Модель и метод идентификации пользователей на основе анализа клавиатурного почерка при вводе свободного текста

Шкляр Евгений Вадимович, старший преподаватель СПбГЭТУ «ЛЭТИ»

Руководитель: Шульженко Анастасия Дмитриевна, к.т.н., доцент каф. ИБ СПбГЭТУ «ЛЭТИ»

Решаемая проблема

Недостаточный уровень развития научно-методического аппарата биометрической идентификации при использовании в качестве идентификатора клавиатурного почерка.

Научная задача

Разработка моделей и методов для повышения эффективности систем идентификации на основе анализа клавиатурного почерка при вводе свободного текста.

Практическая значимость

Разработанный модели и методы могут использоваться для повышения эффективности идентификации в биометрических системах любых типов, перечисленных в ГОСТ Р 54412-2019

Цель исследования

Повысить эффективность биометрической идентификации на основе анализа клавиатурного почерка за счет снижения коэффициента равных ошибок (EER).

Положения, выносимые на защиту

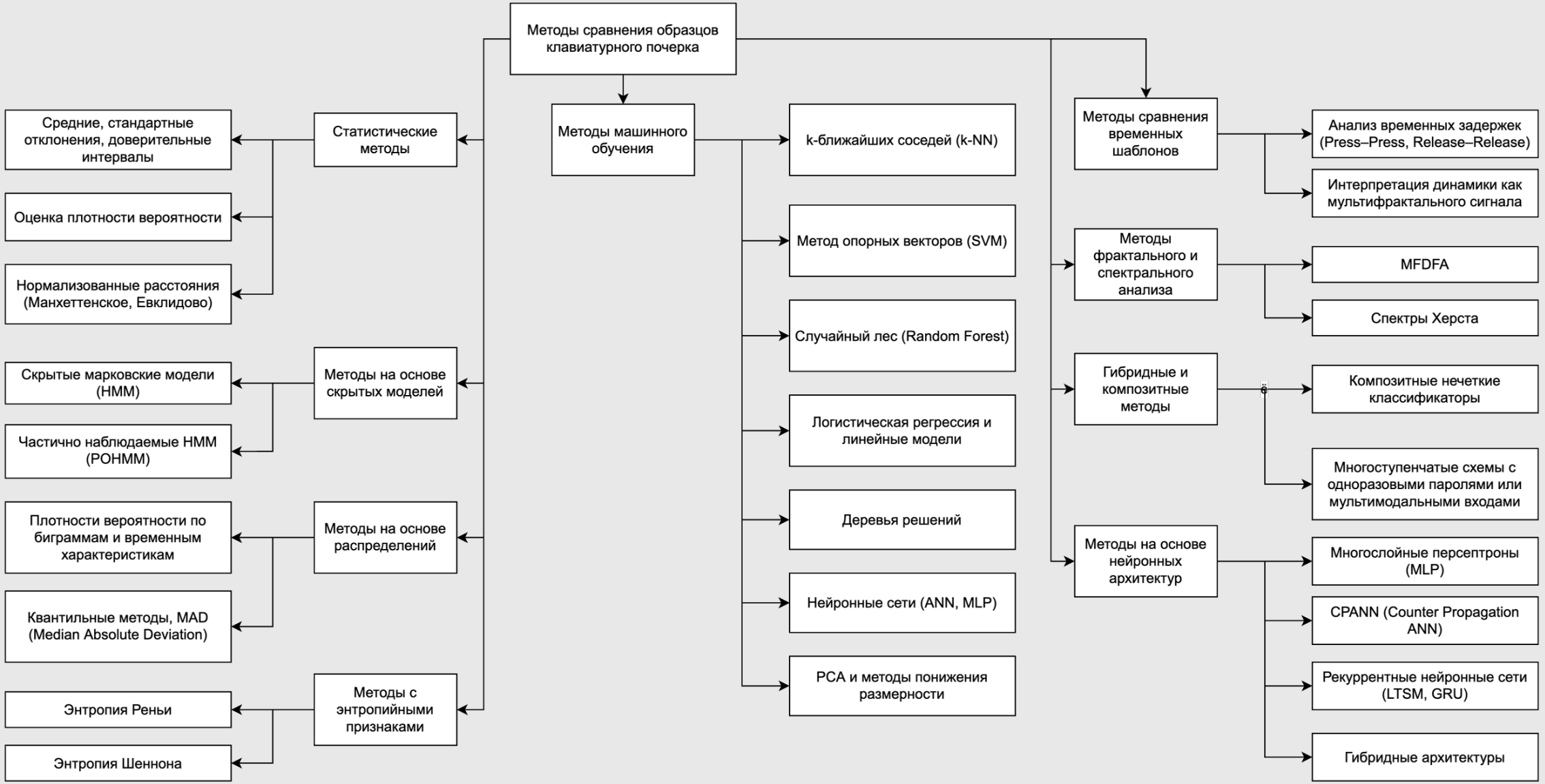
1. Алгоритм формирования списка слов с заданным распределением биграмм.
2. Математическая модель клавиатурного почерка пользователя ИС.
3. Метод биометрической идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм.
4. Алгоритм функционирования программно-аппаратной подсистемы биометрической идентификации пользователей ИС.

Научная новизна

1. Разработан новый, не имеющий аналогов среди известных публикаций, алгоритм формирования списка слов с заданным распределением биграмм на русском языке, обеспечивающий функционирование подсистемы сбора биометрических данных БСОВ.
2. Впервые описана математическая модель клавиатурного почерка как совокупность функций плотности распределения времени ввода всех уникальных биграмм исходного текста.
3. Разработан новый метод биометрической идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм, отличающийся использованием для оценки эффективности комбинированной метрики сходства распределений.
4. Построена новая архитектура функционирования программно-аппаратной биометрической системы идентификации на основе анализа клавиатурного почерка в локальном и онлайн-режимах, реализующая предыдущие положения.

Актуальное состояние исследований клавиатурного почерка

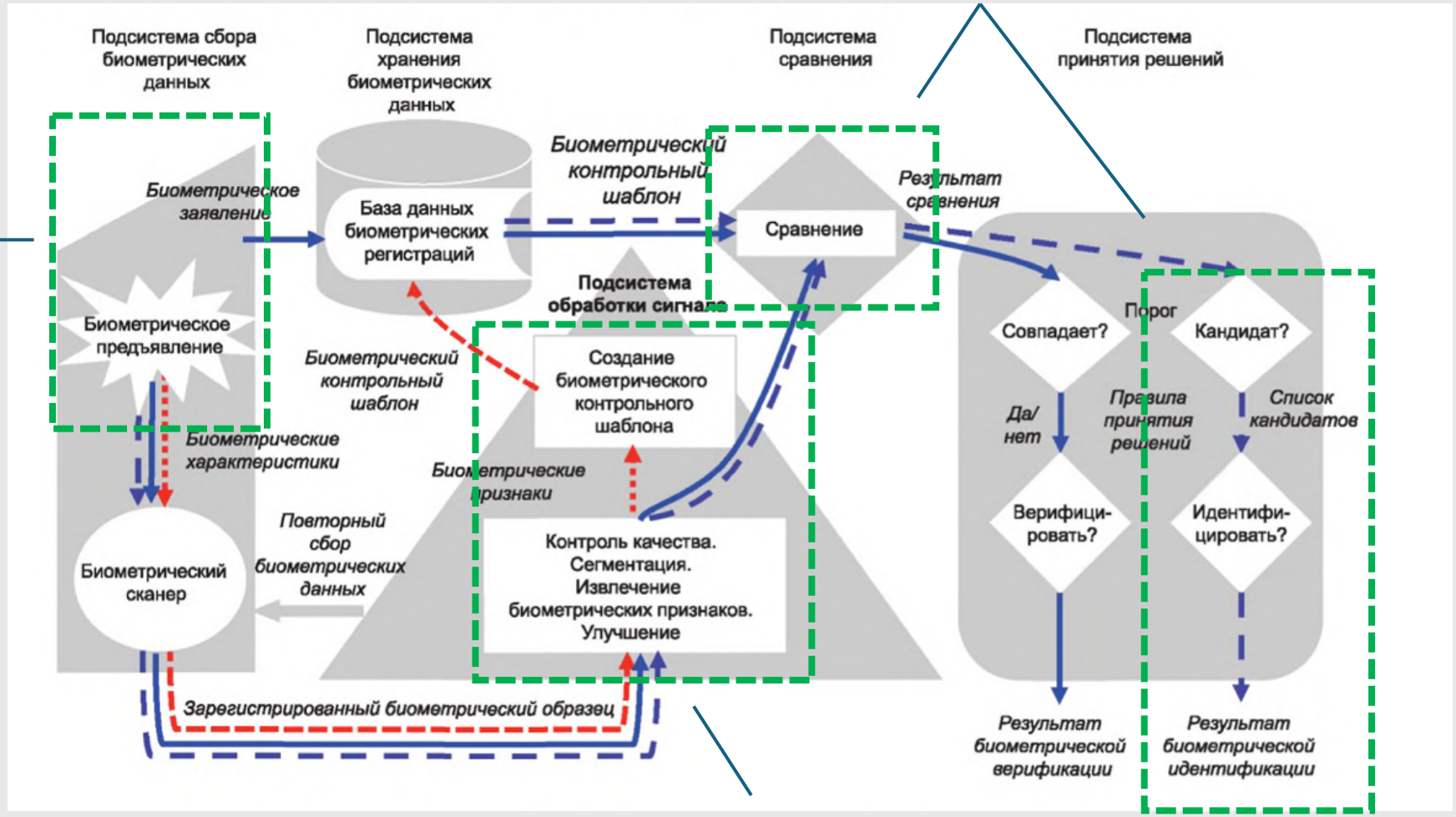
173 статьи с 1998 по 2025 г. из рецензируемых изданий Scopus, WoS, ВАК и РИНЦ.



Взаимосвязь положений в рамках биометрической системы общего вида

Положение 3

Положение 1



Положение 2

Положение 4 реализует все блоки по положениям 1, 2 и 3 в рамках устройств.

1. Алгоритм формирования списка слов с заданным распределением биграмм

Обоснование задачи

Алгоритм необходим для обеспечения функционирования подсистемы сбора биометрических данных при идентификации по клавиатурному почерку на основе ввода свободного текста.

Требования к алгоритму

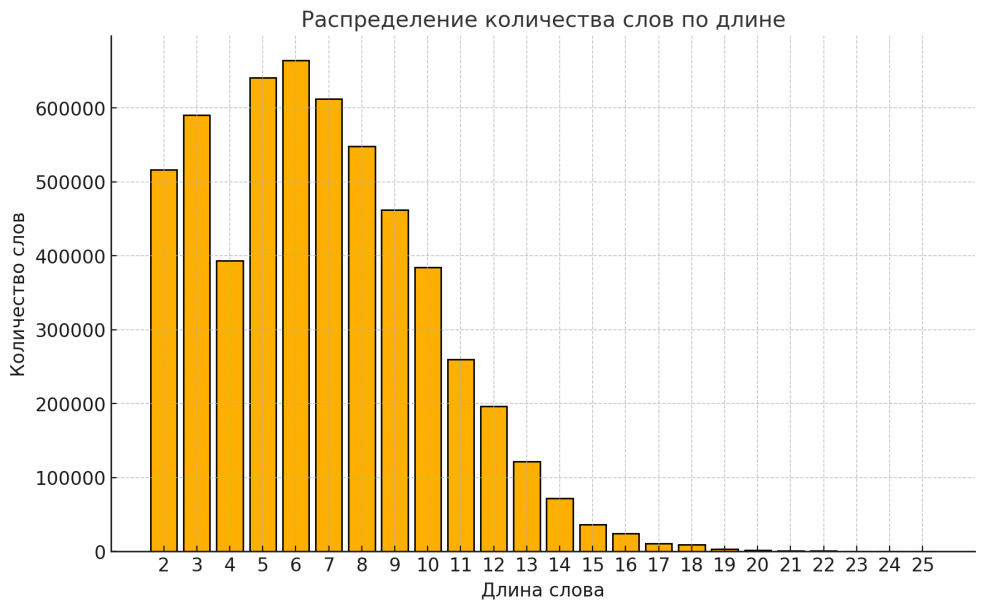
1. **Вариативность и осмысленность списка слов.** Слова должны быть знакомы и реально существовать. Требуется сбор словаря.
2. **Время работы.** Не более 1 секунды при идентификации.
3. **Качество генерации.** Минимум 90% от требуемого распределения биграмм.
4. **Время на набор текста.** Не более 1 минуты, то есть не более 220 символов (средняя скорость печати)

Требование 1: вариативность и осмысленность

Собран словарь **из 209 311 слов** и словоформ, сопоставимо с объемом словаря Даля.

Сплошная выборка: 10 000 статей с «Хабр» с октября 2024 по январь 2025.

950 уникальных биграмм из 1089 возможных.



Распределение количества слов по длине

ро, ов, но, ра, ст, ен, ва, ер, ан, ре, ни, ко, ны, пр, те, по, ал, на, ат, ли, ос, ом, ти, ор, то, ри, нн, та, не, ка, ми, ин, ет, ес.

34 биграммы из 950 составляют 50% всех вхождений биграмм в собранном корпусе

Алгоритм формирования списка слов с заданным распределением биграмм

Входные данные

$$\left(B_{main},\ \{B_1,\ B_2,\ \ldots,\ B_k\},\ D,\ N,\ L_{min},\ L_{max} \right)$$

$\mathbf{B_{main}}$— основная биграмма, обеспечивает хотя бы одно статистически значимое распределение.

$\{\mathbf{B_1},\ \mathbf{B_2}, \ldots,\ \mathbf{B_k}\}$ — набор дополнительных целевых биграмм, где k – количество целевых биграмм**.** Набор $\{B_i\}$ формируется из 34 наиболее значимых биграмм, выявленных ранее.

$\boldsymbol{D}$ **—** словарь**.** Список, который используется для выбора подходящих слов.

$N \in \mathbb{N}$ — количество слов в выходном наборе. Задается с учетом требований биометрической системы и баланса между точностью и удобством. В рамках эксперимента $10 \leq N \leq 40$.

$\boldsymbol{L_{min}}$ **и** $\boldsymbol{L_{max}}$ — минимальная и максимальная длина слов. $L_{min} \leq |w| \leq L_{max}$, где $|w|$ – длина конкретного слова в символах. В словаре D_{habr} $L_{min} = 5$ и $L_{max} = 7$.

Шаг 1. Подготовка данных

$D_{\text{filt}} = \{w \in D \mid L_{min} \leq |w| \leq L_{max}\}$
Разбиваем Dfilt на 3 непересекающиеся группы:
$G_1 = \{w \in D_{\text{filt}} \mid B_{\text{main}} \in w \wedge \forall B_i \notin w\}$
$G_2 = \{w \in D_{\text{filt}} \mid B_{\text{main}} \in w \wedge \exists B_i \in w\}$
$G_3 = \{w \in D_{\text{filt}} \mid B_{\text{main}} \in w \wedge \forall i\ B_i \in w\}$

Шаг 2. Отбор кандидатов по целевым и основным биграммам

Для каждой $B_i \in \{B_1, B_2, \ldots, B_k\}$ формируем

$Q_i = \{w \in D_{\text{filt}} \mid B_{\text{main}} \in w \wedge B_i \in w\}$

$C = \bigcup_{i=1}^{k} Q_i$ — объединенный набор кандидатов

Шаг 3. Формирование предварительного списка и нормализация его размера

Если $|C| < N$, $mo\ C' \subseteq C$, $|C'| = N$
$C' = \text{RandomSample}(C, N)$ — случайное усечение

Шаг 4. Дозаполнение набора при недостаточном количестве слов

Если $|C| < N$, mo
1. Добавляем слова из G_3 в С, пока $|C| < N$
2. Если $|C| < N$, добавляем слова из G_1, пока $|C| < N$

Шаг 5. Определение сходства двух наборов слов

Параметры эталонного набора

$C^{ref} = (c_1^{ref}, c_2^{ref}, \ldots, c_n^{ref})$ — массив частот биграмм
$P_i = \frac{c_i^{ref}}{\Sigma_{j=1}^{n} c_j^{ref}}$ — нормированные распределения частот
$P = (P_1, P_2, \ldots, P_n)$ — нормированный вектор частот

Параметры тестируемого набора

$C^{test} = (c_1^{test}, c_2^{test}, \ldots, c_n^{test})$ — массив частот биграмм
$Q_i = \frac{c_i^{test}}{\Sigma_{j=1}^{n} c_j^{test}}$ —нормированные распределения частот
$Q = (Q_1, Q_2, \ldots, Q_n)$ — нормированный вектор частот

$D(P, Q) = \sqrt{\Sigma_{i=1}^{n} (P_i - Q_i)^2}$ — евклидово расстояние
$S = (1 - D(P, Q)) * 100\%$ — сходство наборов в процентах

Алгоритм итеративной генерации

Основная задача – найти «точку стабилизации», которая отражает момент, после которого увеличение числа попыток генерации перестает оказывать существенное влияние на итоговое качество результата.

