Труды XXV научной конференции по радиофизике

СЕКЦИЯ
«ИНФОРМАЦИОННЫЕ СИСТЕМЫ. СРЕДСТВА, ТЕХНОЛОГИИ, БЕЗОПАСНОСТЬ»

Председатель – Л.Ю. Ротков, секретарь – А.А. Рябов.
Нижегородский государственный университет им. Н.И. Лобачевского.

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ QR-CODE ДЛЯ ПЕРЕДАЧИ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ

К.Д. Холин, С.П. Никитенкова

ННГУ им. Н.И. Лобачевского

QR-код представляет собой двухмерный штрих-код, который может содержать в себе различную информацию. QR расшифровывается как «Quick Response» («быстрый отклик»), что отражает способность устройств и программного обеспечения быстро распознавать код и преобразовывать содержащиеся в нем данные. QR-код был изобретен в 1994 году специалистами подразделения японской корпорации DENSO Wave Incorporated. QR-код пришел на смену штрих-коду. По сравнению со штрих-кодами QR-коды имеют лучшую читаемость, большую емкость и высокую способность обнаружения и исправления ошибок. Сегодня ни одна сфера человеческой деятельности не обходится без этого изобретения: с их помощью шифруются URL-адреса в интернете, осуществляются онлайн-платежи, используются в автоматизированных системах управления техническими процессами и т.д. Технология QR-кодов получила широкое распространение во время эпидемии COVID-19 в качестве пропусков, контроля самоизоляции и т.д.

QR-код может играть роль платежного поручения. Введение банками функции снятия денег с чужой карты через QR-код может открыть новые схемы для мошенничества. QR-коды являются эффективным способом передачи информации по открытому каналу. Однако в качестве машиночитаемого символа QR-код раскрывает хранимую информацию и значительно снижает ее безопасность, что означает, что QR-коды нельзя напрямую использовать для хранения важной конфиденциальной информации.

Для сокрытия данных путем изменения значений пикселей блоков QR-кода могут использоваться стеганографические методы, а также внедрение цифрового водяного знака. В поисках безопасного метода использования QR-кодов может использоваться шифрование. Например, миграционная служба Японии использует зашифрованные QR-коды при выдаче виз в паспортах. В этом случае появляется одна очень важная уязвимость – это компрометация ключа шифрования.

Эффективным методом защиты конфиденциальный информации, передаваемой с помощью QR-кодов, может являться разделение секрета, разрешающее доступ к необходимой секретной информации только при одновременном предъявлении своих частей секрета участниками информационного взаимодействия, не доверяющих друг другу. Схема разделения секрета была впервые предложена Шамиром в 1979 г. [1]. Схема Шамира позволяет реализовать (k,n)-пороговое разделение секретного сообщения между n сторонами так, чтобы только любые k и более сторон $k \leq n$ могли восстановить секрет. При этом любые k-1 и менее сторон не смогут восстановить секрет.

В данной работе в схемах разделения секрета использовались интерполяционные полиномы Лагранжа и Ньютона.

В предложенной схеме сначала выбираем значение k (k <= n), секретный ключ $a0$ и большое простое число p (p> = $a0$). Далее выбираем n участников x_1, x_2, ..., x_n, где n – количество QR-кодов, которые используются для сокрытия информации. Затем полином $f(x)$ степени (k - 1) строится следующим образом:

$$f(x) = a_0 + a_1x^1 + a_2x^2 + \ldots + a_{k-1}x^{k-1} (mod\ p)$$

Таким образом, могут быть сгенерированы пары секретных частей (x_i, $f(x_i) = y_i$) для каждого участника. В процедуре декодирования любой, кто получит k из n секретных частей, восстановит секретные данные $a0$ с помощью интерполяционного полинома Лагранжа

$$f(x) = \sum_{a=1}^{k} y_{ia} \prod_{j=1, j\neq a}^{k} \frac{x - x_{ij}}{x_{ia} - x_{ij}} (mod\ p)$$

При построении схемы разделения секрета может использоваться интерполяционный полином Ньютона. В этом случае коэффициенты существующего секретного полинома не нужно пересчитывать при добавлении нового коэффициента. Система может стать более гибкой и масштабируемой.

$$f(x) = [F_0 + F_1(x - x_0) + \cdots + F_n(x - x_0) \ldots (x - x_{n-1})] (mod\ p)$$

где F_i – разделенная разность i-ого порядка.

$$F_n = \frac{F(x_1, \ldots, x_i) - F(x_1, \ldots, x_{i-1})}{x_{i-1} - x_0}$$

Схема Шамира не создаёт проверяемых фрагментов, поэтому могут быть предъявлены поддельные фрагменты, что препятствует восстановлению правильного секрета. Эта проблема может быть решена с помощью проверяемых схем разделения секрета, таких как схема Фельдмана.

Сгенерированные доли секрета, представляющие собой пару (i, $f(i)$), где i – порядковый номер участника, а $f(i)$ – значение многочлена в этой точке, встраиваются в каждый QR-код. Напрямую считать контент с QR-кодов невозможно, если количество полученных теней не достигает предопределенного порога.

Схема разделения секрета обеспечивает высокий уровень безопасности передаваемых данных и отличается простотой реализации. QR-коды являются удобным способом быстрой доставки информации. Таким образом, комбинация совместного использования схемы разделения секрета и QR-кодов позволяет значительно сэкономить на аппаратных средствах и обслуживании программного обеспечения.

[1] Shamir A. // “How to Share a Secret”. Communication of the ACM. 1979. 22(11). P. 612.
[2] Denso-wave, https://www.qrcode.com/en/

ОПТИМИЗАЦИЯ ПРОЦЕДУРЫ ПОИСКА КРАТКОВРЕМЕННО РАБОТАЮЩИХ ИСТОЧНИКОВ РАДИОИЗЛУЧЕНИЯ В ПАНОРАМНОМ ПРИЕМНИКЕ РАДИОТЕХНИЧЕСКОГО КОНТРОЛЯ

И.Н. Карельский, Л.Ю. Ротков

ННГУ им. Н.И. Лобачевского

Одной из проблемных задач, возлагаемых на панорамные приемники (ПП) оперативного радиотехнического контроля источников радиоизлучений, является своевременное обнаружение радиосигналов с малым временем их присутствия на входе обнаружителя. Ограниченность времени на прием и обнаружение сигнала может быть обусловлена как особенностями работы самого источника (повышенной скрытностью, кратковременностью выполняемой задачи, эпизодичностью излучения в направлении ПП и др.) [1], так и дополнительными затратами времени на подготовку ПП к приему сигнала с заданного направления и на заданной несущей частоте. Кроме того, малое время, отводимое на обнаружение сигналов, может быть задано со стороны потребителя информации ПП, например, от средства постановки радиопомех. Отмеченную особенность функционирования ПП можно обозначить как поиск и обнаружение сигналов кратковременно работающих источников (КИРИ).

Если КИРИ (например, РЛС) излучает импульсные сигналы через направленную вращающуюся антенну, то время присутствия сигналов в точке приема ($t_{\text{изл}}$) определяется длительностью ограниченной серии импульсов не превышающей долей секунд. Время, затрачиваемое на разворот диаграммы направленности (ДН) антенны ПП и настройку рабочей частоты ПП, обуславливает дополнительное сокращение количества импульсов на входе обнаружителя, и снижает эффективность его работы. Поэтому *целью работы* является повышение эффективности поиска и обнаружения КИРИ в условиях ограниченного ресурса времени на их проведение.

При наличии энергетической доступности сигналов в ПП рассматриваются два вида поиска источников [2]: поиск по направлению прихода электромагнитной волны в точку приема и поиск по несущей частоте сигнала.

Применение в ПП антенных решеток с электронным управлением ДН позволяет ускорить поиск и реализовать быстрый гарантированный поиск по направлению. Для этого период вращения ДН антенны ПП $\left(\mathrm{T_p}\right)$ должен быть меньше времени присутствия сигнала в точке приема $\mathrm{T_p} \leq t_{\text{изл}}$. С учетом этого условия можно определить максимально допустимый период перестройки приемника $T_{n\,max}$, обеспечивающий просмотр всего заданного диапазона частот контроля Δf_p на каждом угловом направлении поиска по направлению с помощью ДН шириной $\Delta\beta_p$: $T_{n\,max} \leq \Delta\beta_p T_p/2\pi$. Этот период фактически определяет *ресурс времени* на поиск КИРИ по частоте: $N=T_{n\,max}$. Величина N может быть целочисленной и равной некоторому суммарному числу периодов перестройки с допустимой скоростью, выделяемому для поиска КИРИ в некотором количестве частотных поддиапазонов поиска $(\Delta f_j), j = 1, \dots, n$, при условии, что на каждый поддиапазон (ПД) выделяется некоторое количество x_j единиц указанного ресурса:

$$N = \sum_{j=1}^{n} x_j \; ; \; \Delta f_{\text{р}} = \sum_{j=1}^{n} \Delta f_j \; , j = 1, \dots, n. \quad (1)$$

Основной причиной пропуска сигналов является ограниченное время просмотра приемником элементарных дискретных частотных участков (ДУЧ), равных величине полосы пропускания ПП $\Delta f_{пр}$. Если скорость перестройки: $\gamma = \Delta f_{\text{р}}/\text{Т}_{\text{п}}$, то время просмотра ДУЧ $\Delta t_{пр}=\Delta f_{пр}/\gamma$. Заметим, что допустимая скорость перестройки ПП не должна превышать величину $\gamma_{max} \le \pi \Delta f_{\text{пр}}^2$ [2]. Появление импульсных сигналов на различных ДУЧ фрагментально и может не совпадать с текущим временем просмотра приемником очередного ДУЧ. Поэтому эффективность поиска сигналов по частоте, при заданных полосе пропускания ПП и скорости перестройки, можно оценивать по вероятности совпадения двух импульсных последовательностей, одна из которых определяется характеристиками приемника ($\Delta t_{\text{пр}}, T_n$), а вторая образуется сигнальной последовательностью: периодом повторения импульсов $T_и$ и длительностью импульса $\tau_и$ [2]. При этом вероятность обнаружения сигналов ИРИ по несущей частоте за один период обзора по частоте определяется по формуле:

$$p_1 = \begin{cases} (\tau_{\text{и}} + \Delta t_{\text{пр}})/\text{Т}_{\text{и}}, & \text{если } \text{Т}_{\text{и}} < \text{Т}_{\text{п}} \\ (\tau_{\text{и}} + \Delta t_{\text{пр}})/\text{Т}_{\text{п}}, & \text{если } \text{Т}_{\text{и}} > \text{Т}_{\text{п}} \end{cases} . \quad (2)$$

Если принять за p_{1j} вероятность обнаружения КИРИ в j-м поддиапазоне за один цикл обзора, то за x_j циклов она возрастет до $p_j(x_j) = 1 - (1 - p_{1j})^{x_j}$.

