МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Школа компьютерных наук

Кафедра программного обеспечения

Отчет по дисциплине:

“Интеллектуальные системы поддержки принятия решений”

на тему:

Разработка системы автоматической обработки автоответчиков и кластеризации аудиофайлов для компании ООО «IQTEK»

### 

Выполнили

студенты группы:

МОиАИС 21.03

Жиряков Валентин

Неткачев Павел

Научный руководитель:

Плотоненко Юрий Анатольевич

ОГЛАВЛЕНИЕ

Оглавление

[ВВЕДЕНИЕ 3](#_Toc185588344)

[ГЛАВА 1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ 5](#_Toc185588345)

[1.1 ВВЕДЕНИЕ В ПРЕДМЕТНУЮ ОБЛАСТЬ 5](#_Toc185588346)

[1.2 АКТУАЛЬНОСТЬ ПРОЕКТА 6](#_Toc185588347)

[1.3 ИДЕЯ РЕШЕНИЯ 8](#_Toc185588348)

[2. РЕАЛИЗАЦИЯ СИСТЕМЫ 9](#_Toc185588349)

[2.1. ТРЕБОВАНИЯ К ПРОЕКТУ 9](#_Toc185588350)

[2.2 ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ТЕХНОЛОГИЙ 10](#_Toc185588351)

[2.3 ОПИСАНИЕ ФАЙЛОВОЙ СИСТЕМЫ: 12](#_Toc185588352)

[2.4 АРХИТЕКТУРА СИСТЕМЫ 13](#_Toc185588353)

[2.5 ОПИСАНИЕ МОДУЛЕЙ 14](#_Toc185588354)

[2.6 ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ ПРОЕКТА 16](#_Toc185588355)

[2.7 ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС ПРОЕКТА 18](#_Toc185588356)

[**Описание стартового интерфейса приложения** 18](#_Toc185588357)

[2.8 ОРГАНИЗАЦИЯ ТЕСТИРОВАНИЯ ПРОЕКТА 20](#_Toc185588358)

[ЗАКЛЮЧЕНИЕ 23](#_Toc185588359)

[СПИСОК ЛИТЕРАТУРЫ 24](#_Toc185588360)

[ПРИЛОЖЕНИЕ 25](#_Toc185588361)

## ВВЕДЕНИЕ

Процесс обработки телефонных записей в call-центрах — сложная и важная задача, связанная с необходимостью анализа больших объемов данных. Телефонные записи включают разнообразные ответы автоответчиков, шумы, музыку и реальных операторов. Текущие методы обработки таких данных, используемые в компании ООО «IQTEK», требуют значительных человеческих и временных ресурсов, что приводит к снижению эффективности работы аналитиков.

Основными проблемами при обработке аудиозаписей являются:

1. Высокая трудоемкость — процесс обработки вручную занимает много времени и подвержен ошибкам.
2. Человеческий фактор — ошибки, связанные с субъективностью восприятия и ограниченным временем анализа.
3. Большой объем данных — ежедневное поступление тысяч записей требует автоматизации для их систематизации и анализа.

Целевая аудитория проекта — аналитики компании, которые занимаются прослушиванием записей, разметкой данных, разделением записей на сегменты и их последующей классификацией.

Задача проекта — создать решение рутинного процесса и повысить качество анализа телефонных записей.

Целью проекта является разработка автоматизированной системы обработки записей с функциями:

* **Кластеризации** — группировка похожих записей по эмбеддингам голосовой активности.
* **Сегментации** — выделение речевых фрагментов из аудиозаписей и их сохранение в виде отдельных файлов.
* **Транскрибации** — преобразование аудио в текст для анализа содержания.

Для достижения цели требуется решение следующих задач:

1. Разработка алгоритмов предобработки и анализа аудио.
2. Интеграция предобученных моделей для выделения голосовой активности, эмбеддингов и транскрибации.
3. Создание desktop-приложения для автоматизации задач аналитиков.

Разработка системы обеспечит повышение точности обработки данных, сократит временные затраты и повысит контроль над качеством анализа записей.

# ГЛАВА 1. ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

## 1.1 ВВЕДЕНИЕ В ПРЕДМЕТНУЮ ОБЛАСТЬ

Предметной областью данной системы является автоматизация рабочих задач аналитиков, которые занимаются обработкой телефонных записей, поступающих из call-центров, для своего продукта и для их клиентов. Эти аудиозаписи содержат, как ответы живых людей (операторов и клиентов), так и сообщения автоответчиков и голосовых помощников. Эти аудиофайлы проходят через ручную нормализацию, кластеризацию и сегментацию.

Специфика обработки телефонных записей заключается в следующем:

1. **Разнообразие содержания файлов**:  
   Аудиофайлы отличаются друг от друга не только по продолжительности, но и по содержанию. Обычно, ответы автоответчиков имеют одинаковую структуру, но могут также включать в себя шумы, музыку, различные вставки голосов реальных людей, и так далее.
2. **Большой объем данных**:  
   Call-центры ежедневно создают тысячи записей, что требует эффективных методов обработки и анализа.
3. **Сложность ручной обработки**:  
   Ручное разделение записей на категории занимает значительное время, что делает такой подход неэффективным и времязатратным. При этом большая часть записей может оставаться необработанной из-за недостатка времени и человеческих ресурсов.
4. **Требование к точности**:  
   Неверная кластеризация записей может привести к искажениям в анализе данных, что негативно влияет на итоговый результат работы ПО по отбиванию автоответчиков.

## 1.2 АКТУАЛЬНОСТЬ ПРОЕКТА

Процесс обработки и анализа аудиофайлов, особенно в контексте исследования аудиозаписей для выявления автосообщений (автоответчиков), является весьма трудоемким и требует значительных временных затрат. В компании ООО «IQTEK» аналитики ежедневно обрабатывают большое количество аудиофайлов, что затрудняет выполнение задач в срок и с необходимой точностью. Особенно сложной задачей является ручная кластеризация аудиофайлов, для последующего анализа и отправки результатов анализа обратно call-центрам.