Входные параметры

NS – количество попыток генерации на текущем шаге.

MAX_{NUM_SETS} – максимально допустимое количество генераций.

S_i – значение метрики сходства между i-м сгенерированным набором данных и эталонным распределением биграмм.

$\mathrm{f}(NS) = \max\{S_1, S_2, \ldots, S_{NS}\}$ – максимальное значение метрики сходства для NS генераций.

$\overline{f}(k) = \frac{1}{w}\sum_{i=k-w+1}^{k} f(i)$ — сглаживание методом скользящего среднего., k номер текущего, w — ширина окна

$\varepsilon = 1\%\ (0.01)$— порог стабилизации, SW — длина интервала проверки стабильности метрики.

Условие стабилизации выполняется, если на интервале длиной SW, начиная с позиции $k+1$, значение сглаженной функции $\overline{f}(k)$ находится в границах диапазона:

$\overline{f}(m) \in \left[\overline{f}(k) - \varepsilon,\ \overline{f}(k) + \varepsilon\right]$

Для всех $m \in \left[k+1, k+SW\right]$

Тогда

$k_{stab} = k$ — точка стабилизации

$NS^* = k_{stab}$ — оптимальное количество наборов

Поиск точки стабилизации

Итоговый алгоритм поиска точки стабилизации:

1. Установить $NS = 1$.
2. Сгенерировать NS наборов данных D_i, где $i = 1, \ldots, NS$.
3. Вычислить значения сходства S_i каждого набора данных с эталонным распределением.
4. Найти максимум:
   $f(NS) = \max\{S_1, S_2, \ldots, S_{NS}\}\#$

1. Построить сглаженную функцию $\overline{f}(NS)$.
2. Проверить условие стабилизации:

   $\forall m \in [NS+1, NS+SW] : \overline{f}(m) \in [\overline{f}(NS) - \varepsilon, \overline{f}(NS) + \varepsilon]\#$
3. При выполнении условия зафиксировать $NS^* = NS$.

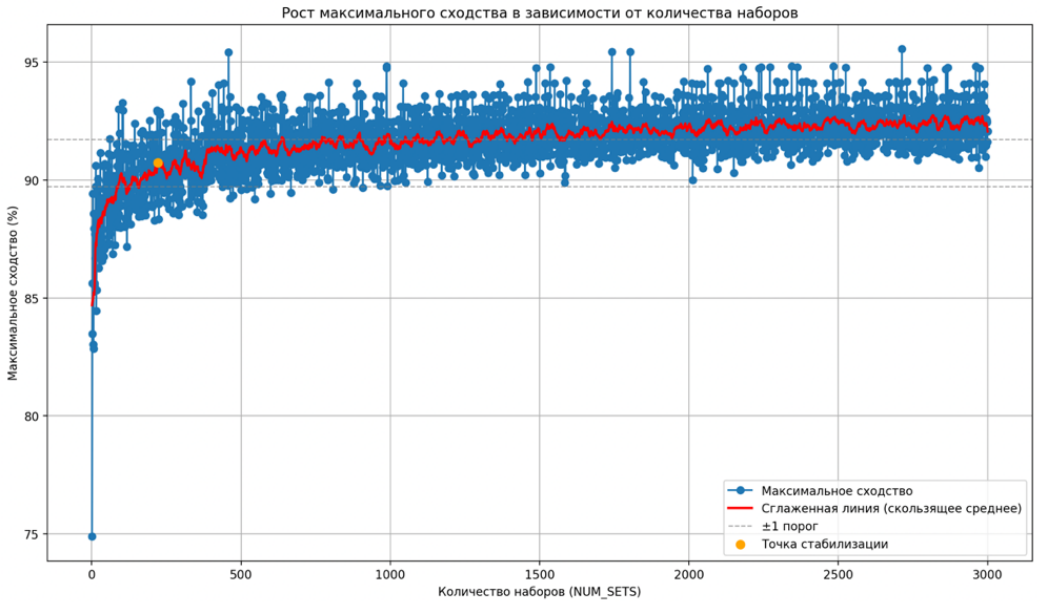
Параметры эксперимента

$NS = 3000$

$\varepsilon = 1\%$

$SW = 400$

Выявлено, что $k_{stab} = 221$



Значение максимального сходства распределений в зависимости от количества наборов

Критерий оптимальности полученного набора слов

$L_{min} \in [4; 21]$ – минимальная длина слова
$L_{max} \in [L_{min}; 21]$ – максимальная длина слова
$tw \in$ [10; 40] — количество слов
$k_{stab} = 221$ попытка генерации

$$S_{max} = \max_{i=1,\ldots,221} S_i$$

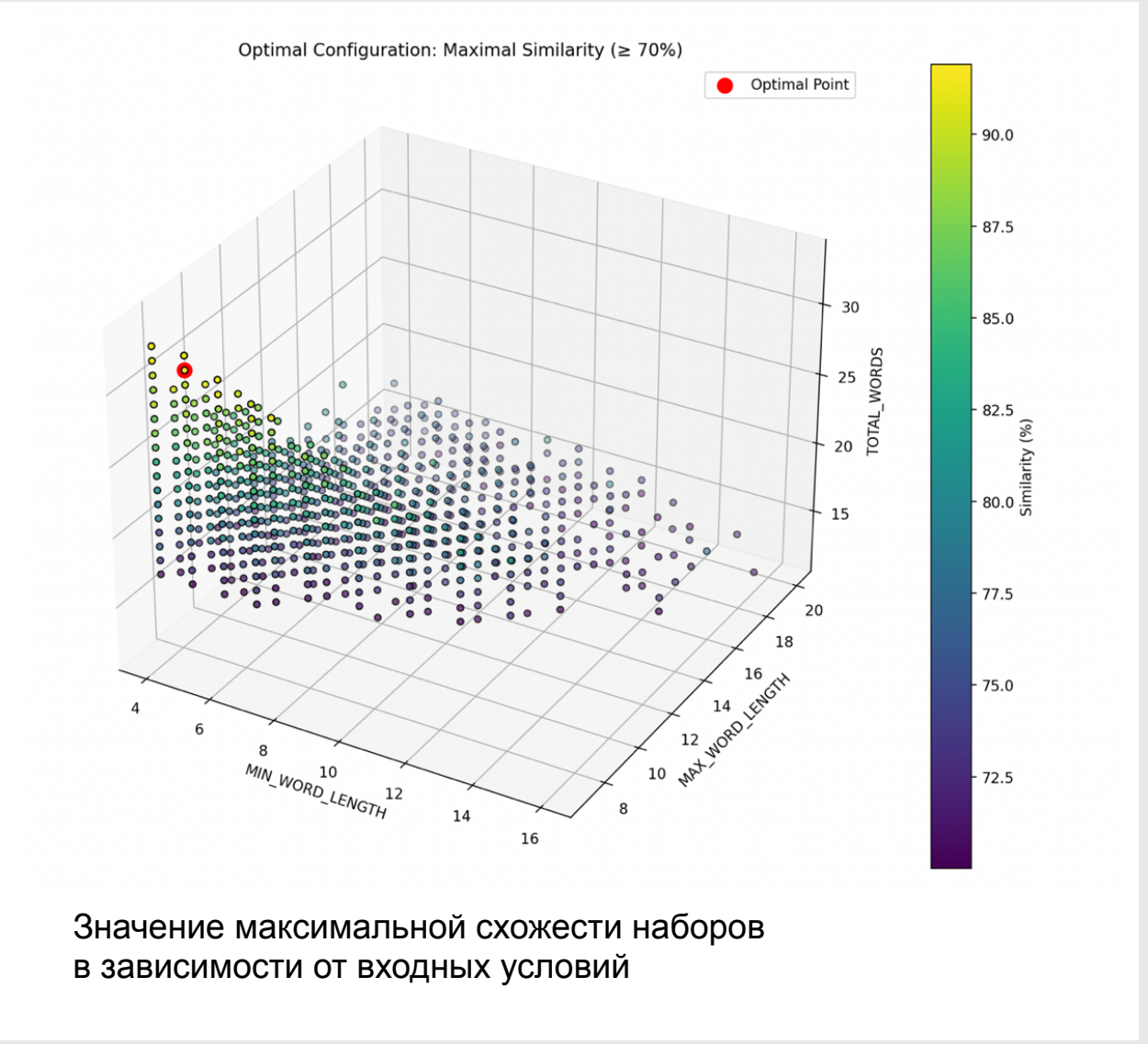
В дальнейшем анализируются только те конфигурации параметров, для которых одновременно выполняются условия:

$S_{max} \geq 70\%$

$$TC = tw \times \frac{L_{min} + L_{max}}{2} \leq 220$$

где TC – оценка общего количества символов в наборе.

Минимальная длина	5 символов
Максимальная длина	7 символов
Общее количество слов в наборе	32
Максимальное сходство	91.9%
Символов в последовательности	213
Приблизительное время печати	58.09 с.



Значение максимальной схожести наборов в зависимости от входных условий

Проверка алгоритма на соответствие требованиям

Параметры генерации тестового набора

Корневая биграмма «ст»
«ор» — 15, «ма» — 10, «ер» — 5, «ко» — 5, «те» — 2, «но» — 2, «ва» — 2.

Сгенерирован тестовый набор из 32 слов от 5 до 7 букв длиной 204 символа, соответствующий заданным характеристикам.

Оценка результативности алгоритма

Критерий	Требование алгоритма	Результат
Время генерации набора	1,0 с.	**0,34 с**
Время печати текста	60 с.	**55,5 с.**
Сходство с исходным распределением	90%	**94.18%**

Распределение частоты биграмм в тестовом наборе

Научная новизна

Разработан новый, не имеющий аналогов среди известных публикаций, алгоритм формирования списка слов с заданным распределением биграмм на русском языке, обеспечивающий функционирование подсистемы сбора биометрических данных БСОВ.

Практическая значимость

Разработанный алгоритм может использоваться в системах идентификации на основе анализа клавиатурного почерка, имеет возможность адаптации для текста на любом языке.

Соответствие паспорту специальности 2.3.6

1. Теория и методология обеспечения информационной безопасности и защиты информации.
12. Технологии идентификации и аутентификации пользователей и субъектов информационных процессов. Системы разграничения доступа.
15. Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности.

Публикации по теме исследования

1. Шкляр, Е. (2025). АЛГОРИТМ ФОРМИРОВАНИЯ СПИСКА СЛОВ С ЗАДАННЫМ РАСПРЕДЕЛЕНИЕМ БИГРАММ ДЛЯ РЕГИСТРАЦИИ БИОМЕТРИЧЕСКИХ КОНТРОЛЬНЫХ ШАБЛОНОВ КЛАВИАТУРНОГО ПОЧЕРКА. *Безопасность информационных технологий, 32*(3), 74-89. doi: http://dx.doi.org/10.26583/bit.2025.3.06
2. Шкляр Е.В. Формирование и оптимизация массива биграмм для задач распознавания клавиатурного почерка. (Измерение, контроль, информатизация)
3. Свидетельство о государственной регистрации программы для ЭВМ № 2025616669 «Программа для формирования списка слов с заданным распределением биграмм»

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025616669

Программа для формирования списка слов с заданным распределением биграмм

Правообладатель: *Шкляр Евгений Вадимович (RU)*

Автор(ы): *Шкляр Евгений Вадимович (RU)*

Заявка № 2025615055

Дата поступления 11 марта 2025 г.

Дата государственной регистрации в Реестре программ для ЭВМ *19 марта 2025 г.*

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат 0692e7c1a6300bf54f240f670bca2026

Владелец Зубов Юрий Сергеевич

Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

2. Математическая модель клавиатурного почерка пользователя ИС

«Динамика работы на клавиатуре является биометрической технологией, построенной на анализе ритма печати.

Динамика работы на клавиатуре человека развивается со временем, так как он учится печатать на клавиатуре, тем самым развивая уникальные навыки печати». *ГОСТ Р 54412-2019*

Клавиатурный почерк — индивидуальный способ печати на клавиатуре, уникальный у разных людей, появившийся в результате **формирования привычки,** навыка.

Предложенное определение клавиатурного почерка

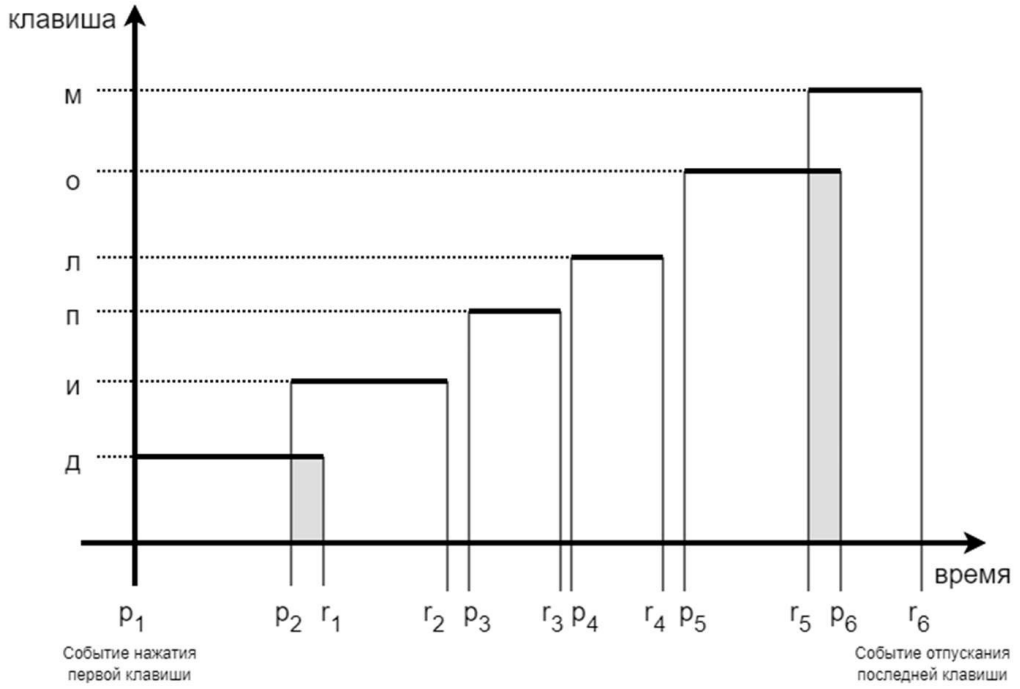


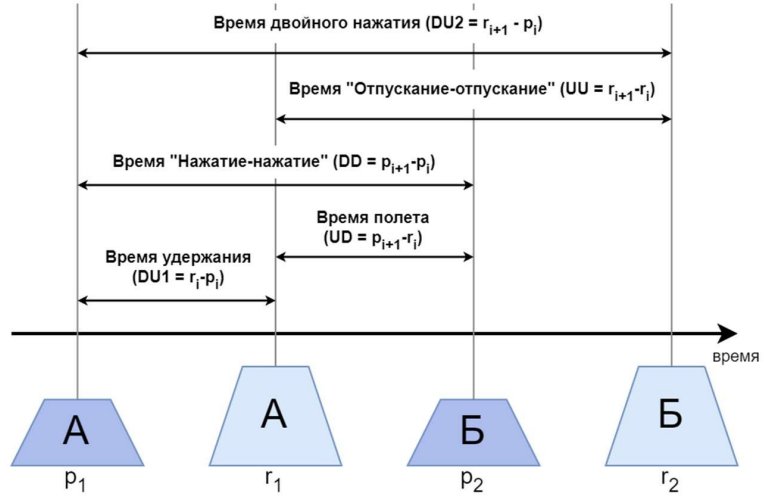
Схема событий клавиатурного почерка

Из определения следует потенциальная возможность проведения идентификации на основе анализа клавиатурного почерка.

