Применение современных языковых моделей для автоматической транскрибации и анализа аудиозаписей телефонных разговоров сотрудников отдела продаж с клиентами

А.А. Буткина, С.А. Фирсова, И.А. Шеметов

МГУ им. Н.П. Огарёва, Саранск

Аннотация: Статья посвящена изучению возможностей автоматической транскрибации и анализа аудиозаписей телефонных разговоров сотрудников отдела продаж с клиентами. Актуальность исследования связана с ростом объема голосовых данных и потребностью в их быстрой обработке в организациях, деятельность которых тесно связана с продажей своих продуктов или услуг клиентам. Автоматическая обработка аудиозаписей позволит провести проверку качества работы сотрудников call-центров, определяя допущенные нарушения в скриптах разговоров с клиентами. Предложенное программное решение основано на использовании модели Whisper для распознавания речи, библиотеки pyannote.audio для диаризации спикеров, а также библиотеки RapidFuzz для организации нечёткого поиска при проведении анализа строк. В ходе экспериментального исследования, проведенного на базе разработанного программного решения, было подтверждено, что использование современных языковых моделей и алгоритмов позволяет добиться высокой степени автоматизации обработки аудиозаписей и может использоваться в качестве инструмента предварительного контроля без участия специалиста. Полученные результаты подтверждают практическую применимость используемого авторами подхода для решения задач контроля качества в отделах продаж или call-центрах.

Ключевые слова: call-центр, аудиофайл, распознавание речи, транскрибация, диаризация спикеров, классификация реплик, обработка аудиозаписей, Whisper, pyannote.audio, RapidFuzz.

Введение

В настоящее время обработка аудиофайлов становится все более актуальной задачей – автоматический перевод голосовых сообщений в текст применяется в самых разных сферах: от обработки звонков в call-центрах до юридических процессов, медицинских консультаций и образовательных сервисов [1, 2]. С ростом объема голосовых данных и потребностью в их быстром анализе традиционные методы транскрибации (например, стенографирование) становятся неэффективными.

Существующие коммерческие сервисы транскрибации, такие как Google Speech-to-Text, IBM Watson Speech-to-Text и Microsoft Azure Speech-

to-Text либо недоступны в России, либо являются слишком дорогостоящими программными решениями для малого и среднего бизнеса [3, 4]. Кроме того, зависимость от зарубежных облачных платформ влечет такие риски их использования, как ограничение доступа к информации или снижение уровня защиты данных. Описанные аспекты делают исследования в области автоматической транскрибации русскоязычных аудиофайлов с использованием бесплатных сервисов и библиотек актуальными.

Целью проводимого исследования является изучение возможностей современных интеллектуальных языковых моделей для автоматической транскрибации и анализа аудиозаписей телефонных разговоров сотрудников отдела продаж с клиентами.

Обзор источников по теме исследования

Транскрибация аудиосообщений – процесс преобразования устной речи в текстовый формат – находит широкое применение в различных сферах, включая журналистику, медицину, образование и бизнес. С развитием технологий автоматического распознавания речи (ASR) и обработки естественного языка (NLP) эффективность и точность транскрибации значительно возросли. В статье [5] автор выделяет, что ASR отвечает за преобразование голосовых команд в текст, а NLP — за смысловую расшифровку уже полученного текста. Как подчёркивает автор, выбор конкретного метода зависит от задачи и условий, в которых работает система.

Обзор современных моделей нейронных сетей для обработки естественного языка представлен в статье [6]. В статье [7] подробно рассмотрены современные opensource-решения и изучены их возможности в решении описанных проблем транскрибации. Описаны четыре наиболее популярные открытые платформы: Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0, проведено сравнение архитектур и особенностей данных

моделей, что дает представление об их возможностях и ограничениях.

Современные системы транскрибации аудио- и видеоконтента рассматриваются в статье [8]. Отмечается, что разработка собственной системы может значительно упростить работу сотрудников, ускорить обработку информации и обеспечить наглядное представление данных. Например, авторами статьи [9] предложен вариант программного комплекса, позволяющий по методу Whisper производить транскрибацию речевых фрагментов регламента служебных переговоров, разработана схема выделения паттернов для аппаратного контроля правильности ведения регламента служебных переговоров на железнодорожном транспорте РФ.