Ниже предлагается алгоритм оптимизации распределения ограниченного временного ресурса при поиске КИРИ по ПД частот с учетом нелинейности функции $p_j(x_j)$, а также с учетом того, что частотные ПД не равнозначны с точки зрения важности получаемой информации о них для потребителя и не одинаковы по вероятности обнаружения.

Эффективность поиска КИРИ по ПД частот можно оценивать по *нормированной суммарной вероятности обнаружения КИРИ* во всем диапазоне контроля $\Delta f_{\text{р}}$ с учетом коэффициента важности каждого поддиапазона (c_j):

$$\text{P}(\text{x}_1, \text{x}_1, \dots \text{x}_n) = \sum_{j=1}^{n} c_j [1 - \left(1 - p_{1j}\right)^{x_j}] \quad (3)$$

Задача достижения максимальной эффективности поиска и обнаружения источников при наличии априорной информации о величинах p_{1j} и c_j может решаться, как задача нахождения оптимального распределения ограниченного временного ресурса по поддиапазонам поиска по критерию максимальной суммарной вероятности:

$$\text{P}(\text{x}_1, \text{x}_1, \dots \text{x}_n) \Rightarrow max, \text{при } \sum_{j=1}^{n} x_j = N; \; x_j \ge 0; \; \Delta f_{\text{р}} = \sum_{j=1}^{n} \Delta f_j; j = 1, \dots, n. \quad (4)$$

Коэффициенты важности c_j определяются методом экспертных оценок (МЭО) на основе имеющихся априорных данных о КИРИ в рассматриваемом ПД, например, по критерию максимального ущерба, наносимого потребителю информации ПП со стороны КИРИ или от объектов, в состав которых они входят.

Полученные МЭО для каждого ПД ненормированные коэффициенты важности $c_{j\text{н}}$ нормируются по отношению к суммарному коэффициенту:

$$c_j = \frac{c_{j\text{н}}}{\sum_{j=1}^{n} c_{j\text{н}}}, j = 1, \dots, n; \sum_{j=1}^{n} c_j = 1. \quad (5)$$

Значения вероятностей p_{1j} находятся по формуле (4). Если в рассматриваемом ПД предполагается нахождение нескольких источников, то определение p_{1j} становится не тривиальной задачей. В частности, может быть принято некоторое усредненное значение, если источники однотипны или наименьшее значение из последовательности значений при разнотипных источниках.

Исходными данными для решения задачи 4 являются: количество ПД; временной ресурс, заданный в дискретных единицах времени – N; значения коэффициентов важности ПД – c_j; значения вероятностей обнаружения источников в ПД – p_{1j}.

Решение задачи производится в соответствии с алгоритмом, предложенным в [3] для оптимального распределения ограниченного запаса однородных ресурсов на основе методов нелинейного программирования.

1) Вводится и вычисляется вспомогательная величина: $\alpha_j = -ln(1 - p_{1j})$, и компоненты разрешающего вектора: $\vec{y} = y_1, y_2, .., y_N$ по формуле:

$$y_j = \overline{y}_j / \sum_{\tau=1}^{n} \overline{y}_\tau, \; j = 1,2, \dots, n, \; где: \; \overline{y}_j = \frac{1}{\alpha_j} = \text{A}_j, \; \text{A} = \sum_{j=1}^{n} \text{A}_j \; . \quad (6)$$

2) Вычисляются величины: $\beta = -ln(c_j \alpha_j), D_j = A_j \, \beta_j, D = \sum_{j=1}^{n} D_j$, и предварительные значения «заданий» на оптимальное распределение ресурса по формуле:

$$v_j(\vec{y}) = v(\vec{y}) - \beta_j, \; где \;\; v(\vec{y}) = \frac{D+N}{A}. \quad (7)$$

3) Проверяется, нет ли среди заданий $v_j(\vec{y})$ величин меньше нуля. Если нет, то осуществляется переход к п. 4. Если есть, то соответствующие поддиапазоны поиска исключаются из рассмотрения и уточняются величины A и D из которых вычитаются соответственно те значения A_j и D_j, для которых $v_j(\vec{y})$ получились отрицательными. Для этих ПД полагаются значения выделяемого ресурса равными 0, после чего осуществляется переход к п. 2.

Итерационный процесс по уточнению «задания» продолжается до тех пор, пока все задания не окажутся положительными. На практике число итераций, как правило, не превышает 3.

4) Рассчитываются компоненты оптимального распределения ограниченного ресурса времени поиска $\vec{x} = (x_1, x_2, .., x_N)$ по формуле:

$$x_j(\vec{y}) = \begin{cases} 0, \text{если } v_j^*(\vec{y}) \le 0 \\ A_j v_j^*(\vec{y}), \text{если } v_j^*(\vec{y}) > 0 \end{cases}, \quad (8)$$

где: $v_j^*(\vec{y})$ – уточненные после итерации значения заданий.

В таблице приведены результаты вычисления нормированной суммарной вероятности обнаружения КИРИ (обзорных РЛС, РЛС обеспечения полевой артиллерии, РЛС ДРЛО типа АВАКС, многофункциональных РЛС зенитных ракетных комплексов, бортовых МФ РЛС и др.), работающих в диапазоне контроля Δf_p =0,1÷12 ГГц, при равномерном [$P_{ср}$ (x_1, x_2,…, x_j)] и оптимальном [$P_{опт}$ (x_1, x_2,…, x_j)] распределении ограниченного запаса временного ресурса (N=70) по семи ПД частот поиска [$x_j(y)_{ср}$ и

$x_j(y)_{опт}$ $j = 1, \dots, 7]$ для заданных априорных значений вероятностей обнаружения и коэффициентов важности по каждому ПД.

Табл.

Поддиапазон	0,1-1,5 ГГц	1,5-3,1 ГГц	3,1-3,5 ГГц	3,5-4,4 ГГц	4,4-4,7 ГГц	4,7-8 ГГц	8-12 ГГц
p_{1j}	0,2874	0,0589	0,116	0,0246	0,6872	0,0475	0,1389
c_j	0,05	0,15	0,2	0,07	0,13	0,15	0,25
$x_j(y)_{опт}$	4,5032	15,4897	18,1681	0	3,4871	19,0165	21,2462
$x_j(y)_{ср}$	10	10	10	10	10	10	10
$p_{опт}(x_j)$	0,7095	0,3025	0,9742	0	0,9967	0,5927	0,9430
$p_{ср}(x_j)$	0,9067	0,3463	0,5783	0,1603	0,9997	0,2887	0,6489
$P_{опт}(x_1, x_2, \dots, x_j)$	0,8796						
$P_{ср}(x_1, x_2, \dots, x_j)$	0,5613						

Вывод: Анализ и сравнение сумм вероятностей обнаружения КИРИ показывают, что оптимальное распределение ограниченного временного ресурса, имеет преимущество по сравнению с равномерным распределением указанного ресурса. Повышение эффективности поиска составляет от 23%, до 52%. Очевидно, что этот выигрыш будет возрастать по мере роста качества имеющейся априорной информации об источниках, находящихся в диапазоне контроля и от степени неодинаковости величин c_j и p_{1j} в контролируемых поддиапазонах.

[1] Ямпольский Л.С. Обобщенный анализ применения средств воздушного нападения ОВС НАТО при проведении военной операции в Югославии «Решительная сила» и в других локальных войнах в 90-х годах – Ульяновск: УлГТУ, 2000, 80 с.
[2] Вакин С. А., Шустов Л. Н. Основы радиопротиводействия и радиотехнической разведки. – М.: Сов. Радио, 1968, 448 с.
[3] Гурин Л.С., Дымарский Я.С., Меркулов А.Д. Задачи и методы оптимального распределения ресурсов. – М.: Сов. Радио, 1968, 464 с.

АНАЛИЗ РЕЗУЛЬТАТОВ РЕАЛИЗАЦИИ ПОДХОДА К ВЫДЕЛЕНИЮ ТЕРМОВ В МОДЕЛИ ЭЛЕКТРОННЫХ ПИСЕМ НА СЛУЧАЙНОСТЬ

С.В. Корелов[1)], А.М. Петров[1)], И.Г. Сидоркина[2)], Л.Ю. Ротков[3)]

[1)] Национальный координационный центр по компьютерным инцидентам
[2)] ПГТУ
[3)] ННГУ им. Н.И. Лобачевского

Введение

В настоящее время одним из важных способов коммуникации является электронная почта, обладающая многочисленными достоинствами, среди основных из которых доступность, оперативность и дешевизна при одновременных больших возможных объемах и разнообразных видах содержимого электронных почтовых сообщений.

Однако ее использование сопровождается некоторыми проблемами. Одной из них является спам, ставший уже классическим бизнес-риском и являющийся проблемой не только для мирового потока электронных писем, но и трафика в общем. Средняя доля спама в почтовом трафике в 2018 году составила 52,48 % [1], в 2019 – 56,51 % [2], а в 2020 – 50,37 % [3]. Также спам является причиной различных негативных последствий для его получателей [например, 4, 5].

В связи с этим исследование, разработка, создание и внедрение новых и совершенствование существующих решений, моделей, алгоритмов, средств, систем и технологий обеспечения безопасности информационных систем, ориентированных на обнаружение (выявление) анонимных массовых непрошенных рассылок электронных писем, является актуальной и практически значимой задачей. При ее рассмотрении важным и актуальным является вопрос выбора значимых (с точки зрения качества обнаружения спама) признаков электронных почтовых сообщений для процесса классификации, что требует проведения соответствующих исследований [6].

В связи с актуальностью и важностью данного направления исследований в задаче обнаружения спама в [7] предложена и в [6] уточнена модель электронных писем, позволяющая специфическим способом выделять текстовые отрезки электронных писем, являющиеся отражением их отличительных признаков, или термы:

$$\Psi_{el} = \langle Prepr, terms, term\_Code \rangle. \quad (1)$$

Модель оперирует с преобразованными в числовой вектор данными, полученными из исходных текстов электронных писем. В качестве ее параметров, оказывающих влияние на выделение термов, в [6, 8-10] обоснованы:

q – количество числовых кодов, сопоставляемых символам текста, в функции преобразования писем в числовой вектор;

Δt – шаг выборки символов текста в функции преобразования писем в числовой вектор;

n – длина выборки (N-граммы – последовательности, порождающей терм).

