ПРОГРАММИРОВАНИЕ, 2020, № 1, с. 15–28

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.56+004.93

ОСОБЕННОСТИ НЕПРЕРЫВНОЙ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ СВОБОДНЫХ ТЕКСТОВ В РЕЖИМЕ СКРЫТОГО МОНИТОРИНГА

© 2020 г. Е. А. Кочегурова[a],*, Ю. А. Мартынова[a],**

[a] *Национальный исследовательский Томский политехнический университет*
634050 Томск, пр. Ленина, 30, Россия
**E-mail: kocheg@mail.r*
***E-mail: martynova@tpu.ru*

Поступила в редакцию 03.06.2019 г.
После доработки 17.06.2019 г.
Принята к публикации 26.06.2019 г.

Данная работа посвящена исследованию особенностей динамической (непрерывной) идентификации пользователей на основе показателей клавиатурного почерка (КП). Клавиатурная динамика отслеживается в режиме скрытого мониторинга при создании пользователем свободного текста в любом приложении. Анализ подходов к статической идентификации не выявил существенных ограничений на их применение к непрерывной идентификации. Главная особенность непрерывной идентификации состоит в способе сбора динамической информации о клавиатурных нажатиях и коррекции шаблонов зарегистрированных пользователей. Показана эффективность включения в алгоритмы распознавания дополнительных классификационных признаков, например, связанных с частотностью использования букв в текстах. Разработано программное приложение для сбора и анализа образцов КП. Исследования, проведенные в домене пользователей с хорошими навыками работы с компьютером, показали вполне удовлетворительную точность распознавания пользователей – в среднем 87%. Причем, точность не зависит от выбранного метрического расстояния при распознавании и несколько повышается при использовании масштабирующих коэффициентов учета частотности букв алфавита.

DOI: 10.31857/S0132347420010033

1. ВВЕДЕНИЕ

Развитие цифровых и сетевых технологий за последнее десятилетие способствовало расширению возможностей для доступа и хранения конфиденциальной информации на цифровых устройствах. Поэтому и защита информации от несанкционированного доступа становится все более актуальной.

По данным аналитического центра InfoWatch число утечек информации в России возрастает год от года. По сравнению с 2016 г. рост числа утечек в 2017 г. резко возрос, почти на 37%, а в 2018 г. по всему миру на 30% больше утечек конфиденциальной информации. Краже чаще всего подвержены персональные и платежные данные и составляют от всего объема утечек за 2017 г. 64.1% и 21.1% соответственно. По-прежнему главным каналом утечек остается Сеть (броузер, cloud) – 69.8% (2016 + 11.6%). Основными виновными утечки информации оказались сотрудники и руководство организаций – 64% и 50% случаев.

В связи с этим повышается необходимость динамического или непрерывного распознавания пользователей, имеющих доступ к общественным и личным информационным ресурсам.

Кроме технических и организационных мер для защиты информации широко используются средства аутентификации (авторизации) и идентификации пользователей. Наряду с парольными и имущественными методами аутентификации, нередко используются биометрические характеристики. В соответствии с Российскими и международными стандартами (ГОСТ Р 54412-2011 и ISO/IEC 24745), биометрия – наука, которая изучает методы определения идентичности человека на основе физиологических и поведенческих особенностей. Биометрический признак (характеристика) – это уникальная, измеримая характеристика, используемая для верификации отдельно взятого человека [1–3]. Выделяют физиологические и поведенческие характеристики. К первой группе относятся уникальные признаки, полученные человеком при рождении. Например, ДНК, отпечатки пальцев, радужная оболочка гла-

за, рисунок вен, форма ушей и др. Поведенческие же характеристики приобретаются со временем и способны меняться с возрастом или в результате внешнего воздействия. В качестве примера можно назвать голос, рукописный и клавиатурный почерк, походку.

Установлено, что динамика нажатия клавиш или клавиатурный почерк (КП) определяет индивидуальный ритм набора текста и может использоваться как биометрическое средство идентификации личности [4–6]. Отмечено, что динамика нажатия клавиш “не то, что вы набираете, а то, как вы печатаете” [Monrose-1997]. Как поведенческая характеристика, КП по своей природе являются динамической. Обычно он формируется через 6 месяцев работы с компьютером [7]. Поведенческие характеристики включают условно постоянную компоненту, обусловленную физиологией пользователя (его способностями и навыками работы с клавиатурой) и случайную компоненту, определяемую психоэмоциональным состоянием человека. Поведенческие характеристики КП в отличие от физиологических сложнее распознать с высокой точностью, но вместе с тем, сложнее и подделать [4]. Основной акцент данного исследования сделан на анализ особенностей динамического распознавания КП.

Статья структурирована следующим образом. В разделах 2–3 рассмотрены существующие подходы и тенденции клавиатурной динамики. Раздел 4 посвящен особенностям сбора данных и показателям КП. В разделах 5–6 обсуждаются показатели эффективности, алгоритмы и методы клавиатурной идентификации. И, наконец, раздел 7 описывает проведенный эксперимент и его результаты.

2. ПОДХОДЫ К КЛАВИАТУРНОЙ ДИНАМИКЕ

Внедрение системы аутентификации по КП не требует значительных материальных вложений. Единственное необходимое оборудование – это стандартная клавиатура, которой оснащены все компьютеры, и высокоэффективное программное обеспечение.

Первые работы по анализу динамики клавиатурных нажатий относятся к 80-м годам прошлого века. В то время была обнаружена связь пользователя и ритма клавиатурных нажатий; были введены первые биометрические характеристики, способы их получения и исследованы статистические методы распознавания пользователей.

Существует несколько классификаций для клавиатурной идентификации пользователей. Одна из них – по типу создаваемого текста.

I. Статическая аутентификация [8–10]. Проверка пользователя осуществляется во время первичной аутентификации на основе структурированного/предопределенного текста. Предопределенный текст обеспечивает более надежную проверку пользователя, чем только ID/пароль при первичной аутентификации. Также предопределенный текст может дополнять проверку ID/пароль в случае возникающих подозрений.

II. Динамическая или непрерывная аутентификация [10–16]. Непрерывная проверка осуществляется на протяжении всего сеанса работы пользователя на основе произвольного текста, вводимого им в любом программном приложении. Такой скрытый мониторинг обеспечивает дополнительную меру безопасности после авторизации пользователя в системе. Свободный текст в большей степени, чем предопределенный, напоминает реальную среду работы пользователя и позволяет выявить поведенческие характеристики.

Хотя аутентификация на основе пароля проста в разработке и обслуживании, но она оказывается бесполезной в случае выявления пароля злоумышленниками. Время набора пароля часто бывает недостаточным, чтобы оценить изменение динамики нажатия клавиш. И если статический режим используется преимущественно при первичной аутентификации, то основной целью скрытого мониторинга является установление подлинности пользователя или его психоэмоционального состояния в процессе работы в корпоративной системе.

Одним из недостатков клавиатурной идентификации является невысокая точность по сравнению с другими биометрическими системами. Поэтому повышение качества распознавания попрежнему является актуальной задачей, особенно для динамической.

Целью данной работы является исследование возможности применения подходов статической идентификации для динамических данных и динамической идентификации.

При автоматической идентификации пользователей на основе КП существует ряд особенностей в организации получения и распознавания входных данных. Основные из них следующие:

– сбор статистики клавиатурных нажатий;

– выбор временных показателей клавиатурной динамики;

– создание шаблонов (профилей) пользователей;

– выбор показателей эффективности распознавания;

– алгоритмы и методы распознавания.

