МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ ЗАПИСИ, ОБРАБОТКИ И КОНВЕРТАЦИИ АУДИО В ТЕКСТОВЫЕ ДАННЫЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Васильева Дмитрия Николаевича

Научный руководитель
ассистент, к. ф.-м. н. _____ Д. В. Мельничук

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы. Разработка приложений для обработки и конвертации аудио в текстовые данные является очень актуальной и востребованной в современном мире. Текстовые данные могут быть использованы в широком спектре задач, таких как создание электронных документов, аннотирование видео и аудио материалов, субтитры для кинофильмов, а также в задачах машинного обучения и искусственного интеллекта.

Приложения для конвертации аудио в текстовые данные имеют большой потенциал в медицинских и правовых сферах, где большое количество голосовых записей должны быть транскрибированы в текстовые документы для дальнейшего анализа. Также данное приложение может использоваться людьми с ограниченными возможностями, которые не могут слышать или кто имеет проблемы со слухом.

Целью бакалаврской работы является изучение сферы ASR моделей и разработка приложения для записи, обработки и конвертации аудио в текстовые данные.

Объектом исследования является процесс записи, обработки и конвертации аудио в текстовые данные с использованием различных технологий и методов программирования.

Предмет исследования – аудио-записи и их конвертация в текстовый формат при помощи разработанного приложения.

Для разработки приложения, выполняющего задачу конвертации аудио в текст, необходимо выполнить следующие задачи:

1. Изучение теоретических основ распознавания речи и акустических моделей.
2. Определение требований к приложению, выбор инструментов разработки и анализ возможных архитектур приложения.
3. Разработка системы записи аудио.
4. Разработка алгоритмов для обработки и анализа аудио в целях распознавания речи.
5. Реализация модели распознавания речи и ее интеграция в приложение.
6. Тестирование и отладка приложения с использованием различных да-

тасетов и аудиофайлов.

Практическая значимость работы заключается в создании инструмента, который позволит автоматически конвертировать аудио в текстовые данные. Это может быть полезно для таких сфер, как медицина, право, образование, бизнес, медиа и другие. Для медицинских заведений приложение может использоваться для транскрибирования голосовых записей диагностических процедур и интервью с пациентами, что позволит ускорить процесс медицинской документации и повысить качество предоставляемых услуг. В правовой сфере приложение может использоваться для транскрибирования судопроизводственных записей, что способствует повышению эффективности судебных процессов и ускоряет разрешение споров. В сфере бизнеса приложение может использоваться для транскрибирования встреч и переговоров, что поможет улучшить процесс принятия решений и повысить производительность.

Таким образом, практическая значимость работы заключается в создании универсального инструмента, который позволит повысить эффективность работы и удобство использования в широком спектре задач.

Структура и содержание работы. Работа состоит из введения, 3 разделов, заключения и списка используемых источников, содержащего 20 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** раскрывается актуальности темы работы, формулируется цель работы, задачи, которые необходимо решить, и отмечается практическая значимость результатов.

В **первом** разделе приводится теоретическая часть работы. Рассматриваются понятия ASR моделей, их применение в различных сферах, обозреваются существующие методы распознавания речи. Как и у любых моделей, у ASR моделей имеются преимущества и недостатки, они также отмечаются в рамках данного раздела.

Стоит упомянуть акустические модели и используемые алгоритмы в системах распознавания речи. Для создания ASR систем используются различные алгоритмы для обработки аудиосигнала, представления его в виде спектрограммы и выделения признаков, которые затем используются для обучения модели. Рассматриваются основные типы акустических моделей, такие как скрытые марковские модели (HMM), условные случайные поля (CRF) и рекуррентные нейронные сети (RNN), а также используемые алгоритмы обучения, такие как алгоритм обратного распространения ошибки (Backpropagation) и алгоритм распространения соответствия (Alignment Propagation).

Скрытые марковские модели (HMM) и динамическое искривление времени (DTW) - два таких примера традиционных статистических методов распознавания речи.

Используя набор транскрибированных аудиообразцов, HMM обучается предсказывать последовательности слов, изменяя параметры модели для максимизации вероятности наблюдаемой аудиопоследовательности.

DTW - это алгоритм динамического программирования, который находит наилучшую возможную последовательность слов путем вычисления расстояния между временными рядами: один представляет неизвестную речь, а другой - известные слова.

