А. В. Еременко, *канд. техн. наук, Омский государственный университет путей сообщения, 4eremenko@gmail.com*

А. Е. Сулавко, *канд. техн. наук, Омский государственный технический университет, sulavich@mail.ru*

Д. В. Мишин, *канд. техн. наук, Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых, mishin.izi@gmail.com*

А. А. Федотов, *Омский государственный университет путей сообщения, fedotov1609@gmail.com*

Идентификационный потенциал клавиатурного почерка с учетом параметров вибрации и силы нажатия на клавиши[1]

Рассматривается проблема защиты данных от неавторизованного доступа посредством идентификации пользователей компьютерных систем по клавиатурному почерку. Произведена оценка информативности различных признаков, характеризующих клавиатурный почерк субъектов, в том числе динамики изменения давления при нажатии на клавиши и параметров вибрации клавиатуры. Для формирования базы биометрических образцов разработана клавиатура с использованием специальных датчиков. Произведена оценка вероятностей ошибок идентификации на основе стратегии Байеса при использовании различных пространств признаков.

Ключевые слова: клавиатурный почерк, сила нажатия на клавиши, датчики, идентификация оператора, биометрический признак.

Введение

Финансовые потери от утечек конфиденциальной информации разного сорта (персональных данных, платежных документов, коммерческой тайны и др.) с каждым годом возрастают. По данным независимого аналитического центра компании Zecurion (одна из крупнейших фирм — производителей систем защиты от утечек), в 2015 г. зарегистрирован рекордный ущерб в мире от утечек информации — более 29 млрд долларов (ранее эта цифра не поднималась выше 25,11 млрд долларов) [1]. Россия оказалась на 4-м месте (после США, Великобритании и Канады) в мире по числу утечек. Финансовые данные физлиц — один из самых востребованных киберпреступниками типов информации (19,1% инцидентов). Чаще всего в 2015 г. информация утекала из госучреждений, предприятий розничной торговли и банков [1]. Представленная информация коррелирует с данными InfoWatch (Россия на 2-м месте по количеству утечек после США, далее следуют Великобритания и Канада) [2] и ряда зарубежных источников (в частности, с результатами исследований PricewaterhouseCoopers [3]).

В настоящее время активно идут процессы информатизации общества. Появляется все

[1] Работа выполнена при финансовой поддержке РФФИ (грант № 16-37-50007).

больше веб-сервисов. Многие государства стремятся создать электронное правительство для оказания услуг гражданам. Доверие к таким веб-сервисам со стороны пользователей должно быть наивысшим [4]. Сформировалась следующая точка зрения: «…роль паспортов и удостоверений личности будут играть личные электронные кабинеты, созданные взаимными усилиями гражданина и организациями, предоставляющими ему значимые интернет-услуги… Проблема безопасного хранения этой личной информации будет усиливаться…» [4]. Актуальность проблемы прослеживается в результатах аналитических исследований, в соответствии с которыми большинство утечек данных приходится именно на веб-сервисы (браузер, облако) [2].

Традиционные средства аутентификации обычно основаны на проверке знания секрета (пароля) или наличия физиологических особенностей субъекта (биометрических признаков). Средства парольной защиты в наибольшей степени подвержены «человеческому фактору», так как пароли являются отчуждаемыми от владельца: их забывают, крадут, теряют, передают третьим лицам. Биометрические системы защиты также не лишены недостатков: статические образы (отпечатки пальца, сетчатка или радужка глаза, лицо) не являются секретными, поэтому их можно скопировать, изготовив физический или цифровой муляж (второй вариант нужен для удаленной аутентификации). Другим вариантом, сочетающим секрет и индивидуальные характеристики субъекта при формировании аутентификатора, является использование тайных биометрических образов. К их числу относится индивидуальный клавиатурный почерк субъекта, проявляющийся при наборе парольной фразы. Однако метод распознавания по клавиатурному почерку пока дает высокие вероятности ошибок (не менее 1% в лучшем случае [5]), что затрудняет его использование в важных информационных системах.

Цель настоящей работы — оценить идентификационный потенциал клавиатурного почерка субъектов с учетом дополнительных данных о давлении (силе нажатия) на клавиши и вибрации клавиатуры при наборе парольных фраз. Для достижения поставленной цели необходимо выполнить следующие задачи:

1) разработать прототип клавиатуры, регистрирующей дополнительные данные о клавиатурном почерке (динамику изменения давления на клавиши и вибрации клавиатуры при вводе текста);

2) сформировать базу образцов клавиатурного почерка для последующего анализа достаточного объема, чтобы делать обоснованные выводы и заключения;

3) произвести анализ образцов и выявить идентификационные признаки — величины, характеризующие пользователей клавиатуры;

4) произвести оценку информативности выявленных признаков;

5) предложить способ идентификации печатающих субъектов с учетом дополнительных признаков клавиатурного почерка;

6) провести оценку эффективности предложенного способа идентификации.

Разработка модифицированной клавиатуры для сбора биометрических данных

Для проведения запланированных исследований необходима клавиатура с возможностью регистрации дополнительных признаков клавиатурного почерка (давления и вибрации). Модели клавиатур с требуемыми функциями существуют либо в виде единичных экспериментальных образцов и не доступны для заказа, либо имеют неподходящий форм-фактор (клавиатуры мобильных устройств). В то же время современные средства разработки программируемых электронных устройств позволяют решить данную задачу самостоятельно с минимальными затратами и за приемлемые сроки.

В качестве платформы разработки программно-аппаратного комплекса для регистрации дополнительных признаков клавиатурного почерка был выбран контроллер Arduino Uno R3, который построен на чипе ATmega328, обеспечивающем преобразование аналогового сигнала в цифровую форму с помощью встроенного АЦП, и может использоваться для разработки интерактивных систем, управляемых различными датчиками и переключателями. Сводные характеристики программируемого контроллера Arduino Uno R3 представлены в табл. 1.

Таблица 1. Характеристики программируемого контроллера Arduino Uno R3

Table 1. Specifications of programmable controller Arduino Uno R3

Технический параметр	Значение
Микроконтроллер	ATmega328
Рабочее напряжение, В	5
Входное напряжение (рекомендуемое), В	7–12
Входное напряжение (предельное), В	6–20
Цифровые Входы/Выходы	14
Аналоговые входы	6
Постоянный ток через вход/выход, мА	40
Постоянный ток для вывода 3.3 В, мА	50
Flash-память, Кб	32 (ATmega328)
ОЗУ, Кб	2 (ATmega328)
EEPROM, Кб	1 (ATmega328)
Тактовая частота, МГц	16

Процесс создания программно-аппаратного комплекса состоял из следующих этапов:

1) проектирование структурной схемы;
2) подбор и приобретение компонентов;
3) инженерные работы.

Структурная схема программно-аппаратного комплекса для регистрации дополнительных признаков клавиатурного почерка от датчиков давления и вибрации представлена на рис. 1.

