МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теоретических основ компьютерной безопасности и криптографии

# Распознавание человека по голосу

АВТОРЕФЕРАТ
дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Слышанкова Никиты Александровича

Научный руководитель _____ А. С. Гераськин
доцент, к.п.н
21.01.2023 г.

Заведующий кафедрой _____ М. Б.Абросимов
д. ф.-м. н., доцент
21.01.2023 г.

Саратов 2023

# ВВЕДЕНИЕ

В нашей жизни остро стоит вопрос защиты информации. Стандартным защитным системам все сложнее и сложнее справиться с распознаванием личности, поэтому вводятся дополнительные меры. В качестве этих мер используют проверки по биометрическим данным человека.

Биометрия фактически является комплексом приемов для распознавания человека, которые основываются на определенных физиологических характеристиках человека. Сейчас в качестве биометрических характеристик выбираются глаза, отпечатки пальцев, вид почерка, узор вен, черты лица или даже печать на клавиатуре. Характеристика позволяет выделить человека. Использование нескольких подобных характеристик позволяет практически со 100-процентной точностью распознать человека. Распознавание происходит считыванием характеристик человека и сравнения с теми, что хранятся в базах с данными пользователей.

Системы распознавания голоса – это вычислительные системы, которые могут определять речь говорящего из общего потока. Эта технология связана с технологией распознавания речи, которая преобразует произнесенные слова в цифровые текстовые сигналы, путем проведения процесса распознавания речи машинами.

Распознавание голоса используется в биометрических целях безопасности, чтобы определить голос конкретного человека. Эта технология стала очень популярной в мобильном банкинге, который требует идентификации подлинности пользователей, а также для других голосовых команд, чтобы помочь им совершать сделки.

Речь человека разбивается на отдельные «звуковые кадры», которые затем преобразуются в цифровую модель. Эти модели принято называть «голосовыми отпечатками». При дальнейшей идентификации сравниваются ранее зарегистрированный и вновь сформированный «голосовые отпечатки».

Для повышения надежности и ускорения распознавания пользователя часто просят отвечать на заранее оговоренные вопросы или произносить пароль. В этом случае распознавание осуществляется в режиме «верификации диктора» (в этой роли выступает сам пользователь, произносящий ответы или пароль).

Вейвлет-преобразование – относительно новая эффективная технология, позволяющая проводить обработку сигналов различного типа. Его свойства обусловливают возможность применения вейвлет-преобразований для анализа звуковых данных с последующим использованием полученных сведений для выделения звукового отпечатка человека[1].

Целью данной дипломной работы является разработка программы распознавания человека по голосу.

Для достижения поставленной цели были сформулированы следующие задачи:

- теоретически изучить методы и алгоритмы распознавания человека по голосу;
- рассмотреть основные методы получения голосового отпечатка;
- рассмотреть принципы работы системы распознавания человека по голосу.
- разработать программу, распознающую человека по голосу.

[1]Гребнов, С. В. Аналитический обзор методов распознавания речи в системах голосового управления [Электронный ресурс] / С. В. Гребнов // ispu.ru [Электронный ресурс] : «Вестник ИГЭУ». – Иваново, Россия, 2009 – . URL: http://ispu.ru/files/%2083-85.pdf (дата обращения 10.01.2023). – Загл. с экрана. – Яз. рус.

## КРАТКОЕ СОДЕРЖАНИЕ

Распознавание пола можно выделить в отдельный тип задач, который довольно успешно решается — при больших объемах начальных данных пол определяется практически безошибочно, а на коротких отрывках вроде ударного гласного звука вероятность ошибки — 5,3 % для мужчин и 3,1 % для женщин.

Сейчас науке известно несколько методов, способных моделировать человеческий голос:

- для класса текстозависимых систем - динамическое преобразование времени (Dynamic Time Warping; DTW) и скрытые марковские модели (Hidden Markov Model; HMM);
- для класса текстонезависимых систем - векторное квантование (Vector Quantification; VQ), модели гауссовой смеси (GMM) и метод опорных векторов (опорная векторная машина (SVM).