Биометрический контрольный шаблон — один или более хранимых биометрических образцов, биометрических шаблонов или биометрических моделей, относящихся к субъекту биометрических данных и используемых в качестве объекта сравнения.

Существующие метрики клавиатурного почерка

Группа метрик	Альтернативные названия
Время удержания клавиши	Hold Time, Dwell Time, Event Duration
Пауза между клавишами	Flight Time, Latency, Up-Down Time, Inter-Key Interval
Интервал между событиями	Down-Down Time, Up-Up Time
Структура ввода	n-Graph Dynamics, Timestamp Sequences
Опечатки и их исправление	Error Rate, Backspace Frequency
Пространственные признаки	Key Distance (DEFT), Key Code / Scan Code
Скорость и ритм	Typing Speed, Pause Duration
Перекрытия во времени	Overlap Ratio
Дополнительные источники данных	IMU Signals (Accel, Gyro)



Существующие способы представления КП

Вектор признаков с фиксированными временным метриками

$$x = \left[H_1,\ F_{1,2},\ H_2,\ F_{2,3},\ \ldots,\ H_n\right]$$

$\boldsymbol{x} \in \mathbb{R}^{2n-1}$ — вектор длины $2n-1$ для строки из n символов;
$\boldsymbol{H_i}$ – время удержания клавиши i
$\boldsymbol{F_{i,i+1}}$ — время между отпусканием клавиши i и нажатием клавиши $i+1$,

Временные ряды

$$S = \left\{ \left(k_i,\ t_i^{down}, t_i^{up} \right) \right\}_{i=1}^{n}$$

$\boldsymbol{k_i}$ — код нажатой клавиши на позиции i;
$\boldsymbol{t_i^{down}}$— момент времени нажатия клавиши k_i;
$\boldsymbol{t_i^{up}}$— момент времени отпускания клавиши k_i;
$\boldsymbol{n}$ — общее число символов в строке;

N-графы (биграммы, триграммы)

Для последовательности

$$K = \left(k_1, k_2, k_3, \ldots,\ k_n\right)$$

$$G^{(2)} = \left\{ \left(k_1, k_2\right), \left(k_2, k_3\right), \ldots, \left(k_{n-1}, k_n\right) \right\}$$

Для каждой пары
H_i — hold time клавиши k_i;
$F_{i,i+1}$ — flight time между клавишами k_i и k_{i+1};

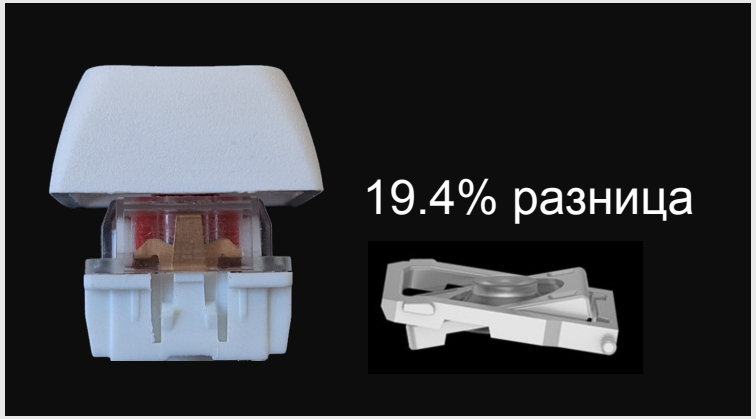
Требования к модели клавиатурного почерка

Независимость от способа сравнения. Результат — структура, пригодная для различных методов сопоставления.

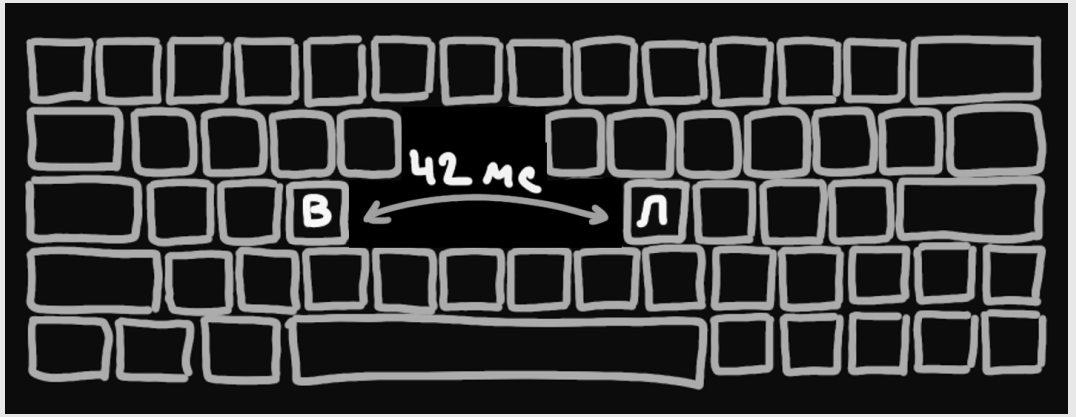
Работа со свободным текстом на любом языке, извлечение признаков вне зависимости от лексического состава или алфавита вводимой последовательности.

Небольшой размер шаблона по сравнению с существующими решениями.

Обоснование выбора решения



Время нажатия на клавишу зависит от типа клавиатуры



Время между нажатиями клавиш отражает привычку

Формальное описание модели

1. Входные данные

$S = (c_i,\ c_j),\ t_{ij} \mid i, j \in \{1, N\},\ t_{ij} \in \mathbb{R}^+$, где

(c_i, c_j) – биграммы,

t_{ij} – временной интервал между нажатиями (c_i, c_j)

N – общее количество троек $(c_i,\ c_j),\ t_{ij}$

2. Выбор уникальных биграмм

Из S отбираются уникальные биграммы: $P = \left(c_i, c_j\right) \mid \exists t_{ij}$

$T_{ij} = t_k \mid \left(c_i, c_j, t_k\right) \in S$ – множество времен для уникальных пар биграмм.

3. Построение ядерной оценки плотности распределения

Для каждой пары $\left(c_i,\ c_j\right)$ на основе T_{ij} строится KDE:

$$\hat{f}_{ij}(t) = \frac{1}{\left|T_{ij}\right| h} \sum_{k=1}^{\left|T_{ij}\right|} K\left(\frac{t - t_k}{h}\right)$$

Где $K(\bullet)$ – ядро, например, Гауссово:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

h – ширина окна сглаживания,

$\left|T_{ij}\right|$ – количество временных интервалов для этой пары букв.

Клавиатурный почерк — совокупность функций плотности распределения для всех уникальных пар символов исходного текста

$$\mathcal{F} = \left\{ \hat{f}_{ij}(t) \;\middle|\; \left(\left(c_i, c_j\right) \in P\right) \right\}$$

Оценка модели на соответствие требованиям. Размер БКШ

- **Функциональное представление.** Возможна адаптация в векторный вид для маломощных компьютеров.
- **Контекстная чувствительность.** Возможен анализ ввода свободного текста.
- **Интерпретируемость.** Существует мат. аппарат для сравнения и интерпретации.
- **Соответствие ГОСТ.**
- **Устойчивость к вариативности и шуму за счет сглаживания.**

Способ формирования БКШ	**Объём БКШ**
Предложенная модель (на основе 8 биграмм)	**~1.6 КБ**
Классический вектор признаков (фиксированный текст)	~0.4–2 КБ
Нейросетевой эмбеддинг	~0.5–2 КБ
KDE по всем биграммам (без отбора)	12–40 КБ

Научная новизна

Впервые описана математическая модель клавиатурного почерка как совокупность функций плотности распределения времени ввода всех уникальных биграмм исходного текста.

Практическая значимость

Разработанная модель используется в методе идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм при сравнении образцов клавиатурного почерка

Соответствие паспорту специальности 2.3.6

1. Теория и методология обеспечения информационной безопасности и защиты информации.
12. Технологии идентификации и аутентификации пользователей и субъектов информационных процессов. Системы разграничения доступа.
15. Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности.

Публикации по теме исследования

Шкляр Е.В. Шульженко А.Д. Математическая модель биометрического контрольного шаблона клавиатурного почерка. *Моделирование, оптимизация и информационные технологии.* 2025;13(*). (принято к публикации)

3. Метод биометрической идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм

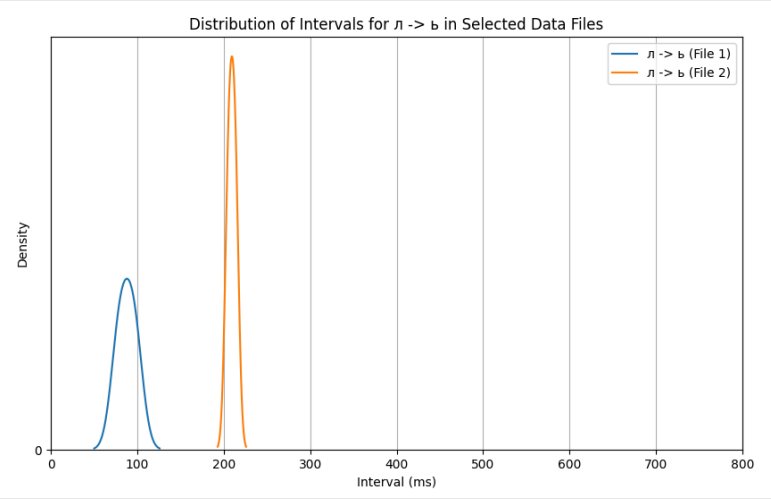
Биометрическая идентификация – автоматическое распознавание индивидов, основанное на их биологических и поведенческих характеристиках (ГОСТ ISO/IEC 2382-37).

Для биометрических систем проверка подлинности пользователей происходит за счёт **сравнения биометрической пробы и биометрического контрольного шаблона** на соответствие порогу. *ГОСТ Р 54412—2019*

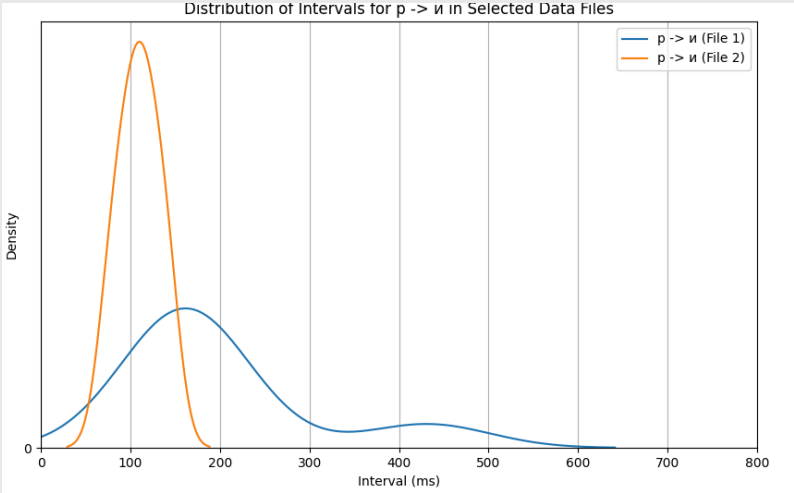
Биометрические характеристики не могут быть использованы при однофакторной аутентификации. Только совместно с другим фактором аутентификации. *ГОСТ Р 58833—2020*

Биометрический контрольный шаблон — один или более хранимых биометрических образцов, биометрических шаблонов или биометрических моделей, относящихся к субъекту биометрических данных и используемых в качестве объекта сравнения. **ГОСТ ISO/IEC 2382-37**

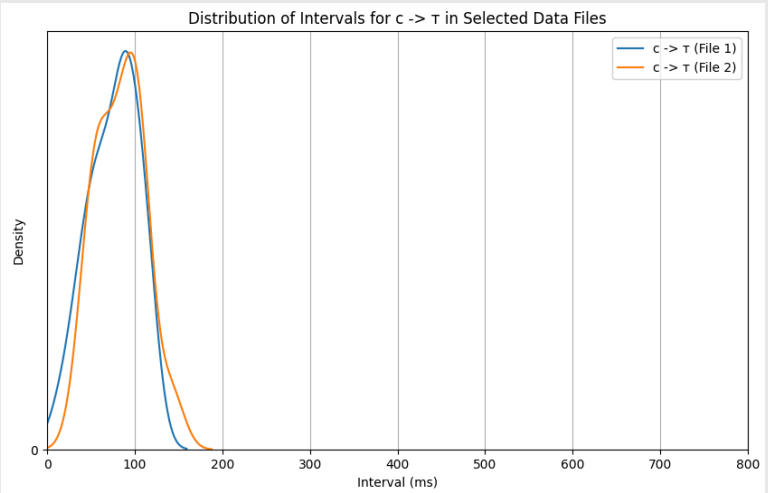
Основная идея предлагаемого решения



Разные люди, не пересекающиеся KDE



Разные люди, пересекающиеся KDE



Один человек, почти полное совпадение KDE

«Наивная» идея решения заключается в вычислении «схожести» распределений плотности вероятности времени набора биграмм.

Анализ эффективности существующих методов сравнения образцов КП

Исследование	EER, %	Точность, %	Датасет	Тип текста
Putra, Chowandra	0,9-12,7	–	собственный (10 пользователей)	индивидуальные пароли)
Senerath, Tharinda et al.	1,8	–	Aalto (168000)	оба (статический + динамический)
Moralez, Fierres et al.	1,85	-	Aalto (78000)	динамический
Dimaratos, Pohn	2,6	97,4	Lee, CMU, Buffalo	фиксированный и свободный
Wyciślik, Wylężek, Momot	2,65	–	Buffalo	фиксированный и свободный
Сулавко	2,65	–	31 / 71 / 32 пользователей	динамический
KVC-onGoing	3,33	–	KVC-onGoing	фиксированный и свободный
Kaluarachchi et al.	3,8	–	SU-AIS BB-MAS, Buffalo	оба (статический + динамический)
Chang, Li, Stamp	3,86	94,6	Buffalo (157 пользователей)	динамический
Acien, Morales	4,8	–	Aalto (168 000)	динамический
Ayotte, Hou	7,8	–	Buffalo (148) / Clarkson II (101)	динамический
Simao, Prado et al.	-	90,14	собственный (137 пользователей)	свободный

Требования к методу

EER < 2%
Точность > 98%
Сравнение на основе метрик расстояния
Работа на свободном тексте

Формализация метода сравнения БКШ клавиатурного почерка

Входные параметры для подсистемы сравнения — два БКШ из ОНР-2

$$\mathcal{F}^{(1)} = \left\{ \hat{f}_{ij}^{(1)}(t) \,\middle|\, \left(\left(c_i, c_j \right) \in P^{(1)} \right\} \right.$$

$$\mathcal{F}^{(2)} = \left\{ \hat{f}_{ij}^{(2)}(t) \,\middle|\, \left(\left(c_i, c_j \right) \in P^{(2)} \right\} \right.$$

Для каждой биграммы $\left(c_i, c_j \right)$ должны выполняться условия $\left| T_{ij}^{(1)} \right| \geq m$ и $\left| T_{ij}^{(2)} \right| \geq m$.

D_k— метрики сходства распределений.

Метрики сходства распределений

$$D_H\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \frac{1}{\left|P\right|} \sum_{\left(c_i, c_j\right) \in P} \left[\frac{1}{\sqrt{2}} \left(\int \left(\sqrt{\hat{f}_{ij}^{(1)}(t)} - \sqrt{\hat{f}_{ij}^{(2)}(t)} \right)^2 dt \right)^{1/2} \right]$$ — расстояние Хеллингера

$$\mathrm{D_{KS}}\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \frac{1}{\left|\mathrm{P}\right|} \sum_{\left(\mathrm{c_i}, \mathrm{c_j}\right) \in \mathrm{P}} \sup_{\mathrm{t}} \left| \mathrm{F}_{\mathrm{ij}}^{(1)}(\mathrm{t}) - \mathrm{F}_{\mathrm{ij}}^{(2)}(\mathrm{t}) \right|$$ — критерий согласия Колмогорова-Смирнова

$$D_{KL}\left(\mathcal{F}^{(1)} \middle| \mathcal{F}^{(2)}\right) = \frac{1}{\left|P\right|} \sum_{\left(c_i, c_j\right) \in P} \int \hat{f}_{ij}^{(1)}(t) \log \left(\frac{\hat{f}_{ij}^{(1)}(t)}{\hat{f}_{ij}^{(2)}(t)} \right) dt$$ — дивергенция Кульбака-Лейблера

$$D_W\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \frac{1}{\left|P\right|} \sum_{\left(c_i, c_j\right) \in P} \int \left| t_1 - t_2 \right| \cdot \pi_{ij}\left(t_1, t_2\right) dt_1 dt_2$$ — расстояние Вассерштейна

$W = w_{ij} \middle| \left(c_i, c_j \right) \in P$ — веса для каждой пары биграмм, назначаются извне.