К Arduino Uno R3 были подключены сенсор вибрации и 5 сенсоров давления. Таким образом, все 6 аналоговых входов были задействованы (см. табл. 1). Для определения силы нажатия на клавиши использован датчик давления Interlink 408 FSR, который представляет собой силоизмерительный резистор, исполненный в виде плоского тонкого пассивного компонента, сопротивление которого пропорционально усилию, действующему на его поверхность. Без нагрузки сопротивление превышает величину 1 МОм и варьируется от 100 кОм до нескольких сотен Ом в зависимости от силы нажатия на поверхность датчика. Для получения данных о вибрации клавиатуры при вводе текста использован пьезоэлектрический датчик вибрации Analog Piezo Disk Vibration Sensor компании DFRobot, способный улавливать даже незначительные колебания, будучи установленным внутри клавиатуры. Пьезоэлемент при подключении к микроконтроллеру выдает сигнал, пропорциональный амплитуде вибрации. Для определения кодов клавиш и моментов их нажатия использован модуль USB Host Shield, предназначенный для подключения HID-устройств и эмуляции их работы в операционной системе. К Arduino Uno R3 через USB Host Shield подключена клавиатура Logitech K120. Корпус клавиатуры был вскрыт, и под ряды ее клавиш установлены датчики давления. Датчики давления также подключаются к Arduino Uno R3.

Частота опроса датчиков микроконтроллером Arduino Uno R3 составляет 3000 Гц, но так как последовательно опрашиваются 6 каналов (датчиков), то реальная частота дискретизации каждого из регистрируемых сигналов составляет 500 Гц. Чтобы оценить максимальную информативную частоту сигналов, формируемых при наборе текста на клавиатуре, нужно ориентироваться на наивысшую возможную скорость печати пользователя на клавиатуре. Используя кла-

Рис. 1. Структурная схема программно-аппаратного комплекса для регистрации дополнительных признаков клавиатурного почерка от датчиков давления и вибрации

Fig. 1. Block diagram of hardware and software for the registration of additional features keyboard handwriting by the pressure and vibration sensors

виатуру Дворака (вариант раскладки клавиатуры, предполагающий более высокую скорость набора текста по сравнению с традиционной раскладкой QWERTY), в 2005 г. Барбара Блэкберн (*Barbara Blackburn*) установила мировой рекорд по скорости набора текста на английском языке, отмеченный в Книге рекордов Гиннесса. Она печатала со средней скоростью 150 слов в минуту на протяжении 50 минут, временами ее скорость поднималась до 170 слов в минуту, а на короткий промежуток времени она достигла скорости 212 слов в минуту. В английском языке средняя длина слова равна 5,2 буквы, однако WPM — количество слов в минуту нередко приравнивается к 5 символам (в других странах скорость набора измеряется также в CPM — символах в минуту или в SPM — ударах в минуту). Таким образом, рекорд Барбары Блэкберн составил около 750 символов в минуту. По другим оценкам, нормальной скоростью набора для клавиатуры с раскладкой QWERTY считается 150–200 символов в минуту, хорошей — 250–300 символов. Максимальной скорости набора на клавиатуре, отмеченной в открытых источниках, соответствует частота 12,5 Гц, норме — 2,5 Гц. Согласно теореме Котельникова частота дискретизации должна быть в два раза выше частоты сигнала. Из этого следует вывод, что частоты дискретизации сигналов 25 Гц вполне достаточно для фиксации всех частотных изменений, происходящих в клавиатурном почерке человека.

Анализ образцов клавиатурного почерка

Для сбора биометрических данных были привлечены 100 испытуемых. Испытуемые подбирались таким образом, чтобы среди них было равное количество представителей всех типов темперамента (холерик, сангвиник, ме-

ланхолик, флегматик), что проверялось тестами Айзенка. Это требовалось для обеспечения чистоты эксперимента (известно, что тип темперамента влияет на скорость реакции и параметры воспроизведения подсознательных движений, к таким параметрам относятся характеристики клавиатурного почерка [6]). Каждый испытуемый осуществлял набор произвольного текста (выдержки из художественных произведений) и производил ввод парольной фразы («прошу разрешить доступ к информации») не менее 120 раз на разработанной клавиатуре. Информативность парольной фразы определяется ее длиной. Парольная фраза должна быть легко запоминаемой и предпочтительно содержать от 21 до 42 нажатий на клавиши [6] (слишком длинные парольные фразы сложно запоминаемы и сложно воспроизводимы, велика вероятность ошибки при наборе фразы на клавиатуре). Каждый фрагмент данных, формируемый при однократном вводе текста или парольной фразы, назовем образцом клавиатурного почерка.

Образцы клавиатурного почерка обрабатывались разработанным программным модулем, в результате каждый образец был преобразован в вектор значений признаков. Вектор значений каких-либо признаков, вычисляемых при обработке одного образца клавиатурного почерка, назовем реализацией клавиатурного почерка (или просто реализацией). Анализируемые в настоящей работе признаки по физическому смыслу можно разделить на несколько условных категорий, которые представлены в табл. 2. Далее приведено более подробное описание признаков.

Базовыми признаками клавиатурного почерка являются времена удержания клавиш и паузы между нажатиями клавиш [5; 7] (категории 1.1–1.2, см. табл. 2). Другим признаком, который встречается в научных работах, является частота или время одновременного нажатия пары клавиш (время перекрытия) в процессе ввода текста или контрольной фразы [8; 9]. Иногда реализуется объединение группы признаков в кластер, описывающий n-грамму — n символов, набираемых на клавиатуре последовательно. В такой кластер могут входить времена удержаний n клавиш, n — 1 пауз между нажатиями клавиш, времена одновременного нажатия 2, 3, n клавиш. В настоящем исследовании регистрировались времена и количество перекрытий клавиш (категория 1.3, см. табл. 2).

Перейдем к описанию дополнительных признаков.

За время нажатия клавиши регистрируется множество показателей моментального давления на клавишу и моментальной вибрации клавиатуры. В качестве признаков могут быть использованы как средние, так и максимальные показатели указанных величин. В настоящей работе решено апробировать максимальный регистрируемый уровень давления и вибрации на клавишу (категория признаков 2.1. и 2.2, см. табл. 2). Корреляция между средним и максимальным значениями давления при нажатии на клавиши очень существенная (более 0,9), при этом различие этих величин для разных испытуемых было менее заметным.

Известно, что анализ в частотной области дает преимущества в оценке зашумленных сигналов [10]. Одним из современных математических аппаратов для анализа спектральных характеристик нестационарных сигналов является вейвлет-анализ. В настоящей работе предложен переход от временного представления функции моментального давления на клавиши $p(t)$ и функции показаний вибрации клавиатуры $vibro(t)$ к частотному, их исследование и поиск динамических характеристик на основе метода многомасштабного анализа (признаки категорий 3.1–3.2, см. табл. 2). Функция $p(t)$ характеризует уровень давления на клавиатуру (при нажатии одной или группы клавиш) в момент времени t, функция $vibro(t)$ ставит в соответствие показатель вибрации моменту времени t.