На данный момент, метод скрытых марковских моделей используют большинство современных систем как метод распознавания.

Для использования СММ для распознавания речи применяются следующие предположения: речь разбивается на состояния, внутри каждого из которых речевой сигнал используется как стационарный, а переход между состояниями осуществляется мгновенно; символ наблюдения, который порождает модель, имеет вероятность, которая зависит от состояния модели в данный момент, но при этом, не зависит от предыдущих.

Алгоритмы распознавания ключевого слова применяют данные модели для определения команд в потоке речи. В большинстве случаев данная задача выполняется с помощью метода скользящего окна (sliding window)3 и метода моделей-заполнителей (filler models)[2].

В этом разделе далее рассматриваются эти методы, описываются их достоинства, недостатки и математические модели.

---

[2]Осетров, В. Перспективы развития систем распознавания речи [Электронный ресурс] / В. Осетров // habr.com [Электронный ресурс] : Коллективный блог. – Киев, Украина, 2014 – . URL: https://habr.com/ru/post/232613/ (дата обращения 10.01.2023). – Загл. с экрана. – Яз. рус.

Рассматривается способ использования DTW алгоритма в распознавании речи при помощи анализа голосового сигнала.

Считаем лучшим алгоритмом отделения слов алгоритм Рабинел-Ламель. Если рассматривать строб-импульсов $\{s_1, s_2, \dots, s_n\}$ $\{s_1, s_2, \dots, s_n\}$, где $n$ – число образов строб-импульсов, а $s_i$ $s_i$, $i = \underline{1,n}$ $i = \underline{1,n}$ – численное выражение образцов, общая энергия строб-импульсов вычисляется по формуле (1)[3]:

$$E(n) = \frac{1}{n}\sum_{i=1}^{p} s_i^2 . E(n) = \frac{1}{n}\sum_{i=1}^{p} s_i^2 . \quad (1)$$

Наконец, рассматривается используемое в настоящей работе вейвлет-преобразование. Вейвлет-преобразование (ВП) широко используется для анализа сигналов. Кроме того, оно имеет значительные применения в сжатии данных: вейвлет-преобразование одномерного сигнала разлагается путем изменения масштаба и сдвигов в базис, состоящий из функций солитонного типа (вейвлетов) с определенными свойствами. Каждая функция в этом базисе характеризует как конкретную пространственную (временную) частоту, так и ее локализацию в физическом пространстве (времени).

В основе вейвлет-преобразования лежит идея многомасштабного анализа, которая заключается в последовательном огрублении исходной информации, содержащейся в процессе. Главные признаки вейвлета:

1. Ограниченность. Квадрат нормы функции должен быть конечным:

$$\|\psi\|^2 = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty$$

2. Локализация. ВП в отличие от преобразования Фурье использует локализованную исходную функцию и во времени, и по частоте. Для этого достаточно, чтобы выполнялись условия:

$$|\psi(t)| \leq C(1 + |t|)^{-1-\varepsilon}$$

$$\left|S_{\psi}(\omega)\right| \leq C(1 + |\omega|)^{-1-\varepsilon}, \varepsilon > 0$$

---

[3]Лукин, А. Введение в цифровую обработку сигналов (математические основы) [Электронный ресурс] / А. Лукин // audio.rightmark.org [Электронный ресурс] Лаборатория компьютерной графики и мультимедиа, МГУ. – Москва, Россия, 2007– . URL: http://audio.rightmark.org/lukin/dspcourse/dspcourse.pdf (дата обращения 30.12.2022). – Загл. с экрана. – Яз. рус.

Например, дельта-функция и гармоническая функция не удовлетворяют необходимому условию одновременной локализации во временной и частотной областях.