На данный момент данный процесс происходит в ручном режиме, что приводит к множеству проблем:

* **Высокая трудозатратность**: Аналитикам приходится тратить большое количество времени на обработку и сегментацию аудиофайлов.
* **Человеческий фактор**: Ручная обработка данных подвержена ошибкам, что может снижать точность и качество анализа.
* **Объем данных**: При увеличении количества аудиофайлов система сталкивается с проблемой обработки больших объемов данных, что делает текущий процесс неэффективным.

Нашей целевой аудиторией, для которой мы делаем проект - это команда аналитиков компании ООО «IQTEK», чья работа напрямую связана с обработкой аудиофайлов. Основные задачи, которые выполняют аналитики:

* Расчет метрик ПО компании;
* Прослушивание и разметка записей разговоров;
* Разделение телефонных записей на отдельные файлы с голосовой активностью;
* Обработка звуковых характеристик аудиофайлов для нормализации;
* Ручное разбиение аудиофайлов с автоответчиками на кластеры;
* Добавление новых файлов в БД.

Основная проблема, с которой сталкивается заказчик, заключается в том, что процесс исследования аудиозаписей и определения автоответчиков требует значительных человеческих и временных затрат. При большом объеме аудиофайлов (сотни или тысячи файлов) текущие методы обработки становятся неэффективными, что приводит к:

* Отставанию от сроков;
* Частично необработанным данным;
* Уменьшению качества анализа из-за человеческих ошибок.

К системе, которая будет решать вышеописанные проблемы, заказчик имеет следующие требования и пожелания:

* Обработка и кластеризация большого объема данных:
  + Система должна быть способна эффективно работать с большим количеством аудиофайлов, обеспечивая кластеризацию на основе схожести содержимого.
* Минимизация времени обработки:
  + Система должна стремиться к минимизации времени обработки данных.
* Достоверность и точность кластеризации:
  + Кластеризация должно допускать минимум либо не допускать ошибок вообще.
* Удобство для аналитиков:
  + Процесс работы с системой должен быть интуитивно понятен и доступен для пользователей с разным уровнем технической подготовки.

Исходя из нашей проблемы и поставленных требований, целью проекта является разработка системы автоматической обработки автоответчиков с последующей их кластеризацией и локальным сохранением отсортированных по кластерам аудиозаписей.

## 1.3 ИДЕЯ РЕШЕНИЯ

Основная идея проекта заключается в разработке системы, которая автоматизирует часть задач, принадлежащих аналитикам. Решение предполагает использование алгоритмов обработки аудио и технологий машинного обучения для кластеризации сегментации аудиофайлов.  
Ключевая задача, которую будет решать наша система:

* Кластеризация и сегментация аудиофайлов с телефонными звонками. Система будет автоматизировать процесс кластеризации и сегментации аудиозаписей, улучшая эффективность работы аналитиков.

Для реализации системы будут использоваться 3 нейронные сети:

* Первая сеть производит предобработку аудиофайлов для выделения речевых фрагментов и удаления тишины или шумов.
* Вторая сеть производит получение эмбеддингов голосовой активности из аудиофайла полученного после работы первой нейронной сети.
* Третья сеть производит транскрибацию аудиофайлов.

Также для реализации проекта необходимо использовать несколько алгоритмов:

* Первый алгоритм производит кластеризацию на основе эмбеддингов полученных от второй нейронной сети.
* Второй алгоритм производит разбиение аудиофайла на сегменты с голосовой активностью с последующей записью в отдельные файлы.
* Третий алгоритм производит генерацию названий для кластера на основе транскрибированного из аудиофайла текста, полученного после работы третьей нейронной сети.

Результатом станет Desktop приложение, которое позволит загружать папки с аудиофайлами, автоматически обрабатывать, кластеризовать, а также сегментировать аудиозаписи с помощью предварительно обученных моделей и алгоритмов.

# 2. РЕАЛИЗАЦИЯ СИСТЕМЫ

## 2.1. ТРЕБОВАНИЯ К ПРОЕКТУ

Проект должен соответствовать ряду технических и функциональных требований, чтобы обеспечить его эффективность и надежность в различных условиях эксплуатации.

1. **Обработка и кластеризация большого объема данных**  
   Система должна иметь возможность обрабатывать сотни и тысячи аудиофайлов, сохраняя результаты кластеризации для дальнейшего анализа. Это требует оптимизированных алгоритмов и эффективного использования вычислительных ресурсов.
2. **Использование технологии CUDA**  
   Программа должна поддерживать вычисления с использованием GPU (графического процессора) от Nvidia через технологию CUDA. Это обеспечивает высокую производительность при выполнении задач, связанных с обработкой аудио и построением эмбеддингов.
3. **Совместимость с системами без GPU**  
   В случае отсутствия видеокарты Nvidia, система должна корректно работать на центральном процессоре (CPU). Однако в таких условиях скорость обработки данных заметно снижается, что следует учитывать при развертывании программы.
4. **Эффективное построение эмбеддингов**  
   Для качественной кластеризации система должна генерировать эмбеддинги аудиофайлов с высокой точностью, что требует достаточной вычислительной мощности и оптимизированного программного обеспечения.
5. **Минимизация времени обработки**  
   При работе с большими объемами данных система должна минимизировать время обработки, обеспечивая оптимальный баланс между точностью кластеризации и скоростью выполнения.
   1. Гибкость при развертывании
   2. Система должна быть готова к запуску на различных аппаратных платформах. Особое внимание уделяется выбору оборудования, если производительность является ключевым фактором для пользователя.

Эти требования и ограничения позволяют определить ключевые аспекты, на которых необходимо сосредоточиться при разработке, тестировании и эксплуатации проекта.

## 2.2 ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ТЕХНОЛОГИЙ

Проект реализуется с использованием современных технологий и инструментов, обеспечивающих надежность и масштабируемость системы. Ниже приведены используемые технологии и инструменты.