Указанные функции отличаются по длительности, поэтому предварительно они при-

Таблица 2. Краткое описание потенциальных идентификационных характеристик клавиатурного почерка

Table 2. Brief description of potential identification characteristics of keyboard handwriting

№ категории	Категория признака	Описание	Ближайшее распределение
1.1	Времена удержания клавиш	Временной промежуток между событиями, при которых определенная клавиша нажата и отпущена (в миллисекундах). С каждой клавишей ассоциирован признак	Нормальный
1.2	Паузы между нажатиями клавиш	Временной промежуток между событиями, при которых одна клавиша нажата и другая клавиша нажата (в миллисекундах). С каждой парой клавиш ассоциирован признак	Логнормальный
1.3	Времена перекрытия клавиш	Временной промежуток, равный длительности одновременного нажатия (удержания) определенной пары клавиш (в миллисекундах). С каждой парой клавиш ассоциирован признак	Логнормальный
2.1	Давление на клавиши	Показатель давления, измеряемый в процессе нажатия на определенную клавишу. С каждой клавишей ассоциирован признак	Нормальный
2.2	Вибрация при нажатии клавиш	Показатель вибрации клавиатуры, измеряемый в процессе нажатия на определенную клавишу. С каждой клавишей ассоциирован признак	Нормальный
3.1	Вейвлет-преобразование функции $p(t)$	Коэффициенты вейвлет-преобразования Добеши D6, вычисляемые из функции давления на клавиши, формируемой в процессе ввода образца клавиатурного почерка	Лапласа (двойное экспоненциальное) / Нормальный
3.2	Вейвлет-преобразование функции $vibro(t)$	Коэффициенты вейвлет-преобразования Добеши D6, вычисляемые из функции вибрации клавиатуры, формируемой в процессе ввода образца клавиатурного почерка	Лапласа (двойное экспоненциальное) / Нормальный

водились к единому временному масштабу. Для этого выполнялось прямое разложение функций $p(t)$ и $vibro(t)$ в ряд Фурье, одновременно вычислялись амплитуда и частота для первых k гармоник, на следующем шаге частоты гармоник масштабируемой функции заменялись частотами соответствующих гармоник, полученных для функции, к которой производится масштабирование. Далее выполнялось обратное преобразование Фурье для k гармоник с измененными характеристиками. Нормированные функции подвергались амплитудно-частотному анализу методом, описанным далее.

Применяемый метод разложения функций $p(t)$ и $vibro(t)$ основан на дискретном вейвлет-преобразовании и использует пирамидальный алгоритм Малла для разложения исходных сигналов на последовательности вейвлет-коэффициентов d_{jk}, характеризующих структуру анализируемого процесса на разных масштабах j. В проводимых исследованиях рассматривались различные базисы вейвлетов Добеши (от D4 до D10). Данные вейвлеты были получены в результате математической процедуры поиска ортонормированных базисов, обладающих конечным носителем, что обеспечивает минимизацию вычислительных затрат при проведении численного анализа экспериментальных данных. Масштабно-временное представление сигнала x получается с использованием методов цифровой фильтрации. Сначала сигнал пропускается через низкочастотный (*low-pass*) фильтр с импульсным откликом g и получается свертка

$$y[n] = (x \cdot g)[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[n-k]. \quad (1)$$

Одновременно сигнал раскладывается с помощью высокочастотного (*high-pass*) фильтра h. В результате получаются детализирующие коэффициенты (на выходе высокочастотного фильтра) и коэффициенты аппроксимации (на выходе низкочастотного фильтра). Так как половина частотного диапазона сигнала была отфильтрована, то согласно теореме Котельникова отсчеты сигналов можно проредить в 2 раза:

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} x[k]g[2n-k];$$

$$y_{high}[n] = \sum_{k=-\infty}^{\infty} x[k]h[2n-k], \qquad (2)$$

где $y_{high}[n]$ и $y_{low}[n]$ — прореженные в два раза выходы высокочастотного и низкочастотного фильтров соответственно; $h[n]$ и $g[n]$ — высокочастотный и низкочастотный фильтры соответственно; $x[k]$ — исходный сигнал; n — номер уровня разложения; k — количество отсчетов в сигнале.

Такое разложение вдвое уменьшило разрешение по времени в силу прореживания сигнала. Однако каждый из получившихся сигналов представляет половину частотной полосы исходного сигнала — частотное разрешение удвоилось.

Так как анализируемые сигналы были дискретизированы на частоте 500 Гц (частота дискретизации процессора Arduino, использовавшегося для сборки экспериментального образца клавиатуры), то согласно теореме Котельникова верхняя частота сигнала, которая может быть найдена в результате частотного анализа, составляет 250 Гц. Таким образом, для сигнала, состоящего, например, из 256 отсчетов, вейвлет-коэффициенты первого уровня разложения (для данного примера их количество составит 128) занимают полосу частот 125–250 Гц. Вейвлет-коэффициенты второго уровня (64 коэффициента в силу двукратного прореживания ряда) описывают гармоники спектра, приходящиеся на полосу частот 62,5–125 Гц. Процедура повторяется до тех пор, пока не останется 1 вейвлет-коэффициент и 1 отсчет аппроксимации на девятом уровне. Всего получается (1+1+2+4+8+16+32+64+128) = 256 коэффициентов. То есть число коэффициентов будет равно числу отсчетов в исходном сигнале. Если основная энергия сигнала была сосредоточена, например, возле частоты 2 Гц, то вейвлет-коэффициенты седьмого уровня будут большими, а вейвлет-коэффициентами младших уровней можно пренебречь.

В табл. 3 приведены расчеты для девяти уровней разложения исследуемых сигналов, применявшихся в экспериментах. Было установлено, что спектр частот для управляющих сигналов подписи находится в пределах 2…12,5 Гц. Таким образом, основная доля мощности сигнала должна быть сосредоточена на уровнях разложения с 5-го по 7-й.

Качество работы алгоритма описанного преобразования и количество получаемых признаков во многом зависят от выбранного вейвлета. D6 показал качественный результат и оптимальное время обработки, поэтому был выбран для вычисления значений признаков. D2 показал самое малое время анализа и ме-

Таблица 3. Масштабные и частотные характеристики вейвлет-коэффициентов на соответствующих уровнях разложения

Table 3. The scale and frequency characteristics of wavelet coefficients at the appropriate levels of decomposition

Уровень разложения	Полоса частот, Гц	Временное разрешение (масштаб), мс
1	125–250	2
2	62,5–125	4
3	31,25–62,5	8
4	15,625–31,25	16
5	7,8125–15,625	32
6	3,90625–7,8125	64
7	1,953125–3,90625	128
8	0,9765625–1,953125	256

нее качественный результат. Также количество признаков данных категорий (3.1–3.2) зависит от уровня дискретизации исходных функций *p* (*t*) и *vibro* (*t*), принимаемого при масштабировании по времени. Эмпирически было определено оптимальное значение получаемых вейвлет-коэффициентов для рассматриваемого случая — 720 из каждой функции. Увеличение числа коэффициентов приводит к повышению времени расчетов, а также модуля коэффициента корреляции между признаками данной категории.

Прежде всего нужно определить законы распределения исследуемых признаков. Был проведен статистический анализ данных, и на основании метода Хи-квадрат Пирсона определены наиболее близкие законы распределения значений признаков из табл. 2 для большинства испытуемых, принимавших участие в эксперименте по сбору данных. Проверка осуществлялась как для каждого субъекта в отдельности, так и для всей совокупности значений признаков испытуемых (рис. 2, 3).

Как видно из табл. 2, часть признаков категории 3.1 имеет распределение значений, близкое к распределению Лапласа (307 из 720), часть признаков близка к нормальному распределению (533 из 720). Аналогично ситуация обстоит с признаками категории 3.2: большинство признаков имеют нормальное распределение (659 из 720), остальные — двойное экспоненциальное (распределение Лапласа у 61 из 720 признаков).