В статье [10] разработан метод создания моделей автоматического распознавания речи для специализированных предметных областей, который включает этап промежуточного обучения лексике предметной области на данных из открытых источников, отобранных с использованием тематического семплирования.

Материалы и методы

В рамках данного исследования используются аудиофайлы телефонных разговоров различной длины, качества и содержания в формате MP3, представляющие собой записи реальных диалогов между сотрудниками отдела продаж и клиентами. Для качественной обработки записей учитывались следующие аспекты:

– частота дискретизации для телефонных линий составляет 8–16 кГц;

– в записи могут присутствовать посторонние шумы, эхо, артефакты сжатия, перекрытие голосов, слабая слышимость речи из-за плохой связи;

– анализируемая речь собеседников может содержать различные акценты, специфическую лексику и разные темпы говорящих.

Используемые методы исследования можно разбить на три основные группы:

1) Предобработка аудиоданных перед выполнением транскрибации:

– фильтрация шума: подавление фоновых помех и улучшение разборчивости речи (методы спектрального вычитания и Wiener-фильтрация);

– выравнивание громкости: нормализация уровня звука для минимизации ошибок распознавания (RMS-нормализация).

2) Распознавание речи и определение участников диалога проводилось с использованием связки нейросетевых решений от OpenAI и Pyannote:

– Whisper (модель large-v2) от OpenAI применяется для высокоточного распознавания русскоязычной речи. Данная модель работает на CPU и GPU, поддерживает сегментацию речи, удаление тишины и возврат таймкодов, а благодаря высокой точности является особенно эффективной при обработке сложных и насыщенных диалогов.

– Библиотека pyannote.audio (speaker-diarization-3.1) используется для диаризации (разделения речи на сегменты по спикерам). В рамках анализа клиентских звонков эта процедура необходима для разделения реплик между сотрудником компании и клиентом, что в дальнейшем позволяет корректно анализировать структуру диалога, стиль общения и поведение каждого участника. Модель загружается через Pipeline.from_pretrained() и позволяет локально или с использованием GPU выделять реплики разных участников.

– Библиотека torchaudio применяется для загрузки и предварительной обработки аудиофайлов перед анализом.

3) Оценка качества транскрибации выполнялась с применением следующих метрик качества:

– WER (Word Error Rate) – доля ошибочно распознанных слов;

– CER (Character Error Rate) – процент ошибок на уровне символов;

– DER (Diarization Error Rate) – процент времени, неправильно определенное как время речи. Данный показатель рассчитывается как сумма трёх типов ошибок (пропущенная речь, ложные срабатывания, ошибки определения спикера) по отношению к общему времени речи.

Описание алгоритма работы программного решения

Перечисленные выше технологии были использованы при реализации программного решения, которое позволило обеспечить автоматизацию проведения исследования. Схематическое представление алгоритма его работы представлено на рис. 1.

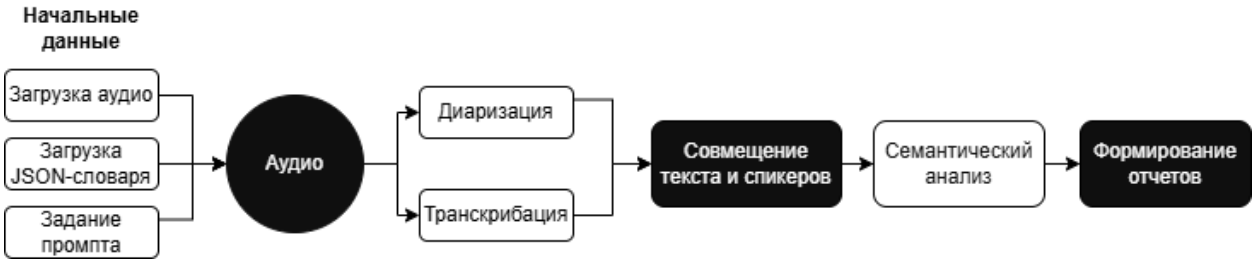

Рис. 1 – Алгоритм работы разработанного программного решения

В качестве **начальных данных** для работы приложения требуется:

1) Загрузка аудиофайла в формате mp3;

2) Загрузка json-словаря для анализа аудиозаписей. По умолчанию используется набор ключевых фраз, разделенных по тематическим категориям (например, «вежливость», «запрещённые выражения», «ключевые слова продаж» и др.), который сохранен в базе данных. Для более точной настройки словаря под индивидуальные требования пользователя предоставляется возможность загрузить собственный словарь с выделением значимых категорий и фраз в них через интерфейс приложения.

3) Для указания контекста транскрибации пользователь может написать промпт, содержащий описание ситуации и тематики разговора, а также ролей его участников. Данный параметр является необязательным.

После загрузки начальных данных выполняется процесс **транскрибации** с помощью OpenAI Whisper large-v2:

1) Загрузка модели. Первоначально происходит загрузка предварительно сохранённой модели Whisper large-v2 с локального диска. Модель запускается с указанием устройства вычислений (например, GPU или CPU), заданного через переменную device.

2) Запуск транскрибации. Выполняется с помощью метода transcribe() с параметром language="ru", указывающем модели на то, что речевые данные представлены на русском языке. Модель выполняет распознавание речи и возвращает полный транскрибированный текст и список временных интервалов (segments) с распознанными фрагментами.

3) Измерение времени выполнения для анализа производительности. Это позволяет при необходимости отследить нагрузку и скорость работы алгоритма на текущей конфигурации оборудования.

На рис. 2 представлен результат транскрибации тестового аудиофайла продолжительностью 61 с., в качестве которого был использован аудиозвонок клиента в клуб «Арена виртуальной реальности».

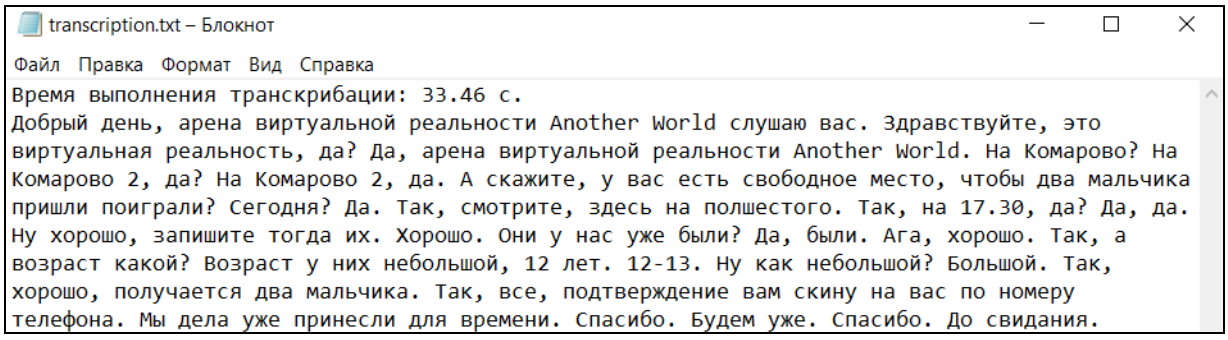

Рис. 2 – Результат транскрибации аудиофайла в формате txt

На следующем этапе выполняется диаризация речи с использованием нейросетевой модели pyannote/speaker-diarization-3.1, загружаемой через библиотеку pyannote.audio. Диаризация проводится с указанием количества участников диалога и словаря соответствий: условный идентификатор → роль («Сотрудник», «Клиент»). Результатом диаризации тестового аудиофайла, фрагмент которого представлен на рис. 3, являются сегменты,

каждый из которых содержит начальное и конечное значение времени, в течение которого говорит конкретный спикер.