Кроме того, также выделяются подходы к обработке шумных акустических сигналов, такие как метод гауссовских смесей (GMM) и глубокое обучение, и применение ансамблей и комбинированных моделей для повышения точности распознавания.

Изучаются алгоритмы обработки и анализа аудио, необходимые для построения систем распознавания речи. В частности, рассматриваются методы предобработки аудиоданных, такие как фильтрация шума и улучшение качества записи, а также алгоритмы извлечения признаков, включая Дискретное Преобразование Фурье (DFT), Мел-частотные кепстральные коэффициенты (MFCC) и Преобразование Вейвлета (Wavelet Transform). Далее, рассматриваются алгоритмы классификации для распознавания речи, включая методы гауссовского смесевого моделирования, скрытых Марковских моделей и нейросетевые подходы (сверточные нейронные сети, рекуррентные нейронные сети и т.д.).

Во **втором** разделе описывается анализ существующих моделей. Были выбраны исходные данные - аудиофайлы, для того чтобы посмотреть как различные модели на рынке смогут транскрибировать аудио в текст. Необходимо было узнать с какой скоростью и с какой точностью справляется та или иная модель.

Производится анализ применимости разработанного приложения в различных областях и сферах применения. Для этого необходимо произвести оценку потенциальных потребностей различных групп пользователей, а также выявить преимущества и недостатки разработанного приложения. Примерами областей применения могут служить обучение и образование, здравоохранение и медицина, продуктивность и бизнес, мультимедиа и развлечения и многие другие. Для каждой из этих областей нужно определить, насколько востребовано приложение для распознавания речи.

Для того чтобы понять, в каких именно сферах и отраслях может быть востребовано приложение для транскрибирования аудио в текст, был проведён опрос целевой аудитории.

Опрос содержал вопросы о том, как часто или в какой степени респонденты работают с аудиофайлами, нужно ли им транскрибировать речь на работе и для каких целей, насколько важны для них точность и скорость распознавания речи.

Анализ результатов опроса помог понять, в каких конкретных областях и сферах наиболее актуальна и востребована технология. Результаты опроса помогут определить приоритетные зоны развития продукта.

Результаты опроса показали, что подобное приложение наиболее востребовано студентами гуманитарных и социальных специальностей. Они отметили, что часто сталкиваются с необходимостью переписывать лекции и выступления научных сотрудников, чтобы лучше запомнить их содержание. Приложение для транскрибации позволило бы им значительно сократить время и уменьшить усилия, затрачиваемые на эту задачу.

Преподаватели также поддержали идею использования данного приложения в учебном процессе. Они указали, что транскрибация аудио в текст может помочь студентам с разным уровнем слуховой памяти более эффективно усваивать информацию и принимать к концу курса.

В целом, опрос продемонстрировал, что приложение для транскрибации аудио в текст будет полезным в образовательном процессе для большинства студентов и преподавателей, и предоставит им возможность сэкономить время и улучшить качество обучения.

В результате исследования удалось узнать способности каждой модели и выделить оптимальный вариант для приложения. В данном случае будем использовать готовую модель whisper, которая будет интегрирована в приложение.

Whisper - это модель, основанная на сверточных нейронных сетях, разработанная для обработки аудиоданных и специализирующаяся на распознавании шепота. Авторы whisper идут по пути компромисса, используя weakly supervised подход. Взяли все доступные аудио с транскрипцями из интернета и пофильтровали их, но не сильно. Получили некий зашумленный датасет, в котором в том числе есть и транскрипции сделанные другими ASR системами, много тишины и шумов, смех, апплодисменты и т.д. Объем получился 680 000 часов на 97 языках, из которых 117 000 часов не на английском. Обучение на таком большом зашумленном датасете, по мнению авторов, дает значительное улучшение обобщающей способности модели и её устойчивости к посторонним звукам.

Почему была выбрана именно эта модель? Библиотека Whisper поддерживает обработку аудиофайлов на более чем 80 языках, включая английский, испанский, китайский, русский, японский, корейский и т.д. Кроме того, библиотека имеет возможность работать с различными диалектами языков,

что делает ее еще более универсальной. Модель преобразования последовательности в последовательность обучается на различных задачах обработки речи, включая распознавание многоязычной речи, перевод речи, идентификацию разговорного языка и определение голосовой активности. Эти задачи совместно представляются в виде последовательности дескрипторов, которые должны быть предсказаны декодером, что позволяет одной модели заменить многие этапы традиционного конвейера обработки речи. Формат многозадачного обучения использует набор специальных дескрипторов, которые служат в качестве спецификаторов задач или целей классификации.