3. СОСТОЯНИЕ ВОПРОСА И АКТУАЛЬНОСТЬ КЛАВИАТУРНОЙ ИДЕНТИФИКАЦИИ

Клавиатурная динамика в последние годы является областью активных исследований и ис-

пользования в системах информационной безопасности вследствие низкой стоимости и простоты сопряжения с существующими системами. Впервые работы по использованию клавиатурных нажатий, как средства идентификации пользователей, были связаны с измерением (изучением) моторных навыков нажатия клавиш. И только в 90-е годы [17] клавиатурную динамику стали рассматривать, как поведенческую характеристику почерка.

Первые обзоры исследований КП [17, 18] относятся к началу 2000 годов. В частности, в работах 2004–2010 гг. [1, 4, 11, 19, 20] были проанализированы актуальные классификаторы на основе статистических методов и нейронных сетей. Было установлено, что вероятностные методы более привлекательны в вычислительном плане, но имеют более низкую точность аутентификации. Появились рекомендации по созданию публичных наборов данных для аутентификации. Также в обзорах тех лет отмечено, что большинство работ относится к статической аутентификации пользователей, т.е. на основе пароля или фиксированного текста. Большинство исследователей подтвердили, что системы аутентификации на основе клавиатурной динамики имеют потенциал в области кибербезопасности и биометрического мониторинга.

В обзорах последних лет 2011–2018 [5, 9, 16, 21–27] кроме исследования развития методов классификации и распознавания пользователей, проанализированы факторы, влияющие на производительность систем аутентификации. В эти годы помимо использования стандартных клавиатур персонального компьютера, активно развивается аутентификация пользователей сенсорных клавиатур, мобильных устройств [28–31], и на основе web-сервисов [32]. Количество публикаций по клавиатурной биометрике достаточно быстро увеличивается, особенно в последнее десятилетие. На рис. 1 показано число публикаций и тенденции роста: в период 1998–2012 данные взяты из [23] и [9], 2013–2017 гг. данные получены нами на основе анализа баз данных (БД) Scopus, IEEEXplore, Science Direct, электронной библиотеки ACM. В [9] изложена технология выявления прямых ссылок на наиболее значимые результаты в этой области. И из более, чем 2000 тематических ссылок на клавиатурную динамику выделены 16 прямых ссылок на работы, опубликованные в 2004–2009 гг. Именно в эти годы были заложены теоретические основы клавиатурной динамики. Этот вывод подтверждают и анализ опубликованных патентов [3].

Исследования нынешнего десятилетия значительно расширили прикладной характер клавиатурной биометрии. Кроме аутентификации на базе мобильных устройств, web-приложений, появились работы по гендерной и возрастной

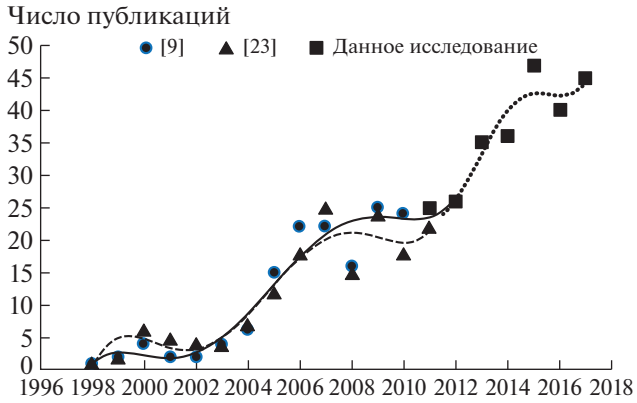

Рис. 1. Тенденции роста числа публикации по клавиатурной динамике.

идентификации на базе клавиатурной динамики [33]. Перспективна линия работ, связанная со смешанной аутентификацией [24, 34]. Например, на базе КП и особенностей лица [34]. Как и динамическая идентификация КП более перспективна, чем статическая, при решении современных задач биометрического распознавания: гендерного, психоэмоционального состояния.

Необходимо отметить, что клавиатурное распознавание не может быть единственным средством биометрической аутентификации, как и не может полностью заменить парольную систему аутентификации. Однако несомненные достоинства, такие как в скрытый мониторинг, низкая стоимость внедрения, высокая степень идентификации и простота интеграции с другими системами безопасности, делают клавиатурную аутентификацию достойным дополнительным средством безопасности.

В реализации системы клавиатурного распознавания участвуют три подсистемы:

а) сбор данных и извлечение характеристик КП;

б) формирование клавиатурного профиля;

в) идентификация пользователя.

4. СБОР ДАННЫХ И ХАРАКТЕРИСТИКИ НАЖАТИЯ КЛАВИШ

Сбор данных о динамике нажатия клавиш – это начальный этап реализации клавиатурной идентификации. Его эффективность полностью определена специальным программным обеспечением и не требует дополнительного оборудования. При клавиатурных нажатиях операционная система фиксирует код ANSI, связанный с нажатой клавишей, время ее нажатия и отпуска. Как событие нажатие клавиши можно измерить с точностью до миллисекунды. На этом и основан сбор первичных данных по динамике нажатия клавиш пользователями.

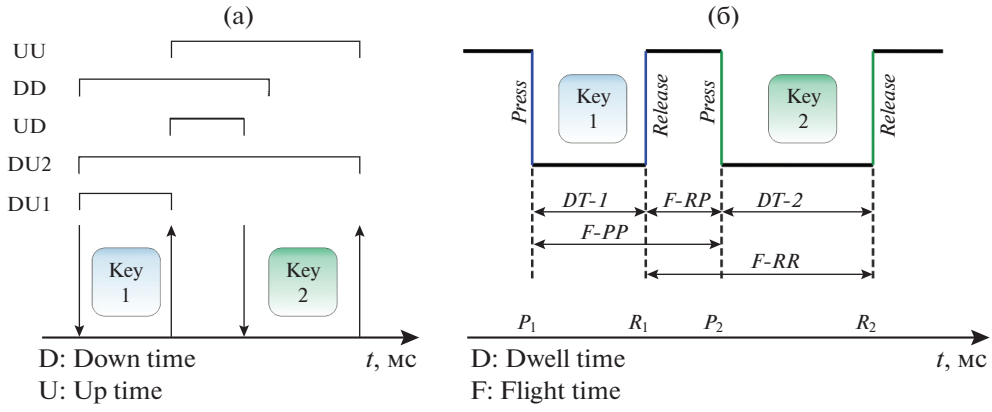

Рис. 2. Характеристики клавиатурного почерка.

В целом динамика нажатия клавиш определяет процесс оценки ритма набора текста и формирует собственный шаблон (профиль, образ) отдельного человека. Шаблон пользователя, как и его рукописная подпись или текст, создается под действием определенных нейрофизиологических факторов. Это делает шаблоны КП уникальными [9]. При использовании клавиатуры компьютера у человека задействовано до 140 мышц [36, 37], 20% от которых управляют нажатием клавиш. Соответственно 28-мерная задача управления позволяет говорить об уникальности КП-пользователей и его потенциальной значимости для распознавания личности.

Шаблон нажатия клавиши способен обеспечить уникальную функцию аутентификации. Он основан на достаточно большом наборе показателей, извлекаемых из статистики о клавиатурных нажатиях, которые характеризуют длительность, задержку нажатия клавиш и скорость набора текста. Большинство исследователей [9, 10, 23, 35] используют следующие показатели:

– время удержания клавиши;

– паузы между нажатиями;

– число ошибок при вводе;

– степень аритмичности при наборе;

– скорость набора;

– особенности использования служебных клавиш.

В некоторых предыдущих исследованиях также учитывалось использование давления на клавиши [24], однако для этого требуется специализированная клавиатура. Но ввиду повсеместного использования устройств с сенсорным экраном интерес к таким работам повышается. При этом часть характеристик: аритмичность, скорость набора, исправление ошибок чаще используются, как дополнительные к основным временным параметрам.

Впервые обоснование для использования временного шаблона нажатий клавиш было выполнено еще в 1980 г. [38]. И на протяжении десятилетий актуальность использования шаблона остается по-прежнему высокой. Для формирования шаблона существует необходимость извлекать эффективные характеристики из зафиксированных операционной системой нажатий.

