-3.

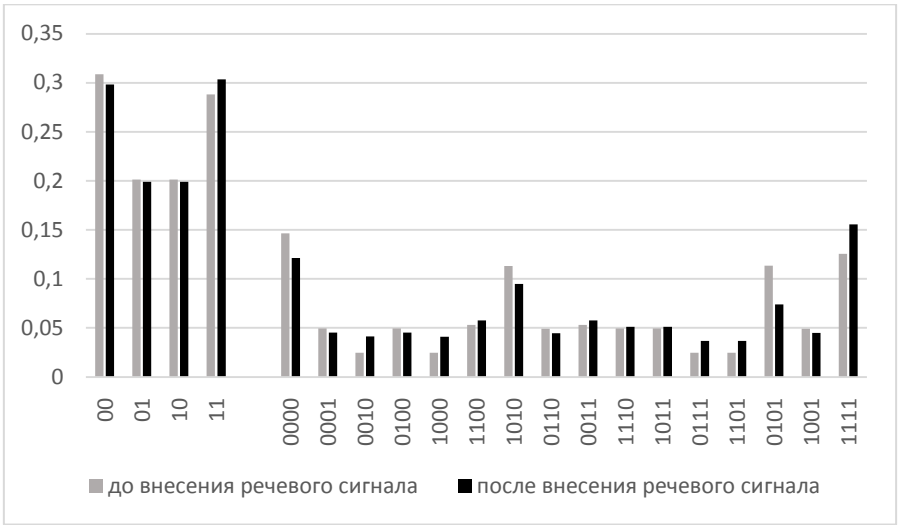

Рис. 1. Статистическая информация аудио файла vivaldi_sinfonia_in_c_major_presto.wav до и после внесения речевого сигнала

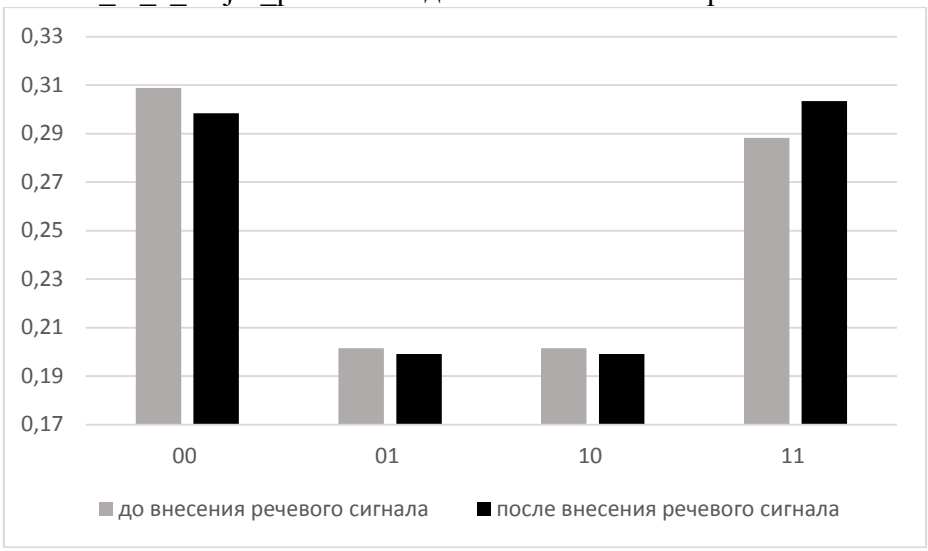

Рис. 2. Статистическая информация аудио файла vivaldi_sinfonia_in_c_major_presto.wav до и после внесения речевого сигнала

Рис. 3. Статистическая информация аудио файла vivaldi_sinfonia_in_c_major_presto.wav до и после внесения речевого сигнала

В общем виде обнаруженная закономерность может быть записана в виде

$$\nu(00,11) > \nu(10,01), \text{ а также}$$
$$\nu(000) > \nu(100,010,001); \nu(111) > \nu(110,101,011) \text{ и}$$
$$\nu(100,001) > \nu(010); \nu(011,100) > \nu(010),$$

где символом ν обозначена относительная частота, с которой встречаются в битовой последовательности соответствующие соседние пары (тройки) битов, а в скобках указаны значения этих битов.

Аналогичные корреляционные зависимости наблюдаются и для частот появления фрагментов соседних битов большей длины, постепенно ослабевая с удалением во времени для аудио файлов и пространстве для изображений.

Таким образом, критерием наличия скрытого вложения в контейнер может служить нарушение закономерностей, задаваемых приведенными выше неравенствами и полученными на основе статистического анализа области **Q**. В качестве эталона (условного), по-видимому, может быть использовано распределение, полученное на основе анализа *всего контейнера*, включая область **Q**. Хотя, строго говоря, вопрос эквивалентности статистик, полученных на всем контейнере и отдельно на материале области **Q**, остается дискуссионным [4].