3. Нулевое среднее. График исходной функции должен осциллировать (быть знакопеременным) вокруг нуля на оси времени и иметь нулевую площадь:

$$\int_{-\infty}^{\infty} |\psi(t)|dt = 0$$

Из этого условия становится понятным выбор названия «вейвлет» – маленькая волна.

Равенство нулю площади функции $\psi(t)$, т.е. нулевого момента, приводит к тому, что фурье-преобразование $S_{\psi}(\omega)$ этой функции равно нулю при $\omega = 0$ и имеет вид полосового фильтра. При различных значениях $a$ это будет набор полосовых фильтров.

Часто для приложений бывает необходимо, чтобы не только нулевой, но и все первые n моментов были равны нулю: $\int_{-\infty}^{\infty} t^n \psi(t)dt = 0$

Вейвлеты $n$-го порядка позволяют анализировать более тонкую (высокочастотную) структуру сигнала, подавляя медленно изменяющиеся его составляющие.

4. Автомодельность. Характерным признаком ВП является его самоподобие. Все вейвлеты конкретного семейства $\psi_{ab}(t)$ имеют то же число осцилляций, что и материнский $\psi(t)$, поскольку получены из него посредством масштабных преобразований $a$ и сдвига $b$.

На основе изученных методов и алгоритмов было принято использовать в практической части преобразование на основе вейвлета Морле, так как он является эффективным и действенным методом перевода сигнала из временного представления в частотно-временное.

Глава 2 посвящена обзору систем распознавания человека по голосу.

13 февраля 2020 года компания ORBL начала коммерческую эксплуатацию своего b2b-продукта - высокоточной системы распознавания речи. Она представляет собой дескриптор речи, который может использоваться для преобразования потоковой спонтанной речи в текст[4].

Технология идентификации по голосу Voice Key компании «Речевые Технологии» позволяет организовать регламентированный доступ пользователей по заданной парольной фразе к ресурсам предприятия, телефонным и WEB-сервисам. Использование технологии Voice Key позволяет существенно повысить защищенность систем и, в то же время, упростить процесс идентификации пользователя. Технология Voice Key обеспечит высокую надежность и стабильность работы системы, а также поможет повысить качество обслуживания клиентов.

Сбербанк 14 мая 2021 года запустил SmartSpeech — сервис, который позволит бизнесу без специального оборудования подключать речевые технологии, например, в интерактивном голосовом меню (IVR), автоответчике, чатах, телемаркетинговых кампаниях или в голосовых интерфейсах взаимодействия.

SmartSpeech можно использовать для голосового ввода контента и команд на веб-сайтах, в приложениях и интеллектуальных устройствах. Технология синтеза и распознавания речи SmartSpeech также используется для создания IVR (интерактивного голосового меню) и автоответчиков для оптимизации работы колл-центров. Помимо распознавания и синтеза речи, сам сервис может использовать различные "подсказки" в зависимости от ситуации, чтобы помочь пользователю точно понять ситуацию.SmartSpeech также используется самим Сбербанком, например, в качестве основы семейства виртуальных помощников "Салют" Он используется в качестве основы семейства виртуальных помощников "Салют". Пользователи также могут проверить баланс своей

---

[4]ORBL Биометрическая система распознавания лиц и речи [Электронный ресурс]: Журнал TAdviser. Государство. Бизнес. Технологии. – URL: https://www.tadviser.ru/index.php/Продукт:ORBL_Биометрическая_система _распознавания_лиц_и_речи (дата обращения 13.12.2022). – Загл. с экрана. – Яз. рус.

банковской карты в любое время дня и ночи, набрав 900, не дожидаясь оператора.

Группа компаний ЦРТ (входит в экосистему Сбербанка) — глобальный разработчик интеллектуальных речевых технологий, распознавания лиц, технологический эксперт в области искусственного интеллекта и машинного обучения. Одна из немногих компаний в мире, которая создает и развивает обе биометрические модальности: лицо и голос. Технологии выявления подделок голоса и распознавание речи от группы ЦРТ занимают лидирующие позиции в мировых рейтингах NIST, ASVspoof Challenge, VOiCES, CHiME Challenge. Решения ЦРТ востребованы в 70 странах мира.