Большая часть исследователей КП анализируют в своих работах [22, 9, 35, 10, 39] временные характеристики между двумя последовательными нажатиями клавиш, так называемые диграфы (Di-graph). Существует два основных диграфа: время ожидания (Dwell time DT, Hold time HD) и время задержки (Flight time FT или keystroke latency). Времена ожидания и задержки имеют наглядную геометрическую иллюстрацию. На рисунках 2а, 2б приведены эти характеристики в двух основных нотациях графического изображения временных характеристик.

На рисунке 2-а) представлены следующие характеристики клавиатурной динамики в терминах: Down/UP :

DU1: временной интервал между моментами, в которые нажата и отпущена клавиша, т.е. время удержания клавиши (Dwell time или Hold time).

DU2: временной интервал между моментами нажатия одной клавиши и отпусканием следующей клавиши.

UD: временной интервал между моментами, в которых одна клавиша отпущена, а другая нажата, т.е. пауза. (Flight time).

DD: временной интервал между моментами нажатия двух клавиш.

UU: временной интервал между моментами отпуска двух клавиш.

На рисунке 2б в терминах Press/ Release приведены следующие времена задержек:

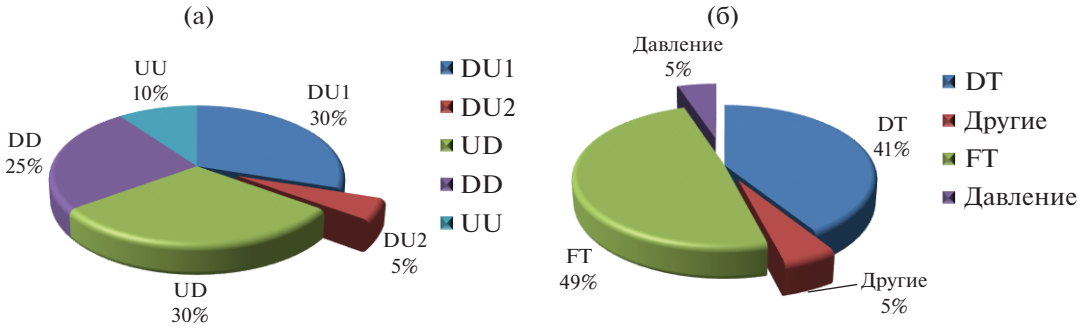

Рис. 3. Процентное распределение временных признаков.

F-RP=UD;

F-PP=DD;

F-RR=UU.

DU – важнейшая характеристика, под которой понимается промежуток времени, в течение которого клавиша находится в нажатом состоянии. Как правило, диапазон изменения удержания клавиши (15–150) миллисекунд.

Манеру создания текстов во многом определяет число перекрытий (наложений) нажатия клавиш. Наложение происходит тогда, когда одна клавиша еще не отпущена, а другая уже нажата. С повышением скорости набора текста увеличивается число наложений. В случае наложений пауза UD принимает отрицательное значение.

4.1. Функции динамики нажатия клавиш

Хотя большинство исследователей при анализе динамики КП считают важнейшими характеристиками время удержания клавиши (DT/DU) и ожидания нажатия (UD), определенный вклад в профиль пользователя вносит анализ задержек (keystroke latency) [11, 23, 45]. Кроме указанных выше диграфов существуют аналогичные характеристики для трех клавиш – триграммы. И они показали неплохие результаты для аутентификации [4]. В обзорах [9, 22, 23] приведен анализ частоты использования временных меток разными исследователями. На рисунках 3 приведены диаграммы процентного соотношения основных временных меток.

Общее число исследователей для представленных данных на рис. 3а и 3б различно (40 [9] и 187 [23]), но общая закономерность отчетливо прослеживается. Основными временными характеристиками клавиатурной динамики являются удержания клавиши (DT/DU) и обобщенная группа времени задержек (FT), которая на рис. 3-а) разбита на 3 части (UD + DD + UU).

Выбранные показатели о клавиатурных нажатиях пользователя собираются, обрабатываются и на этой основе формируется эталонный профиль пользователя. Большинство исследований в области анализа нажатия клавиш собирают данные из структурированного и предопределенного текста.

При непрерывном (динамическом) сборе данных необходимо возникает ряд дополнительных вопросов [6]:

– объем данных для формирования шаблонов. Условно объемы данных при непрерывном исследовании клавиатурной динамики принято разделять на небольшой (менее 1000 нажатий), средний (менее 6000 нажатий) и большой (более 6000 нажатий) на каждого пользователя [14, 23, 24]. Иногда в свободных текстах анализируется количество слов. По данным [23] 57% исследователей анализируют короткие тексты, 24% – длинные, оставшиеся 19% распределены между текстами, содержащие только цифры и тексты неизвестного содержания;

– формирование и использование общедоступных БД профилей пользователей. Первые БД клавиатурных нажатий появились в 2009 года. БД содержат разное количество пользователей (30–300) и разное число повторений в сессию. Но абсолютное большинство БД содержат клавиатурные данные, собранные на основании паролей и фиксированных текстов и, следовательно, малопригодны для скрытого и непрерывного мониторинга;

– полное отсутствие подобных БД с шаблонами пользователей на русском языке. А это приводит к необходимости сбора данных каждым исследователем. И соответственно развитие теории распознавания клавиатурной динамики и затрудняет сопоставление результатов в равных условиях.

5. ОЦЕНКА ЭФФЕКТИВНОСТИ АУТЕНТИФИКАЦИИ

Эффективность идентификации показывает способность метода отличить подлинного пользователя домена и предотвратить несанкционированный доступ.

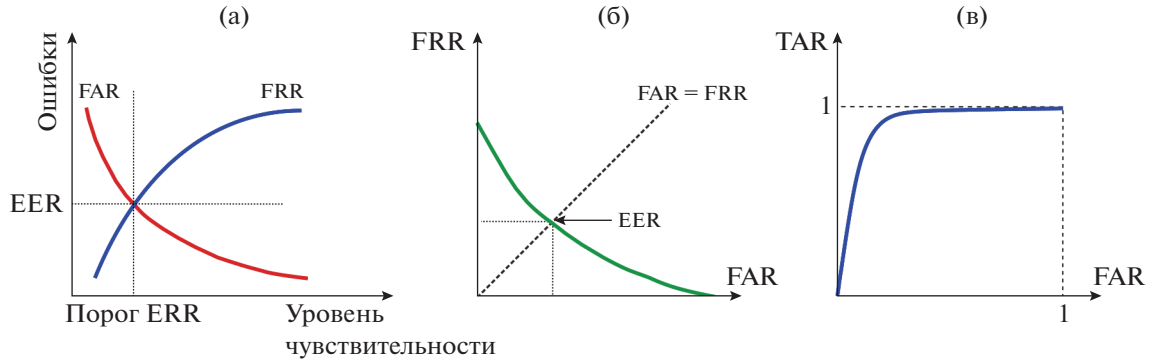

Рис. 4. Показатели эффективности клавиатурной идентификации.

При сравнении двух образцов КП возможны следующие варианты развития событий [40, 41]:

True Accept (TA): образцы принадлежат одному и тому же пользователю, и система определяет образцы как схожие – это ожидаемое событие;

True Reject (TR): образцы принадлежат разным пользователям, и система определяет их как не схожие – также ожидаемое событие;

False Reject (FR): образцы принадлежат одному и тому же пользователю, но система определяет их как несхожие – т.е. опровергается верная гипотеза;

False Accept (FA): образцы принадлежат разным пользователям, но система определяет их как схожие – принимается ложная гипотеза.

Показатели эффективности, используемые в разных клавиатурных исследованиях, можно обобщить в четыре основные показателя частоты ошибок.