Корректность, практическая применимость и результативность модели (1) продемонстрированы в [6-9], а в [6, 8-10] обоснован выбор значения параметра n.

Настоящая статья посвящена анализу результатов реализации подхода к выделению термов в модели электронных писем на случайность.

Основная часть

В [8, 9] показано, что применение модели (1) на англоязычном наборе электронных писем (сформирован и описан в [11] с дополнительными изменениями в соответствии с [9], состоит из 6 групп легальных писем общим количеством 16100 писем и 6 групп спамовых писем общим количеством 16420 писем) дает наилучшие результаты обнаружения при численном значении ключевого параметра $n = 1$ и $n = 2$.

При этом очевидно, что выделение термов происходит на основе всего лишь одного или последовательности из двух символов (начало и окончание терма). В связи с этим авторами сделано предположение о возможном наличии случайности результатов предложенного подхода к выделению термов в соответствии с моделью (1). Для проверки данной гипотезы поставлен эксперимент, в качестве значений параметров модели электронных писем для которого приняты следующие:

$q = 256$ – соответствует количеству символов кодировки Windows-1251;

$\Delta t = 1$ – шаг дискретизации равен одному символу;

$n = 1 \ldots 3$.

На основании полученных и обоснованных в [6] результатов все письма указанного англоязычного набора прошли следующую предварительную предобработку:

- удаление повторений символов пробелов, табуляции и переносов строк;
- перевод всех букв в верхний регистр.

Эксперимент и оценка его результатов проводились аналогично [6, 8, 9].

На первом шаге для каждой категории (класса) писем (легальные и спамовые) каждой группы писем были рассчитаны наборы термов в соответствии с моделью (1).

На втором шаге для этих же писем наборы термов были рассчитаны следующим «псевдослучайным» образом. Для каждого письма на основе результатов предыдущего шага определялось множество, элементами которого являлось количество термов (безотносительно к их символьному составу), сгруппированными по их длинам (например, письмо – множество из 5 термов по 10 символов и 3 терма по 5 символов). Далее случайным образом из полученного множества выбиралась длина терма и, начиная с первого символа письма, выделялся первый «псевдослучайный» терм, а значение количества термов с данной длиной уменьшалось на единицу (например, выбран терм длиной 5 символов, тогда множество на следующем шаге примет вид: 5 термов по 10 символов, **2** терма по 5 символов). Данное действие повторялось до тех пор, пока не исчерпывались все длины всех термов исходного множества, а выделение каждого последующего «псевдослучайного» терма начиналось с символа, следующего в письме за последним символом предыдущего «псевдослучайного» терма.

На третьем шаге отдельно для каждого из полученного на шагах один и два наборов термов (обычного и «псевдослучайного») определялись коэффициенты принадлежности каждого письма к легальным или спамовым письмам, за который принято суммарное количество содержащихся в письме термов, встретившихся в соответствующих категориях всех групп.

Решение о принадлежности письма к спамовым или легальным принималось с использованием простейшего решающего правила – по принципу большего суммарного количества термов соответствующей категории.

В качестве мер оценки результатов эксперимента использованы полнота R, точность P обнаружения (классификации) и сбалансированная -мера [6, 8-10, 12-15]. Результаты эксперимента по каждой из этих мер представлены на рисунках 1-3.

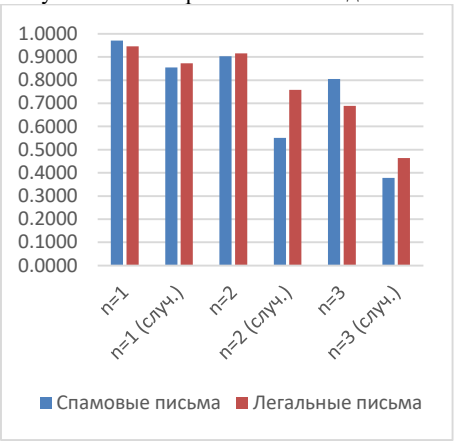

Рис. 1

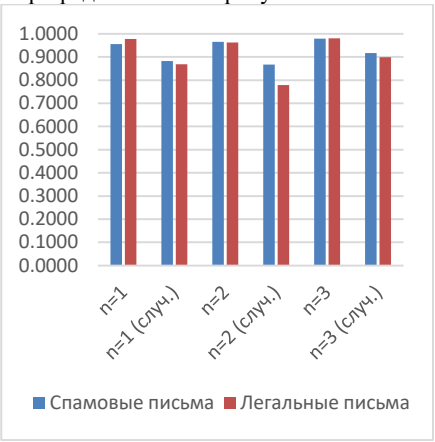

Рис. 2

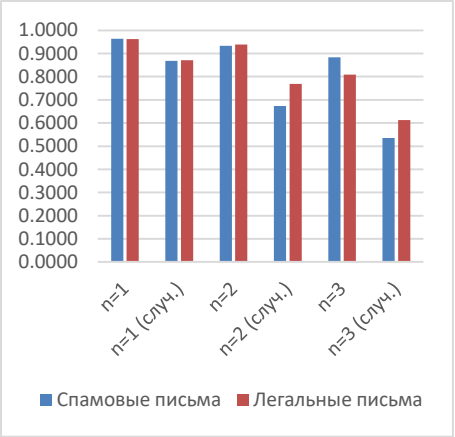

Рис. 3

Анализ полученных результатов показывает значимое различие в полученных значениях при реализации подхода к выделению термов в модели (1) с «псевдослучайным» выделением термов. При этом необходимо отметить, что при $n = 1$ данное различие не столь значительно (около 10 %), как при $n = 2$ и $n = 3$ (более 20 %). Это объясняется применяемым в эксперименте подходом к классификации (учитываются все термы с суммарным единичным весом), а также тем, что возможное максимальное (предельное) число уникальных термов в наборах соответствующих классов при бесконечном увеличении числа писем в обучающих наборах при $n = 1$ достигается гораздо быстрее, чем при $n = 2$ и $n = 3$. Дополнительно проведенные эксперименты с «псевдослучайным» выделением термов показали, что данная разница остается практически одинаковой.

Таким образом, проведенный эксперимент не подтверждает выдвинутую выше авторами гипотезу.

Заключение

Таким образом, результаты эксперимента свидетельствуют о неслучайности результатов реализации подхода к выделению термов в модели электронных писем и подтверждают его обоснованность. При этом косвенно подтверждается сделанный в [10] вывод о целесообразности использования комбинированного подхода при использовании комбинаций численных значений ключевого параметра n модели электронных писем (1) в задаче обнаружения спама.

Одновременно результаты эксперимента показывают целесообразность применения весовых коэффициентов термов и процедур снижения размерности признакового пространства с целью исключения фактора случайности в процессе классификации, обусловленного возможным достижением максимального (предельного) числа уникальных термов в наборах соответствующих классов при бесконечном увеличении числа писем в обучающих наборах.

[1] Вергелис М., Щербакова Т., Сидорина Т. Спам и фишинг в 2018 году [Электронный ресурс] // Securelist. – 2019. Режим доступа: https://securelist.ru/spamand-phishing-in-2018/93453 (дата обращения 19.01.2021).

[2] Вергелис М., Щербакова Т., Сидорина Т., Куликова Т. Спам и фишинг в 2019 году [Электронный ресурс] // Securelist. – 2020. Режим доступа: https://securelist.ru/spam-report-2019/95727 (дата обращения 19.01.2021).

[3] Куликова Т., Щербакова Т., Сидорина Т. Спам и фишинг в 2020 году [Электронный ресурс] // Securelist. – 2021. Режим доступа: https://securelist.ru/spam-and-phishing-in-2020/100408/ (дата обращения 21.04.2021).

[4] Abdulhamid Sh.M., Shuaib M., Osho O., Ismaila I., Alhassan J.K. Comparative Analysis of Classification Algorithms for Email Spam Detection // International Journal of Computer Network and Information Security (IJCNIS). 2018. Vol. 10. No. 1. PP. 60-67. DOI:10.5815/ijcnis.2018.01.07.

[5] Sharaff A., Nagwani N., Dhadse A. Comparative Study of Classification Algorithms for Spam Email Detection // Shetty N., Prasad N., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications. New Delhi: Springer, 2016. PP. 237-244. DOI:10.1007/978-81-322-2553-9_23.

[6] Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Предобработка текстов электронных писем в задаче обнаружения спама // Труды учебных заведений связи. 2020. Т. 6. № 4. С. 80–90. DOI: https://doi.org/10.31854/1813-324X-2020-6-4-80-90.

[7] Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Модель электронных писем в задаче обнаружения спама // Вестник Поволжского государственного технологического университета. Сер.: Радиотехнические и инфокоммуникационные системы. 2020. № 2 (46). С. 44-54. DOI: https://doi.org/10.25686/2306-2819.2020.2.44.

[8] Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. К вопросу об определении численного значения параметра в модели электронных писем // Труды XXIV научной конференции по радиофизике, посвященной 75-летию радиофизического

факультета (Нижний Новгород, 13-31 мая 2020 г.). – Нижний Новгород: ННГУ, 2020. С. 471-474.

[9] Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Определение длины выборки в модели электронных писем // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2020. № 4 (36). С. 31-47. DOI: 10.15593/2224-9397/2020.4.02.

[10] Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Комбинирование значений параметра модели электронных писем // Материалы XII Международной Интернет-конференции молодых ученых, аспирантов и студентов «Инновационные технологии: теория, инструменты, практика» (InnoTech-2020). – Пермь: ПНИПУ. Принята к публикации 14.01.2021.

[11] Metsis V., Androutsopoulos I., Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? // Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006, Mountain View, USA, 27–28 July 2006). 2006. PP. 28–69.

[12] Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. 2002. Vol. 34, No. 1. Pp. 1-47. DOI: https://doi.org/10.1145/505282.505283.

[13] Sebastiani F. Text Categorization // Zanasi A. (ed.). Text Mining and its Applications. Southampton: WIT Press, 2005. Pp. 109-129.

[14] Aas K., Eikvil L. Text Categorisation: A Survey // Norwegian Computing Center. Tech. Report number: 941, 1999.

[15] Manning C., Raghavan P., Shütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. DOI: https://doi.org/10.1017/CBO9780511809071.