Время перекрытия клавиш (категория 1.3, см. табл. 2) является крайне нестабильным

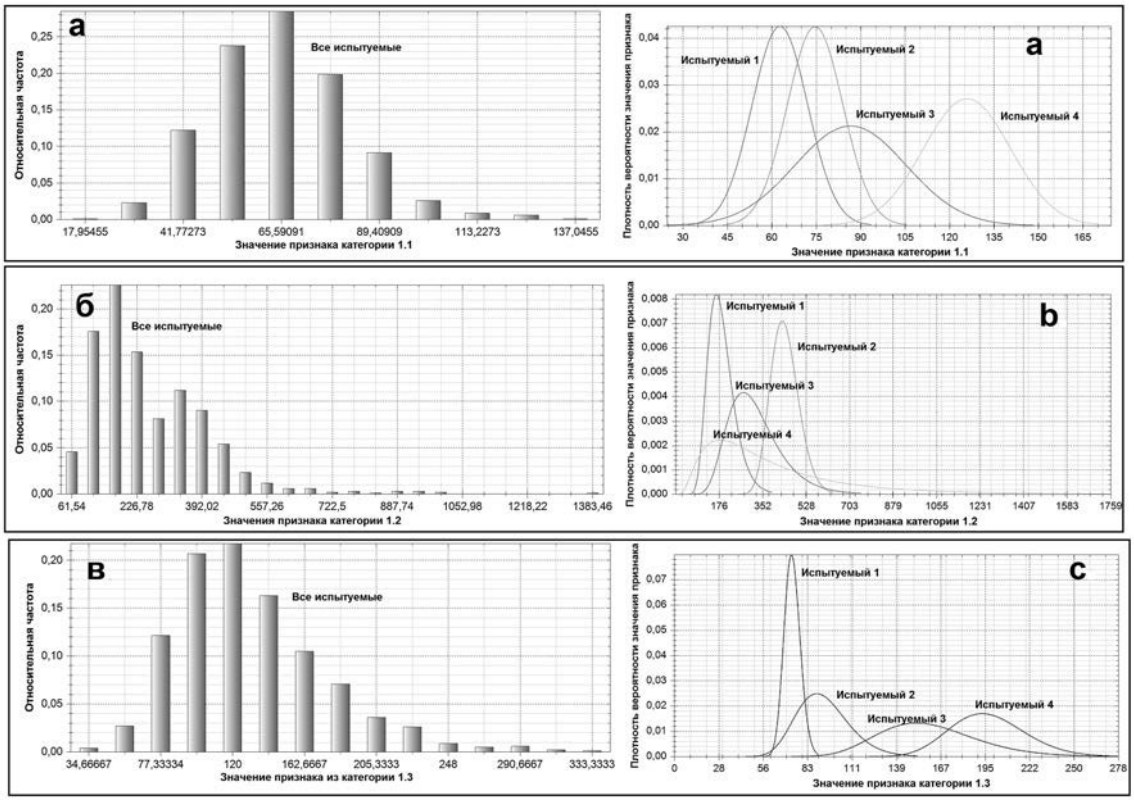

Рис. 2. Гистограмма относительных частот (для 100 испытуемых, слева) и плотности распределения значений признаков (справа): а — категории 1.1; б — категории 1.2; в — категории 1.3

Fig. 2. A histogram of the relative frequencies (for 100 subjects, left) and the density distribution of feature values (right): a — category 1.1; b — category 1.2; c — category 1.3

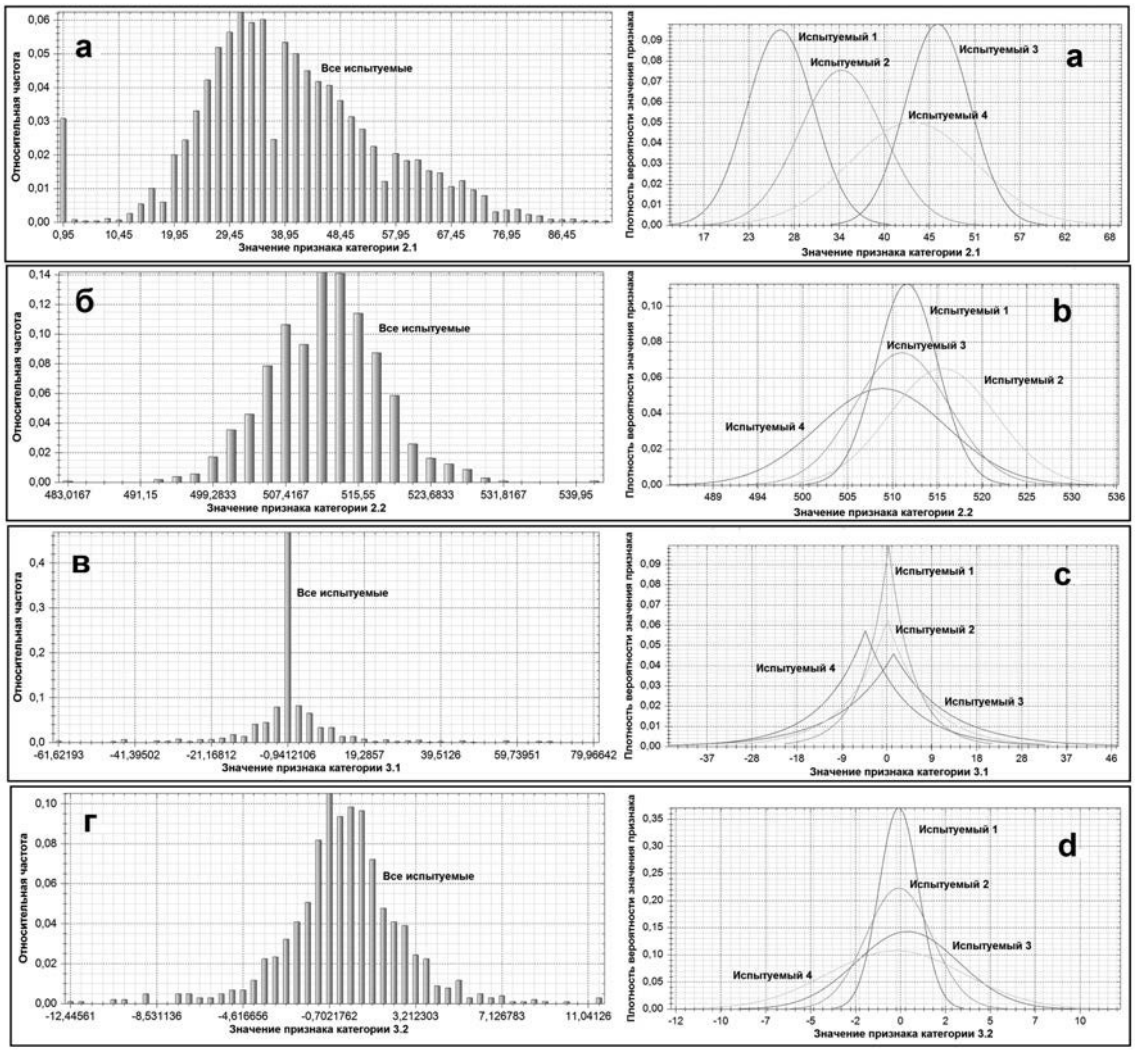

Рис. 3. Гистограмма относительных частот (для 100 испытуемых, слева) и плотности распределения значений признаков (справа): а — категории 2.1; б — категории 2.2; в — категории 3.1; г — категории 3.2

Fig. 3. A histogram of the relative frequencies (for 100 subjects, left) and the density distribution of feature values (right): a — category 2.1; b — category 2.2; c — category 3.1; d — category 3.2

признаком, так как данное событие может присутствовать либо отсутствовать при повторном воспроизведении парольной фразы одним и тем же субъектом.