Технология диаризации и распознавания речи, созданная группой компаний ЦРТ, признана лучшей на международном конкурсе CHiME Speech Separation and Recognition Challenge (CHiME-6)[5].

Раздел 3 описывает практическую часть дипломной работы, а именно создание программного обеспечения, производящего вейвлет-анализ записей голоса. Использованное программное обеспечение: Windows 10 x64, язык программирования Python 3.9, библиотеки: tkinter, tkinter.ttk, numpy as np, os, sys, pyaudio, matplotlib, pywt, scaleogram as scg, time. Главное окно созданного программного продукта указано на рисунке 8. При нажатии на кнопку “Запись” производится запись голоса, длительностью 3 секунды, создается новый пользователь и отображается в левой верхней части окна приложения.

[5]Российская технология распознавания речи группы компаний ЦРТ признана лучшей в мире [Электронный ресурс]: Группа компаний ЦРТ. Новости и события. – URL: https://www.speechpro.ru/media/news/07-05-2020 (дата обращения 12.11.2022). Загл. с экрана. Яз. рус.

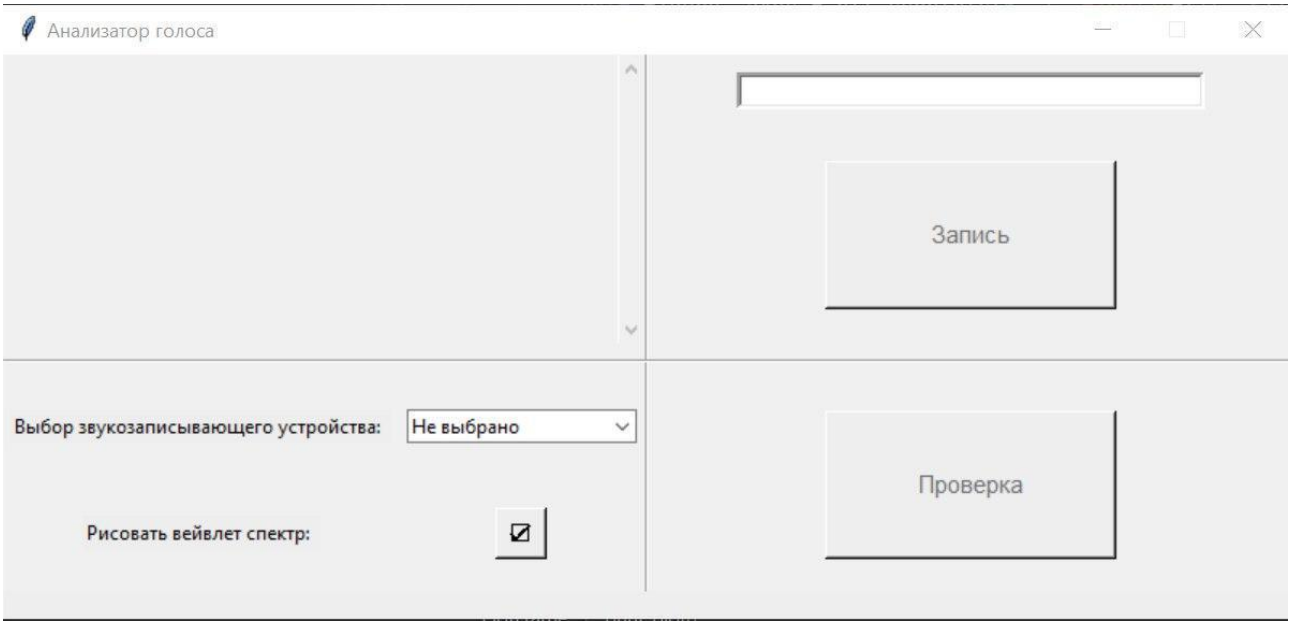


Рисунок 8 – Главное окно программного продукта

При нажатии кнопки “Проверка” пользователь Никита воспроизвел свою трехсекундную фразу и приложение опознало его корректно, в соответствии с рисунком 11.