False Rejection Rate (FRR) – оценка ложного отклонения, известна в статистической радиотехнике, как ошибка I рода. FRR определяет процент случаев, когда законный пользователь ошибочно отклоняется. Вычисляется, как отношение ложно отклоненных пользователей к общему количеству пользователей системы.

$$\text{FRR} = \frac{\text{Количество ложных отказов}}{\text{Общее количество попыток}} \quad (1)$$

Более низкая FRR подразумевает меньшее отклонение пользователей и более легкий доступ для них.

False Acceptance Rate (FAR) – ошибка ложного принятия. FAR или ошибка II рода определяет процент случаев принятия нелегальных пользователей. Вычисляется, как отношение ложно принятых неавторизованных пользователей к общему количеству пользователей системы.

$$\text{FAR} = \frac{\text{Количество ложных совпадений}}{\text{Общее количество попыток}} \quad (2)$$

Гипотетически ошибки FRR и FAR варьируются в зависимости от уровня чувствительности алгоритма (порогового значения) как показано на рис. 4а: когда одна ошибка уменьшается, другая увеличивается.

Более высокие значения FAR обычно предпочтительнее в системах, где безопасность не имеет первостепенной важности, тогда как более высокие значения FRR являются предпочтительным в приложениях с высокой степенью защиты [1]. Компромисс между FAR и FRR должен определяться целями конкретной прикладной задачи. Например, низкий пропуск нелегальных пользователей (FAR) соответствует высокому пороговому значению (чувствительности), но приводит к большому отклонению зарегистрированных пользователей (FRR), т.е. к их низкому пропуску.

Между ошибками FAR и FRR существует компромисс, т.е. можно уменьшить одну погрешность за счет другой, регулируя порог принятия решения. На рисунке 4б представлена кривая DET (Detection Error Trade-off) [41, 42], позволяющая определить компромисс между ошибками I рода (FRR) и II рода (FAR). По ней можно установить одну из желаемых ошибок и увидеть жертву другой.

Часто используемая Equal Error Rate (EER) – равная частота ошибок – для определения общей точности системы распознавания. EER представляет значение ошибки, когда FAR и FRR принимают равные значения [10]. В отличие от FAR и FRR, ошибка EER не зависит от уровня чувствительности алгоритма аутентификации. Чем ниже значение EER, тем эффективнее система распознавания при заданном пороговом значении. Независимый от порога EER более подходит для оценки эффективности алгоритмов распознавания.

Показатели FAR, FRR и EER весьма популярны и перспективны в системах клавиатурной аутентификации. Менее часто используемая характеристика ROC (Receiver Operating Characteris-

Рис. 5. Частота использования методов клавиатурного распознавания.

tic) [41, 42], позволяет увидеть предельные значения показателей эффективности. Характеристика ROC отображает компромисс между верным (TAR) и ошибочным (FAR) пропуском пользователя при различных пороговых значениях. Верхний левый угол графика представляет собой идеальную точку, где TAR равно единице, а FAR равно нулю.

6. АЛГОРИТМЫ РАСПОЗНАВАНИЯ ПОЛЬЗОВАТЕЛЕЙ

В разделе 2 данной работы описано состояние вопроса, актуальность и подходы к клавиатурной идентификации в историческом аспекте. Большинство алгоритмов распознавания клавиатурных профилей, используемых на протяжении трех десятилетий, относятся к статической идентификации пользователей. Обзорные работы [3, 6, 9, 14, 16, 22–24], опубликованные за последние 5 лет, отражают современное состояние клавиатурного распознавания. Практически каждое опубликованное исследование включает подробные таблицы с перечислением авторов, алгоритмов классификации и распознавания, временных показателей и достигнутых показателей эффективности. Резюмируя результаты этих исследований можно отметить:

– количество исследований по динамической аутентификации очевидно мало, не более 10% [23];

– основные исследования по-прежнему посвящены использованию стандартной клавиатуры;

– 80% исследований основаны на коротких или предопределенных текстах.

Подходы к динамической и статической идентификации принципиально не отличаются и с позиции распознавания условно разделены на 3 основные группы:

– оценка метрических расстояний;

– статистические методы;

– методы машинного обучения.

Хотя статистические методы и были первыми в задаче распознавания КП, их использование актуально и сейчас во всех задачах клавиатурной идентификации. Вероятностный подход оценивает вероятность того, что клавиатурный профиль принадлежит пользователю некоторого домена (БД). Используемые методы вероятностного моделирования включают байесовские методы, скрытую Марковскую модель, функцию гауссовой плотности и взвешенную вероятность. Анализируемыми статистическими показателями являются: средние, медианные, стандартные отклонения, статистический t-критерий для оценки подобия [10, 13, 25] и др. Однако разделение на статистические и вероятностные методы условно, и они могут быть объединены в один класс методов. Используя данные [23, 39] по методам распознавания КП на рис. 5 показано процентное распределение методов распознавания и классификации КП.

Из рисунка 5 следует, что оценка расстояния (метрики) является самым популярным методом распознавания и достигает по частоте использования 23%. В рамках этого подхода вычисляется расстояние между эталонным и текущим профилем пользователя, которое затем сопоставляется с пороговым значением. Наиболее часто используемыми метрическими расстояниями являются меры Евклида, Манхэттенская, Хэмминга, Махаланобиса [11, 24, 26]. В том числе на основе относительных расстояний для длинного и свободного текста [14].

Следующим по частоте использования (16%) является метод искусственных нейронных сетей (ИНС), относящийся к категории методов распознавания образов на основе машинного обучения. Предполагается, что ИНС более эффективна, чем статистические методы [15, 24, 43]. Однако сложность использования ИНС, как классификатора КП, состоит в необходимости иметь для обучения

Таблица 1. Распределение букв алфавита по частотным диапазонам

Русский алфавит		Английский алфавит	
Частотный диапазон %	Буквы алфавита	Частотный диапазон %	Буквы алфавита
[10.98–5.47]	о/е/а/и/н/т/с	[12.02–6.02]	e/t/a/o/i/n/s/r
[4.75–1.59]	р/в/л/к/м/д/п/у/я/ы/ь/г/з/б	[5.92–2.09]	h/d/l/u/c/m/f/y/w
[1.45–0.013]	я/й/х/ж/ш/ю/ц/щ/э/ф/ъ/ё	[2.03–0.07]	g/p/b/v/k/x/q/j/z

сети, образцы легальных и нелегальных пользователей. Общие архитектуры ИНС представляют собой многослойный персептрон (MLP) [45], радиальную базовую функциональную сеть (RBFN), квантование векторного обучения (LVQ) и самоорганизующуюся карту (SOM) [43]. Однако, непрерывная идентификация пользователей предполагает обновление шаблонов пользователей, что приводит к переобучению сети и увеличению времени идентификации.

К группе распознавания КП на основе машинного обучения также относятся ряд известных алгоритмов, включая дерево решений, нечеткую логику и эволюционные вычисления.

Дерево принятия решений весьма популярный метод, благодаря своей низкой вычислительной сложности. Но ввиду рекуррентности процедуры распознавания, метод эффективен лишь при небольших множествах легальных и нелегальных пользователей [13, 41].

Нечеткая логика использует многозначную логику для моделирования задач с неоднозначными данными. Основная идея заключается в построении границ области принятия решений на основе данных обучения с функциями принадлежности и нечеткими правилами. После выделения пространства признаков и вычисления значений принадлежности производится идентификация категории, которой принадлежит исследуемый шаблон.

Эволюционные вычислительные методы, основанные на идее естественного отбора, в задаче клавиатурной идентификации включают ряд известных подходов: генетические алгоритмы (GA), алгоритмы роевого интеллекта и ряд других [5]. Данные оптимизационные алгоритмы позволяют осуществить направленный поиск максимального совпадения анализируемых клавиатурных шаблонов, тем самым повышая точность распознавания.