**Библиотеки:**

* **numpy**: Библиотека для высокопроизводительных вычислений и работы с многомерными массивами данных.
* **nltk**: Библиотека для обработки естественного языка, предоставляющая инструменты для работы с текстами и их анализом.
* **torch и torchaudio**: Библиотеки для работы с нейронными сетями. Обеспечивает выполнение вычислений на GPU и поддержку глубокого обучения.
* **Scikit-learn (DBSCAN, PCA)**: Для кластеризации аудиоэмбеддингов. PCA используется как метод снижения размерности для визуализации многомерных данных;
* **Plotly.graph\_objs, Dash**: Для создания интерактивных графиков и визуализации в 3D пространстве;
* **Pydub**: Используется для разделения аудиофайлов на сегменты.

**Модели для разработки**:

* **Silero VAD**: Предобученный детектор голосовой активности для удаления тишины между голосовыми участками;
* **EncDecRNNTModel (TitaNet Large)**: Предобученная модель от NVIDIA для извлечения аудиоэмбеддингов.
* **Whisper**: Модель от OpenAI для автоматической транскрибации аудиофайлов, которая позволяет с высокой точностью преобразовывать аудио в текст и давать названия кластерам.

**Подходы к разработке и организации кода:**

* **Модульность:** Код разделен на независимые компоненты, что упрощает тестирование, обновление и добавление нового функционала.
* **Повторное использование кода:** Использование функций и классов, которые можно многократно применять в различных модулях проекта, минимизируя дублирование логики.
* **Документированность:** Каждый модуль содержит комментарии и документацию, упрощающие понимание структуры и логики работы системы.

**Практическое применение технологий:**

* **Предобработка аудиофайлов:**Удаление шума, тишины из аудиофайлов, повышение частоты дискретизации и соединение участков голосовой активности в один файл с помощью модели **Silero VAD**.
* **Выделение эмбеддингов:**Получение векторного представления аудиофайлов (эмбеддингов), которое сохраняет важные характеристики звука с помощью модели **TitaNet Large**.
* **Кластеризация аудиофайлов:**

Автоматическая кластеризация аудиофайлов, на основе эмбеддингов, c помщью алгоритма **DBSCAN** из библиотеки **Scikit-learn**.

* **Генерация названий:**

Автоматическая генерация названий для кластеров на основе транскрибированного текста, полученного с помощью модели **Whisper.**

## 2.3 ОПИСАНИЕ ФАЙЛОВОЙ СИСТЕМЫ:

Файловая система представляет собой систему папок, предназначенную для хранения промежуточных и итоговых результатов работы системы

**Входные данные:**

На вход подается папка с неструктурированными и необработанными аудиофайлами. Каждый файл содержит запись телефонного разговора, в котором присутствует только канал с ответами автоответчика. Также могут присутствовать файлы с различными шумами и тишиной. Пример входных данных представлен в Приложении 1.

**Сегментация:**

Во время сегментации аудиофайлы с голосовой активностью сохраняются в папку “clustered audio”, где к исходному имени файла добавляется номер сегмента. Пример, итогов сегментации представлен в Приложении 2.

**Предобработка:**

Во время предобработки аудиофайлов, обработанные результаты сохраняются в промежуточную папку “tmp\_vad” с теми же названиями, что и у оригиналов. Также на данном этапе происходит фильтрация файлов не содержащих голосовой активности, такие файлы сохраняются сразу в паки “Шум” и “Тишина” . Пример предобработанных аудиофайлов представлен в Приложении 3

**Кластеризация:**

Во время работы кластеризации, оригинальные аудиофайлы сохраняются в соответствующие им пронумерованные папки кластеров (Например, Cluster 5). Пример содержания кластера представлен в Приложении 4. Все папки с кластерами сохраняются в общую папку “clustered audio”. Пример представлен в Приложении 5. После завершения работы кластеризации, происходит процесс генерации названий для кластеров, с последующим переименованием кластеров, на основе транскрибированного текста аудиофайлов в соответствующем кластере. Пример итогового результата представлен в Приложении 6.

## 2.4 АРХИТЕКТУРА СИСТЕМЫ

**Описание схемы архитектуры приложения (Приложение 7)**:

1. **Ввод аудиофайлов.** Приложение получает аудиофайлы (формата WAV, MP3 или FLAC) из указанной папки пользователем. Для каждого файла будет выполняться дальнейшая обработка.

2. **Выбор действия.** Пользователь выбирает, что он хочет сделать с файлами: выполнить кластеризацию или сегментацию. При выборе сегментации начнется непосредственно обработка файлов. При выборе кластеризации будет выполняться следующий блок программы, начиная с предобработки.

3. **Сегментация аудиофайлов (Pydub).** Если пользователь выберет данный модуль обработки, то каждый файл будет разделен на отрезки предложений, который обнаружит алгоритм библиотеки Pydub. Он разделяет аудиофайлы на основе тишины, порог которой он определяет на основе средней громкости аудио. Полученные отрезки сохраняются локально в новую папку.

4. **Предобработка аудио (Silero VAD).** В этом блоке используется модель VAD (Voice Activity Detection), которая обрабатывает аудиофайл и разделяет его на сегменты речи и тишины. Сегменты с низким уровнем амплитуды (тишина) или слишком короткой продолжительностью (менее 0.3 секунды) удаляются. Если во время обработки голос не был обнаружен в аудиофайле, то он классифицируется как «Тишина», либо «Шум» и он перемещается в новосозданные папки внутри проекта соответственно.

5. **Извлечение эмбеддингов (NVIDIA TitaNet).** Каждый сегмент речи из аудиофайлов обрабатывается для получения эмбеддингов с помощью модели для верификации говорящих. Эмбеддинги — векторные представления аудиофайлов, которые будут использоваться для дальнейшей кластеризации.

6. **Кластеризация (DBSCAN).** Полученные эмбеддинги используются для кластеризации аудиофайлов с помощью алгоритма DBSCAN. Кластеризация позволяет группировать похожие по характеристикам аудиофайлы.

7. **Обработка кластеров и присваивание им новых названий (Whisper).** Выполняется транскрибация аудиофайлов, извлекая из них текст с помощью модели Whisper и назначая кластерам названия на основе наиболее часто встречающихся фраз в файлах. В кластер сохраняется текстовый документ с результатами транскрибации и новым названием кластера (Рисунок 1).