ВИЗУАЛИЗАЦИЯ ОБЪЕКТОВ С ДИНАМИЧЕСКИМИ ПАРАМЕТРАМИ, ЗАВИСЯЩИМИ ОТ ГЕОИНФОРМАЦИОННЫХ ДАННЫХ

А.А. Коротышева, С.Н. Жуков

ННГУ им. Н.И. Лобачевского

Введение

Одним из интересных и перспективных направлений информационных технологий в сфере автотранспорта является технология дополненной реальности (augmented reality, AR). Основная идея такой технологии - предоставление водителю навигационной и другой необходимой информации в виде интерактивной проекции на лобовое стекло автомобиля. Так как проецируемая информация находится на уровне глаз водителя, он воспринимает её, не отвлекаясь от вождения и отслеживания ситуации на дороге, что обеспечивает повышение безопасности дорожного движения. Разработки в этом направлении в настоящее время активно проводятся во всем мире [1].

Актуальной задачей в этой области представляется разработка и совершенствование алгоритмов реализации технологии визуализации дополненной реальности для навигационного оснащения автомобиля, повышающих эффективность обработки и отображения объектов с динамическими параметрами, зависящими от геоинформационных данных.

Алгоритм построения маршрута

В качестве источника геоинформационных данных при реализации системы дополненной реальности был выбран OpenStreetMap – проект с открытым исходным кодом, которой является аналогом плиточных картографических сервисов, используемых такими системами, как OpenLayers. OpenStreetMap имеет глобальные векторные данные на уровне улиц и других пространственных объектов [2].

Оптимальным маршрутом в ГИС является маршрут с минимальным расстоянием или временем. Дорожную сеть можно представить в виде графа, где ребра - дороги, а вершины - перекрестки и конечные точки. В ГИС используются алгоритмы с выполнением предварительной обработки графа для ускорения работы отдельных запросов и более эффективным использованием памяти.

Для построения маршрутов в данной работе был выбран маршрутный сервис OSRM (Open Street Routing Machine) – открытый проект с http-сервисом [3], использующий для оптимизации маршрута алгоритм Contraction Hierarchies [4]. В основе этого алгоритма лежат известные алгоритмы поиска маршрута Дейкстры [5], определяемые выражениями (1) и (2) и A* алгоритма, с использованием выражения (3).

$$d[v] = \min_{p:\, u[p]=false} d[p], \tag{1}$$

$$d[to] = min(d[to], d[v] + len), \tag{2}$$

где $d[v]$ – текущая длина для v – вершины кратчайшего пути из s в v. Изначально $d[s] = 0$, а для всех остальных вершин эта длина равна бесконечности.

$$f(n) = g(n) + h(n), \tag{3}$$

где $f(n)$ — минимальная стоимость перехода в соседний узел; $g(n)$ — стоимость пути от начальной вершины до любой другой; $h(n)$ — эвристическое приближение стоимости пути от узла n до конечного узла.

При задании точек маршрута выполняется пересчет координат из широты/долготы в проекцию Меркатора/WGS84 по формулам (4, 5).

$$X = a * long, \tag{4}$$

$$Y = a * ln\left[tan\left(\frac{\pi}{4} + \frac{lat}{2}\right) * \left(\frac{1-e*sin\,lat}{1+e*sin\,lat}\right)^{\frac{e}{2}}\right], e = \sqrt{1-\left(\frac{b}{a}\right)^2}, \tag{5}$$

где $long/lat$ - долгота/широта в радианах ($lat * \pi/180$, $lon * \pi/180$); e - эксцентриситет эллипса, $f = \frac{a-b}{a}, e = \sqrt{2f - f^2}$; a и b - большая и малая полуоси эллипса.

Алгоритм визуализации

Алгоритм визуализации дополненной реальности при построении маршрута с использованием сервиса OSRM и OpenStreetMap представлен на рис. 1.

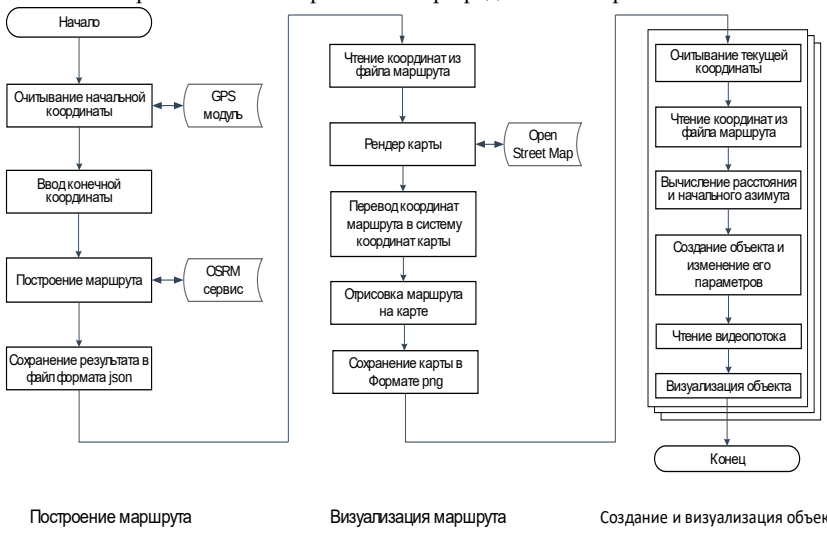

Рис. 1

При построении маршрута определяются начальная и конечная координаты маршрута, их значения обрабатываются и передаются в маршрутный сервис OSRM. Для определения координат используется приемник GPS спутниковой системы навигации. Скаченные с серверов Open Street Map тайлы объединяются в карту, полученные координаты переводятся в систему координат карты и визуализируются посредством 2D-графики.

Видеопоток с камеры поступает в функцию отрисовки объекта, в которой производится обработка и анализ каждого кадра видеопотока. Вычисляется маска объекта, его растяжение и поворот, что позволяет однозначно задать положение объекта в пространстве. Затем объект выводится на экран при помощи графической библиотеки.

Программная реализация и результаты моделирования

Алгоритм визуализации динамических объектов дополненной реальности и обработки геоданных были реализованы в программном коде на языке Python. При проведении моделирования программа, реализующая предложенный алгоритм, в режиме реального времени производила рендеринг карты с отрисовкой маршрута и выводила на экран с изображением дороги, полученным от видеокамеры, дополнительный слой с визуализацией направления движения по заданному маршруту в виде «подсказок» стрелок-указателей (рис. 2).

Рис. 2

Маршрут сохранялся в отдельном файле формата json. При проведении моделирования с gps приемника были записаны данные по скорости (рис. 3), которые, наряду со стрелками-указателями, также могут быть представлены водителю. Сравнение полученных результатов моделирования с результатами работы штатного навигатора автомобиля показало их сопоставимость и подтвердило эффективность разработанного алгоритма.

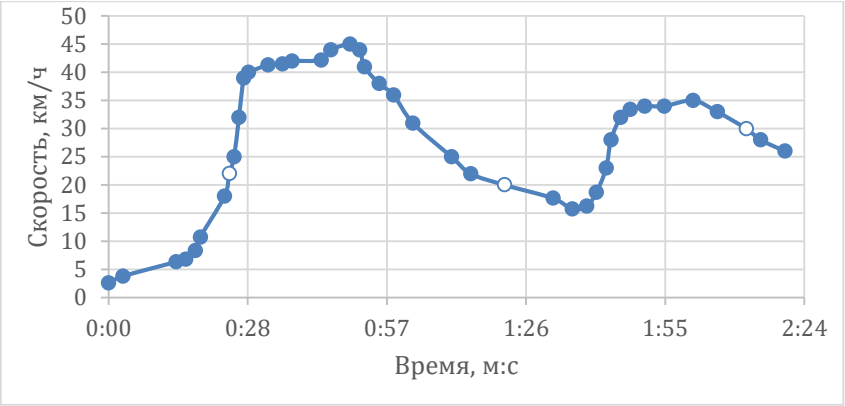

Рис. 3

Заключение

Создан алгоритм визуализации объектов с динамическими параметрами, зависящими от геоинформационных данных с использованием широкодоступных сервисов OSRM и Open Street Map, разработан интерактивный интерфейс, обладающий интегрированным эффектом от совмещения преимуществ навигационных систем и сервисов геоинформационных данных. Разработанный программный код на языке Python может быть использован в системе навигационного оснащения автомобиля. Совершенствование алгоритмов и дальнейшее развитие системы может представлять интерес по следующим направлениям: повышение быстродействия исполнения программного кода путем разработки его части на языке C++, повышение информативности интерфейса (вывод скорости движения, расхода топлива и т.п.), дополнение моделей визуализации голосовыми сообщениями (подсказками).

[1] Charissis V., Papanastasiou S. Human–machine collaboration through vehicle head up display interface // Cogn Tech Work. 2010. Vol. 12. P. 41.
[2] OpenStreetMap - wiki-карта мира [Электронный ресурс] // OpenStreetMap: [сайт]. URL: https://www.openstreetmap.org/ (дата обращения: 13.04.2021).
[3] OSRM API Documentation [Электронный ресурс] // Project OSRM: [сайт]. URL: http://project-osrm.org/docs/v5.22.0/api/#general-options (дата обращения: 31.03.2021).
[4] Robert Geisberger, Peter Sanders, Dominik Schultes, and Daniel Delling, Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks, WEA, 2008, pp. 319-333.
[5] Dijkstra E. W. A Note on Two Problems in Connexion with Graphs // Numerische Mathematik, 1959. Vol. 1. P. 269.

МОДЕЛИРОВАНИЕ СИСТЕМЫ ОБНАРУЖЕНИЯ СЕТЕВЫХ ВТОРЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

В.Д. Мышленник, С.П. Никитенкова

ННГУ им. Н.И. Лобачевского

Одной из актуальных задач в сфере информационной безопасности является создание системы обнаружения аномалий. Под аномалией понимается явное отклонение текущего состояния информационной системы от наблюдаемых ранее и находящихся вне соответствия политике безопасности. Применение технологий искусственного интеллекта (ИИ) и машинного обучения в защите от кибератак становится одним из ключевых направлений в информационной безопасности. Технологии искусственного интеллекта и машинного обучения помогают прослеживать связь событий и применять эвристику для выявления аномалий.

Актуальность решения задачи выбора моделей глубокого обучения для обнаружения аномалий связана с необходимостью анализа большого числа событий безопасности для обнаружения сетевых вторжений.

В данной работе рассматриваются различные алгоритмы машинного обучения в обнаружении и классификации атак.