Другой известный классификатор – метод опорных векторов (SVM) [10, 13, 12, 24]. Концепция этого метода заключается в том, чтобы сначала определить, как два класса функций легальных и нелегальных пользователей отличаются друг от друга. Затем создается граница, которая лучше всего разделяет эти классы функций. И по расположению проверяемого шаблона относительно выделенной границы выделяют законных и незаконных пользователей. SVM обладает показателями эффективности распознавания сопоставимыми с нейронной сетью при меньших вычислительных затратах. Однако производительность заметно снижается, когда набор временных показателей велик.

Кластерный анализ при анализе КП – это метод объединения похожих клавиатурных профилей и образцов. Цель состоит в том, чтобы сгруппировать образец с аналогичными клавиатурными признаками для формирования однородного кластера. Профили из разных кластеров очень разнородны, но очень похожи между собой в одном кластере. К этой группе методов относятся K-means, K-star, DBScan, KNN и др. [10, 14, 24, 45].

Разделение методов распознавания КП по группам и подходам весьма условно. Разные подходы могут приводить к одной и той же модели, но при этом методы ее обучения могут быть разными.

7. ОПИСАНИЕ ЭКСПЕРИМЕНТА И РЕЗУЛЬТАТЫ

Описанные выше подходы к статической идентификации не выявили существенных ограничений на их применение к непрерывной (динамической) идентификации. Главная особенность непрерывной идентификации состоит в способе сбора динамической информации о клавиатурных нажатиях и коррекции шаблонов зарегистрированных пользователей. В связи с этим при проведении эксперимента и его обработке были поставлены следующие задачи:

– сбор и актуализация динамических наборов данных;

– расширение известных подходов статической идентификации для случая динамического распознавания личности на основе свободных и длинных текстов;

– повышение эффективности процедуры непрерывной идентификации в пространстве выделенных клавиатурных признаков.

Анализ подходов к статической идентификации и особенностей непрерывной идентификации позволил сформировать методику автоматического распознавания пользователей некоторого домена при создании им произвольных текстов. В основе

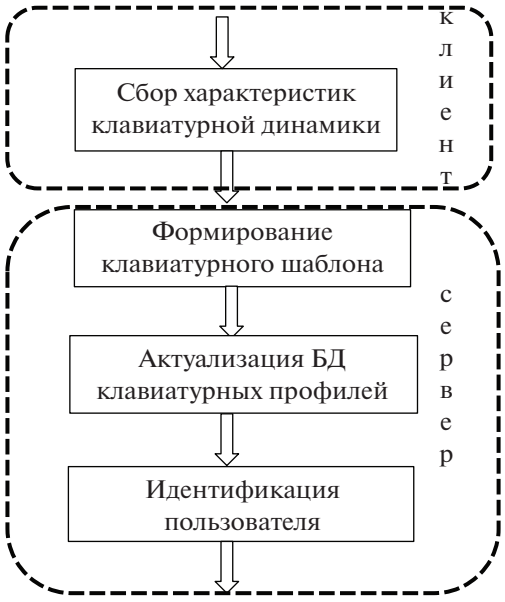

Рис. 6

методики лежит расширение проанализированных подходов к статической идентификации. Основные этапы автоматического распознавания следующие.

I. Непрерывный сбор информации о клавиатурных нажатиях.

Эта информация собирается на основе работы пользователя в любом приложении Windows: в окне браузера, тестовом редакторе и других приложениях. ОС с помощью механизма перехвата сообщений Windows-hook позволяет зафиксировать любое нажатие (отпускание) клавиш функцией события до того, как оно дойдет до приложения [12]. При этом сохраняется информация о том, какая клавиша была нажата или отпущена и время возникновения события клавиатуры.

Для поддержания обычного психоэмоционального состояния пользователь, как правило, не информирован о мониторинге своих действий на клавиатуре. При этом режим скрытого мониторинга не нарушает конфиденциальности персональных данных, т.к. не производит семантический анализ вводимых данных.

II. Актуализация динамических наборов данных.

В процессе работы на клавиатуре и при создании текстов несколько изменяется ритм набора. Поэтому при динамической идентификации личности имеет смысл анализировать не весь поток данных, а определенный объем последних символов. В данной работе минимально допустимый объем свободного текста составляет 200 символов. Далее происходит накопление объема клавиатурных нажатий до 1000 символов и последние 1000 символов постоянно обновляются на протяжении сеанса работы пользователя. Данные объемы статистических данных гарантируют получение несмещенных и состоятельных оценок для временных характеристик КП, а также репрезентативность (частотность) букв алфавита.

III. Формирование временного показателя клавиатурной динамики.

Анализ показателей, характеризующих динамику нажатия клавиш показал, что наиболее часто в прикладных задачах рассматриваются время удержания (DU) и пауза (UD), рис. 3.

В данном исследовании для повышения эффективности сырых данных было принято связать временные характеристики с частотой использования букв алфавита. Частотный диапазон использования букв достаточно широк: [0.013%–10.98%] для букв русского алфавита и [0.07%–12.02%] для английского.

Клавиатурный ритм в значительной мере зависит от частоты появления букв алфавита в тексте. Буквы, редко используемые в текстах, требуют большего внимания и времени при нажатии,

Таблица 2. Характеристики клавиатурного профиля

Клавиша	а	б	о	п	я	а	б	о	п	я
Пользователь	nvb16					ksk12				
Мат. ожидание	76.	85.5	77.5	74.0	81.4	83.9	81.7	67.5	88.8	95.1
Ср. отклонение	2.7	8.3	3.1	5.5	6.1	4.1	17.6	3.8	7.9	9.3
Max	82.3	106	83.2	89.5	95.5	91.8	135	74.6	107	117.3
Min	71.9	60	70.3	65.8	71	77.2	64.4	62.3	73.6	82.2
Пользователь	lrd1					vve15				
Мат. ожидание	99.6	102.8	100.6	100.4	116.5	75.4	58.0	57.5	75.4	87.1
Ср. отклонение	6.7	7.9	6.4	9.1	13.4	8.5	5.0	4.9	8.3	11.4
Max	113.2	117.3	113.72	119.1	143.2	98.4	65	68.3	101	121.8
Min	90.8	89.4	89.4	81.1	90.6	63.5	43.5	50.5	64.2	68.1

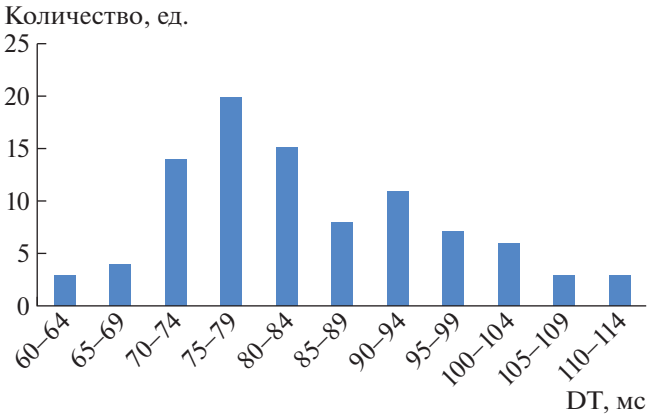

Рис. 7. Распределение времени удержания клавиши, мс.

так как не ассоциированы с мышечной памятью. Поэтому было предложено разделить частотные диапазоны русского/английского алфавита в соотношении 0.5/0.3/0.2. В таблице 1 представлено это разделение. Такие же нормирующие коэффициенты введены для временных характеристик каждой группы букв.

IV. Создание шаблонов (профилей) пользователей.

В БД хранятся клавиатурные шаблоны каждого известного системе пользователя домена, иногда за несколько последних сеансов работы пользователя.