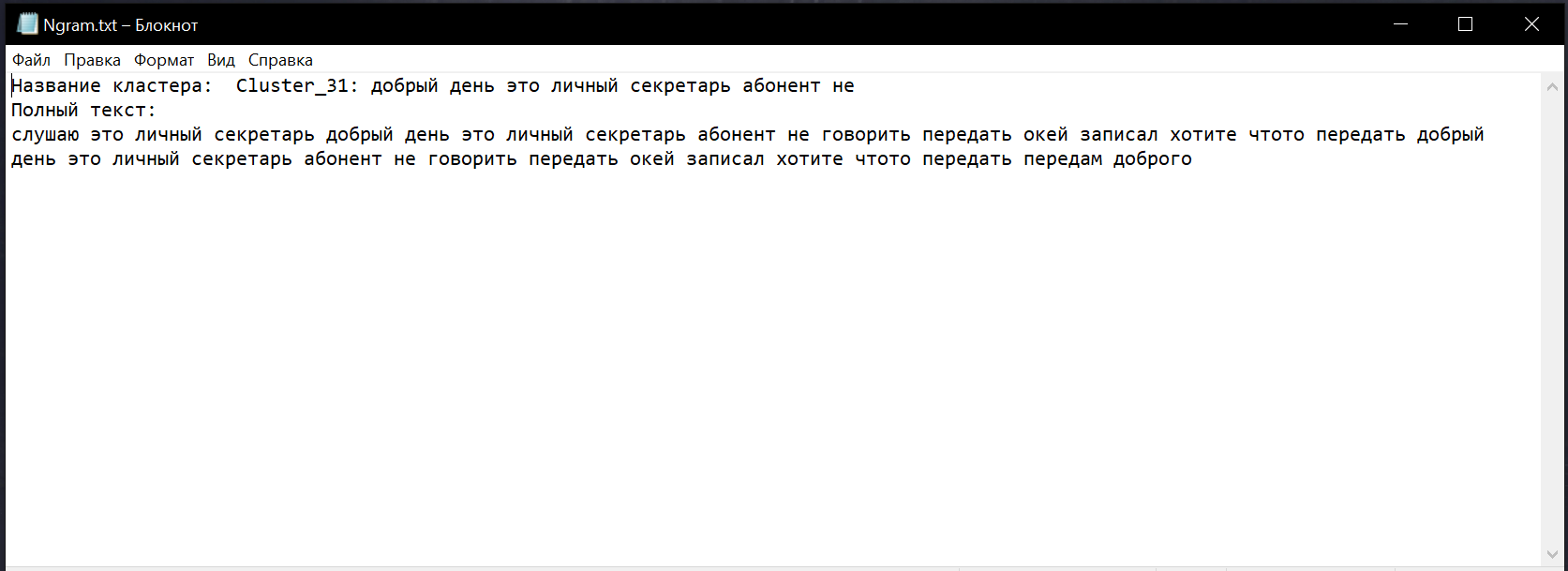


Рис. 1. Текстовый файл с результатами обработки кластера

8. **Сохранение кластеров.** После кластеризации файлы копируются в отдельные папки для каждого кластера.

## 2.5 ОПИСАНИЕ МОДУЛЕЙ

Для реализации функциональности системы разработаны отдельные модули, каждый из которых выполняет свою важную задачу. При реализации функций использовались доверенные источники информации с подробным описанием моделей и алгоритмов для облегчения работы с кодом. Ключевые модули, используемые в проекте:

* **process\_audio\_files**Модуль, отвечающий за предобработку аудиофайлов. Его основная задача состоит в очистки аудиофайла от шума и тишины между участками голосовой активности. Кроме того, модуль используется для фильтрации входных данных от аудиозаписей, не содержащих голосовой активности. Модуль имеет основное влияние на точность. Основываясь на статье[1, 2], было принято решение о выборе модели для данного модуля Silero VAD.
* **get\_all\_embeddings**

Модуль, использующий модель TitaNET, для получения эмбеддингов из всех аудиофайлов. Кроме того, в этом модуле производится кластеризация с помощью алгоритма DBSCAN. На качество работы данного модуля, ключевое влияние оказывает модуль предобработки, хоть в статье[3] и говорится о том, что модель TitanNET была предобучена на телефонных записях с шумами, но все же лучшую точность она показывает на очищенных данных.

* **print\_clustering\_metrics**

Данный модуль производит вычисление метрик, таких как “Индекс Дейвиса-Боулдина” и “Силуэтный коэффициент”, использующихся для оценки качества работы системы.

* **segment\_audio\_files**Модуль выполняет обработку аудиофайлов, разделяя их на сегменты на основе тишины, усиливая громкость и сохраняя полученные сегменты в соответствующую папку. Аудиофайл делится на несколько сегментов, основываясь на уровне тишины. Порог тишины рассчитывается динамически, в зависимости от средней громкости аудиофайла.
* **plot\_3d\_clusters\_plotly**

Модуль, отвечающий за визуализацию кластеров, он понижает размерность векторного представления аудиофайла до трёх и создает 3D пространство для наглядной демонстрации результатов кластеризации.

* **process\_clusters**

Модуль в основе своей использующий модель Whisper[4] для транскрибации аудиофайлов, находящихся в кластерах. К транскрибированным текстам применяется алгоритм N-gramm для выделения ключевых фраз, динамической длины, которая зависит от средней длины текста аудиофайла в кластере.

## 2.6 ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ ПРОЕКТА

1. Silero VAD (Voice Activity Detection)

Модель Silero VAD используется для детекции речи в аудио сигнале, что позволяет выделить сегменты, содержащие только речь, и исключить тишину или шум.

Принцип работы:

Входной аудиофайл преобразуется в спектрограмму. Модель Silero VAD использует предобученную сеть, чтобы классифицировать каждое окно в аудио как "речь" или "тишина".

Речь классифицируется как 1, а тишина как 0.

Модель VAD анализирует аудио-сигнал в виде последовательности окон, где каждое окно xt​ классифицируется с помощью функции активации:

где — вероятность того, что в окне присутствует речь W — обучаемый вес, b — смещение, σ — функция активации.

1. **DBSCAN**

Алгоритм DBSCAN используется для кластеризации аудиофайлов на основе эмбеддингов, полученных из модели Speaker Verification.