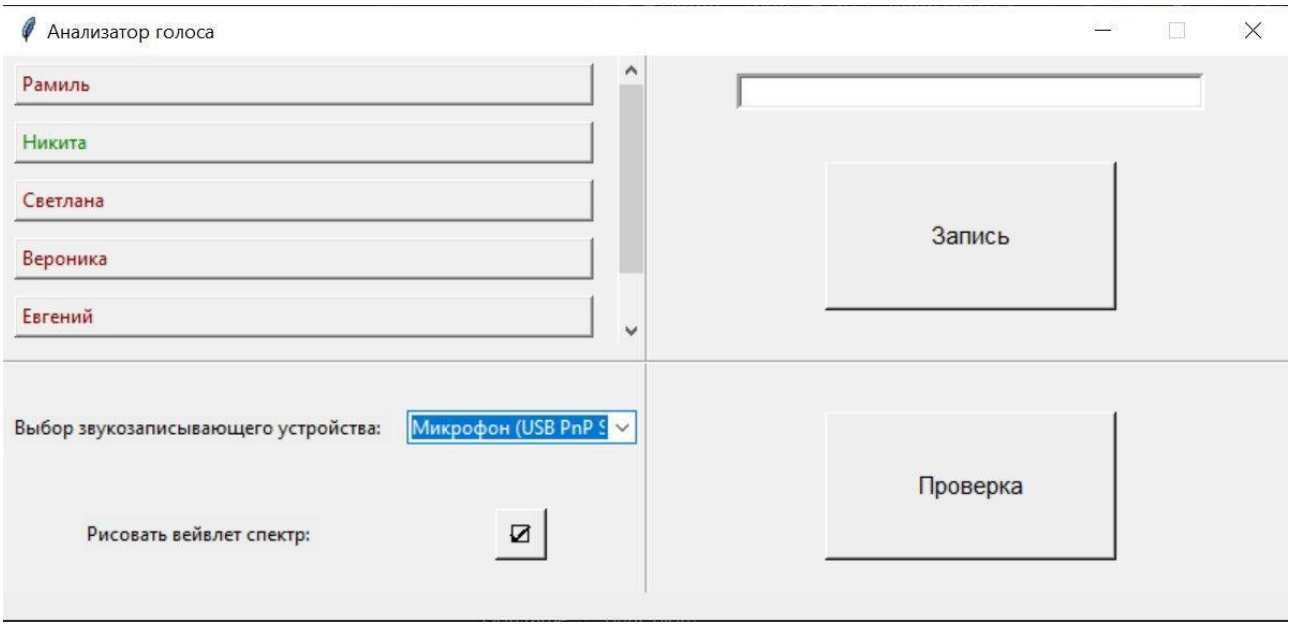


Рисунок 11 – Проверка голоса пользователем Никита

Вместе с нажатием кнопки пользователя выводятся графики аудиосигналов, проанализированных параметров и схожести (третий ряд), в соответствии с рисунком 13. Для верхнего ряда, значения по оси X - номер кадра, а по оси Y - значение сигнала. Для второго и третьего ряда, значения по оси X - номер параметра, а по оси Y - число от 0 до 1.