Для каждой клавиши в текущем сеансе пользователя вычисляются средние значения основных временных характеристик клавиатурной динамики. Полученная информация формирует клавиатурный профиль (шаблон, почерк) пользователя и записывается в БД пользователей.

При динамическом распознавании требуются по каждому пользователю обновлять стек сеансовых шаблонов. В работе принято, что, если список шаблонов конкретного пользователя становится слишком длинным (более десяти образцов), самый старый образец удаляется. Так происходит актуализация клавиатурного профиля отдельных пользователей и БД домена.

V. Алгоритмы и методы распознавания.

Проведенные исследования показали, что самым популярным методом распознавания пользователя по его КП по-прежнему является оценка расстояния (метрики) между текущим и эталонным образцами КП. И в соответствии с рисунком 5 частота использования метода в прикладных исследованиях КП составляет 23%. Традиционно используются следующие меры близости: Евклидова, Манхэттенская, Махаланобиса и Хемминга. Евклидова метрика весьма популярна, так как является естественным расстоянием в Евклидовой геометрии. Манхэттенская или метрика городских кварталов, введена Г. Минковским, по сравнению с Евклидовой уменьшает влияние отдельных выбросов, так как не возводит разность в квадрат. Расстояние Махаланобиса применяется в случае ненулевой корреляции переменных и эта метрика инвариантна к масштабу. В данной работе алгоритм распознавания исследован с использованием Евклидовой и Манхэттенской метрики.

Для идентификации пользователя, каждый введенный образец почерка сравнивался с клавиатурными шаблонами из БД. При этом возможны две ситуации.

■ Введенный образец аналогичен одному из имеющихся в БД шаблонов почерка. Образцы считаются аналогичными, если расстояние между векторами характеристик этих образцов не превышают некоторого порогового значения.

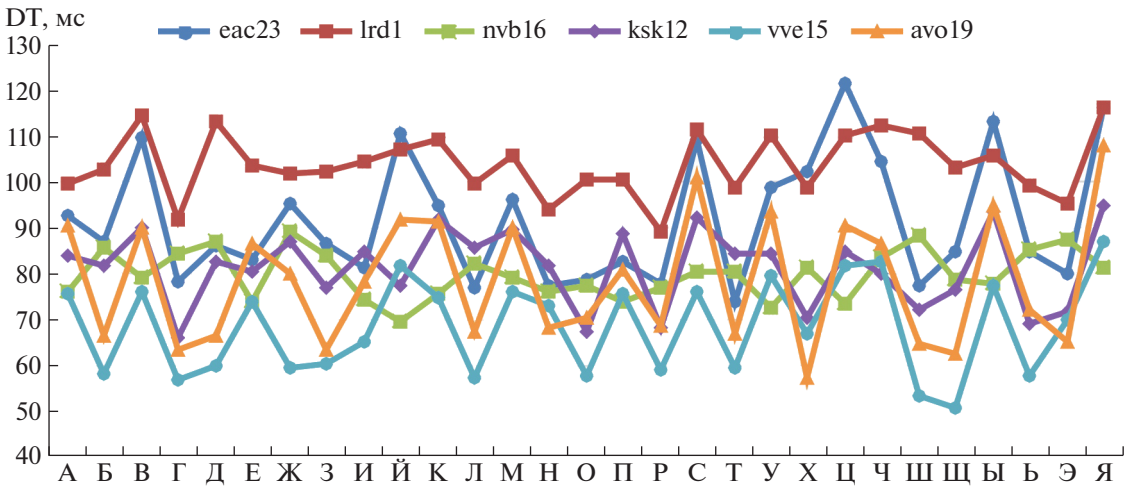

Рис. 8. Среднее время удержания клавиш пользователями домена, мс.

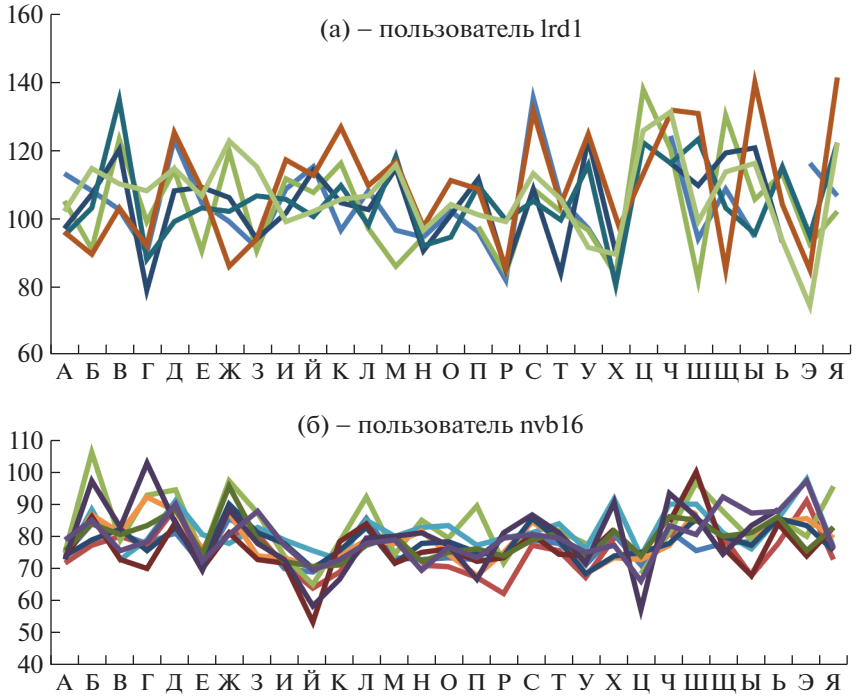

Рис. 9. Среднее время удержания клавиш в разных сеансах, мс.

Иначе образцы – непохожи. И, если образец аналогичен одному из имеющихся, система идентифицирует пользователя и добавляет новый образец почерка в стек сеансов в БД.

■ Новый образец не похож ни на один из имеющихся образцов. В этом случае фиксируется ошибка аутентификации и пользователь идентифицируется, как неопознанный.

Для реализации предложенной методики распознавания пользователя было создано программное приложение и проведен эксперимент. В эксперименте приняли участие в общей сложности тридцать пользователей ПК, использующих стандартную клавиатуру компьютера. Исследование было проведено на основе свободно создаваемых пользователями текстов в режиме скрытого мониторинга.

Пользователи домена имели разные уровни навыков набора текста, которые варьировались между умеренными и очень хорошими, что характерно для студентов и преподавателей университета.

Разработанное программное приложение установлено в корпоративном домене университета. Функционально приложение предназначено для непрерывного сбора информации о клавиатурных нажатиях на русском/английском языках и идентификации санкционированного или несанкционированного пользователя. Основные этапы работы приложения приведены на рис. 6.

После накопления достаточного количества данных, они отправляются на сервер программы для дальнейшей обработки. Для повышения надежности передача данных происходит посредством TCP-сокетов. Серверный компонент вычисляет средние значения основных временных характеристик клавиатурной динамики для каждой клавиши в текущем сеансе пользователя.

Непрерывный сбор и анализ клавиатурных данных позволил оценить идентификационные возможности пользовательских шаблонов. В таблице 2 в качестве примера приведен шаблон КП для 4-х пользователей с указанием их логинов в системе. В шаблоне представлены статистические характеристики времени удержания (DT) в миллисекундах для нескольких букв русского алфавита.

Анализ полученных образцов КП показал, что временная характеристика DT для каждой клавиши имеет бимодальный закон распределения. На рисунке 7 гистограмма распределения приведена для клавиши А. Бимодальная форма объясняется наличием наложений при нажатии клавиш, то есть вторая клавиша уже нажата, а первая еще не отпущена. При этом время удержания увеличивается. Именно поэтому имеет смысл учитывать число наложений при распознавании КП.