Параметры распределения признаков, вычисленные по значениям, характеризующим испытуемого, являются эталонным описанием идентифицируемого субъекта (или эталоном).

Оценка корреляционной зависимости признаков клавиатурного почерка

Чтобы понять, сколько новой информации содержится в дополнительных признаках, определим корреляционную зависимость между временами удержания клавиш, давлением на них и вибрацией клавиатуры, созда-

ваемой при нажатии на клавиши, а также паузами до нажатия следующих клавиш фразы. Для этого достаточно вычислить коэффициенты парной корреляции между соответствующими участками реализации субъекта. Вычислим данные коэффициенты корреляции по всем реализациям всех испытуемых и построим гистограммы относительных частот этих коэффициентов (рис. 4).

Также определим, насколько зависимы признаки категории 3.1 между собой. Для этого вычислим коэффициенты парной корреляции между когерентными сечениями соответствующих признаков. Сечением назовем совокупность значений признака (по аналогии со случайной величиной). Когерентные сечения двух признаков содержат последовательности их значений, совпадающие по фазе, т. е. порядок следования реализаций в массиве значений признака одинаков для 2 сечений. По результатам оценки корреляционная взаимная зависимость 720 признаков категории иногда превышает 0,3 и по шкале Чеддока соответствует слабой. Аналогичный результат был получен относительно признаков категории 3.2. Поэтому дальнейшее увеличение количества вейвлет-коэффициентов возможно, однако время на обработку образцов клавиатурного почерка с добавлением этих признаков существенно возрастает.

Как видно из рис. 4, корреляция между признаками различных категорий почти во всех случаях не превышает 0,7 и в основном соответствует слабой зависимости по шкале Чеддока либо отсутствует вовсе, в некоторых случаях (не более 15%) зависимость является умеренной, крайне редко заметной (менее 3%). Вывод: информация в признаках из различных категорий не дублируется, каналы получения признаков можно считать слабо зависимыми.

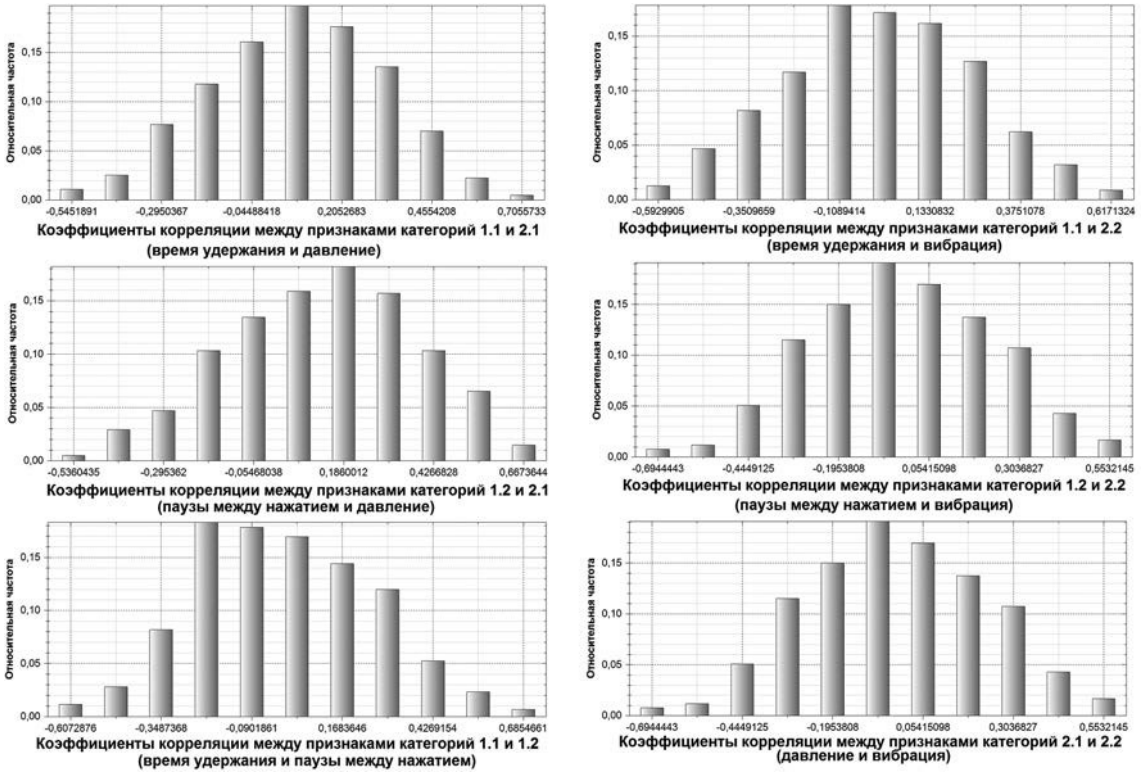

Рис. 4. Взаимная зависимость между признаками из различных категорий

Fig. 4. Mutual dependence between the features of the different categories

Оценка информативности признаков клавиатурного почерка

Об информативности признака можно судить по площадям пересечения плотностей вероятностей его значений, характеризующих различных субъектов [11]. Площадь пересечения S_i функции плотности вероятности признака, характеризующая *i*-го испытуемого, с аналогичными плотностями, характеризующими других испытуемых, является суммой вероятностей ошибок 1-го и 2-го рода, т. е. вероятностью ошибочной идентификации данного субъекта по данному признаку. При этом для различных субъектов данная S_i-площадь может существенно различаться, интегральную оценку информативности можно получить исходя из распределения S_i-площадей. Было установлено, что S_i-площади имеют распределение, близкое к нормальному. Построив графики параметров распределений S_i-площадей по всем имеющимся эталонам, можно оценить общую информативность признаков. Наиболее информативные признаки имеют наименьшее математическое ожидание $Mx(S_i)$ площадей S_i, напротив, признаки с наименьшей информативностью обладают наивысшими оценками $Mx(S_i)$. Чем больше информативность признака различается для различных субъектов, тем выше среднеквадратичное отклонение $Sx(S_i)$ площадей S_i. Признаки с высокими оценками $Sx(S_i)$ являются информативными для одних испытуемых и малоинформативными для других.

Несложно заметить (рис. 5, а), что при увеличении количества идентифицируемых образов S_i-площадь любого признака, характеризующего эти образы, возрастает и стремится к единице, т. е. вероятность ошибочного распознавания любого образа по определенному (одному) признаку при достаточно большом числе эталонов субъектов становится близкой по значению к единице. Более объективную оценку информативности признака дает подсчет средних парных площадей $Sp_{i,j}$ пересечения плотностей вероятности его значений ($Sp_{i,j}$-площадь вычисляется как площадь пересечения только двух функций плотности вероятности признака, характеризующих *i*-го и *j*-го субъектов, как показано на рис. 5, б). Чтобы увидеть разницу между S_i- и $Sp_{i,j}$-площадями, достаточно взглянуть на рис. 5, который демонстрирует их физический смысл. Аналогичный график параметров нормального распределения $Sp_{i,j}$-площадей приведен на рис. 6. $Sp_{i,j}$-площади не возрастают при увеличении количества исследуемых субъектов, их значения стремятся к некоторому числу, характеризующему информативность признака. Обобщенные оценки информативности признаков можно видеть в табл. 4.