Для обучения в работе используется набор данных UNSW-NB15 [1,2]. Атаки были выбраны с постоянно обновляемого сайта CVE, при этом нормальное поведение не моделировалось. Трафик на уровне пакетов перехватывался через утилиту TCPdump. К признакам, таким как название протокола, сервис передачи данных, применен метод One-Hot Encoding для представления категориальных переменных в виде двоичных векторов. Далее произведена нормализация значений всех признаков. Для обучения использовалось 75% случайных записей преобразованного для классификации набора данных, оставшиеся 25% для тестирования моделей. Для обнаружения аномалий в системе используется бинарная классификация. Для классификации данные были предварительно разделены на два класса:

- Normal – отсутствие атаки
- Attack – наличие, одной из размеченных в наборе UNSW-NB15, атаки.

В качестве основных метрик оценки качества моделей использовались:

- balanced_accuracy_score – сбалансированная точность
- recall - полнота
- F1-score – F-мера
- ROC AUC – площадь под ROC-кривой.

Анализ показывает, что наилучшие значения качества на обучающем наборе обеспечивает алгоритм AdaBoost. Стоит отметить, что в качестве базового решающего дерева в данном алгоритме использовалось лучшее, полученное для Decision tree. Данный метод классификации обеспечивает 2,7% ложноотрицательных срабатываний и 1,5% ложноположительных срабатываний, что является приемлемым результатом. В таблице приведено сравнение моделей машинного обучения в случае бинарной классификации.

Модель	Оценка			
	balanced accuracy score	recall	F1-score	ROC AUC
Decision tree	0,9791	0,9846	0,9848	0,9791
Random Forest	0,9526	0,9348	0,9526	0,9580
Gaussian NB	0,6181	0,2470	0,3942	0,6181
AdaBoost	0,9806	0,9846	0,9848	0,9791
KNeighbors	0,9791	0,9846	0,9848	0,9791

В качестве диагностического инструмента интерпретации вероятностного прогноза для задач прогнозирующего моделирования двоичной (двухклассовой) классификации использовались кривые ROC, представленные на рисунке, где False Positive Rate – доля ложных положительных классификаций. True Positive Rate – доля верных положительных классификаций. False Positive Rate показывает сколько из обшего числа негативных существующих значений предсказаны как ложно позитивные. True Positive Rate – доля верных положительных классификаций, т.е. насколько алгоритм определяет правильные решения из всех возможных.

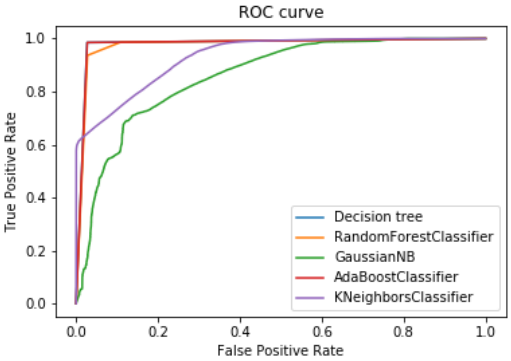

Рис.

Далее рассмотрим классификацию атак по категориям. Результаты представлены в таблице. В размеченном наборе данных UNSW-NB15 выделялись 9 классов атак:

- Fuzzers – генерация случайных данных, чтобы вызвать отказ программы или сети,
- Analysis – содержит различные атаки, связанные со сканированием портов, спамом и внедрением в HTML-файлы,
- Bakdoors – обход механизмов защиты с целью скрытого доступа к данным или программам,
- DoS – отказ в обслуживании сервера или сетевого ресурса,

- Exploits – эксплуатация известных атакующему уязвимостей в операционной системе или программе,
- Generic – техника обнаружения трафика, шифрованного блочным шифром,
- Reconnaissance – разведывательные атаки,
- Shellcode – передача небольших частей кода, используемых для эксплуатации уязвимостей программ,
- Worms – атаки, связанные с самореплицируемыми вирусами.

Модель	Оценка		
	balanced accuracy score	recall	F1-score
Decision tree	0,5713	0,7849	0,7849
RandomForestClassifier	0,4696	0,7469	0,7598
GaussianNB	0,1762	0,1528	0,2348
AdaBoostClassifier	0,5713	0,7853	0,7849
KNeighborsClassifier	0,3348	0,7030	0,6778

Наилучшим образом среди алгоритмов многоклассовой классификации как и на предыдушем этапе является – AdaBoost, он немного улучшает результаты Decision Tree, на основе которого построен.

Анализ матрицы ошибок позволяет сделать следующие выводы: алгоритм AdaBoost достаточно точно выявляет атаки типа Generic(98,5%), Fuzzers (85,8%), Reconnaissance (77,2%) и Exploits (71,1%), Reconnaissance (77,2%). Алгоритм довольно часто считает, что атака относится к классу Exploits вместо реального класса DoS (43,7%),Analysis (31,8%), Worms(38,2%), Backdoor (27,8%). Предположительно, ошибки на данном классе связаны со схожестью по последствиям и признакам атаки типа Fuzzers с другими типами атак.

На основе сравнительного анализа качества алгоритмов бинарной и многоклассовой классификации можно сделать вывод, что для выявления атак в сетях компании можно использовать алгоритм машинного обучения AdaBoost.

Достоинством алгоритмов машинного обучения перед стандартными сигнатурными методами является то, что они могут быть использованы для выявления новых разновидностей и типов атак в пределах описанных классов.

В целом для корпоративных сетей рекомендуется использовать сочетание сигнатурных методов анализа и методов машинного обучения.

[1] Moustafa N., Slay J. // "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS). 2015. P. 1.

[2] Moustafa N., Slay J. // The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set, Information Security Journal: A Global Perspective, 25:1-3. P. 18.

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ КЛАССИФИКАТОРА СЕТЕВОГО ТРАФИКА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ.

Р.Г. Нужный, Л.Ю. Ротков, В.А. Мокляков

ННГУ им. Н.И. Лобачевского

Приложения, использующие разные протоколы прикладного уровня, генерируют потоки транспортного уровня с различными статистическими характеристиками. Для того чтобы правильно классифицировать потоки, необходимо правильно соотнести статистические метрики для приложений, генерирующих эти потоки. Тогда по значениям входных меток можно будет с высокой точностью определить, какое приложение сгенерировало тот или иной поток, следовательно – классифицировать трафик.

Следующие основные показатели характеризуют транспортные потоки сетевого трафика:

- Последовательность размеров сегментов транспортного уровня (TCP или UDP), отправленных со стороны клиента;
- Последовательность размеров сегментов транспортного уровня, отправленных со стороны сервера;
- Последовательность размеров порций данных, отправленных со стороны клиента;
- Последовательность размеров порций данных, отправленных со стороны сервера.

В виду того, что характеристики первых переданных/полученных IP-пакетов при обмене данными могут нести много информации об используемом протоколе, имеет смысл анализировать не весь поток целиком, лишь срез первых N сегментов потока.

Поскольку релевантность статистических характеристик может меняться от различных параметров трафика, алгоритм машинного обучения «Random Forest» идеально подойдет для их классификации, так как он слабо чувствителен к шумам и корреляции признаков [1].

Данный алгоритм основан на принципе обучения с учителем, соответственно для его работы необходима некоторая выборка объектов, для которых уже точно определены метки классов. Соответственно все трассировки, которые будут подаваться на вход классификатора, будем делить на обучающие и проверочные выборки: на обучающей выборке будет проводиться обучение модели, а на проверочной оценка выполнения классификации. Обычно трассировки разделяются на подвыборки пропорционально 1 к 3.

Для корректной работы классификатора, данные, подаваемые на его вход, должны быть соответствующим образом подготовлены. В первую очередь необходимо разобрать «сырой» дамп сетевой трассировки на потоки транспортного уровня, рассчитать статистические характеристики и определить протоколы прикладного уровня для обучающей выборки.

Для тестирования классификатора использованы собственные сетевые трассировки, полученные с помощью экспериментального стенда для пассивного прослушивания трафика, схема подключения которого изображена на рисунке.

Подключение стенда осуществлялось в точках выхода в сеть «Интернет», перед пограничным маршрутизатором, за которым работала сеть из персональных компью-

теров, а также подключались клиенты, использующие мобильные приложения. Длительность записей сетевых трассировок составляет от 1 до 3-х суток.

В качестве параметров алгоритма Random Forest после нескольких экспериментов были выбраны следующие значения:

- Число деревьев: 27;
- Критерий: энтропия;
- Максимальная глубина дерева: 9.

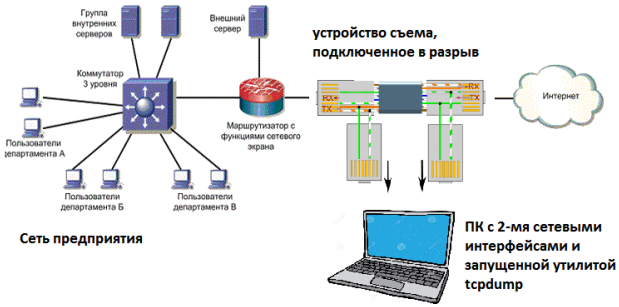

Рис.

Перед непосредственной подачей трассировок на вход классификатора, файлы дампов «.pcap» были обработаны утилитой «editcap».

Для подачи на вход классификатора подготовлен дамп сетевой трассировки, состоящий преимущественно из HTTP-трафика, состоящий по данным анализатора трафика Wireshark из 20007 HTTP-потоков.

Классификация показала следующие результаты. Анализировались только первые 10 пакетов из каждого потока:

	BitTorrent	DNS	HTTP	SSL
BitTorrent	0	1	0	1
DNS	0	32	0	0
HTTP	0	0	19823	21
SSL	0	0	80	3094

Классификатор распознал 19823 HTTP-потоков, что говорит о его высокой точности, при том, что по данным анализатора Wireshark, на его вход подавалось чуть более 20000 потоков HTTP-трафика. Как видно по диагонали, классификатор также распознал 32 потока DNS, и 3094 SSL-рукопожатий, использующихся для передачи сертификата. Все значения вне диагонали – различного рода ошибки.

Ниже представлены данные точности и полноты предсказаний по каждому классу. Точность для класса K – это доля предсказаний вида «объект X принадлежит классу K», которые оказались верными. Полнота для класса K – количество объектов X, которые распознаны классификатором как принадлежащие классу K, делённое на общее количество принадлежащих классу K объектов. Ф-мера – среднее гармоническое полноты и точности:

	precision	recall	f1-score	support
BitTorrent	0.00	0.00	0.00	2
DNS	0.97	1.00	0.98	32
HTTP	1.00	1.00	1.00	19844
SSL	0.99	0.97	0.98	3174
accuracy			1.00	23052
macro avg	0.74	0.74	0.74	23052
weighted avg	1.00	1.00	1.00	23052

Появление ошибок классификации в первую очередь связано с недостаточным обучением алгоритма классификации, что требует более тщательного исследования данного вопроса. Для классификации обучающей выборки предложенного классификатора используется модуль nDPI, который может неверно идентифицировать некоторые протоколы прикладного уровня, соответственно необходимо совершенствовать данный инструмент, дополняя его другими методами анализа, в том числе ручного, что касается новых и нестандартизированных протоколов.