$\tau_к$ — пороговые веса для каждой функции сходства

Формализация процесса сравнения

1. Фильтрация биграмм по числу вхождений. Наборы пересечений пар символов, удовлетворяющие условию:

$$P^* = \left(c_i, c_j\right) \mid \left(c_i, c_j\right) \in P^{(1)} \cap P^{(2)}, \left|T_{ij}^{(1)}\right| \geq m, \left|T_{ij}^{(2)}\right| \geq m$$

2. Вычисляется сходство каждой пары биграмм $\left(c_i, c_j\right) \in P^*$ на основе каждой d_k.

$$S_{ij}^{(k)} = d_k\left(\hat{f}_{ij}^{(1)}, \hat{f}_{ij}^{(2)}\right)$$

$S_{ij}^{(k)} \in [0,1]$ — нормализация значений вычисленных свойств

3. Нормализованные веса как доля веса каждой биграммы $\left(c_i, c_j\right)$ относительно суммарного веса всех биграмм, входящих в множество P^* (то есть в набор биграмм, используемый в сравнении).

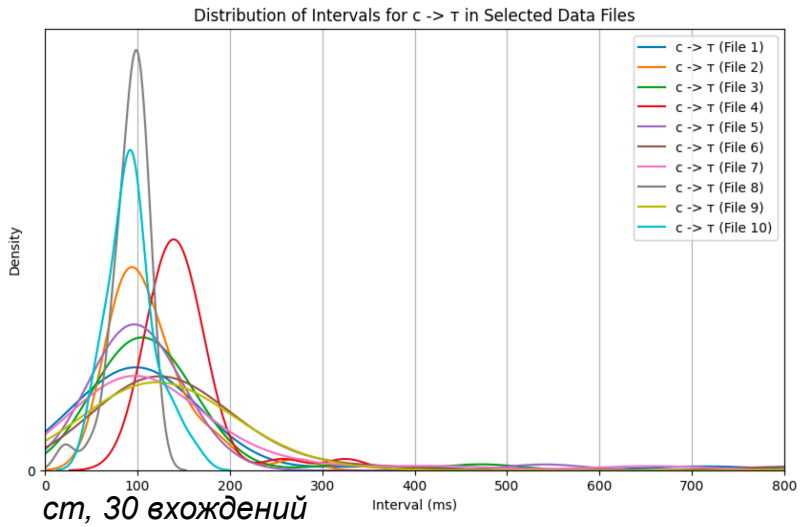
$$\omega_{ij}^{norm} = \frac{\omega_{ij}}{\sum_{\left(c_i, c_j\right) \in P^*} \omega_{ij}}$$

4. $S_{total} = \sum_{\left(c_i, c_j\right) \in P^*} \omega_{ij}^{norm} \sum_k \alpha_k S_{ij}^{(k)}$ — итоговое сходство с учетом нормализованных весов и функций.

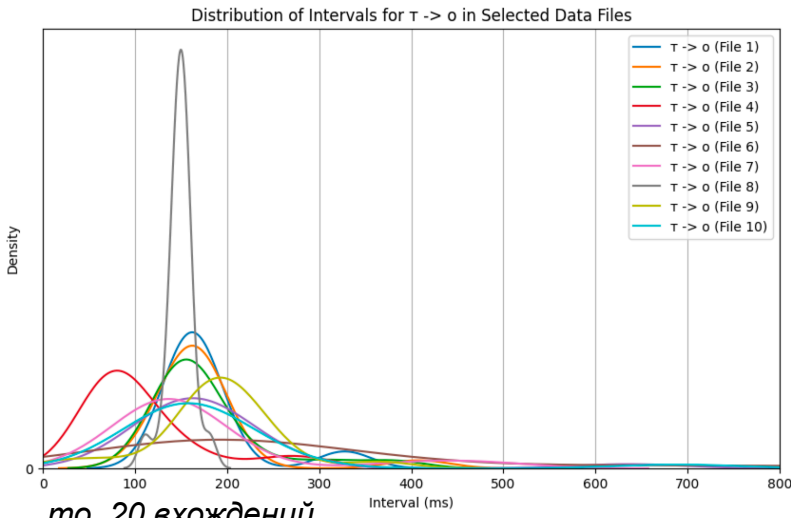
Где α_k – коэффициент влияния каждой функции сходства.

5. Если $S_{total} \geq \tau$, то считаем, что БКШ принадлежат одному человеку.

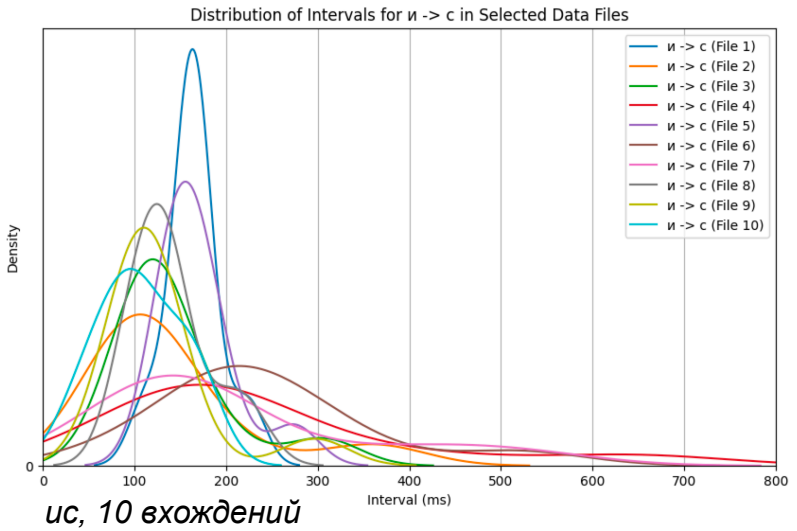
Плотности распределений времени набора биграмм. 10 разных пользователей



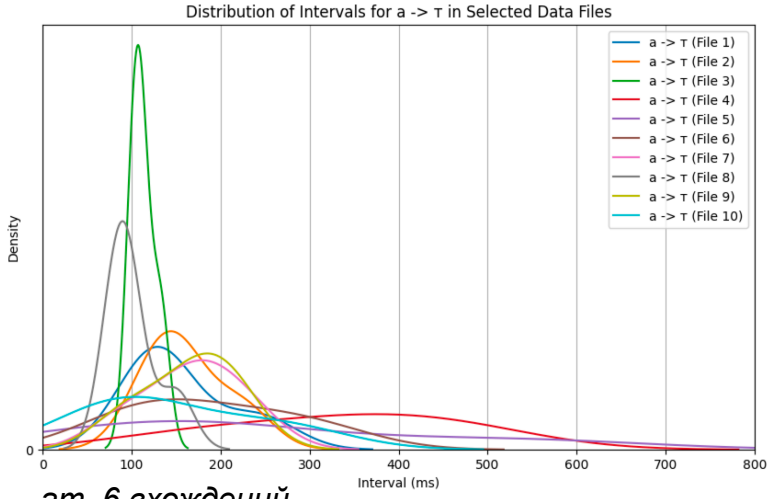
ст, 30 вхождений



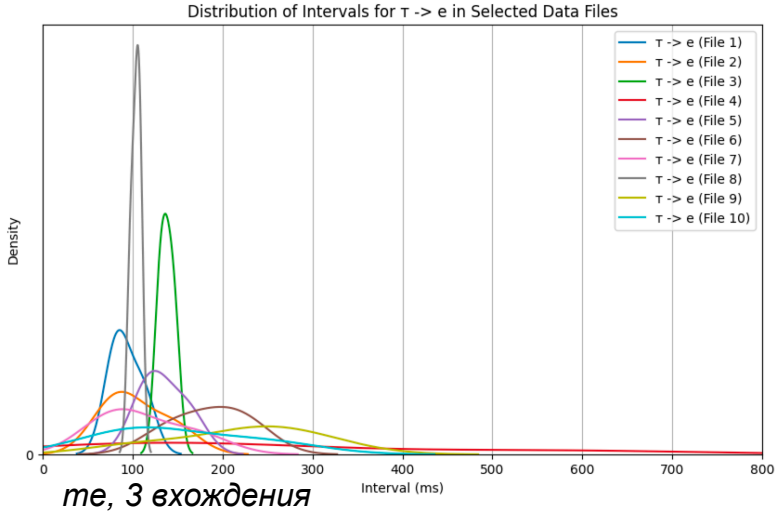
то, 20 вхождений



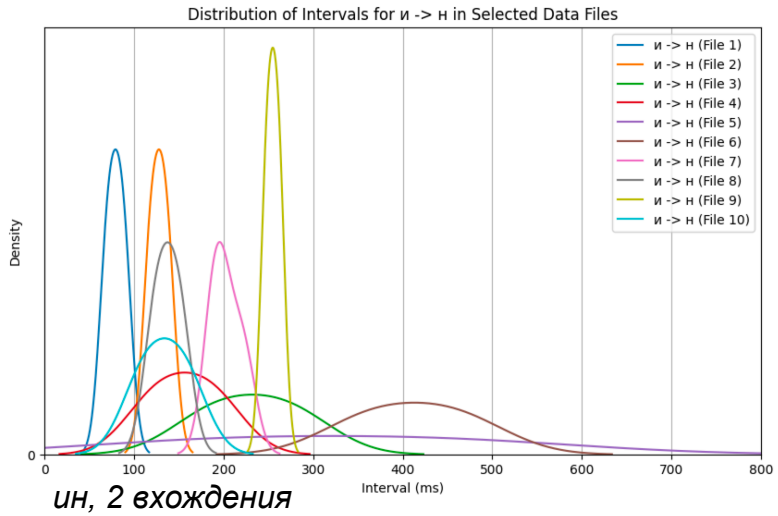
ис, 10 вхождений



ат, 6 вхождений

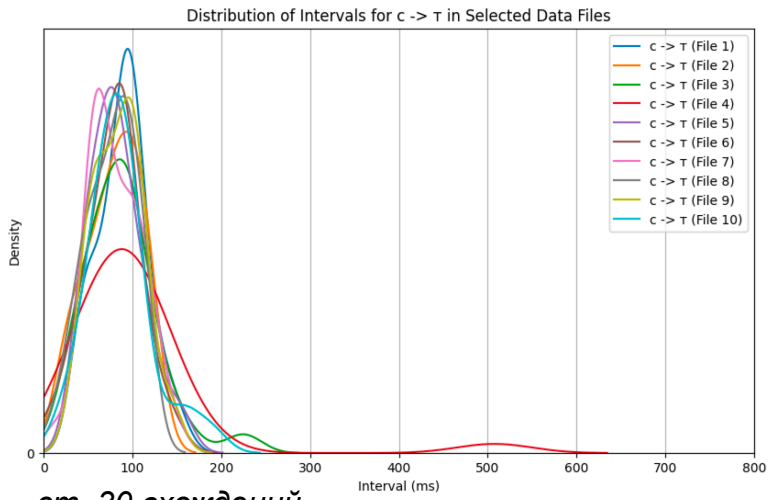


те, 3 вхождения

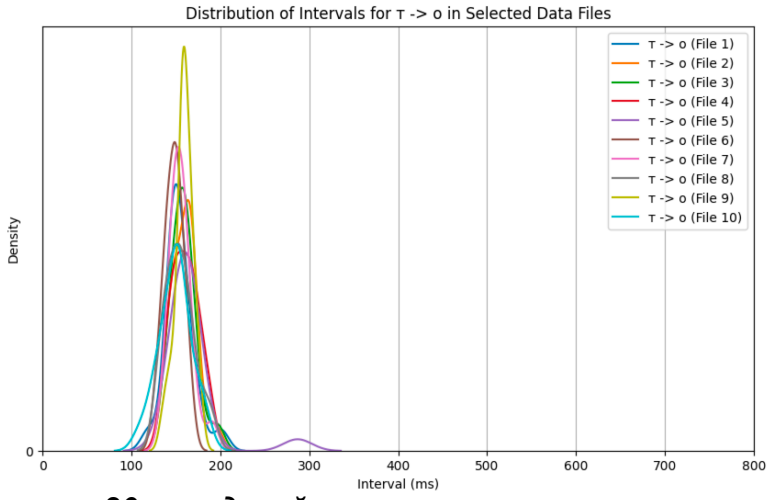


ин, 2 вхождения

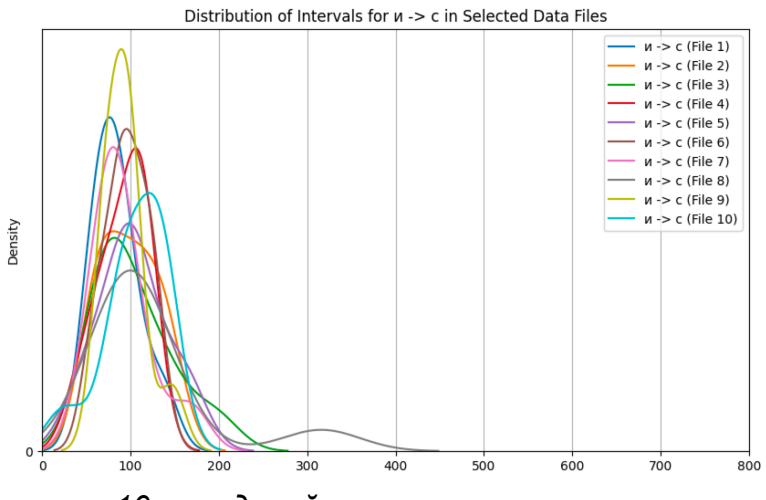
Плотности распределений времени набора биграмм. Один и тот же пользователь 10 раз