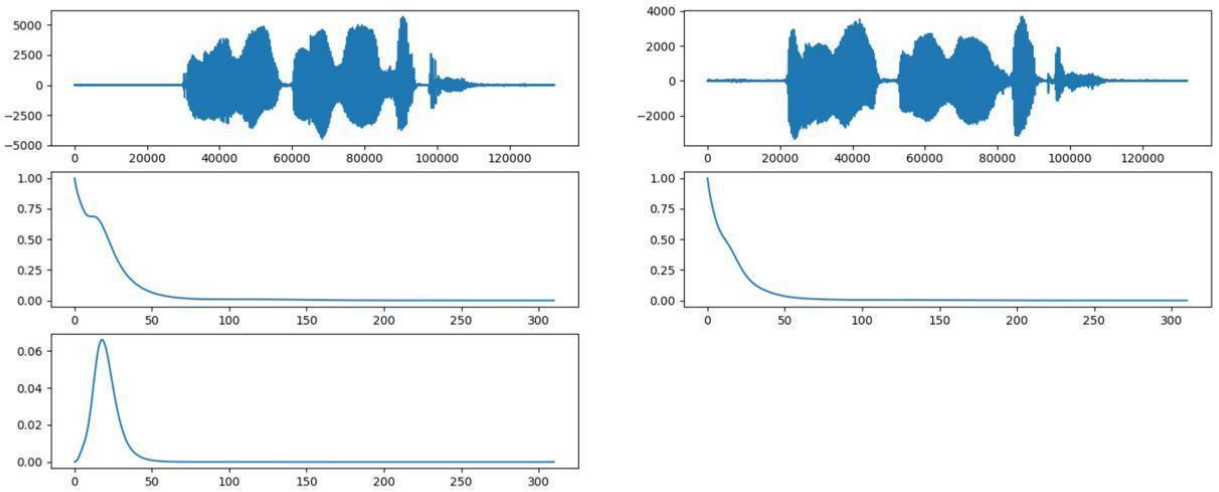


Рисунок 13 – Графики характеристик сигнала

В таблице 1 указаны значения отношения количества успешных вхождений под пользователем П2 всеми представленными пользователями к количеству попыток произнесения фраз.

Таблица 1 - Отношение успешных вхождений к количеству попыток

| Исследование \ Участник, произносящий фразу | П1 | П2 | П3 | П4 | П5 | П6 |
|---|---|---|---|---|---|---|
| Фраза П2, без помех | 0 | 1 | 0 | 0 | 0 | 0 |
| Фраза П2, с помехами | 0 | 1 | 0 | 0 | 0 | 0 |
| Фраза П2, с речью на фоне | 0,1 | 0,9 | 0 | 0 | 0,1 | 0 |

Таким образом, на основании данных из таблицы 1, можно утверждать, что разработанный метод анализа вейвлет коэффициентов в разы улучшает как опознавание соответствующего пользователя, так и в большинстве случаев предотвращает доступ для нежелательных пользователей. Процент верно положительных срабатываний примерно 97%. Процент ложноположительных срабатываний не превышает 2%.

# ЗАКЛЮЧЕНИЕ

В данной работе были изучены алгоритмы и методы распознавания человека по голосу, некоторые из которых до сих пор являются передовыми средствами для решения данной задачи. Рассмотренные основные методы распознавания речи на основе скрытых марковских моделей: метод скользящего окна и метод моделей заполнителей, – применяются в системах голосового управления и несмотря на то, что они имеют свои недостатки, до сих пор используются во всех крупных системах, целью которых является распознать человека по голосу.

В ходе работы была разработана программа, которая позволяет распознать человека по голосу, с использованием вейвлета Морле. Данный способ является наиболее эффективным в решении задачи распознавания человека по голосу, как было показано в главе 1, поэтому именно он и был использован в разработке программы.

Разработанный метод анализа вейвлет коэффициентов в разы улучшил опознавание пользователя по голосу, в сравнении с обычным методом корреляции вейвлет коэффициентов, позволил предотвратить доступ для нежелательных пользователей. По итогам тестирования программы можно сделать вывод, что выполнить распознавание человека по голосу является довольно сложной задачей, которая включает в себя затраты многих ресурсов, однако, выполнение данной задачи позволяет наиболее эффективно защищать конфиденциальную информацию, является хорошим средством аутентификации, особенно во взаимодействии с другими подобными средствами.