Ошибки классификации также могут быть вызваны недостаточной производительностью аппаратно-программной платформы вычислительного стенда, использующегося в данном эксперименте, что потребует для дальнейших исследований более ресурсоемкого оборудования.

Ниже представлена классификация сетевой трассировки, собранной на живой сети в течении 3-х суток и состоящей из 122852 различных потоков трафика. При записи данных в файл трассировки, на клиентских компьютерах помимо прочего, производилась загрузка и раздача файлов через приложение uTorrent, использующее протокол прикладного уровня BitTorrent:

	BitTorrent	DNS	Google	HTTP	QUIC	RTP	SNMP	SOCKS	SSL	Tor	Viber
BitTorrent	9122	5	0	8	0	0	0	0	1	0	0
DNS	2	248	0	1	0	0	0	0	0	0	0
Google	0	1	12	1	0	0	0	0	0	0	0
HTTP	1	0	0	107446	0	0	0	0	335	0	0
QUIC	0	2	0	0	20	0	0	0	0	0	0
RTP	1	3	0	0	0	0	0	0	0	0	0
RX	1	0	0	0	0	51	0	0	0	0	0
SNMP	0	0	0	0	0	0	52	0	0	0	0
SOCKS	0	0	0	0	0	0	0	1	0	0	1
SSL	2	0	0	125	0	0	0	0	5373	0	0
STUN	1	0	0	0	0	0	0	0	0	0	0
Tor	0	0	0	0	0	0	0	0	2	0	0
Viber	2	7	0	1	0	0	0	2	0	0	8

Как видно классификатор практически без ошибок определил потоки трафика BitTorrent, это также видно из данных точности и полноты предсказаний по этому классу. Вместе с тем, классификатор довольно точно идентифицировал потоки HTTP-трафика, не считая некоторых ошибок вне диагонали, пересекающихся с классом SSL.

Таблица точности предсказаний:

	precision	recall	f1-score	support
BitTorrent	1.00	1.00	1.00	9138
DNS	0.93	0.98	0.96	252
Google	1.00	0.86	0.92	14
HTTP	1.00	1.00	1.00	107782
QUIC	1.00	0.91	0.95	22
RTP	0.67	0.33	0.44	6
RX	1.00	0.98	0.99	52
SNMP	1.00	1.00	1.00	52
SOCKS	0.33	0.50	0.40	2
SSL	0.94	0.98	0.96	5500
STUN	0.75	0.75	0.75	8
Tor	0.00	0.00	0.00	2
Viber	0.89	0.40	0.55	20
WhatsAppVoice	0.00	0.00	0.00	2
accuracy			1.00	122852
macro avg	0.75	0.69	0.71	122852
weighted avg	1.00	1.00	1.00	122852

На основании полученных данных из эксперимента можно сделать вывод о достаточно полезной перспективе применяемого метода, так как полученные результаты оказались довольно точными, за исключением некоторых погрешностей классификации. Основное достоинство данного подхода перед другими методами заключается в использовании для анализа статистических характеристик потока трафика, без исследования полезной нагрузки, что делает возможным его применение в классификации шифрованного трафика.

Вместе с тем, в программной среде Python уже имеются нужные библиотеки и модули для реализации данного алгоритма машинного обучения, что упрощает его построение и дает возможность для совершенствования метода любому исследователю по данному направлению.

Что касается реализованного метода классификации на основе алгоритма Random Forest, то его основное преимущество в том, что он не требует глубокого анализа полезной нагрузки и не нарушает конфиденциальности пользовательских данных, а использует лишь статистические характеристики потока. Данный подход можно применять для идентификации используемых протоколов прикладного уровня в тех условиях, когда полезная нагрузка зашифрована. Также его можно использовать и для превентивного сбора статистики при анализе трафика, так как для определения протокола прикладного уровня достаточно иметь всего 256 байт информации на каждый поток.

[1] Машинное обучение вместо DPI. Строим классификатор трафика. https://habr.com/ru/post/304926/.

[2] Нужный Р.Г., Ротков Л.Ю., Мокляков В.А. Обзор статистических методов классификации сетевого трафика. Труды XXIV научной конференции по радиофизике, посвященной 75-летию радиофизического факультета, 2020 год.

ПРИМЕНЕНИЕ МЕТОДА БАЗОВЫХ ПАРАМЕТРОВ ДЛЯ ПОИСКА АНОМАЛЬНОЙ АКТИВНОСТИ В СЕТЕВОМ ТРАФИКЕ

А.А. Горбунов, Д.В. Смирнов

ННГУ им. Н.И. Лобачевского

Компьютерные сети являются основой современного общества и задачи передачи информации являются одними из первостепенных. Помимо проблемы непосредственно методов передачи информации, критической областью является безопасность канала сообщения.

Существует множество методов для обеспечения безопасности данных, одним из которых является анализ сетевого трафика. Анализ может производиться как на основе статистической информации о получаемых данных, методами кратномасштабного или интеллектуального анализа [1]. В настоящей работе для анализа была использована статистическая информация о получаемых пакетах и их размере. Такой подход позволяет выявлять ряд аномальных активностей, которые могут служить маркерами стороннего вмешательства в сеть.

Для того чтобы произвести анализ сетевого трафика был использован алгоритм на основе определения базовых параметров [2]. В первую очередь изучаемую последовательность сетевых пакетов необходимо разделить на участки стационарности для проведения исследования. Проводилось сравнение типов разбиения ряда данных для трех различных методов.

- Тип 1. Разбиение по предлагаемому алгоритму. Данный алгоритм основан на выделении наименьшей уникальной последовательности и разбиении трафика на элементы различной длины, требующих различной сложности источника данных.
- Тип 2. Разбиение, предложенное в работе [2]. Суть данного разбиения заключается в разбиении трафика на участки стационарности с равной сложностью источника данных. Для этой цели была выбрана сложность источника $n = 3$.
- Тип 3. Разбиение по фиксированному размеру. Данный тип разбиения предполагает разделение последовательности на равные по числу пакетов участки, каждый по 40 пакетов.

При обнаружении аномалий будем основываться на процентном отклонении PD величины x от медианы ее распределения m_x:

$$PD = \frac{(x - m_x)}{m_x} \times 100\%,$$

а также среднепроцентного отклонения PD_{avg} для участка длины n:

$$PD_{avg} = \frac{1}{n} \times \sum_{i=1}^{n} PD_i .$$

Поиск аномалий проводился по двум видам данных: по размерам пакетов и по размеру участка стационарности. Поиск по размерам пакетов позволяет обнаружить участки стационарности с завышенным объемом пакетов, что является аномальной ситуацией, в то время как поиск по размеру участка стационарности позволяет обнаружить попытки флуда или сканирования сети [3]. Критерием для признания участка

стационарности аномальным при этом считалось превышение параметром X^2 граничного значения

$$X^2 = \sum_{b=1}^{B} \frac{(Y - y_b)^2}{y_b} \geq Y^2,$$

где Y – среднепроцентное отклонение величины от медианы распределения по всему ряду данных за исключением проверяемого участка, y_b – среднепроцентное отклонение величины для исследуемого участка, а B – размер участка стационарности.

В ходе настоящей работы было проведено исследование последовательностей пакетов сетевого трафика, полученного из сети радиофизического факультета. Информация о длинах пакетов получалась с помощью программы-анализатора Wireshark, а также с приведением в текстовую форму для удобства дальнейшего анализа. На рис. 1 и 2 представлены результаты, полученные при использовании предлагаемого метода разбиения. По оси абсцисс отложены номера участков стационарности, а по оси ординат – полученные для них величины отношения X^2/Y^2. Аномалии выделены красным цветом, а нормальные элементы – синим.

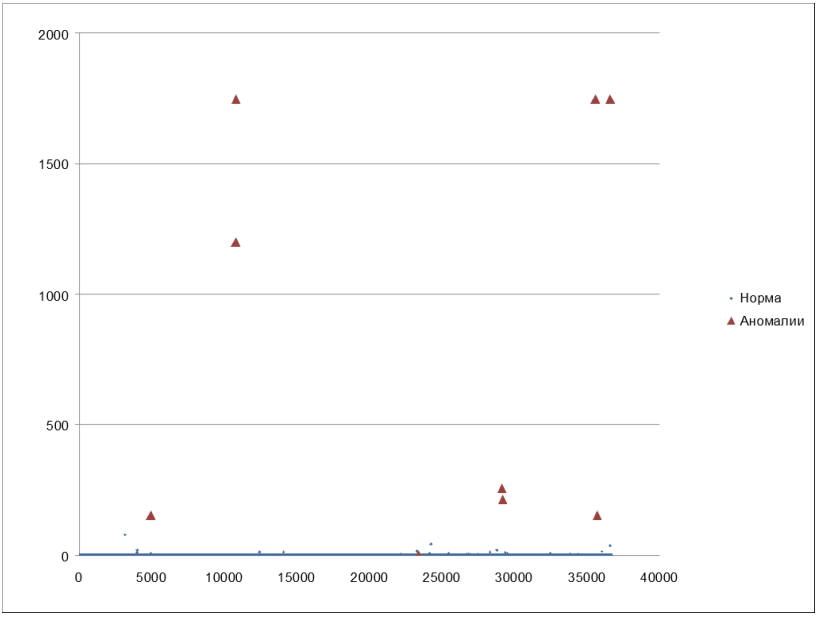

Рис. 1

На рис. 1 представлены результаты для размеров участков стационарности, полученных с помощью предлагаемого разбиения; чем выше значение по оси ординат, тем выше вероятность того что аномалия является признаком атаки.