Пример представления данных о КП пользователей, приведенный в табл. 1, демонстрирует определенное расхождение поведенческих характеристик пользователей. Визуально такое рас-

Таблица 3. Сравнительный анализ показателей эффективности

	Евклидово расстояние		Манхэттенское расстояние	
	без учета частотности	с учетом частотности	без учета частотности	с учетом частотности
FRR	2.8%	2.7%	2.6%	2.4%
FAR	0%	0%	0%	0%
FRR + FAR	13.6%	12.7%	13.3%	12.3%
Точность алгоритма	86.4%	87.3%	86.9%	87.7%

хождение между шестью пользователями видно на рис. 8 для всех букв русского алфавита, исключая редко используемые ф, ъ, ё.

Как видно из рис. 8 форма кривых, отражающая рисунок клавиатурного профиля разных пользователей, отличается друг от друга. Это подтверждает возможность распознавания ритма ввода пользователей, а именно ритм и определяет КП.

Также имеются незначительные расхождения между сеансами в клавиатурном профиле для отдельного пользователя. На рисунке 9а и 9б представлены DT для 10 сеансов двух пользователей домена.

Отчетливо видно, что в отдельных реализациях профиля есть визуальные отличия, но в целом КП каждого пользователя узнаваем, а, следовательно, отражает его поведенческую характеристику.

Отличия обусловлены тем, что клавиатурные нажатия производятся людьми, которые подвержены усталости, психоэмоциональным эмоциям или двигательным проблемам. Такие изменения указывают на то, что данные нажатия клавиш могут содержать шум и выбросы.

В ходе эксперимента текущие образцы КП, собранные клиентом программы, сопоставлялись с эталонными шаблонами БД на сервере. Для сравнения были вычислены приведенные метрические расстояния с учетом весовых коэффициентов частотности и при их отсутствии.

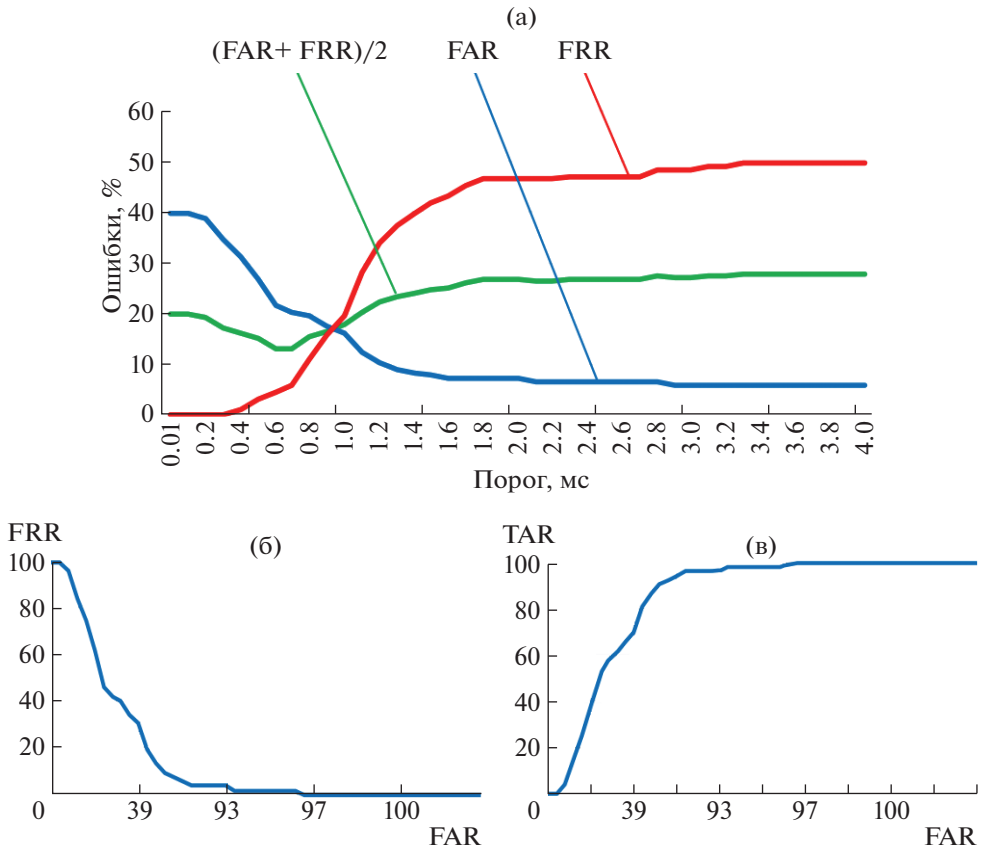

Рис. 10. Показатели эффективности распознавания пользователей.

Качество распознавания пользователей было оценено при помощи показателей эффективности, принятые в клавиатурной динамике и проанализированные в разделе 4. Часто используемые показатели FRR, FAR, и DET приведены соответственно на рисунке 10а, б, в. Также на рисунке 10а приведена полусумма показателей $\frac{FRR + FAR}{2}$, весьма привлекательная в ряде задач идентификации. Для установления значений порога сходства шаблонов в ходе эксперимента его значение варьировалось в диапазоне от 0.1 до 20 мс. На рисунке 10а приведена только информативная часть диапазона.

На основании проведенных исследований были получены значения минимальных критериев эффективности с использованием разных метрических расстояний и векторного критерия частотности букв алфавита, табл. 3.

В целом результаты получились довольно схожими для всех вариантов. Некоторое улучшение принесло использование масштабирующих коэффициентов учета частотности букв алфавита. В среднем суммарная ошибка уменьшилась на 1%, что отвечает основной цели исследования – повышение точности распознавания.

8. ЗАКЛЮЧЕНИЕ

Анализ исследований, проведенных нами и другими авторами в области идентификации пользователей на основе свободных текстов, позволил сделать ряд выводов.

1. Идентификация пользователя по его КП в режиме скрытого мониторинга возможна на основе произвольного текста, создаваемого в любом приложении.

2. Динамическая идентификация кроме особенностей, характерных для любого типа идентификации (создание шаблонов пользователей, выбор показателей, выбор алгоритмов распознавания) имеет ряд дополнительных:

– непрерывная корректировка динамических шаблонов, определяющая психоэмоциональное состояние человека;

– формирование векторного показателя клавиатурных нажатий, наиболее полно отражающего изменчивость клавиатурного почерка;

– включение в алгоритмы распознавания дополнительных классификационных признаков, например, связанных с частотностью использования букв в текстах или другими национальными особенностями текстов.

3. Исследования, проведенные в домене пользователей с хорошими навыками работы с компьютером, показали вполне удовлетворительную точность распознавания пользователей – в среднем 87%. Причем, точность не зависит от выбранного метрического расстояния при распознавании и несколько повышается при использовании масштабирующих коэффициентов учета частотности букв алфавита.

9. БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РФФИ (№ 18-07-01007).