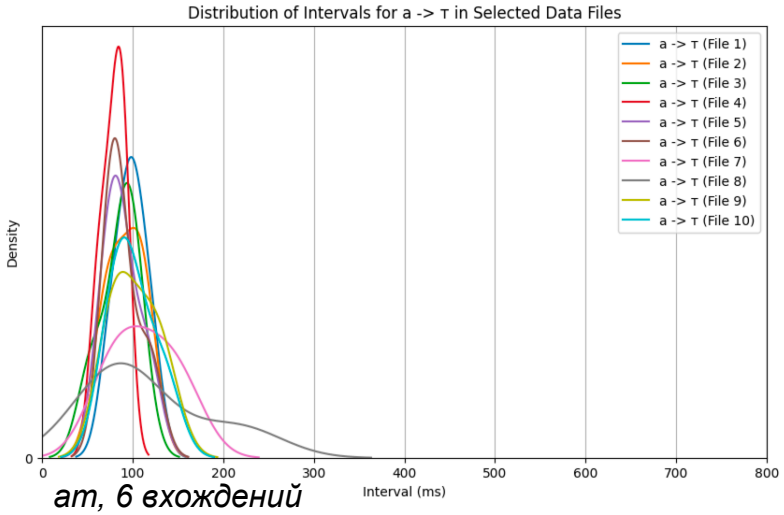
ст, 30 вхождений



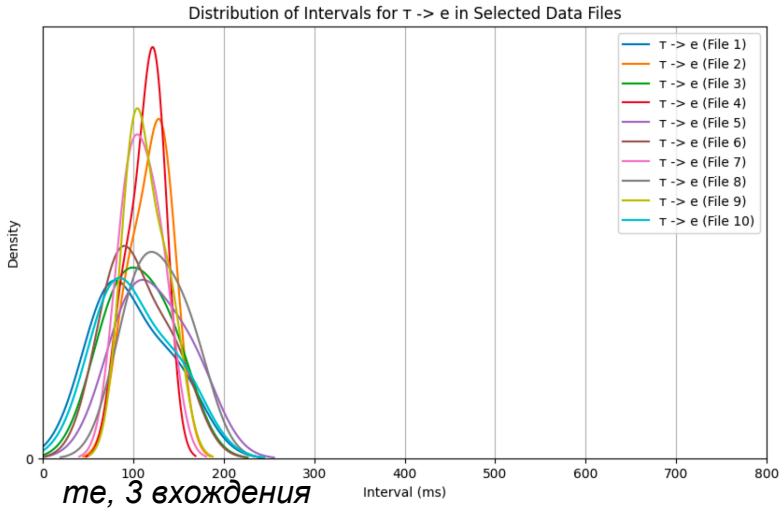
то, 20 вхождений



ис, 10 вхождений



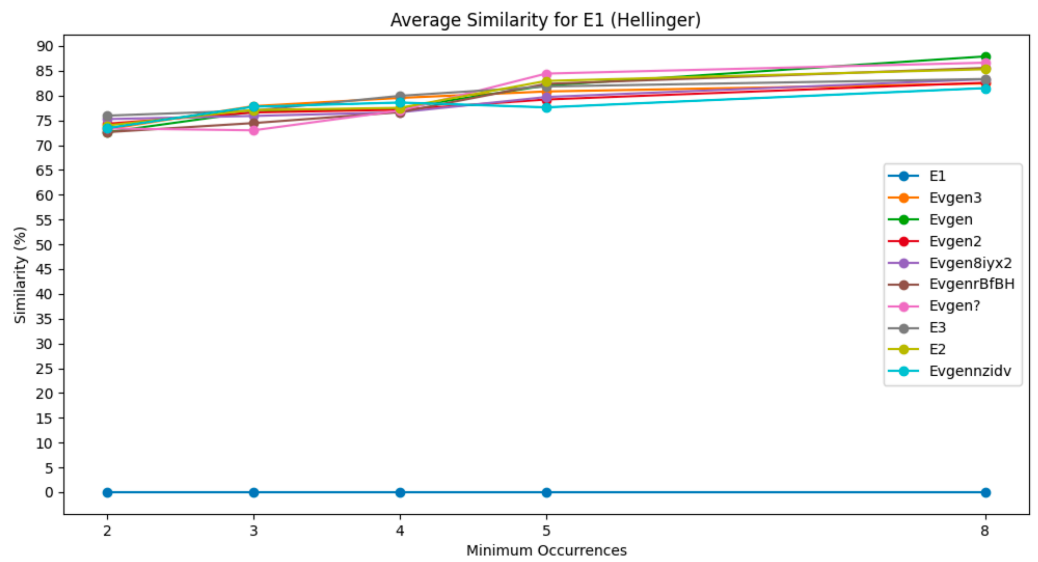
ат, 6 вхождений



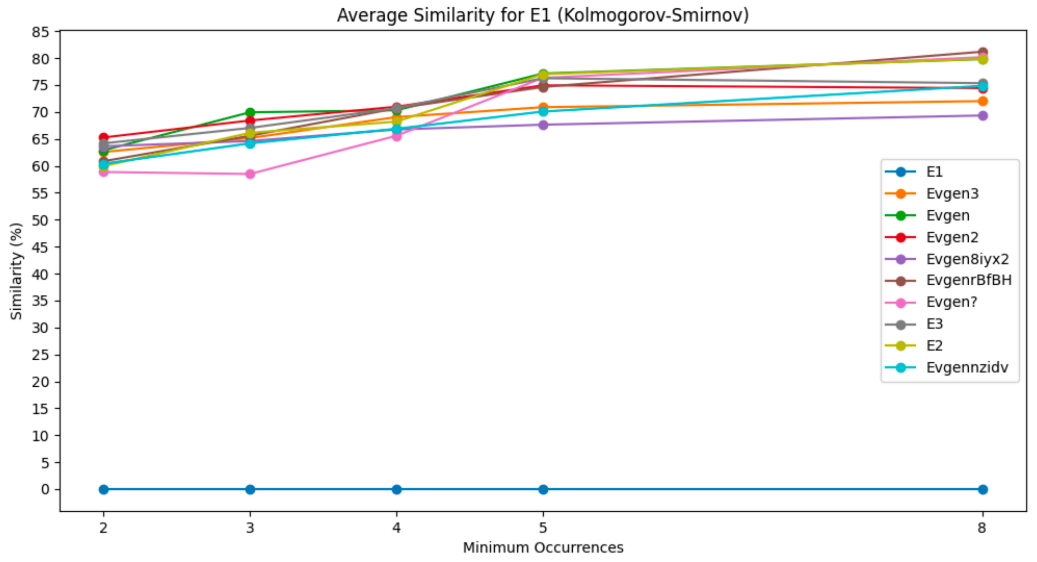
те, 3 вхождения

Различение пользователей происходит за счёт накопления разницы расстояний между распределениями

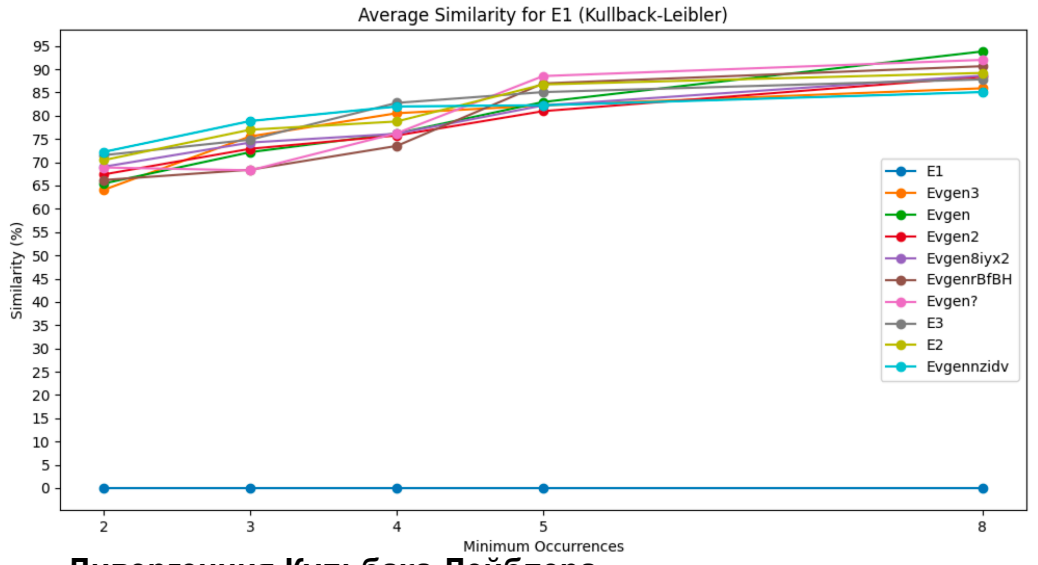
Среднее сходство в зависимости от количества вхождений биграмм (один пользователь)



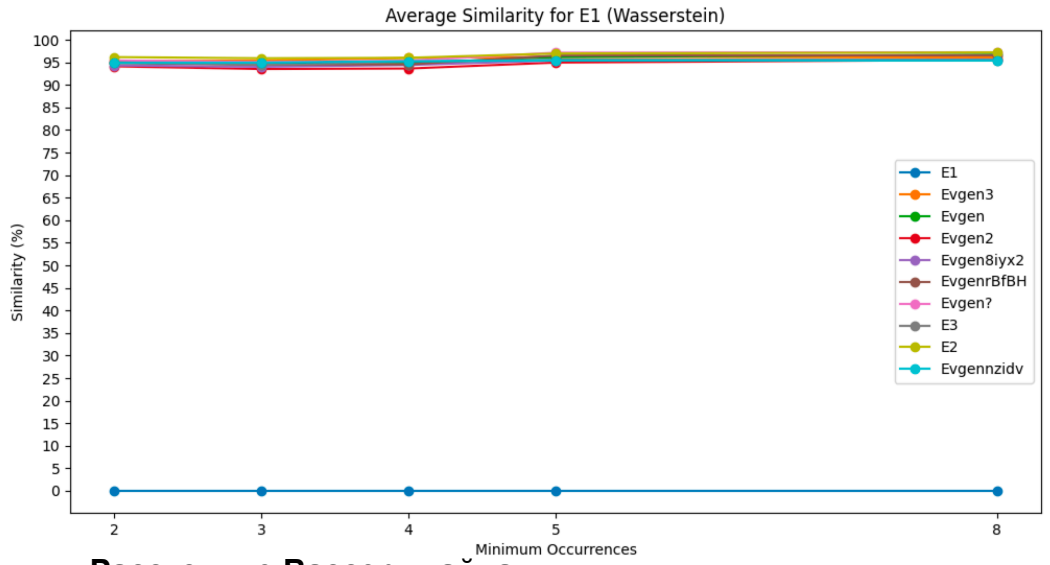
Расстояние Хеллингера



Критерий согласия Колмогорова-Смирнова

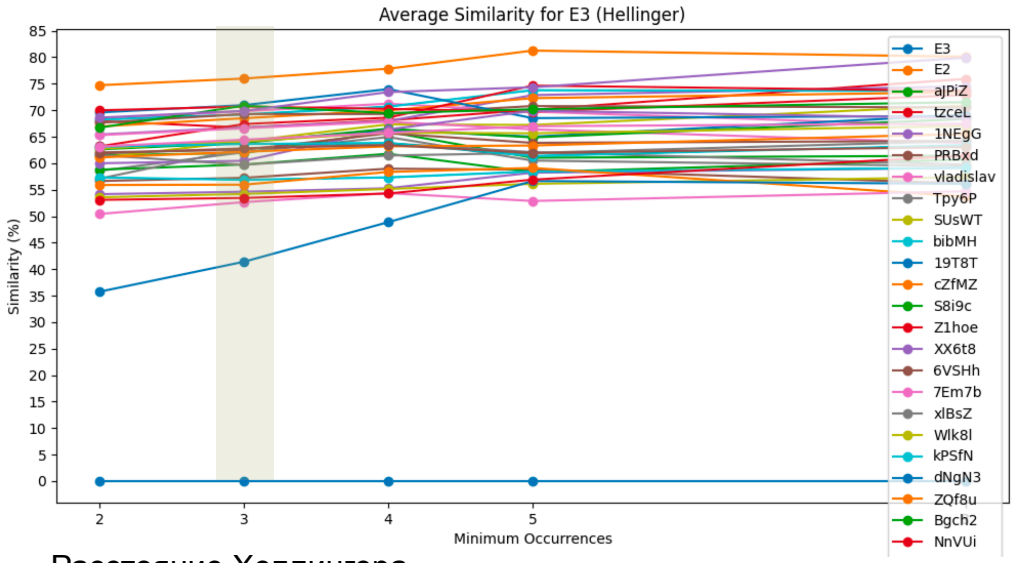


Дивергенция Кульбака-Лейблера

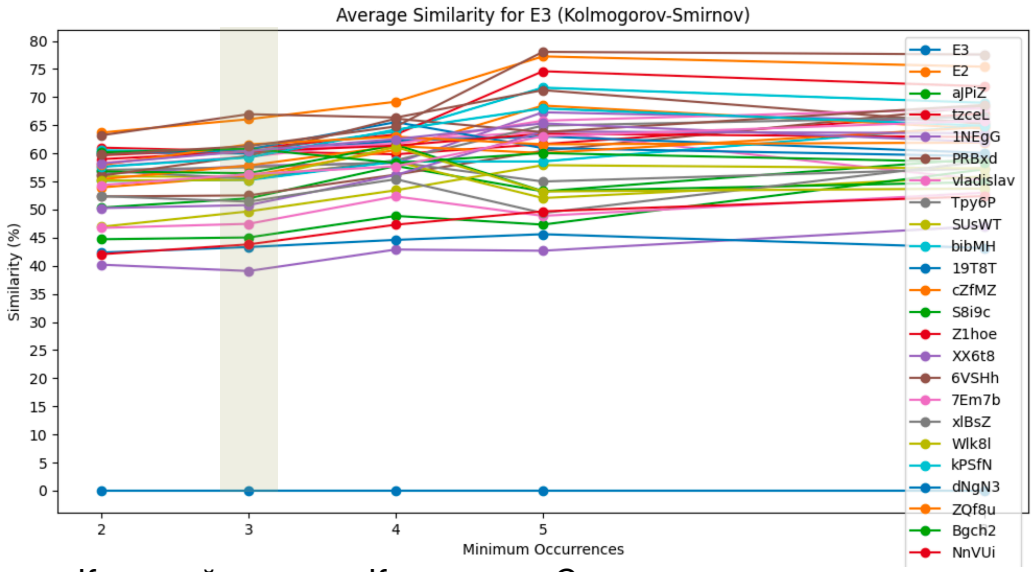


Расстояние Вассерштайна

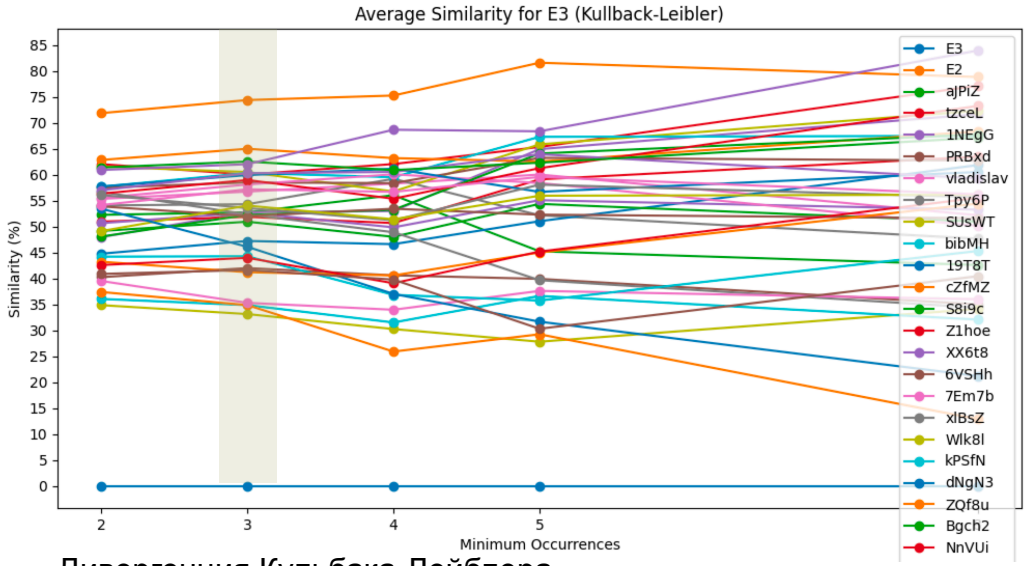
Среднее сходство в зависимости от количества биграмм (идентификация)



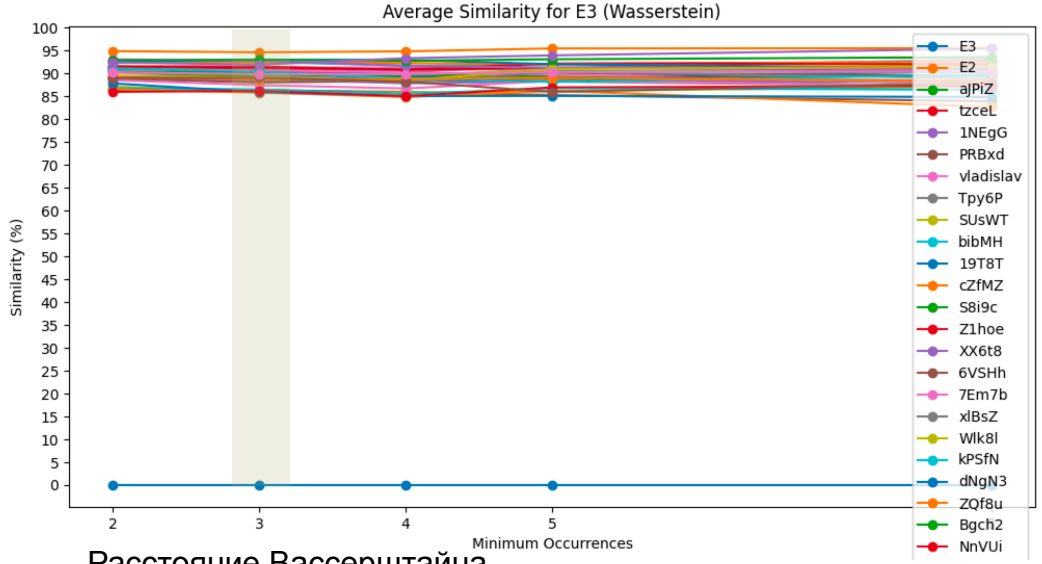
Расстояние Хеллингера



Критерий согласия Колмогоров-Смирнова



Дивергенция Кульбака-Лейблера



Расстояние Вассерштайна

Формализация оптимизационной задачи по уменьшению ошибок 1 и 2 рода с помощью метода Grid Search

Искомые параметры оптимизационной задачи

m — минимальное количество вхождений биграммы

τ — пороговое значение

$\alpha_к$ — коэффициент весов метрик расстояния

Входные параметры:

$\mathscr{F}^{(ref)} = \left\{ \hat{f}_{ij}^{(ref)}(t) \;\middle|\; \left(\left(c_i, c_j \right) \in P^{(ref)} \right\}$ — контрольный БКШ.