Исследуется разработка системы записи аудио. Это включает в себя выбор и настройку аудиоустройств, используемых для записи, определение формата записываемых файлов, установку частоты дискретизации и битовой глубины, а также разработку программного интерфейса для управления записью и сохранения аудиофайлов в нужном формате. Разработка системы записи аудио является важным этапом при создании систем распознавания речи, так как от качества записи зависит точность распознавания.

В качестве разработки был выбран язык программирования Python, так как этот язык имеет множество инструментов и библиотек для обработки речи и его распознавания. Кроме того, Python имеет простой синтаксис, что упрощает написание кода, и обширное сообщество, которое быстро поможет решить любую проблему в процессе разработки. Также Python является многоплатформенным языком и может работать на любой операционной системе, что делает его универсальным инструментом для разработки приложений.

В качестве инструмента записи звука используется библиотека PyAudio. Она обеспечивает возможность записи и воспроизведения аудио, а также предоставляет доступ к определенным параметрам звуковой карты, таким как частота дискретизации, битрейт, количество каналов и т.д.

В приложении библиотека PyAudio используется для записи аудиофайла, который будет проходить траскрибирование. Она обеспечивает возможность записи звука с микрофона, сохранения и чтения аудиофайлов, а также применения различных фильтров и определения различных параметров аудио-потока. PyAudio является одной из наиболее популярных библиотек для работы с аудио в Python благодаря своей открытой и простой в исполь-

зовании структуре и большому количеству доступных функций.

Далее происходит разработка пользовательского интерфейса приложения. Здесь мы будем создавать графический интерфейс, который будет взаимодействовать с моделью распознавания речи.

Для создания интерфейса приложения была использована библиотека PyQT5, которая предоставляет множество инструментов для создания пользовательских интерфейсов в Python. PyQT5 обеспечивает интеграцию Python со стандартной библиотекой Qt, что позволяет создавать красивые, функциональные, быстрые и масштабируемые пользовательские интерфейсы.

PyQt5 был выбран по нескольким причинам:

1. Большое сообщество пользователей и разработчиков: существует множество форумов, сообществ и ресурсов, ориентированных на Pyqt5, что позволяет быстро получить ответ на любой вопрос и решить проблемы, связанные с разработкой.
2. Возможность создавать интерфейс с помощью "визуального"редактора: Pyqt5 предлагает бесплатный инструмент Qt Designer для создания графического интерфейса. Это упрощает процесс разработки и позволяет быстро создавать и изменять интерфейс без необходимости переписывать код.
3. Мультиплатформенность: приложения, созданные с помощью Pyqt5, могут работать на операционных системах Windows, MacOS, Linux, а также на Android и iOS.
4. Широкий выбор возможностей: Pyqt5 позволяет создавать приложения с разнообразными функциями и элементами управления.
5. Большое количество материалов и документации: по Pyqt5 есть множество статей, книг и документации, что делает процесс изучения и использования библиотеки более удобным и доступным.

При проектировании было важно сделать приложение минималистичным, поэтому на главном экране для управления записью располагается всего 2 кнопки - record, stop. С уже траснрибированным файлом можно произвести несколько операций: удаление, переименование, чтение.

Также в приложении была добавлена возможность загружать свой файл для транскрибации, в случае если пользователь не хочет записывать аудио-

файл в ту же секунду, а использовать готовый вариант записи. Важно было определить, как ведёт себя приложение в реальных условиях. Для тестирования моделей были взяты аудиофайлы с речью. Было проведено несколько тестов с разным уровнем сложности whisper. После проведения тестирования были сделаны выводы об эффективности приложения и его конкурентноспособности.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Разработана модель для распознавания речи с использованием существующих методов ASR и алгоритмов обработки и анализа аудио;
2. Решена задача интеграции модели в приложение;
3. Создан пользовательский интерфейс приложения с помощью библиотеки PyQt5 для удобства работы с распознаванием речи;
4. Проведено тестирование и отладка приложения;
5. Оценена качества полученной модели с помощью соответствующих метрик (WER, WRR);
6. Проведен анализ применимости разработанного приложения в различных областях.