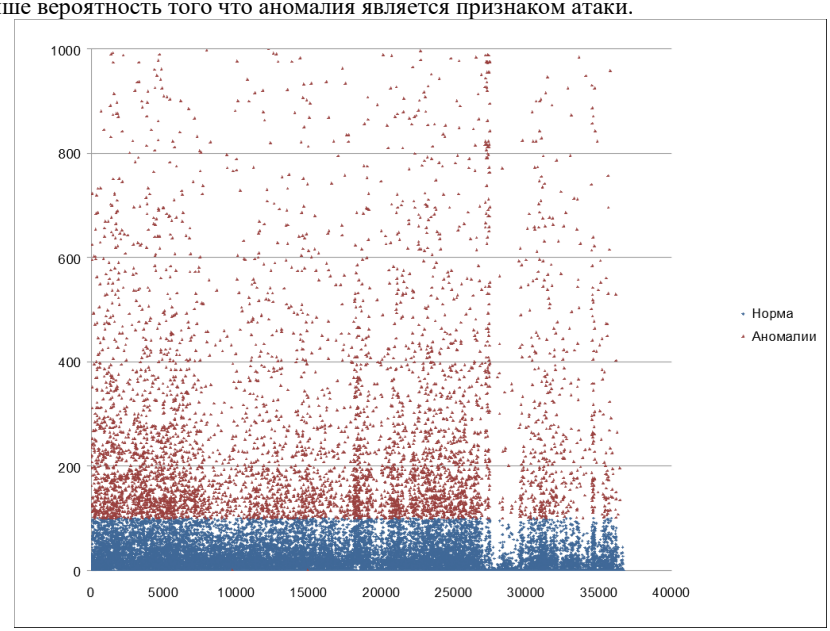

Рис. 2

На рис. 2 представлены результаты для среднепроцентного отклонения размера пакетов, а значение на оси ординат показывает, на сколько размер участка стационарности больше среднестатистического.

В заключение можно отметить, что количественно аномалий, обнаруживаемое при применении предлагаемого метода, уступает по общим данным аналогичному методу по разбиению второго типа, однако превосходит его в процентном соотношении. Соответственно, можно полагать, что поиск аномалий с применением разбиения первого типа является более тонким.

[1] Лукацкий А.В. Обнаружение атак. – СПб: БХВ-Петербург, 2001.
[2] Кирьянов К.Г. Выбор оптимальных базовых параметров источников экспериментальных данных при их идентификации. // Труды III Международной конференции "Идентификация систем и задачи управления SICPRO'04". – М.: ИПУ РАН, 2004, с. 187-208.
[3] Шелухин О.И. Обнаружение вторжений в компьютерные сети (сетевые аномалии). – М.: Горячая линия – Телеком, 2013.

АДАПТАЦИЯ МЕТОДА БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ ПО ГОЛОСУ К ТИХОМУ ПРОИЗНЕСЕНИЮ ПАРОЛЬНЫХ ФРАЗ ДЛЯ ПРОТИВОДЕЙСТВИЯ УТЕЧКИ РЕЧЕВОЙ ИНФОРМАЦИИ ПО АКУСТИЧЕСКИМ КАНАЛАМ

Р.А. Васильев, Л.Ю. Ротков

ННГУ им. Н.И. Лобачевского

Общие положения

В настоящее время многие финансовые операции, идентификация в системах разграничения доступа, запросы конфиденциальной информации по телефону, управление различными устройствами, основывается на применении систем биометрической идентификации по голосу (БИГ) [1].

БИГ существенно отличается от стандартных систем идентификации и систем контроля управления доступом, использующих символьные пароли и ключи. БИГ производится по уникальным и индивидуальным признакам личности и практически исключает вероятность несанкционированных действий, связанных с потерей, кражей или передачей пароля третьим лицам [2].

Широкое применение БИГ систем влечет за собой повышенный интерес со стороны злоумышленников. Наиболее частыми являются атаки с применением ранее применяемых биометрических признаков, например, аудио - запись парольной фразы.

Системы БИГ необходимо проектировать так, чтобы свести к минимуму указанные выше атаки. В данной статье описан адаптированный к тихому произнесению парольных фраз (ТПФ) метод БИГ, основанный на применении метода обеляющего фильтра (МОФ), реализованный в «Информационной системе идентификации дикторов по голосу» (ИС ИДГ) [3], доработанной для решения задачи защиты речевой информации (парольные фразы) от утечки по акустическим каналам [4].

Теоретический анализ

Общая формулировка задачи БИГ сводится к тому, что требуется отнести выборку из речи пользователя X к одному из R > 1 неопределенных пользователей или отдельных слов (классов) [5]. Каждому классу соответствуют образцы речи конкретного пользователя, обладающие общими признаками, образующими образ голоса пользователя (ОГП). Каждый ОГП обладает определенным набором устойчивых признаков P_r, $r = \overline{1,R}$. В данном случае решение задачи БИГ сводится к установлению соотношения $P_x = P_v$, $v \le R$ между набором признаком пользователя X и одним из ОГП в базе голосовых данных.

Это стандартная задача статистической классификации. Ее решение обычно основывается на критерии максимального правдоподобия. Применительно к нашей модели речевых сигналов в виде L независимых отрезков (массивов), длиной n отсчетов каждый, такая задача подробно рассматривалась в работе [6-8].

Главная идея метода состоит в существенной (в десятки раз) редукции или сжатии данных за счет того, что в базе априорных данных хранятся не сами отрезки речи

длиной $nL = 10^2 ... 10^3$ отсчетов каждый, а их образы в виде набора из R векторов АР-коэффициентов (их также часто называют коэффициентами линейного предсказания - КЛП), размерность которых $M = 10...30 < nL$ в реальных условиях ограничивается конечной степенью сложности спектрального состава человеческого голоса. Причем в отличие от известных алгоритмов автоматического распознавания речи на основе КЛП [9] в рассматриваемом МОФ применяется принципиально иной критерий для оценивания рассогласования между различными речевыми образами, уходящий своими корнями в теоретико-информационный подход и информационную метрику Кульбака-Лейблера [10].

Экспериментальные исследования

Для экспериментальных исследований была использована ранее разработанная и проверенная ИС ИДГ [11-13], адаптированная к ТПФ посредством доработки модуля регулировки чувствительности к уровню речевого сигнала пользователя (диктора) *X*. Главное окно ИС ИДГ представлено на рис. 1.

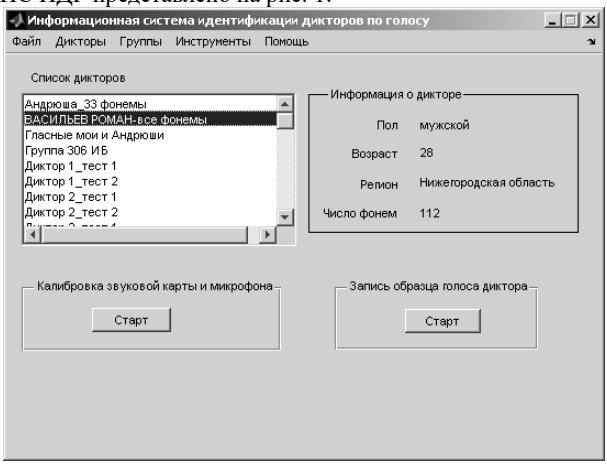

Рис. 1

Экспериментальные исследования, состоящие из трех этапов, представлены в табл. 1.

Табл. 1

Номер эксперимента	Цель эксперимента	Задачи эксперимента
1	Выявление эталонного уровня сигнала речевой идентификации пользователя по голосу, анализ недостатков и преимуществ сигналов разного уровня в системе голосовой идентификации.	Запись в БД эталонного уровня сигнала для дальнейших экспериментов, выявление среднего, низкого и высокого уровня сигнала для голосовой идентификации, определение оптимального уровня сигнала для голосовой идентификации, основываясь на преимуществах и недостатках различного уровня сигнала.
2	Определить вероятность правильной идентификации пользователей по голосу.	Определение вероятности правильной идентификации трех пользователей по голосу при условии разделения уровня сигнала на 3 интервала (до 45дБ, от 45-70дБ, более 70дБ)
3	Определение октавных коэффициентов звукоизоляции в типовом помещении, определить выполняются ли нормы защищенности в данном помещении для работы с конфиденциальной информацией.	Определить назначение объекта контроля, выбрать требуемый уровень защиты от утечки речевой информации по акустическим каналам. Определение контролируемой зоны, определение смежных помещений и ограждающих конструкций, определение контрольных точек для замера уровня акустического сигнала (шума). Проверить – выполняются ли нормы защищенности

В **первом эксперименте**, для выявления эталонного уровня сигнала *Uc* для речевой идентификации пользователя, проведено 9 опытов с различным уровнем сигнала (от 35 дБ до 84 дБ). В процессе опытов уровни сигналов разбиты на 3 группы: высокие (*Uc*~78-88дБ), средние (*Uc*~58-65дБ) и низкие (*Uc*~35-40дБ).

Проведен анализ каждой из групп сигналов и определена вероятность идентификации пользователя по голосу (табл. 2).

Табл. 2

Уровень сигнала	Вероятность идентификации	Вывод
Uc~78-88дБ	Более 90%	Идентификация в 90% проходит успешно, но вероятность утечки информации по акустическому каналу высокая.
Uc~58-65дБ	Более 70%	Идентификация в более70% проходит успешно, вероятность утечки информации по акустическому каналу в пределах нормы
Uc~35-40дБ	Менее 50%	Идентификация проходит не успешно, вероятность утечки информации по акустическому каналу ничтожно мала.

На рис. 2 показана успешная идентификация по голосу в ИС ИДГ при *Uc*=65дБ. В модуле ИС ИДГ «подсчет фонем» выделено и опознано 789 фонем, из которых 467 принадлежат конкретному диктору (пользователю). В модуле «Идентификация» показано, что диктор успешно определен.

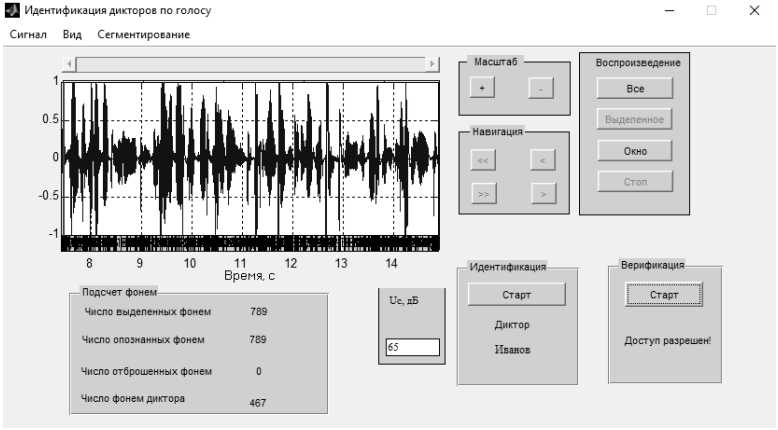

Рис. 2

Далее, на рис. 3 представлен случай с неуспешной идентификацией по голосу в ИС ИДГ при *Uc*=40дБ. В модуле ИС ИДГ «подсчет фонем» выделено 759 фонем, из которых опознано только 325 фонем и не определено, к какому диктору (пользователю) они относятся. В модуле «Идентификация» показано, что диктор не определен.