$\mathbb{F} = \{\mathscr{F}^{(1)}, \mathscr{F}^{(2)}, \dots\}$ — множество тестируемых образцов.

$m \in \left[2, max\left(\left| P^{(ref)} \cap P^{(test)} \right| \right) \right]$ — количество общих биграмм в двух наборах , где $P^{(ref)} \cap P^{(test)}$ – пересечение биграмм в выборках.

$\tau \in [0,1],\ \Delta\tau = 0.01$ — диапазон изменения порога.

d_k — набор функций расстояния.

α_k — веса функций расстояния, определяющие вклад каждой метрики в итоговую оценку. $\sum_k \alpha_k = 1,\ \alpha_k \geq 0$

Целевая функция — совокупность взвешенных ошибок первого и второго рода с заданным набором параметров:

$$\mathscr{L}(\mathrm{m}, \tau, \alpha) = \lambda_1 \mathrm{FRR}(\mathrm{m}, \tau, \alpha) + \lambda_2 \mathrm{FAR}(\mathrm{m}, \tau, \alpha).$$

Здесь FRR – ошибка первого рода, FAR – ошибка второго рода, λ_1и λ_2 – коэффициенты важности ошибок первого рода.

Алгоритм решения оптимизационной задачи по уменьшению ошибок 1 и 2 рода с помощью метода Grid Search

1. Подготовка данных

$\mathscr{F}^{(\text{ref})}$ — контрольный образец, $\mathbb{F}$ — тестовые образцы

Определить $P^{(ref)} \cap P^{(test)}$, m.

2. Перебор параметров методом Grid Search

$$\forall\, m \in \left[2; \max\left(\left|P^{(ref)} \cap P^{(test)}\right|\right)\right], \text{ если } \left|\text{T}_{\text{ij}}^{(\text{ref})}\right| \geq \text{m},$$

анализ $\left(\text{c}_\text{i}, \text{c}_\text{j}\right)$

2.1 $\tau \in [0; 0.1; 1]$

$\alpha_\text{k} = <\text{d}_\text{h},\ \text{d}_\text{ks},\ \text{d}_\text{kl},\ \text{d}_\text{w}>:$
$\text{d}_\text{h} + d_{ks} + d_{kl} + d_w = 1,$
$\text{d}_\text{h},\ \text{d}_\text{ks},\ \text{d}_\text{kl},\ \text{d}_\text{w} \in [0; 1]$

2.2. Вычисление сходства S_{total}

2.2.1 $\forall \left(\text{c}_\text{i}, \text{c}_\text{j}\right)$ и $\forall \left(\hat{\text{f}}_{\text{ij}}^{(\text{ref})}, \hat{\text{f}}_{\text{ij}}^{(\text{test})}\right)$:

$$\text{S}_{\text{ij}}^{(\text{k})} = \text{d}_\text{k}\left(\hat{\text{f}}_{\text{ij}}^{(\text{ref})}, \hat{\text{f}}_{\text{ij}}^{(\text{test})}\right)$$

2.2.2 Нормализация $S_{ij}^{(k)}$

2.2.3 $$\text{S}_{\text{total}} = \sum_{\left(\text{c}_\text{i},\text{c}_\text{j}\right) \in \text{P}^*} \omega_{\text{ij}}^{\text{norm}} \sum_{\text{k}} \alpha_\text{k}\, \text{S}_{\text{ij}}^{(\text{k})}$$

Если $\text{S}_{\text{total}} \geq \tau \Rightarrow$ два образца принадлежат одному пользователю.

Если $\text{S}_{\text{total}} < \tau \Rightarrow$ два образца принадлежат разным пользователям.

Оценка FAR и FRR

$\forall\, (\text{m}, \tau, \alpha)$:

$$\text{FRR}(\text{m}, \tau, \alpha) = \frac{количество\ ложных\ недопусков}{общее\ количество\ подлинных\ образцов}$$

$$\text{FAR}(\text{m}, \tau, \alpha) = \frac{количество\ ошибочных\ принятий}{общее\ количество\ неподлинных\ образцов}$$

2.3 Решение оптимизационной задачи

$$\mathscr{L}(\text{m}, \tau, \alpha) = \lambda_1 \text{FRR}(\text{m}, \tau, \alpha) + \lambda_2 \text{FAR}(\text{m}, \tau, \alpha)$$

3. O: $\left(\text{m}'\tau'\alpha^*\right) = \text{argmin}\mathscr{L}(\text{m}, \tau, \alpha)$

Выходные параметры:

$\text{m}'\tau'\alpha^*$ – оптимальные параметры

Эффективность биометрической системы на основе анализа клавиатурного почерка (при использовании одной метрики расстояния)

EER при использовании метрик без комбинирования

Метрика расстояния	**EER**
Расстояние Хеллингера	**0.0374**
Расстояние Вассерштайна	0.4628
Критерий согласия Колмогорова-Смирнова	0.2129
Дивергенция Кульбака-Лейблера	0.2092

Только расстояние Хеллингера

EER	**0.037433**
Порог	80
Точность	**0.965347**
Precision	0.977778
Recall	0.970588

Только расстояние Вассерштейна

EER	**0.462790**
Порог	33
Точность	**0.539604**
Precision	0.704762
Recall	0.544118

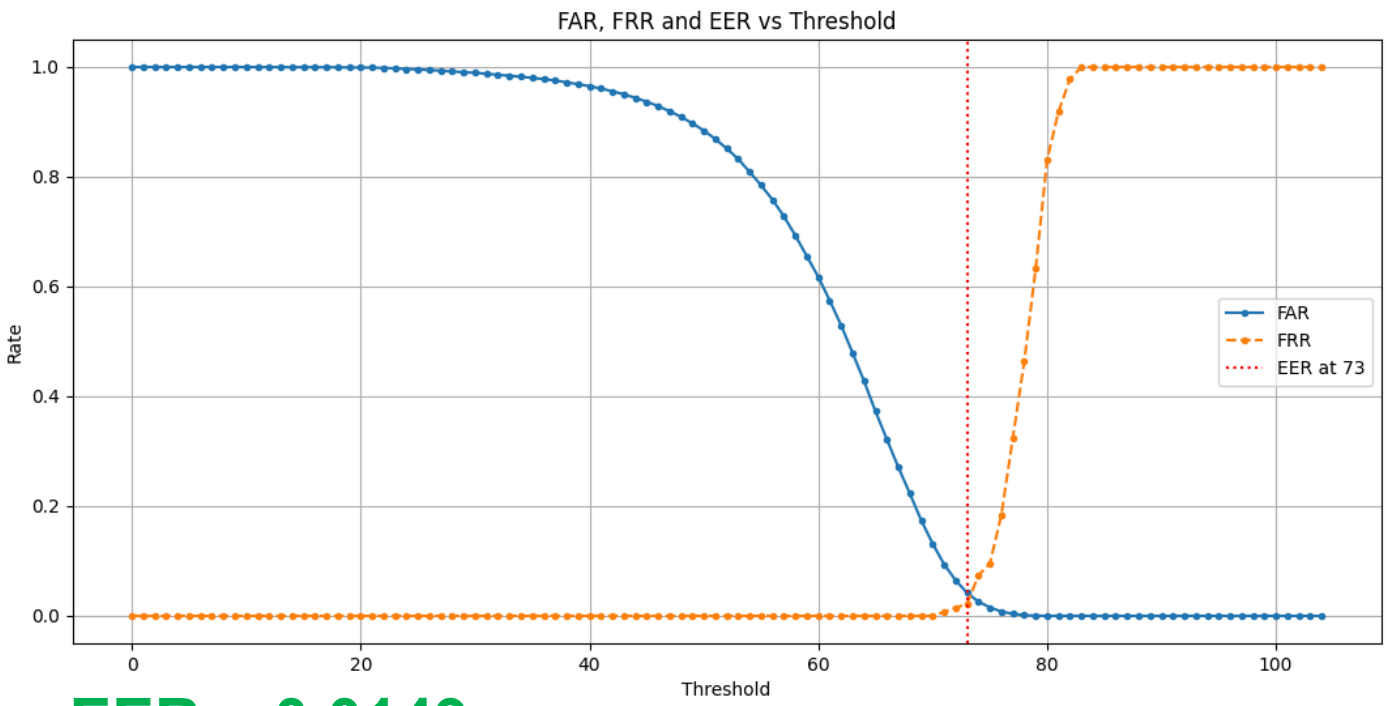
Только критерий согласия Колмогорова-Смирнова

EER	**0.212901**
Порог	37
Точность	**0.792079**
Precision	0.879032
Recall	0.801471

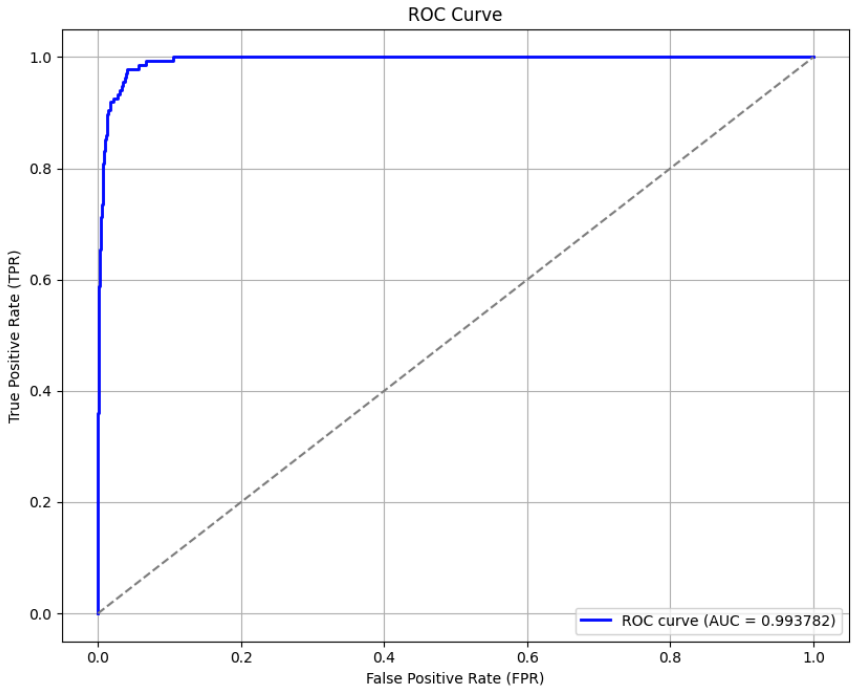
Только дивергенция Кульбака-Лейблера

EER	**0.209225**
Порог	96
Accuracy	**0.797030**
Precision	0.880000
Recall	0.808824

Итоговая эффективность метода идентификации на основе анализа клавиатурного почерка



EER = 0.0149



AUC = 0.9938

Оптимальные веса метрик расстояния

Метрика	Вес метрики
Расстояние Хеллингера	0.4
Расстояние Вассерштейна	0.05
Критерий согласия Колмогорова-Смирнова	0.25
Дивергенция Кульбака-Лейблера	0.3

Дополнительные показатели эффективности

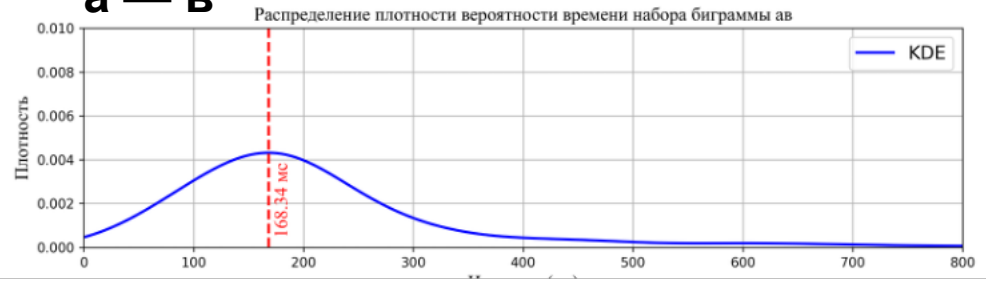
ZMFR	0.9852
Точность	0.9851
Precision	0.9926
Recall	0.9853

Сравнение эффективности разработанного метода с существующими

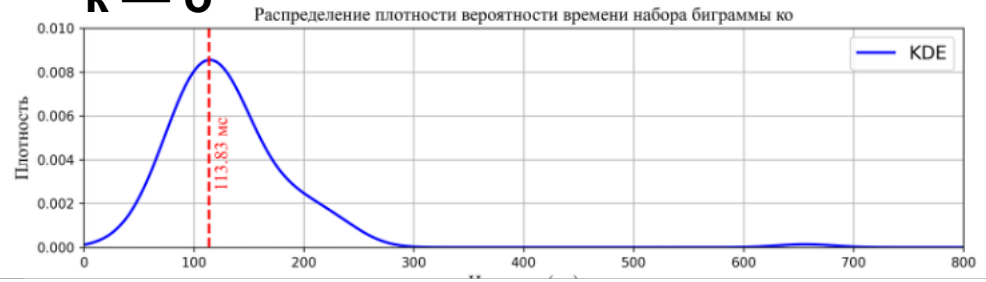
Исследователи	**EER, %**	**Точность, %**	**Датасет**	**Тип текста**
Putra, Chowandra	0,9-12,7	–	собственный (10 пользователей)	индивидуальные пароли
Данное исследование	**1,49**	**98,51**	**собственный (160 пользователей)**	**свободный текст на основе алгоритма**
Senerath, Tharinda et al.	1,8	–	Aalto (168000)	оба (статический + динамический)
Moralez, Fierres et al.	1,85	-	Aalto (78000)	динамический
Dimaratos, Pohn	2,6	97,4	Lee, CMU, Buffalo	фиксированный и свободный
Wyciślik, Wylężek, Momot	2,65	–	Buffalo	фиксированный и свободный
Сулавко	2,65	–	31 / 71 / 32 пользователей	динамический
KVC-onGoing	3,33	–	KVC-onGoing	фиксированный и свободный
Kaluarachchi et al.	3,8	–	SU-AIS BB-MAS, Buffalo	оба (статический + динамический)
Chang, Li, Stamp	3,86	94,6	Buffalo (157 пользователей)	динамический
Acien, Morales	4,8	–	Aalto (168 000)	динамический
Ayotte, Hou	7,8	–	Buffalo (148) / Clarkson II (101)	динамический
Simao, Prado et al.	-	90,14	собственный (137 пользователей)	свободный