Для каждой точки p в наборе данных вычисляется её окрестность N(p) радиусом ε. Окружность определяется как множество точек, которые находятся на расстоянии, меньшем или равном ε от точки p:

где distance(p, q) — это расстояние между точками p и q.

Для двух точек расстояние между ними рассчитывается как евклидово расстояние:

где — это векторы признаков (эмбеддинги аудиофайлов), а

n — размерность. Если ≤ ϵ,  то эти точки принадлежат одному кластеру.

Точка p является ядром кластера, если количество точек в её окрестности

N(p, ε) больше или равно MinPts:

где MinPts: минимальное количество точек в окрестности для того, чтобы область считалась кластером.

1. **PCA (Principal Component Analysis)**

Для визуализации и уменьшения размерности эмбеддингов используется метод PCA (анализ главных компонент). Он позволяет проецировать многомерные данные на меньшую размерность (например, 3D) В PCA для каждого вектора признаков x ищется проекция на новое пространство с помощью матрицы собственных векторов V:

где — новый вектор признаков в пространстве главных компонент,

V — матрица собственных векторов.

1. **Метрики качества**

Для оценки качества кластеризации применяются следующие метрики:

* Силуэтный коэффициент:

где — среднее расстояние до остальных точек в том же кластере, а — минимальное среднее расстояние до точек в другом кластере.

* Индекс Дейвиса-Боулдина:

где — внутрикластерное расстояние для точки i, а — межкластерное расстояние для точки i, а — расстояние между центрами кластеров.

## 2.7 ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС ПРОЕКТА

### **Описание стартового интерфейса приложения**

На начальном экране приложения расположен **основной интерфейс**, который прост и интуитивно понятен для пользователя. Интерфейс включает следующие элементы:

1. **Начальное окно с кнопкой "Выбрать папку" (Рисунок 2):**

При нажатии на эту кнопку открывается стандартное **диалоговое окно выбора папки**. Пользователь выбирает папку, в которой находятся аудиофайлы (например, .wav или .mp3), которые будут обрабатываться. После успешного выбора папки интерфейс обновляется, и появляются новые кнопки для выбора дальнейших действий.

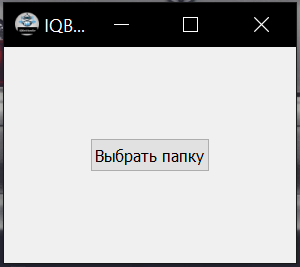


Рис. 2 - Начальное окно с кнопкой выбора папки с файлами

1. **Окно с выбором действия и полем логов (Рисунок 3):**

После выбора папки пользователь видит две активные кнопки:

**Кнопка "Кластеризация"**: Запускает процесс предобработки и кластеризации аудиофайлов. В результате кластеризации пользователю будут выданы кластеры с аудиофайлами.

**Кнопка "Сегментация"**: Запускает процесс сегментации аудиофайлов. Каждый отрезок с голосовой активностью будет извлечен и сохранен в отдельный файл, что позволяет работать с фрагментами аудио, где присутствует речь.

При нажатии одной из кнопки, запустится скрипт выполнения функций. Ниже, если нажать на кнопку “Показать подробностей”, откроется поле логов. В нём можно увидеть, на какой стадии сейчас выполняется процесс обработки аудиофайлов.

В течении обработки будет заполняться шкала прогресса, чтобы пользователь знал, сколько ему осталось ждать.

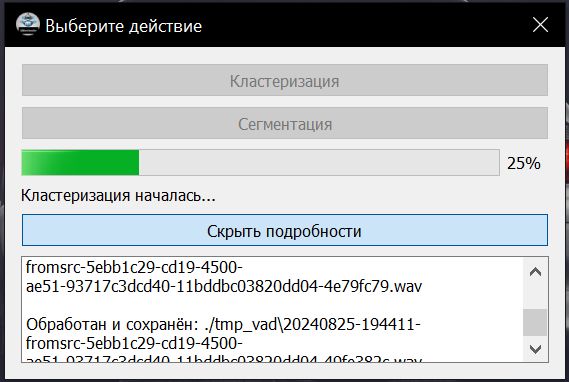


Рис. 3. Окно с выбором обработки аудиофайлов

1. **Визуализация (Рисунок 4):**

После завершения работы, пользователю выводится окно с визуализацией выполненной кластеризации дополнительных возможностей:

В окне располагается 3D-визуализация кластеров на локальном сервере. Визуализация строится с использованием данных, полученных в процессе кластеризации, и отображает разноцветные фигуры в трехмерном пространстве. Пользователь может перемещаться в 3D визуализации при помощи настроек управления DBSCAN через компьютерную мышь. Это позволяет легко сделать анализ файлов и выделить файлы с низкой точностью кластеризации.