СПИСОК ЛИТЕРАТУРЫ

1. *Yampolskiy R.V.* Behavioural biometrics: a survey and classification / R. V. Yampolskiy, V. Govindaraju // International Journal of Biometrics. 2008. V. 1. № 1. P. 81–113.
2. *Jain A.* Handbook of biometrics / A. Jain, P. Flynn, A. Ross. N.Y.: Springer, 2007. 556 p.
3. *Васильев В.И., Ложников П.С., Сулавко А.Е., Еременко А.В.* Технологии скрытой биометрической идентификации пользователей компьютерных систем (обзор) // Вопросы защиты информации. 2015. № 3 (110). С. 37–47.
4. *Bergadano F., Gunetti D., Picardi, C.* User authentication through keystroke dynamics // ACM Transactions on Information and System Security. 2002. V. 5. № 4. P. 367–397.
5. *Karnan M., Akila M., Krishnaraj N.* Biometric personal authentication using keystroke dynamics: A review// Applied Soft Computing. 2011. V. 11. № 2. P. 1565–1573.
6. *Pisani P.H., Lorena A.C.* Emphasizing typing signature in keystroke dynamicsusing immune algorithms // Applied Soft Computing. 2015. V. 34. P. 178–193.
7. *Иванов А.И.* Биометрическая идентификация личности по динамике подсознательных движений. Пенза: ПГУ. 2000. 187 с.
8. *Chang T.Y.* Dynamically generate a long-lived private key based on password keystroke features and neural network // Information Sciences. 2011. V. 211. P. 36–47.
9. *Pisani P.H., Lorena A.C.* A systematic review on keystroke dynamics // Journal of the Brazilian Computer Society. 2013. V. 19. № 4. P. 573–587.
10. *Kim J., Kim H., Kang P.* Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection // Applied Soft Computing. 2018. V. 62. P. 1077–1087.
11. *Gunetti D., Picardi C.* Keystroke analysis of free text // ACM Transactions on Information and System Security. 2005. V. 8. P. 312–347.
12. *Messerman T., Mustafić S., Camtepe A., Albayrak S.* Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics // Proceedings of the International Joint Conference on Biometrics (IJCB 2011). 2011. P. 1–8.
13. *Alsultan A., Warwick K., Wei H.* Non-conventional keystroke dynamics for user authentication // Pattern Recognition Letters. 2017. V. 89. P. 53–59.
14. *Kang P., Cho S.* Keystroke dynamics-based user authentication using long and free text strings from various input devices // Information Sciences. 2015. V. 308. P. 72–93.

15. *Ahmed A.A.* Biometric recognition based on free-text keystroke dynamics // IEEE Transactions on Cybernetics. 2014. V. 44. № 4. P. 458–472.
16. *Alsultan K., Warwick H.* Keystroke dynamics authentication: a survey of free-text methods // International Journal of Computer Science Issues. 2013. V. 10. № 4. P. 1–10.
17. *Joyce R., Gupta G.* Identity authentication based on keystroke latencies // Communications of the ACM. 1990. V. 33. № 2. P. 168–176.
18. *Spillane R.J.* Keyboard Apparatus for Personal Identification // Technical Disclosure Bulletin. 1975. V. 17. № 3346.
19. *Crawford H.* Keystroke dynamics: Characteristics and opportunities // Proceedings of the Eighth Annual International Conference on Privacy Security and Trust (PST). 2010. P. 205–212.
20. *Peacock A., Ke X., Wilkerson M.* Typing patterns: A key to user identification // IEEE Security and Privacy. 2004. V. 2. № 5. P. 40–47.
21. *Shanmugapriya D., Padmavathi G.* A survey of biometric keystroke dynamics: Approaches, security and challenges // International Journal of Computer Science and Information Security. 2009. V. 5. № 1. P. 115–119.
22. *Banerjee S.P., Woodard D.L.* Biometric authentication and identification using keystroke dynamics: a survey // Journal of Pattern Recognition Research. 2012. V. 7. № 1. P. 116–139.
23. *Teh P.S., Teoh A.B., Yue S.* A survey of keystroke dynamics biometrics // The Scientific World Journal. 2013. P. 1–24.
24. *Mondal S., Bours P.* A study on continuous authentication using a combination of keystroke and mouse biometrics // Neurocomputing. 2016. V. 230. P. 1–22.
25. *Крутохвостов Д.С., Хиценко В.Е.* Парольная и непрерывная аутентификация по клавиатурному почерку средствами математической статистики // Вопросы кибербезопасности. 2017. Т. 24. № 5. С. 91–99.
26. *Kochegurova E.A., Luneva E.E., Gorokhova E.S.* On continuous user authentication via hidden free-text based monitoring // Advances in Intelligent Systems and Computing. 2019. V. 875. P. 66–75.
27. *Vinayak R., Arora K.* A Survey of User Authentication using Keystroke Dynamics // International Journal of Scientific Research Engineering & Technology (IJSRET). 2015. V. 4. № 4. P. 378–384.
28. *Teh P.S.* A survey on touch dynamics authentication in mobile devices. Review article / P.S. Teh, N. Zhang, A.B. Teoh, K. Chen // Computers & Security. 2016. V. 59. P. 210–235.
29. *Mahfouz A., Eldin A.S., Mahmoud T.M.* A survey on behavioral biometric authentication on smartphones // Research article Journal of Information Security and Applications. 2017. V. 37. P. 28–37.
30. *Corpus K.R., Gonzales R.J., Morada A.S., Vea L.A.* Mobile user identification through authentication using keystroke dynamics and accelerometer biometrics // Proceedings of the International Conference on Mobile Software Engineering and Systems (MOBILESoft 16). 2016. P. 1–12.
31. *Соколов Д.А.* Использование клавиатурного почерка для аутентификации в распределенных системах с мобильными клиентами // Безопасность информационных технологий. 2010. № 2. С. 50–53.
32. *West A.G.* Analyzing the Keystroke Dynamics of Web Identifiers // Proceedings of the 2017 ACM on Web Science Conference (WebSci'17). 2017. P. 181–190.
33. *Pentel A.* Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns // Proceedings of UMAP'17 Adjunct, Publication of the 25th Conference on User Modeling. Adaptation and Personalization. 2017. P. 381–385.
34. *Ложников П.С., Сулавко А.Е., Бурая Е.В., Писаренко В.Ю.* Аутентификация пользователей компьютера на основе клавиатурного почерка и особенностей лица // Вопросы кибербезопасности. 2017. Т. 21. № 3. С. 24–34.
35. *Morales A., Fierrez J., Tolosana R., Ortega-Garcia J., Galbally J., Gomez-Barrero M., Anjos A., Marcel S.* KBOC: Keystroke Biometrics OnGoing Competition // Proceedings 8th IEEE International Conference on Biometrics: Theory, Applications and Systems. 2016. P. 1–6.
36. *Ворона В.А., Тихонов В.А.* Системы контроля и управления доступом. М.: Горячая линия – Телеком, 2010, 274 с.
37. *Berthold M., Borgelt C., Hopner F., Klawonn F.* Guide to Intelligent Data Analysis. Texts in Computer Science, London: Springer, 2010. V. 42. 394 p.
38. *Gaines R.S., Lisowski W., Press S.J., Shapiro N.* Authentication by Keystroke Timing: Some Preliminary Results. Technical Report R-2526-NSF. Santa Monica. CA: Rand Corporation, 1980. 41 p.
39. *Ali M.L., Monaco J.V., Tappert C.C., Qiu M.* Keystroke Biometric Systems for User Authentication // J Sign Process Systems. 2017. V. 86. P. 175–190.
40. *Kochegurova E.A., Gorokhova E.S., Mozgaleva A.I.* Development of the Keystroke Dynamics Recognition System // Journal of Physics: Conference Series. 2017. V. 803. № 1. P. 1–6.
41. *Alpar O.* Frequency spectrograms for biometric keystroke authentication using neural network based classifier // Knowledge-Based Systems. 2017. V. 116. P. 163–171.
42. *Goodkind A., Brizan D.G., Rosenberg A.* Utilizing overt and latent linguistic structure to improve keystroke-based authentication // Image and Vision Computin, 2017. V. 58. P. 230–238.
43. *Dozono H., Ito S., Nakakuni M.* The authentication system for multi-modal behavior biometrics using concurrent Pareto learning SOM // Proceedings of the 21st International Conference on Artificial Neural Networks. 2011. Part II. P. 197–204.
44. *Popovici E.C., Guta O.G., Stancu L., Arseni S.C., Fratu O.* Mlp neural network for keystroke-based user identification system // Proceeding 11th international conference on telecommunication in modern satellite, cable and broadcasting services (TELSIKS). 2013. V. 1. P. 155–158.
45. *Maxion R.A., Killourhy K.S.* Keystroke biometrics with number-pad input // Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN '10). 2010. P. 201–210.