Далее, используя инструментарий математической статистики и теории вероятностей, попытаемся получить приемлемые для практического применения аналитические соотношения, определяющие наличие (или отсутствие) скрытого вложения в аудио или графический файл.

Далее, используя инструментарий математической статистики и теории вероятностей, попытаемся получить аналитические соотношения, определяющие наличие (или отсутствие) скрытого вложения в аудио или графический файл.

Полученная информация позволяет рассматривать задачу обнаружения "вложения" как статистическую задачу выявления несоответствия распределения по группам символов типу распределений, предполагаемому характерным для данного типа контейнеров. Может быть предложено несколько подходов к решению такой задачи. Наиболее естественным в данном случае является применение критерия согласия Пирсона, статистика которого имеет вид

$$\chi^2 = \sum_{k=1}^{m} \frac{(n_k - n_k^*)}{n_k^*}$$

где $m = 2^n$ - количество возможных позиций в распределении, n_k - наблюдаемые частоты, n_k^* - частоты, характерные для «пустых» контейнеров.

Нулевая гипотеза о совпадении распределений отвергается, если значение статистики превышает соответствующий выбранному уровню значимости квантиль распределения хи-квадрат. Однако, в случае распределений разреженного типа (т.е.когда многие из вероятностей либо относительных частот близки к нулю), а именно такими являются анализируемые распределения, асимптотика статистики хи-квадрат оказывается непредсказуемой. В работе [5] предлагается построение эмпирической функции распределения статистики путем сэмплирования

$$\tilde{F}(x^2) = \frac{1}{N} \sum_{j=1}^{n} I[x^2 > x_j^2],$$

где N - количество сэмплов, x_j - значение статистики для j-го сэмпла, $I[...]$ - индикатор события.

Отметим, что применение такой методики в нашем случае вполне оправдано при условии замены сэмплов случайно выбраными файлами исследуемого типа. Представляется перспективным также использование специфического вида эмпирической функции

распределения, построеной по приведенным выше данным. Хорошо известно, что, например, критерии Колмогорова и омега-квадрат, опирающиеся на статистики

$$K_n = \sup\left[F^*(t) - F_n(t)\right], \omega_n^2 = \int_{-\infty}^{+\infty}\left[F^*(t) - F_n(t)\right]^2 dt,$$

где n - объем выборки, $F^*(t)$ - функция распрелеления для пустого контейнера, $F_n(t)$ - функция распределения для исследуемого файла.

Указанные статистики являются более мощными ,чем критерий хи-квадрат. К сожалению, эти критерии имеют обоснованую асимптотику статистик лишь в случае непрерывных распределений. Однако, в работе [6] предложена методика, позволяющая с помощью преобразования Смирнова эмпирической функции распределения использовать их и в случае дискретных распределений. В этом случае генерируется выборка $y_1,...,y_m$ объема m значений равномерно распределенной на интервале [0,1] случайной величины. Если γ_k - эмпирические относительные частоты, то функция распределения $F_n(t)$ имеет скачки величины γ_k в точках с абсциссами k. Построив значения $Z_k = F_n(k-1) + y_k\gamma_k$, сведем указанные выше статистики к виду

$$K_n = \max\left[F_n(z) - z\right], 0 \le z \le 1$$

$$\omega^2 = \int_0^1 [F_n(z) - z]^2\, dz$$

Вопрос сходимости распределений таких статистик нуждается в дополнительном изучении. Отметим, наконец, что в приведенном примере явно просматривается в загруженном контейнере наличие смеси распределений различных типов

$$F_n(t) = \alpha F^*(T) + (1-\alpha)F_1(t),$$

где $F_1(t)$ - функция распределения для загруженного контейнера.Это позволяет использовать процедуру последовательного перебора (с определенной дискретностью) значений параметра смеси до достижения максимального уровня согласия, что дает возможность оценить объем скрытого сообщения.

Литература

1.Ю. П.Буценко, Г.Н.Розоринов, Ю.Г.Савченко, Общее и селектиное тестирование псевдослучайных битовых последовательностей, Сучасний захист інформації, №2, 2014, с.55-62

2. В. Г. Грибунин, И. Н. Оков, И. В. Туринцев. Цифровая стеганография. – М.: Солон-Пресс, 2002. – 272с.

3. К.Шеннон, Математическая теория связи\\ сб. «Работы по теории информации и кибернетике», - М., ИЛ, - 1963, с.243-332

4. Г.Є.Кузьменко, В.А Литвинов, С.Я.Майстренко. Алгоритм і моделі автоматичної ідентифікації та корекції типових помилок користувача наоснові природної надмірності – Математичні машини і системи =2004, №2, с.134-148

5. В.Р.Целых, К.В.Воронцов, Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании, Машинное обучение и анализ данных,-Т.1,№4, 2012, с.437-447

6.Ю.Б.Лемешко Непараметрические критерии согласия. Руководство по применению.-М. Инфра-М. 2014,-163с.

Надійшла 24.02.2015 р. Рецензент: д.т.н., проф. Розорінов Г.М.