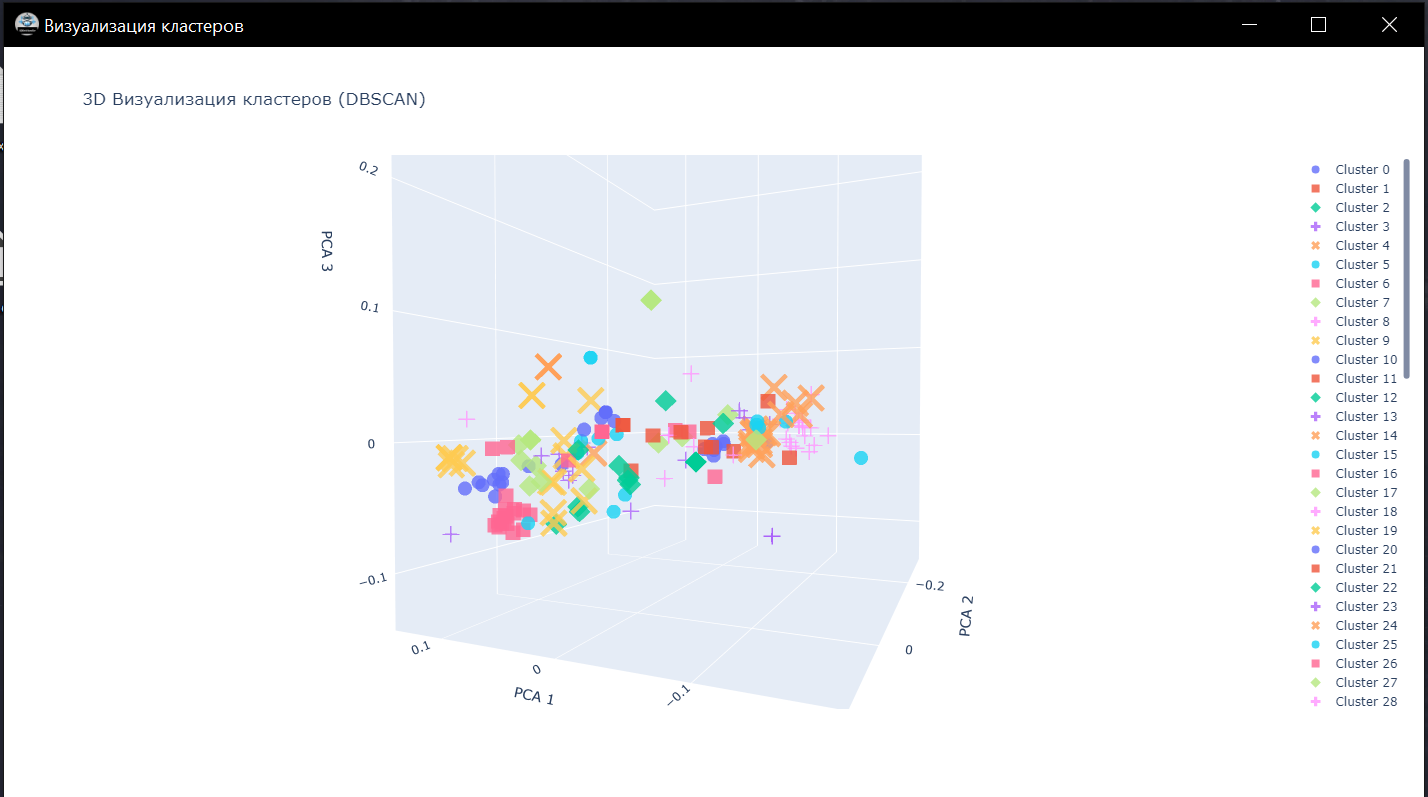


Рис. 4. Окно с визуализацией

1. **Итоговый результат**

В конце выполнения кластеризации все оригинальные аудиофайлы сохраняются локально в папку “clustered audio”. Кластеры имеют новые названия для удобного понимания содержимого внутри этих кластеров (Рисунок 5).

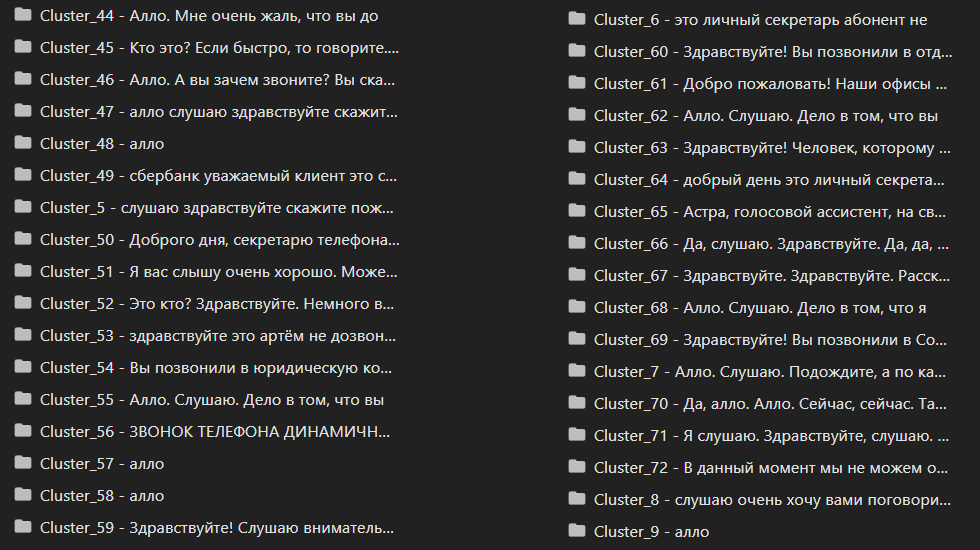


Рис. 5. Сохраненные кластеры с аудиофайлами

## 2.8 ОРГАНИЗАЦИЯ ТЕСТИРОВАНИЯ ПРОЕКТА

Тестирование проводилось на 400 аудиофайлах, которые были прослушаны и вручную размечены по кластерам. Совпадение либо приближение к истинному значению кластеров было основным показателем точность работы системы. Тестирование проводилось в четыре этапа**.**

**Каждый этап включает:**

* Предобработку данных с различными подходами.
* Извлечение эмбеддингов с помощью модели Titanet.
* Кластеризация данных.
* Замеры метрик качества кластеризации.
* Ручная проверка всех кластеров на правильность.

Таблица 1

График проведения тестирования

|  |  |  |
| --- | --- | --- |
| Этап | Дата | Действия |
| Первый этап | 10.09.2024 – 15.09.2024 | Проведение базового тестирования. Использование необработанных данных (частота семплирования аудио 8 000 Гц). Измерение базовых метрик кластеризации. |
| Второй этап | 03.10.2024 – 06.10.2024 | Добавление ресемплинга до 16 000 Гц. Проведение тестирования и замер результатов. |
| Третий этап | 01.11.2024 – 05.11.2024 | Внедрение базового VAD для удаления тишины и неречевых сегментов. Тестирование с минимальной настройкой VAD. |
| Четвертый этап | 11.11.2024 – 15.11.2024 | Проведение тонкой настройки параметров VAD. Измерение конечных метрик. |

**Оценка результатов тестирования**

Для оценки качества кластеризации использовались следующие способы и метрики:

* **Средний силуэтный коэффициент:**  
  Оценивает плотность кластеров и расстояние между ними. Значение от -1 до 1, где ближе к 1 — лучше.
* **Индекс Дейвиса-Боулдина**:  
  Оценивает компактность кластеров и их разделение. Меньшее значение означает лучшее качество кластеризации.
* **Количество кластеров**:  
  Сравнивается с истинным количеством классов, известным из ручной разметки.
* **Визуализация кластеров**:  
  Построение 3D-графиков с использованием PCA для анализа качества группировки файлов визуально.
* **Ручная проверка результатов:**Проверка результатов кластеризации, заключалась в ручном прослушивании кластеров, на предмет выявления аудиофайлов, которые находились в неправильных кластерах.