Представленные в табл. 4 данные относительно S_i-площадей были получены при количестве эталонов 20, так как если задействовать все 100 имеющихся эталонов испытуемых, все площади превысят 0,98, что не является наглядным результатом для со-

Таблица 4. Средние оценки информативности признаков
Table 4. Average ratings of informativeness of features

№ категории	Категория признака	*Mx* (*S*)	*Sx* (*S*)	*Mx* (*Sp*)	*Sx* (*Sp*)
1.1	Время удержания клавиш	0,96060	0,05377	0,62099	0,08352
1.2	Паузы между нажатием клавиш	0,94460	0,06887	0,57107	0,11538
1.3	Время перекрытия клавиш	Значений признаков недостаточно			
2.1	Давление на клавиши	0,92078	0,10518	0,53589	0,10424
2.2	Вибрация при нажатии клавиш	0,98560	0,01986	0,80117	0,03526
3.1	Вейвлет-преобразование функции $p(t)$	0,98010	0,02189	0,77147	0,05706
3.2	Вейвлет-преобразование функции $vibro(t)$	0,98729	0,01756	0,78358	0,04579

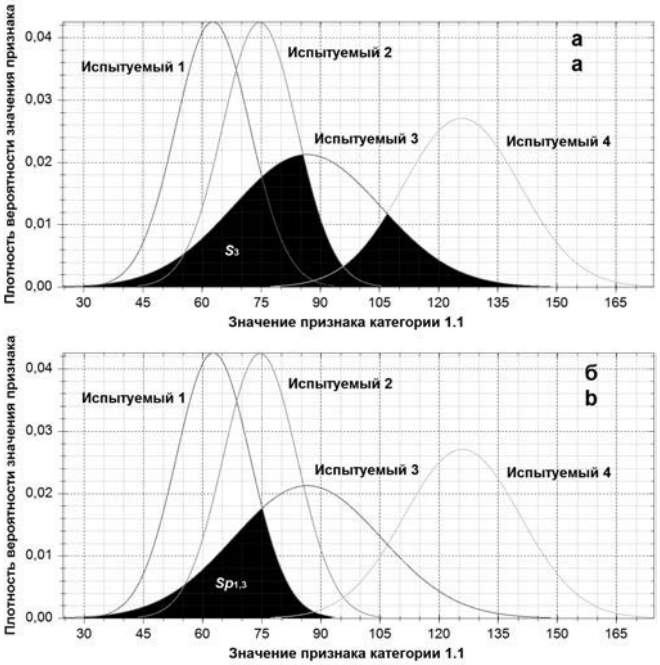

Рис. 5. Подходы к оценке информативности признаков (площади закрашены):
а — через вычисление S_i-площадей; б — через вычисление $Sp_{i,j}$-площадей

Fig. 5. Approaches to the evaluation of features informative (squares shaded):
a — through the calculation of S_i-squares; b — through the calculation of $Sp_{i,j}$-squares

поставления. Несмотря на сравнительно низкие средние оценки информативности признака категорий 3.1–3.2, общий потенциал этих признаков очень высок, что обусловлено их большим количеством. Из парольной фразы порядка 30–40 символов за приемлемое время можно получить 1440 признаков с преимущественно слабой взаимной корреляцией по шкале Чеддока.

Как видно на рис. 6, наибольшей информативностью обладают вейвлет-коэффициенты с высокой частотой. При переходе к следующему уровню разложения разрешение по времени повышается, а диапазон частот снижается (см. табл. 3), наибольшее суммарное количество информации о печатающем субъекте содержится в высокочастотной части функций $p(t)$ и $vibro(t)$. В целом вибрация клавиатуры менее информативна, чем давление на клавиши для целей идентификации оператора.

Идентификация субъектов по клавиатурному почерку

Реализовать алгоритм идентификации субъектов можно при помощи метода последовательного применения формулы гипотез Байеса. Данный метод также называют стратегией Байеса [11; 12], и он заключается в вычислении интегральных апостериорных вероятностей гипотез за некоторое число шагов, равное количеству признаков, при помощи формулы гипотез Байеса (3). Каждая гипотеза подразумевает, что предъявляемая реализация клавиатурного почерка принадлежит определенному субъекту, т. е. каждая гипотеза ассоциируется с определенным эталоном субъекта. На каждом шаге за априорную вероятность принимается апостериорная вероятность, вычисленная на предыдущем шаге, в качестве условной вероятности подается на вход плотность вероятности значения оче-