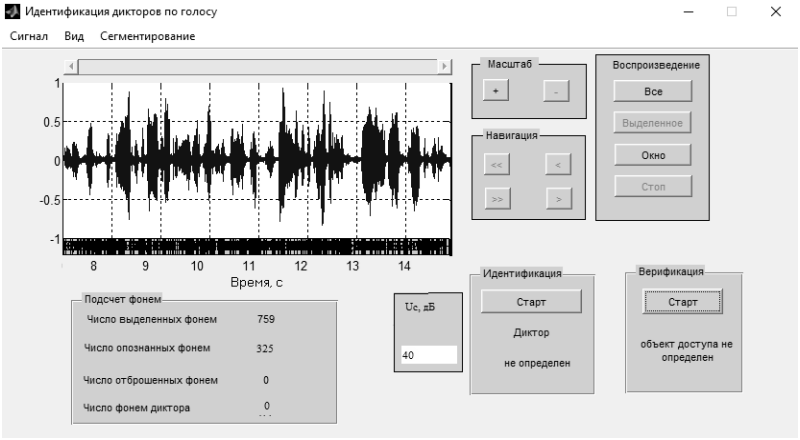

Рис. 3

Выводы по первому эксперименту:

- Определен уровень эталонного сигнала речевой идентификации *Uc*=60дБ;
- Сигналы от 35-40дБ воспринимаются ИС ИДГ гораздо хуже, в связи с чем процент успешной голосовой идентификации очень маленький, но преимуществом тихого воспроизведения парольных фраз является отсутствие акустического канала утечки информации;
- Сигналы от 58-65дБ ИС ИДГ воспринимает гораздо лучше, нежели сигналы от 35-40дБ, и как следствие, процент успешной идентификации выше и достигает более 70%, однако вероятность утечки речевой информации возрастает и необходимо проводить дополнительную оценку защищенности по акустическим каналам.
- Сигналы от 78-88 дБ ИС ИДГ воспринимает лучше, чем сигналы двух других групп, и процент успешной идентификации пользователя по голосу достигает более 90%. Несмотря на высокий процент успешной идентификации, этой группе сигналов присуща высокая вероятность утечки речевой информации по акустическим каналам.

Во **втором эксперименте** определена вероятность правильной идентификации трёх пользователей (Иванов, Петров, Сидоров) по голосу при условии разделения уровня сигнала для каждого из пользователей на 3 интервала: до 45дБ, от 45-70дб, более 70 дБ (табл. 3).

Табл. 3

Uc~более 70дБ									
ФИО	Uc1,дБ	Uc2,дБ	Uc3,дБ	Uc4,дБ	Uср,дБ	Идентификация прошла успешно			
						Р1	Р2	Р3	Р4
Иванов	80	83	90	86	84,75	Да	Да	Да	Да
Петров	90	85	93	98	91,5	Да	Да	Да	Да
Сидоров	88	82	84	79	83,25	Да	Да	Да	Да
Uc~45-70дБ									
ФИО	Uc1,дБ	Uc2,дБ	Uc3,дБ	Uc4,дБ	Uср,дБ	Идентификация прошла успешно			
Иванов	61	65	63	58	61,75	Да	Да	Да	Да
Петров	68	70	69	65	68	Да	Да	Да	Да
Сидоров	48	52	46	49	48,75	Нет	Да	Нет	Да
Uc~до 45Дб									
ФИО	Uc1,дБ	Uc2,дБ	Uc3,дБ	Uc4,дБ	Uср,дБ	Идентификация прошла успешно			
Иванов	43	45	40	42	42,5	Нет	Нет	Нет	Нет
Петров	44	38	42	41	41,25	Нет	Нет	Нет	Нет
Сидоров	40	38	35	37	37,5	Нет	Нет	Нет	Нет

Рассмотрим каждый из трех диапазонов отдельно. Анализируя 1-й диапазон (*Uc*~более 70дБ), можно сказать, что идентификация пользователей по голосу проходит во всех случаях успешно. Средний уровень сигнала для пользователя Иванов равен 84,75 дБ, идентификация проходит успешно во всех опытах. Средний уровень сигнала для пользователя Петров 91,5 дБ, идентификация проходит успешно во всех опытах. И наконец, пользователь Сидоров имеет средний уровень голосового сигнала 83,25 дБ, для которого идентификация по всех опытах также проходит успешно.

Проводя анализ 2-го диапазона (*Uc*~ 45-70 дБ) средний уровень сигнала для пользователя Иванов равен 61,75 дБ, идентификация в 4-х опытах успешно завершена. Пользователь Петров имеет уровень сигнала 68 дБ, соответственно идентификация во всех опытах прошла успешно. Что касается пользователя Сидорова, то его средний уровень сигнала составляет 48,75 дБ, и как следствие, из четырех экспериментов – в двух идентификация проходит успешно, в двух программа не может распознать голос.

Рассматривая 3-й диапазон (*Uc*~до 45 дБ) для трех пользователей системы, уровень голосового сигнала каждого из них приблизительно равен 40 дБ. Вероятность

идентификации для пользователей с таким низким уровнем сигнала стремиться к нулю.

Выводы ко второму эксперименту:

- Выбрав в 1-м опыте эталонный уровень сигнала Uэс=60 дБ, определно, что идентификация пользователя по голосу проходит только тогда, когда уровень голосового сигнала пользователя Uс>49 дБ;
- Уровень сигнала более 70 дБ во всех опытах прошел идентификацию, но вероятность утечки информации по акустическим каналам возрастает с увеличением Uс;
- Оптимальным уровнем сигнала для успешной идентификации пользователя по голосу в ИС ИДГ является Uс=50дБ-60дБ;

В соответствии с методикой оценки [14], в **третьем эксперименте** проведена оценка защищенности речевой информации (парольные фразы) от утечки по акустическим каналам в типовом офисном помещении (ТОП) с помощью системы оценки защищенности выделенных помещений по виброакустическому каналу «ШЕПОТ» (Сертификат ФСТЭК №643 от 05.07.2002г.).

Границей контролируемой зоны ТОП являются наружные и внутренние стены ТОП, а также перекрытие пола и потолка.

В результате анализа возможных каналов утечки речевой информации, инженерно-строительных и организационно-режимных мер, применяемых в ТОП, произведен расчет значений октавных коэффициентов звукоизоляции ограждающих конструкций *Qi* в выбранных контрольных точках (КТ).

Вывод по третьему эксперименту:

В результате проведенных измерений и расчетов установлено, что при уровне сигнала в 60дБ значения октавных коэффициентов звукоизоляции ограждающих конструкций и инженерно-технических систем во всех контрольных точках соответствуют значениям, приведенным в [14], что обеспечивает защищённость данного ТОП от утечки речевой конфиденциальной информации по акустическому каналу.

Выводы

В настоящей статье рассмотрены особенности БИГ при условии ТПФ, выявлен оптимальный уровень сигнала *Uс*~60дБ для адаптации метода идентификации в ИС ИДГ, что позволило получить высокую вероятность правильной БИГ, по сравнению с другим методом БИГ [15]. С учетом эталонного уровня сигнала были сделаны замеры октавных коэффициентах звукоизоляции для ТОП, и установлено, что при уровне сигнала *Uс*~60дБ утечки речевой информации по акустическим каналам не происходит.

[1] Николаев Д. Б., Васильев Р. А. Анализ возможности применения голосовой идентификации в системах разграничения доступа к информации // Научный результат // Серия: Информационные технологии. Белгородский государственный университет. 2016. Вып. 1. С 48-57.

[2] Савченко В. В., Васильев Р. А. Анализ эмоционального состояния дикторов по голосу на основе фонетического детектора лжи // Научные ведомости Белгородского государственного университета. 2014. Вып. № 21(192)32\1. С. 186-195.
[3] Васильев Р. А. Свид. о гос. регистрации программы для ЭВМ №2015663306 Программа идентификации дикторов по голосу // Васильев Р.А. Зарег. 15.12.2015г. – М.: Роспатент, 2015.
[4] Бузов Г. А. Защита от утечки информации по техническим каналам. – М.: 2005г.
[5] Цыпкин Я. 3. Адаптация и обучение в автоматических системах. – М.: Наука, 1968.
[6] Савченко В. В. Информационная теория восприятия речи. // Изв. вузов России. Радиоэлектроника. 2007. Вып. 6. С. 3–9.
[7] Савченко В. В. Теоретико-информационное обоснование гауссовой модели сигналов в задачах автоматического распознавания речи. // Изв. вузов России. Радиоэлектроника. 2008. Вып. 1. С. 24–33.
[8] Савченко В. В., Акатьев Д. Ю., Карпов Н. В. Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра // Изв. вузов России. Радиоэлектроника. 2007. Вып.4. С. 11-19.
[9] Потапова Р. К. Речь: коммуникация, информация, кибернетика: Учеб, пособие. 2-е изд. – М.: Эдитори- ал УРСС, 2001.
[10] Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
[11] Васильев Р.А. Биометрическая идентификация пользователей информационных систем на основе кластерной модели элементарных речевых единиц: Дис. … к-та тех. наук. – М., 2017. 153 с.
[12] Васильев Р.А. Исследование фонетического строя речи и идентификация дикторов по голосу // Безопасность информационных технологий. 2013. Т. 20, № 1. С. 85-86.
[13] Васильев Р.А. Исследование особенностей фонетического строя речи и текстонезависимая идентификация дикторов по непрерывной речи // Информационная безопасность регионов. 2012. № 2 (11). С. 57-63.
[14] Волобуев С. В. Оценка акустической защищенности с применением инструментальных средств. // Системы безопасности связи и телекоммуникаций. 1999. № 25. С. 38 -45.
[15] Аграновский А. В., Леднов Д. А. Метод текстонезависимой идентификации пользователя на основе индивидуальности произношения гласных звуков // Акустика и прикладная лингвистика: Ежегодник РАО. Вып. 3. – М.: 2002. С. 103-115.

Секция «Информационные системы.
Средства, технологии, безопасность»

Заседание секции проводилось 18 мая 2021 г.
Председатель – Л.Ю. Ротков, секретарь – А.А. Рябов.
Нижегородский государственный университет им. Н.И. Лобачевского.