Таблица 2

Результаты тестирования системы

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Этап | Средний силуэтный коэффициент | Индекс Дейвиса-Боулдина | Количество кластеров | Истинное количество кластеров |
| Первый этап | 0,2945 | 0,5842 | 124 | 70 |
| Второй этап | 0,2810 | 0,5746 | 133 | 70 |
| Третий этап | 0,4059 | 0,5349 | 101 | 70 |
| Четвертый этап | 0,4793 | 0,5336 | 73 | 70 |

После анализа результатов представленных в Таблице 1, можно сделать следующие выводы:

* Использование необработанных аудиофайлов, ведет к плохим итоговым результатам.
* Улучшение метрик на третьем и четвёртом этапах показало, что использование VAD положительно влияет на качество кластеризации.
* На четвертом этапе достигнуты лучшие результаты благодаря тонкой настройке алгоритмов предобработки.
* Подход с использованием DBSCAN оказался наиболее эффективным при предварительной очистке данных от тишины и шума.

# ЗАКЛЮЧЕНИЕ

В проекте реализовано прикладное ПО для автоматической обработки автоответчиков с кластеризацией, сегментацией и локальным сохранением отсортированных по кластерам аудиозаписей. В работе было использовано несколько нейронных сетей, в совокупности которые показали великолепный результат. Для анализа текстов, которые были транскрибированы из аудиофайлов, используется алгоритм определения фраз с высокой частотностью появления. Команда выполнила все поставленные задачи, благодаря чему, проект полностью функционирует, а люди, для которых писался данный продукт, используют его для выполнения своих рабочих обязанностей.

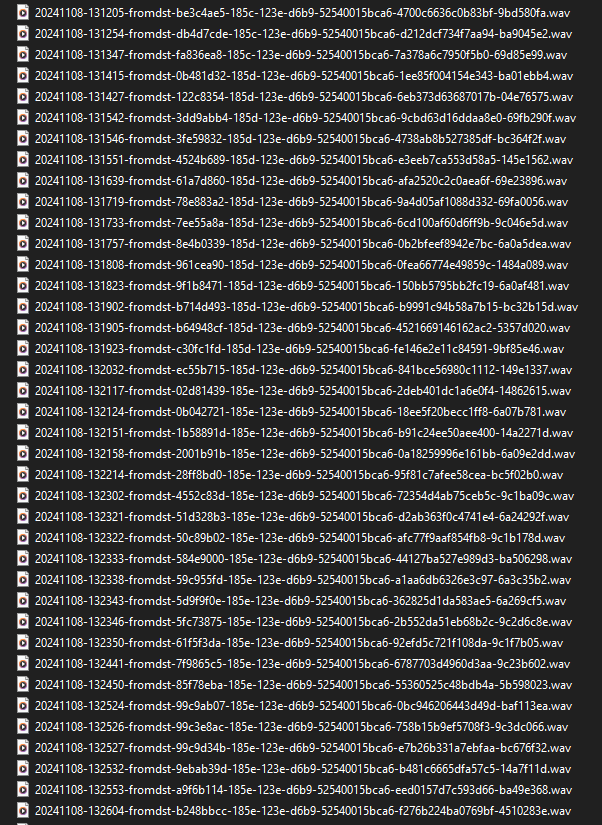
# СПИСОК ЛИТЕРАТУРЫ

1. А ты используешь VAD? Что это такое и зачем он нужен // Хабр URL: https://habr.com/ru/articles/594745/ (дата обращения: 12.10.2024).
2. Исследование звука: удаление шумов // Хабр URL: https://habr.com/ru/articles/793608/ (дата обращения: 22.09.2024).
3. TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context // arXiv URL: https://arxiv.org/abs/2110.04410 (дата обращения: 10.09.2024).
4. Распознавание речи (транскрибация) по аудиозаписям диалогов. Whisper. Личный опыт // Хабр URL: https://habr.com/ru/articles/768562/ (дата обращения: 10.11.2024).
5. Желтова К.А. Филимонов И.С. Егорова Д.В. Сравнение алгоритмов кластерного анализа для решения задачи поиска схожих аудиозаписей // Решетневские чтения. - 2018