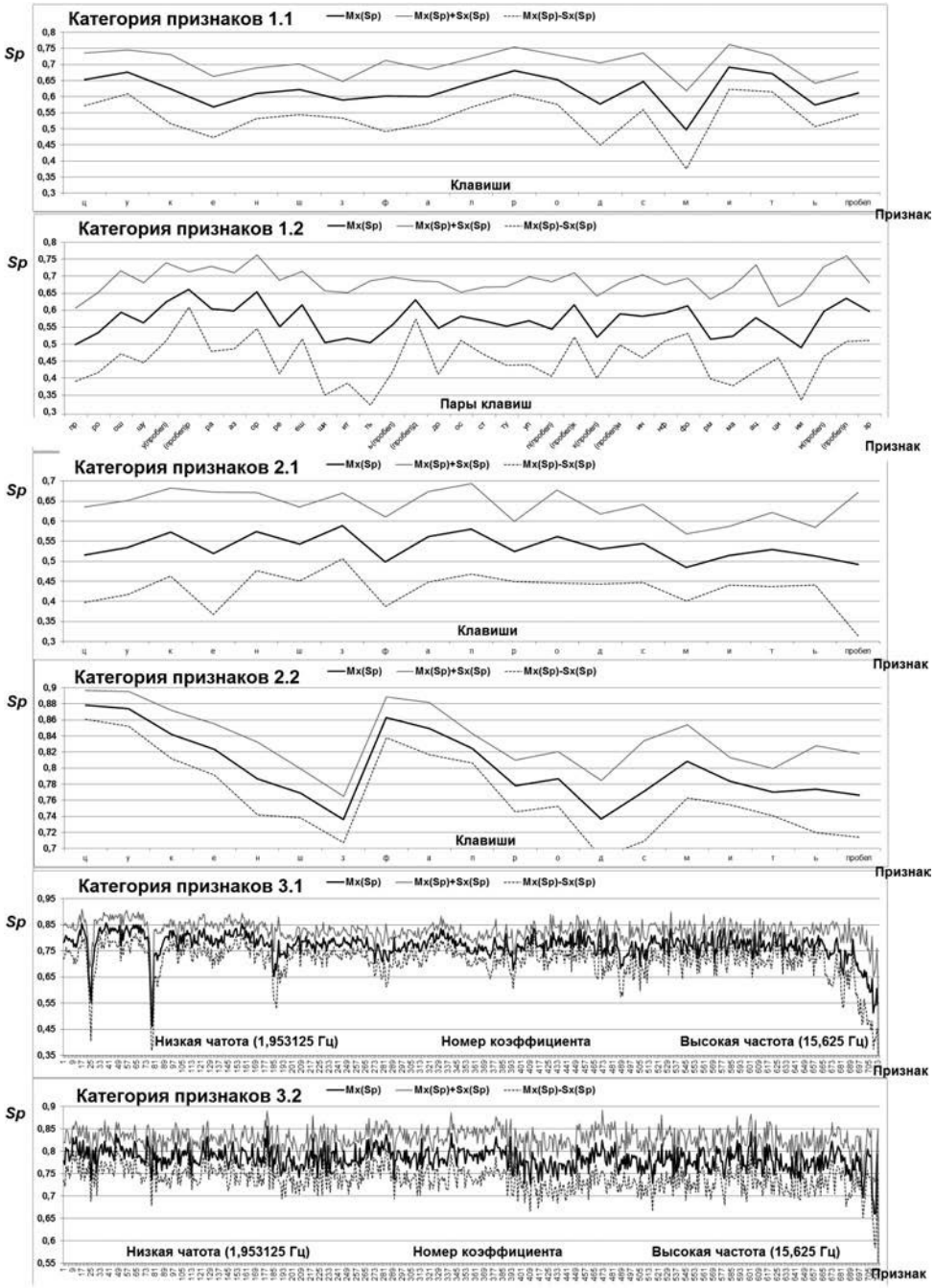

Рис. 6. Оценка информативности признаков через $Sp_{i,j}$-площади (100 субъектов)

Fig. 6. Evaluation of features informative through the $Sp_{i,j}$-squares of 100 subjects

редного признака. На первом шаге все гипотезы (субъекты) считаются равновероятными, т. е. $P_0(H_i/A) = 1/n$, где n — количество гипотез (пользователей). Условные вероятности вычисляются исходя из закона распределения значений признаков. В [12] показано, что стратегия Байеса дает лучшие результаты при идентификации произвольных образов в про-

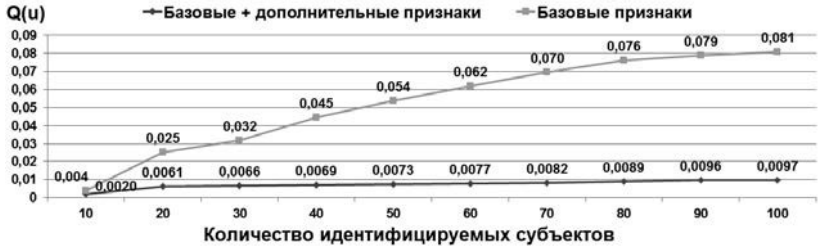

Рис. 7. Вероятности ошибок идентификации $Q(u)$
при количестве идентифицируемых образов субъектов u

Fig. 7. Probability of identifying errors $Q(u)$, where u is the quantity of identifiable images of subjects

странстве малоинформативных признаков по сравнению с некоторыми другими методами из статистической теории принятия решений.

$$P(H_i|A_j) = \frac{P_{j-1}(H_i|A)P(A_j|H_i)}{\sum_{i=1}^{n} P_{j-1}(H_i|A)P(A_j|H_i)}, \quad (3)$$

где $P(A_j|H_i)$ — условная вероятность гипотезы H_i о том, что предъявленные данные принадлежат эталону i-го испытуемого, равная плотности вероятности значения j-го признака A_j; $P_{j-1}(H_i|A)$ — апостериорная вероятность i-й гипотезы на j — 1 шаге при поступлении значения j-го признака.

Для создания эталонов испытуемых (вычисления параметров распределения признаков) использовалось по 21 реализации парольной фразы от каждого испытуемого по аналогии с тем, какое количество рекомендуется использовать для обучения нейронных сетей (сетей персептронов) в ГОСТ Р 52633.5–2011 [13]. Остальные реализации использовались для проведения идентификации. Всего проведено более 10 000 опытов. Достоверность всех результатов составила более 0,99 при доверительном интервале вероятности 0,003.

Результаты эксперимента иллюстрирует рис. 7. Появление события одновременного нажатия на клавиши характерно не для всех попыток ввода и не для всех пользователей. Использование этой категории признаков при идентификации на основе стратегии Байеса затруднительно, так как не ясно, чему должна быть равна условная вероятность $P(A_j|H_i)$ в случае, если у субъекта j-й признак отсутствует (при $P(A_j|H_i) = 0$, количество ошибок возрастает). Поэтому вероятности ошибок идентификации $Q(u)$, представленные на рис. 7, получены без использования признаков категории 1.3.

Заключение

Сформулируем основные полученные результаты.

1. Разработана клавиатура с использованием специальных датчиков давления и вибрации для снятия дополнительных характеристик клавиатурного почерка.

2. Предложена категория признаков, основанная на применении вейвлет-преобразования Добеши D6 к функции давления пальцев на клавиши и функции вибрации клавиатуры при наборе текста.

3. Определены законы распределения базовых и дополнительных признаков клавиатурного почерка.

4. Произведена оценка корреляционной зависимости признаков: определено, что зависимость базовых (временные характеристики нажатий клавиш) и дополнительных признаков (давление на клавиши и вибрация клавиатуры) в более чем 80% случаев является слабой (по шкале Чеддока). Таким образом, в предложенных признаках содержится новая информация о субъекте.

5. Определена информативность признаков, наивысшей информативностью обладают признаки давления на клавиши.

6. По результатам эксперимента вероятность ошибок при количестве идентифицируемых субъектов 100 составила 0,0097 при использовании как базовых, так и дополнительных признаков. Вероятность ошибок при использовании только базовых признаков составила 0,081.

7. Установлено, что дополнительные признаки в среднем могут снизить количество ошибок более чем в 7 раз.

При осуществлении вейвлет-преобразования на уровнях разложения, соответствующих скорости печати пользователей, возможно получить большее количество слабо и умеренно коррелирующих признаков, чем использовалось в настоящей работе (по предварительным оценкам, порядка 3000–4000) из функций давления и вибрации по одной парольной фразе, состоящей из 30–40 символов. При этом вероятность ошибки может снизиться в разы, однако время на обработку такой транзакции (и эксперимента) экспоненциально повышается. Данную категорию признаков предлагается использовать в целях идентификации, аутентификации пользователей важных информационных систем, в том числе осуществляемую в скрытом режиме, а также для распознавания психофизиологического состояния оператора в системах производственной и информационной безопасности, по аналогии с тем, как это предлагается делать в работе [14].