Графики плотностей распределения вероятности по биграммам по всему набору данных

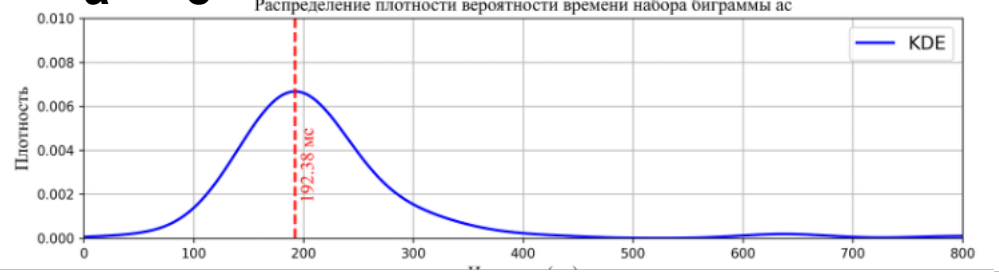
а — в



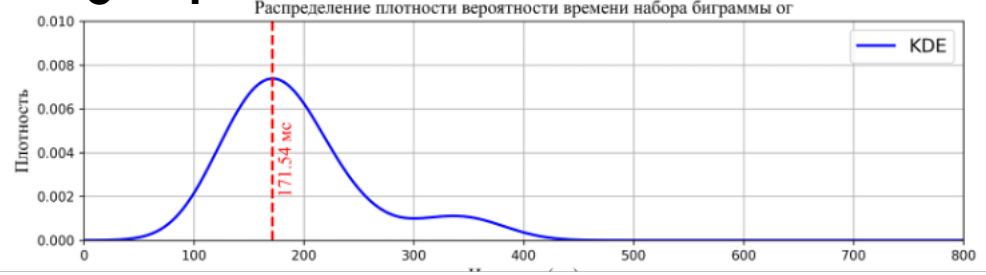
к — о



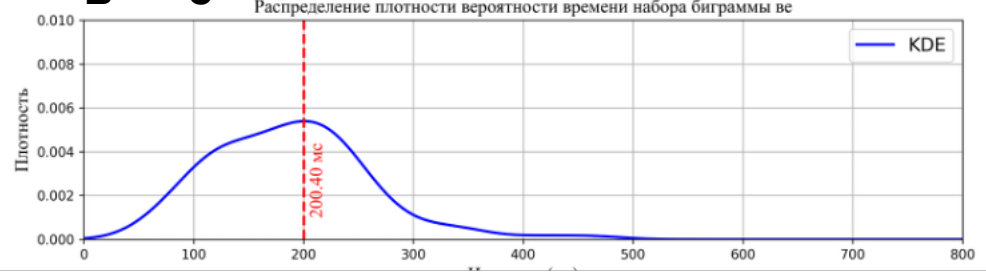
а — с



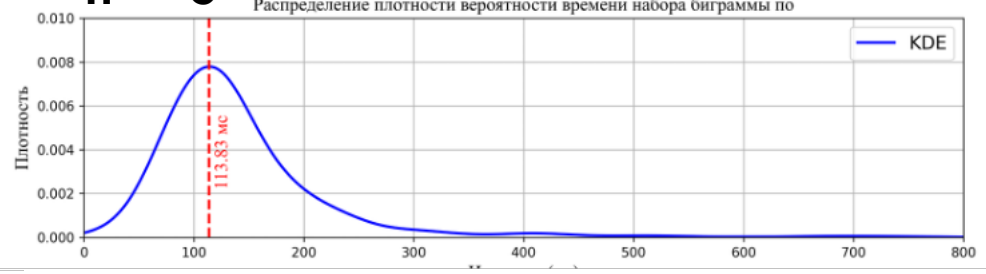
о — г



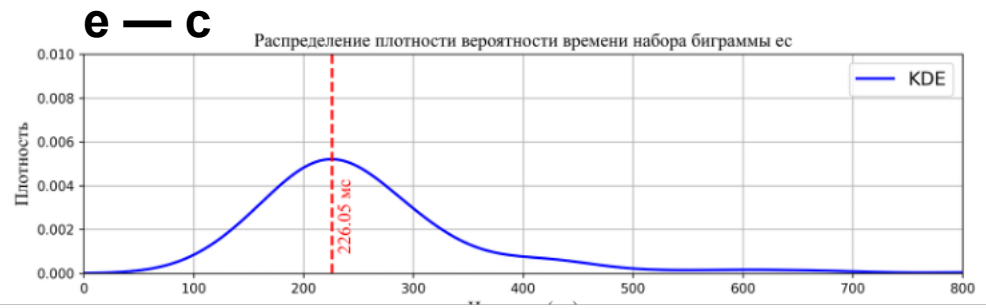
в — е



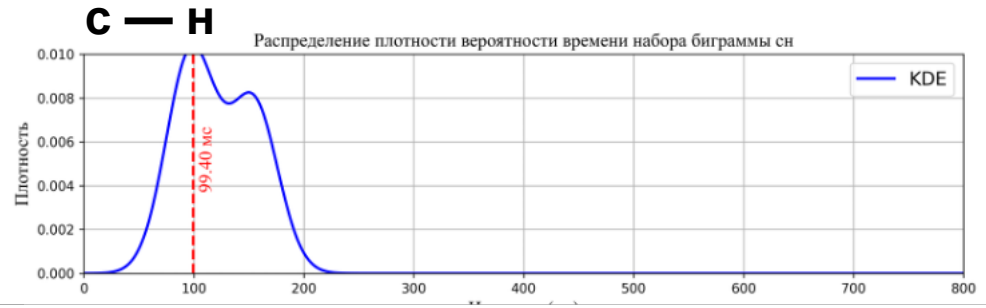
п — о



е — с



с — н



Научная новизна

Разработан новый метод биометрической идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм, отличающийся использованием для оценки эффективности комбинированной метрики сходства распределений.

Практическая значимость

Математическая модель и метод положены в основу архитектуры функционирования программно-аппаратной подсистемы биометрической идентификации пользователей ИС.

Соответствие паспорту специальности 2.3.6

1. Теория и методология обеспечения информационной безопасности и защиты информации.
12. Технологии идентификации и аутентификации пользователей и субъектов информационных процессов. Системы разграничения доступа.
15. Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности.

Публикации по теме исследования

1. Шкляр Е.В., Воробьев Е.Г., Савельев М.Ф. Распознавание клавиатурного почерка в браузере (2017, Известия СПбГЭТУ ЛЭТИ, перечень ВАК)
2. Шкляр Е.В. ON USING KEYSTROKE DYNAMICS TO PROTECT AGAINST BADUSB ATTACK, Материалы XVII Всероссийской научно-практической конференции с международным участием на английском языке «Диалог культур». В 3 ч. / Минобрнауки РФ; ФГБОУ ВО «С.-Петерб. гос. ун-т промышленных технологий и дизайна»; сост. К. А. Сечина, М. С. Липатов; под общ. ред. В. В. Кирилловой. — СПб.: ВШТЭ СПбГУПТД, 2024. — Ч. I. — 228 с.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025686747

Программа для биометрической идентификации пользователей ИС на основе геометрического сходства распределений вероятности времени набора биграмм

Правообладатель: *Шкляр Евгений Вадимович (RU)*

Автор(ы): *Шкляр Евгений Вадимович (RU)*

Заявка № 2025685766
Дата поступления 25 сентября 2025 г.
Дата государственной регистрации
в Реестре программ для ЭВМ *06 октября 2025 г.*

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1ce300b154f240f670bcca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

4. Алгоритм функционирования программно-аппаратной подсистемы биометрической идентификации пользователей

Решаемая проблема

В настоящее время неизвестно о существовании эффективно функционирующего устройства для биометрической идентификации на основе анализа клавиатурного почерка, соответствующего ГОСТ 54412-2019.

Требования к функциональности программно-аппаратного комплекса

1. **Регистрация субъекта.** При первом включении устройство должно запрашивать и обрабатывать данные для создания биометрического контрольного шаблона.
2. **Сохранение состояния.** При последующих включениях устройство должно использовать ранее полученный биометрический шаблон.
3. **Питание только по USB.** Устройство не должно требовать дополнительных источников питания для себя и клавиатуры, кроме питания по USB от целевой системы.
4. **Идентификация.** После успешной биометрической идентификация устройство должно вводить целевой пароль в систему путем эмуляции нажатий клавиатуры.
5. **Стационарность.** Устройство не требует отключения после идентификации и не блокирует дальнейшее использование клавиатуры.
6. Соответствие **структуре и компонентам БСОВ** и требованиям ГОСТ 54412-2019.

Схема компонентов программно-аппаратного комплекса в режиме онлайн-аутентификатора

1. Подсистема сбора биометрических данных

Биометрическое предъявление: 30 слов по алгоритму из ОНР-1 с клавиатуры.

Извлечение биометрических характеристик: биграммы и время между ними по ОНР-2.

2. Подсистема хранения биометрических данных

База шаблонов хранится на сервере.

На стороне устройства происходит только сбор и первичная обработка данных КП,

3. Подсистема обработки сигнала

Формирование БКШ происходит на сервере по модели из ОНР-2.

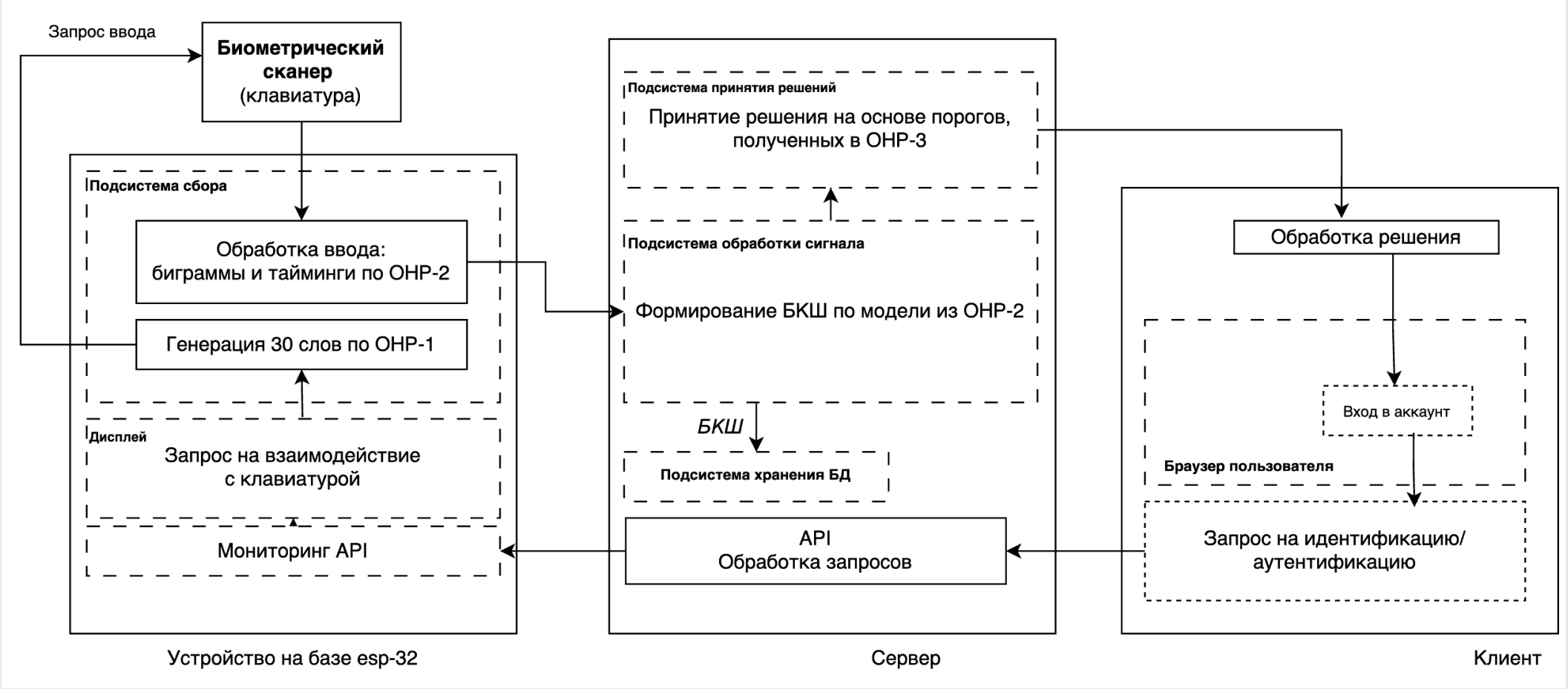
4. Подсистема сравнения

Сравнение происходит на сервере по методу из ОНР-3.

5. Подсистема принятия решений

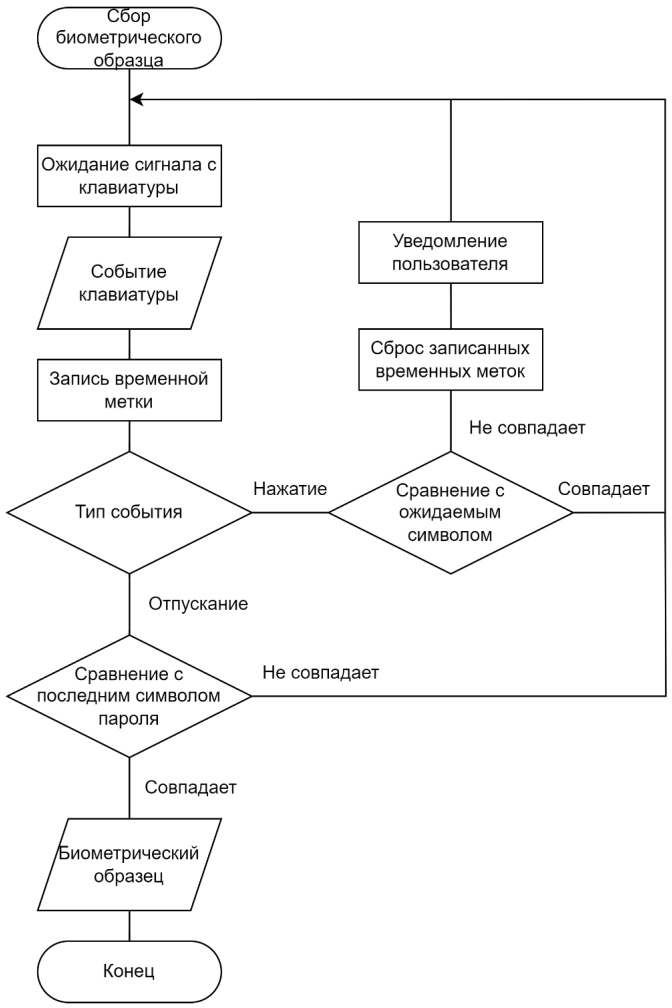
Принятие решений происходит на основе пороговых значений из ОНР-3.

Решение зависит от типа запроса: аутентификация или идентификация.