# ПРИЛОЖЕНИЕ

Приложение 1

Пример входных данных в систему



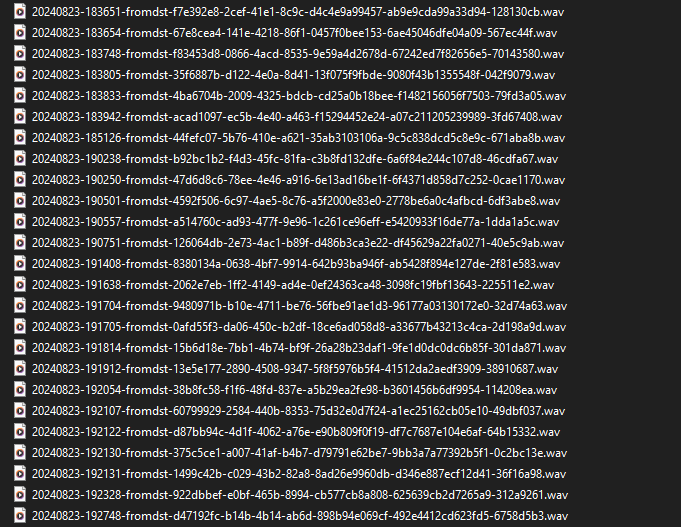
Приложение 2

Пример сегментированных аудиофайлов



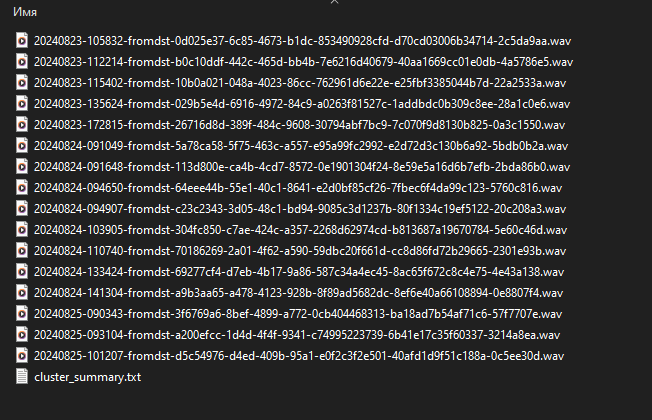
Приложение 3

Пример предобработанных аудиофайлов



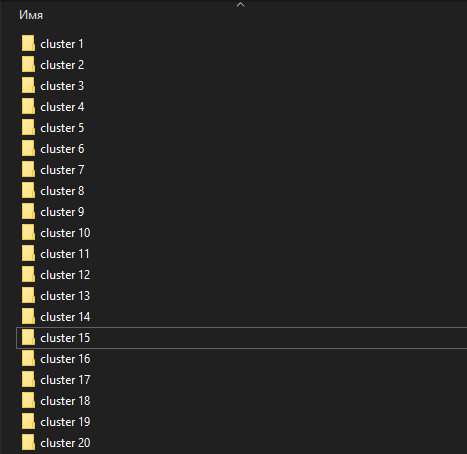
Приложение 4

Пример содержания одного кластера



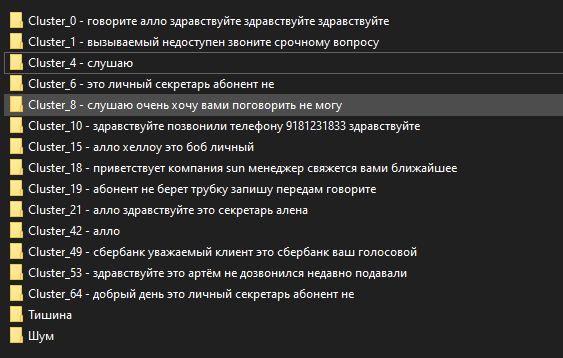
Приложение 5

Пример итогов работы кластеризации

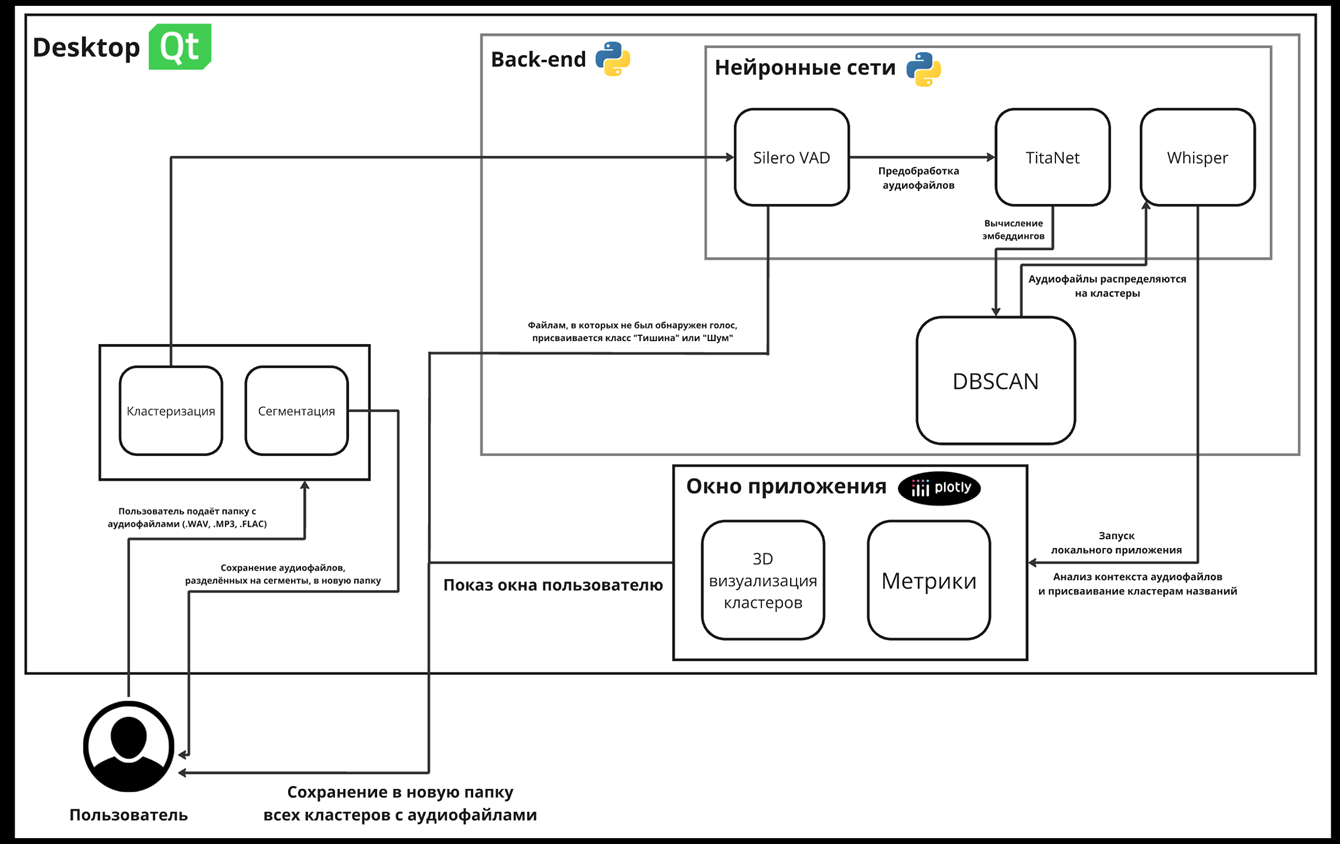


Приложение 6

Пример итогов работы проекта



Приложение 7

Схема архитектуры приложения