Список литературы

1. Утечки конфиденциальной информации в России и в мире. Итоги 2016 года. Zecurion Analytics. URL: http://www.zecurion.ru/upload/iblock/1e5/Zecurion_Data_Leaks_2016_full. pdf (дата обращения: 06.07.2016).
2. Утечки данных. Банки. Мир, Россия. 2015 год. InfoWatch. URL: https: //www.infowatch.ru/analytics/reports (дата обращения: 06.07.2016).
3. The Global State of Information Security® Survey 2016. PricewaterhouseCoopers URL: http://www.pwc.com/gx/en/issues/cyber-security/information-security-survey/download.html (дата обращения: 27.06.2016).
4. *Иванов А. И.* Нейросетевая защита конфиденциальных биометрических образов гражданина и его личных криптографических ключей: монография. Пенза, 2014. — 57 с.
5. *Васильев В. И., Ложников П. С., Сулавко А. Е., Еременко А. В.* Технологии скрытой биометрической идентификации пользователей компьютерных систем // Вопросы защиты информации. 2015. №3. С. 37–47.
6. *Иванов А. И.* Биометрическая идентификация личности по динамике подсознательных движений. Пенза: Изд-во Пенз. гос. ун-та, 2000. — 188 с.
7. *Еременко А. В., Сулавко А. Е.* Двухфакторная аутентификация пользователей компьютерных систем на удаленном сервере по клавиатурному почерку // Прикладная информатика. 2015. Т. 10. №6 (60). С. 48–59.
8. *Salil P. Banerjee, Damon L.* Woodard. Biometric Authentication and Identification using Keystroke Dynamics: A Survey // Journal of Pattern Recognition Research. 2012. №7. P. 116–139.
9. *Pisani Paulo Henrique, Lorena Ana Carolina.* A systematic review on keystroke dynamics // Journal of the Brazilian Computer Society. 2013. 19 (4).
10. *Graps A.* An Introduction to Wavelets // IEEE Computational Sciences and Engineering. 1995. Vol. 2. №2. P. 50–61.
11. *Епифанцев Б. Н., Ложников П. С., Сулавко А. Е.* Алгоритм идентификации гипотез в пространстве малоинформативных признаков на основе последовательного применения формулы Байеса // Межотраслевая информационная служба. 2013. №2. С. 57–62.
12. *Епифанцев Б. Н., Ложников П. С., Сулавко А. Е.* Сравнение алгоритмов комплексирования признаков в задачах распознавания образов // Вопросы защиты информации. 2012. №1. С. 60–66.
13. ГОСТ Р52633.5–2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия — код доступа. М.: Стандартинформ, 2011. — 20 с.
14. *Епифанцев Б. Н., Ложников П. С., Сулавко А. Е., Жумажанова С. С.* Идентификационный потенциал рукописных паролей в процессе их воспроизведения // Автометрия. 2016. №3. С. 28–36.

References

1. *Utechki konfidentsialnoy informatsii v Rossii i v mire. Itogi 2016 goda. Zecurion Analytics*. Available at: http://www.zecurion.ru/upload/iblock/1e5/Zecurion_Data _Leaks_2016_full. pdf.
2. *Utechki dannyih. Banki. Mir, Rossiya. 2015 god. InfoWatch*. Available at: https: //www.infowatch.ru/analytics/reports).
3. *The Global State of Information Security® Survey 2016. PricewaterhouseCoopers*. Available at: http://www.pwc.com/gx/en/issues/cyber-security/ information-security-survey/download.html).

4. Ivanov A. I. *Nejrosetevaja zashhita konfidencial'nyh biometricheskih obrazov grazhdanina i ego lichnyh kriptograficheskih kljuchej*. Monografija, Penza, 2014. 57 p.
5. Vasil'ev V. I., Lozhnikov P. S., Sulavko A. E., Eremenko A. V. Tehnologii skrytoj biometricheskoj identifikacii pol'zovatelej komp'juternyh system. *Voprosy Zashhity Informacii*, 2015, no. 3, pp. 37–47.
6. Ivanov A. I. *Biometricheskaja identifikacija lichnosti po dinamike podsoznatel'nyh dvizhenij*. Penza, Izdatelstvo Penzenskogo gosudarstvennogo universiteta, 2000. 188 p.
7. Eremenko A. V., Sulavko A. E. Dvuhfaktornaja autentifikacija pol'zovatelej komp'juternyh sistem na udalennom servere po klaviaturnomu pocherku. *Prikladnaja Informatika* — Journal of Applied Informatics, 2015, vol. 10, no. 6 (60), P. 48–59.
8. Salil P. Banerjee, Damon L. Woodard. Biometric Authentication and Identification using Keystroke Dynamics: A Survey. *Journal of Pattern Recognition Research*, 2012, no. 7, pp. 116–139.
9. Pisani Paulo Henrique, Lorena Ana Carolina. A systematic review on keystroke dynamics. *Journal of the Brazilian Computer Society*, 2013, 19 (4).
10. Graps A. An Introduction to Wavelets. *IEEE Computational Sciences and Engineering*, 1995, vol. 2, no. 2, pp. 50–61.
11. Epifancev B. N., Lozhnikov P. S., Sulavko A. E. Algoritm identifikacii gipotez v prostranstve maloinformativnyh priznakov na osnove posledovatel'nogo primenenija formuly Bajesa. *Mezhotraslevaja Informacionnaja Sluzhba*, 2013, no. 2, pp. 57–62.
12. Epifancev B. N., Lozhnikov P. S., Sulavko A. E. Sravnenie algoritmov kompleksirovanija priznakov v zadachah raspoznavanija obrazov. *Voprosy Zashhity Informacii*, 2012, no. 1, pp. 60–66.
13. GOST R52633.5–2011. *Zashhita informacii. Tehnika zashhity informacii. Avtomaticheskoe obuchenie nejrosetevyh preobrazovatelej biometrija-kod dostupa*. Moscow, Standartinform, 2011. 20 p.
14. Epifantsev B. N., Lozhnikov P. S., Sulavko A. E., Zhumazhanova S. S. Identification Potential of Online Handwritten Signature Verification. *Optoelectronics, Instrumentation and Data Processing*, 2016, no. 3 (52), pp. 238–244. DOI: 10.3103/S8756699016030043.

A. Eremenko, *Omsk Transport University, Omsk, Russia, 4eremenko@gmail.com*

A. Sulavko, *Omsk State Technical University, Omsk, Russia, sulavich@mail.ru*

D. Mishin, *Vladimir State University named after Alexander and Nikolay Stoletovs, Vladimir, Russia, mishin.izi@gmail.com*

A. Fedotov, *Omsk Transport University, Omsk, Russia, fedotov1609@gmail.com*

Identification potential of keyboard handwriting considering vibration parameters and force keystrokes[1]

The article considers the problem of data protection from unauthorized access by means of user identification by keyboard handwriting. The estimation of informativeness of different features that characterize the keyboard handwriting of subjects, including the dynamics of change in pressure when you press the keys and keyboard settings vibration. The category of new features, based on using of wavelet transform Daubechies D6 to function of the pressure fingers on the keys and keyboard functions of vibration while typing, was proposed. The laws of distribution of basic and additional features of keyboard handwriting were determined. To form the base of biometric samples a keyboard was designed with the use of special sensors. The estimation of the correlation dependence of features was made. It is determined that the correlation between basic features (temporal characteristics of keystrokes) and additional features (pressure on the keys and the keyboard vibration) in more than 80% of cases is weak. Thus, in the proposed new attributes contain information about the subject. An assessment of the probability of identification errors based on the Bayesian strategy using the various features of the spaces was made. It is found that additional features can reduce the average number of errors is more than 7 times.

Keywords: keyboard writing, force of pressure on keys, sensors, operator identification, biometric feature.

About authors: A. Eremenko, *PhD in Technique*; A. Sulavko, *PhD in Technique*; D. Mishin, *PhD in Technique*

For citation: Eremenko A., Sulavko A., Mishin D., Fedotov A. Identification potential of keyboard handwriting considering vibration parameters and force keystrokes. *Prikladnaya Informatika* — Journal of Applied Informatics, 2017, vol. 12, no. 1 (67), pp. 79–94 (in Russian).

[1] The reported study was partially supported by RFBR, research project No. 16-37-50007.