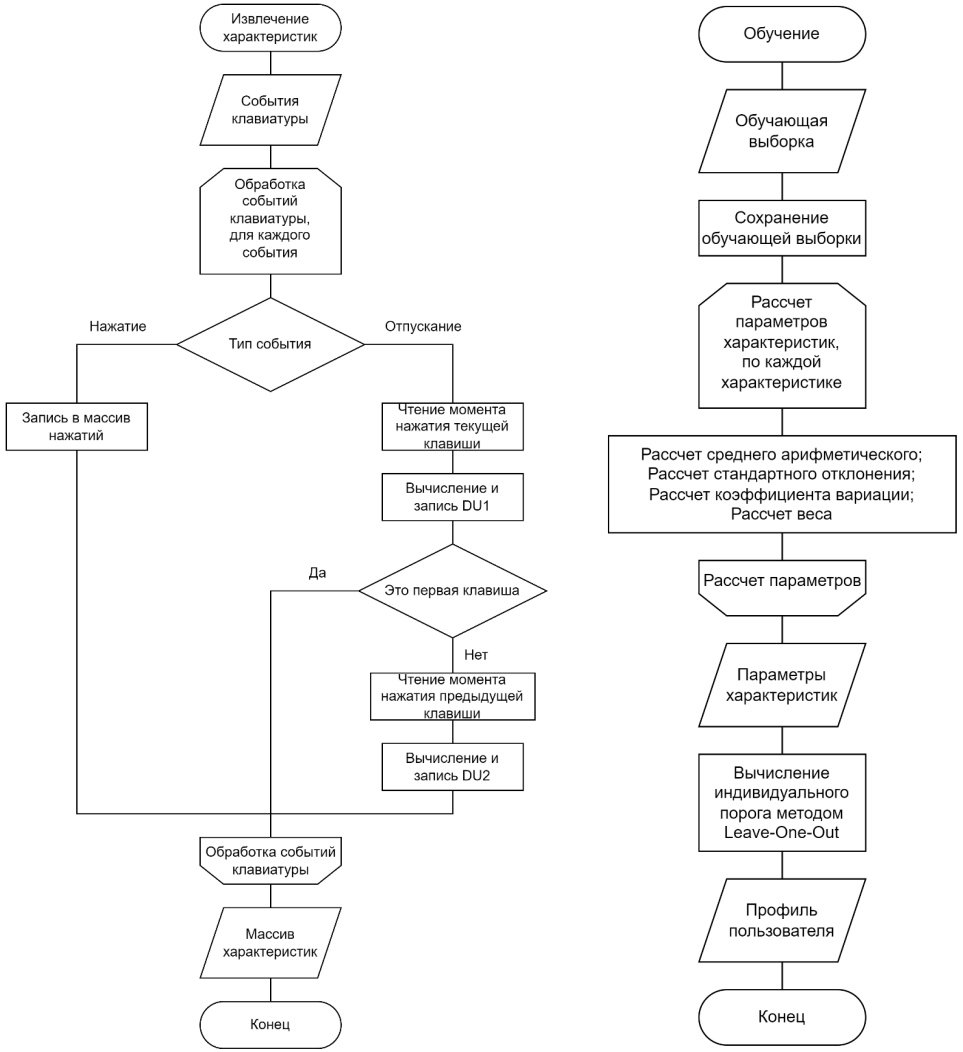


Алгоритм функционирования подсистем в рамках БСОВ в режиме локального аутентификатора

Подсистема сбора данных



Подсистема обработки сигнала



Подсистемы сравнения и принятия решений

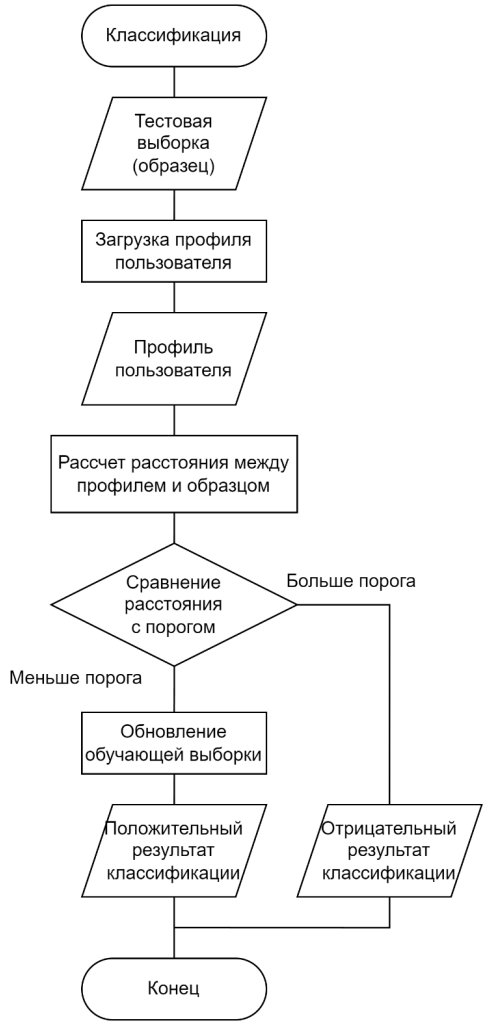


Схема и прототип аппаратного анализатора клавиатурного почерка

Функционирование в режиме онлайн-идентификации

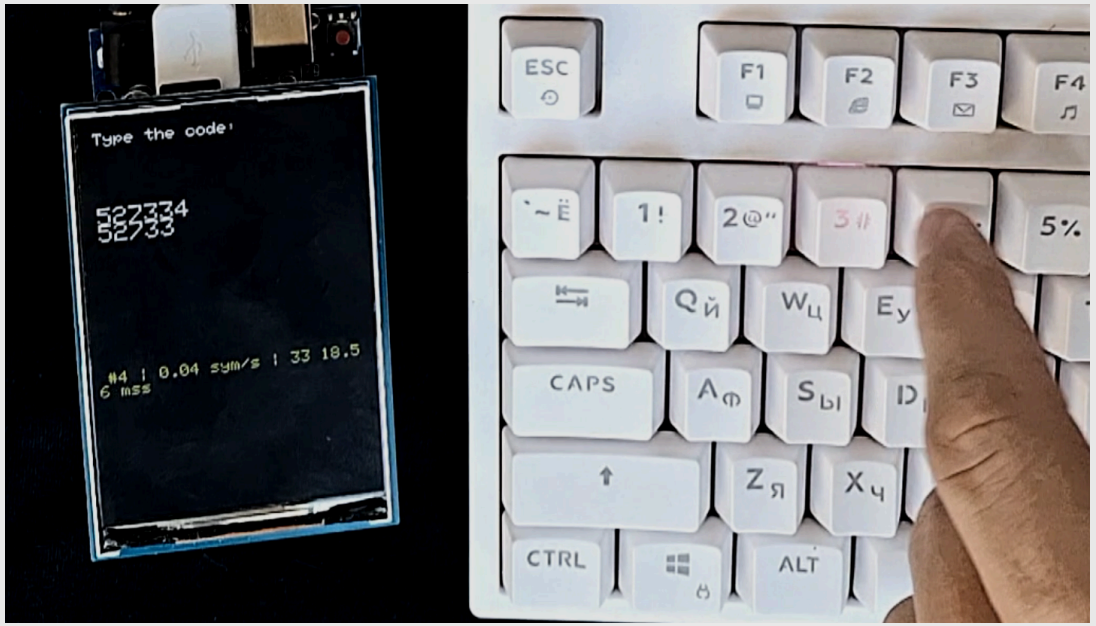
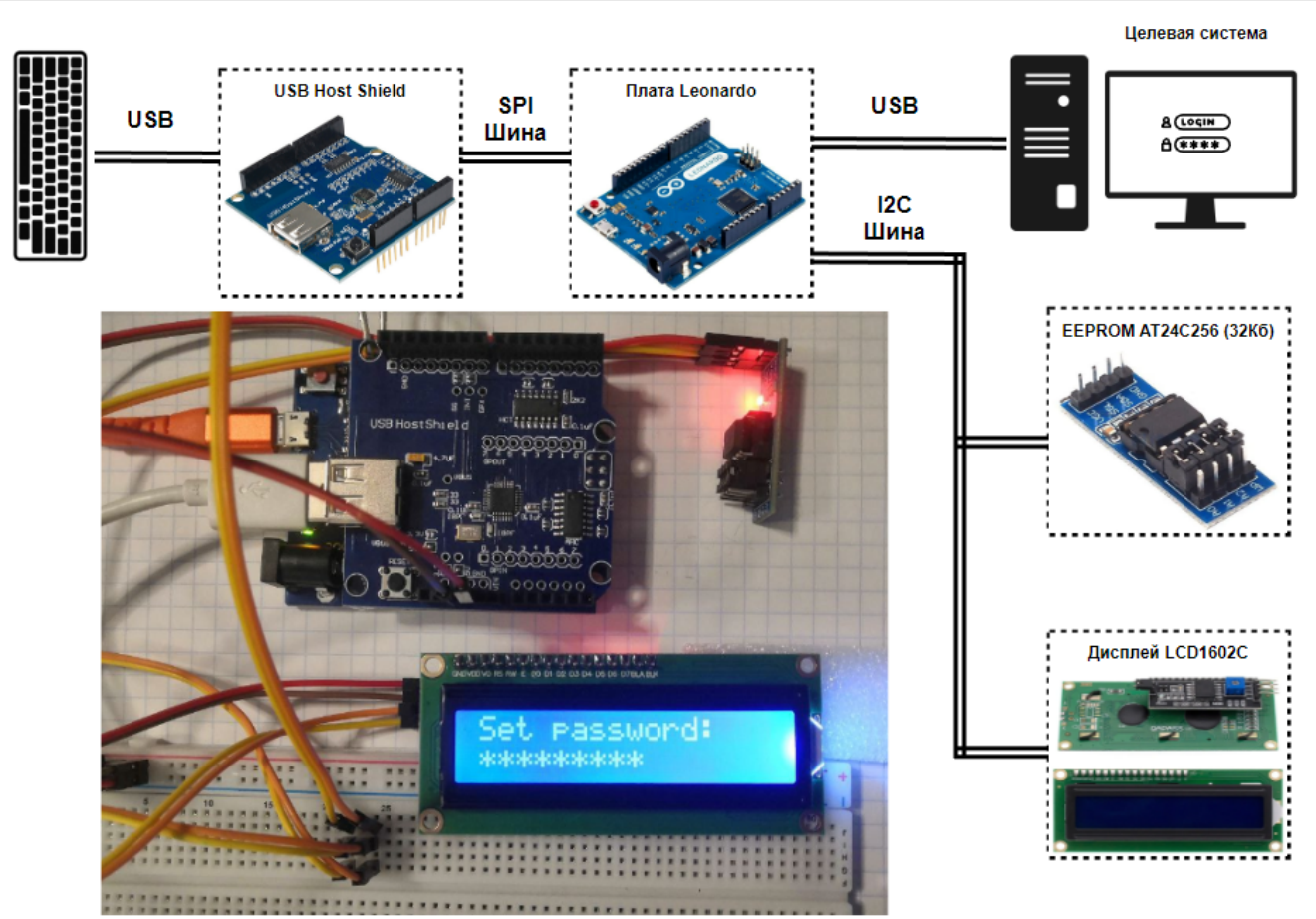


Схема функционирования в режиме локальной идентификации



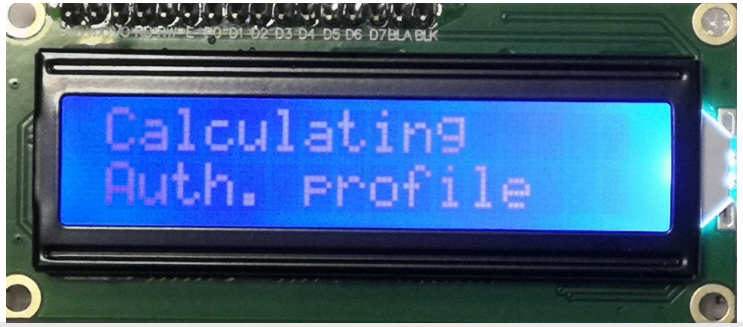
Функционирование программно-аппаратного комплекса в режиме локальной идентификации

Регистрация целевого пароля и пароля для анализа КП

Накопление обучающей выборки и извлечение характеристик


```
Training sample #1 collected
69.00 66.00 73.00 66.00 43.00 83.00 65.00 76.00
Training sample #2 collected
63.00 76.00 73.00 76.00 53.00 73.00 63.00 86.00
Training sample #3 collected
73.00 66.00 63.00 66.00 63.00 83.00 63.00 73.00
Training sample #4 collected
34.00 66.00 63.00 68.00 65.00 73.00 63.00 73.00
```

Создание и сохранение биометрического клавиатурного шаблона




```
### Biometric authentication template builded ###
Features count: 17
Threshold: 16.97
Means:
65.70 74.00 72.10 72.00 56.30 79.20 67.30 80.70 66.20 186.90
Stds:
19.76 7.48 9.34 9.02 6.60 6.62 4.90 7.23 4.92 17.87 16.25 18.
```

Схема подключения анализатора

Научная новизна

Построены новый алгоритм функционирования программно-аппаратной биометрической системы идентификации на основе анализа клавиатурного почерка в локальном и онлайн-режимах, реализующий предыдущие положения.

Практическая значимость

Обеспечивает безопасную и масштабируемую реализацию онлайн-идентификации на основе клавиатурного почерка.

За счёт распределения функций между микроконтроллером (сбор и первичная обработка), сервером (моделирование и сравнение биометрических признаков) и клиентом (управление сессией) достигаются высокая устойчивость к атакам, минимизация передачи чувствительных данных и возможность интеграции в существующие информационные системы с низкими вычислительными затратами на стороне пользователя.

Соответствие паспорту специальности 2.3.6

2. Методы, аппаратно-программные средства и организационные меры защиты систем (объектов) формирования и предоставления пользователям информационных ресурсов различного вида.

12. Технологии идентификации и аутентификации пользователей и субъектов информационных процессов. Системы разграничения доступа.

15. Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025682743

Программа идентификации пользователей на основе анализа клавиатурного почерка при вводе свободного текста

Правообладатель: ***Шкляр Евгений Вадимович (RU)***

Автор(ы): ***Шкляр Евгений Вадимович (RU)***

Заявка № **2025681941**
Дата поступления **20 августа 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***27 августа 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

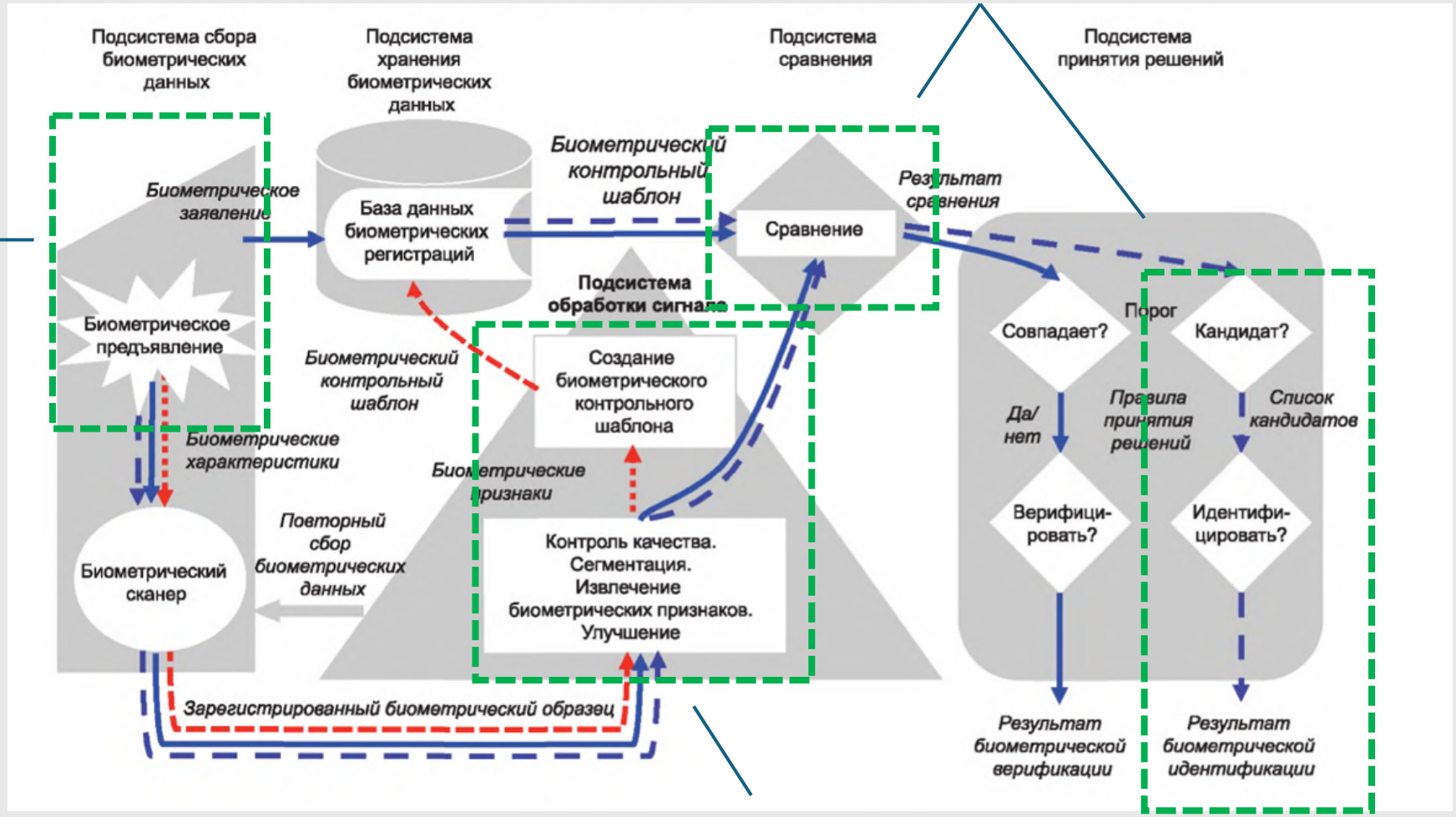
ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e761a3300ff54f240f670bcca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

Взаимосвязь положений в рамках биометрической системы общего вида

Положение 3

Положение 1



Положение 2

Положение 4 реализует все блоки по положениям 1, 2 и 3 в рамках программно-аппаратных комплексов.

Спасибо за внимание!