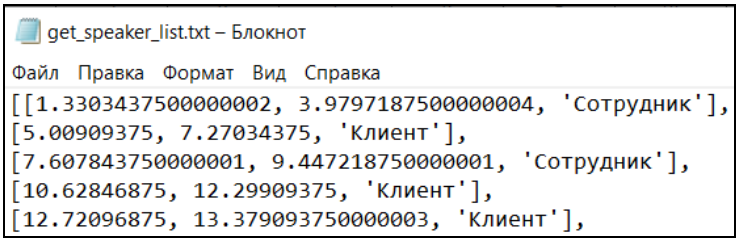

Рис. 3 – Фрагмент вывода результата диаризации

Далее происходит совмещение результатов двух предыдущих шагов, целью которого является получение удобного для анализа диалога между участниками с явной маркировкой реплик каждого. Для этого выполняется:

1) Сопоставление текста и спикеров на основе временных меток из результатов диаризации и транскрибации. Были предусмотрены случаи неточного наложения сегментов, для чего использовалась гибкая система поиска соответствующего спикера в расширенном диапазоне времени.

2) Формирование отчета, содержащего структурированную информацию, собранную по всем этапам обработки. Данный отчет может быть использован как для внутреннего анализа, так и для внешнего аудита или мониторинга. Предусмотрены две формы отчета: табличная – с ролями, таймкодами и текстом, и диалоговая – в виде последовательных реплик участников с временными отметками (рис. 4).

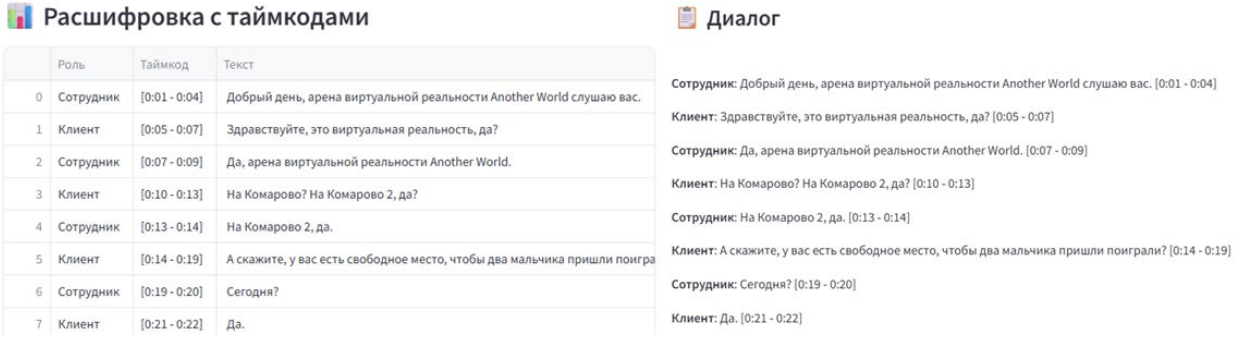

Рис. 4 – Вывод отчета по аудиозаписи в виде таблицы и диалога

На следующем этапе выполняется семантический анализ транскрибированного текста, который выявляет особенности речевого поведения, например, наличие слов-паразитов, соблюдение стандартов общения, использование недопустимых формулировок или запрещённой лексики. Для этого используются JSON-словари, в которых слова могут быть разбиты по следующим категориям: приветствие и прощание, слова-паразиты, скриптовые фразы (например, «могу ли я вам помочь»), ключевые фразы клиента (например, «хочу оформить», «меня интересует»), фразы, нарушающие скрипт и другие.

Для поиска ключевых фраз в тексте применяется метод нечёткого сравнения (fuzzy matching), реализованный с помощью библиотеки rapidfuzz. Такой подход позволяет учитывать возможные отклонения от точной формулировки – в том числе орфографические ошибки, синонимы и особенности произношения. Анализ осуществляется раздельно для реплик сотрудника и клиента. Работа алгоритма строится следующим образом: транскрибированный текст разбивается на отдельные слова, с помощью скользящего окна, длина которого соответствует количеству слов в искомой фразе, проводится сравнение с эталонными выражениями. Если совпадение превышает заданный порог (85%), фраза считается найденной. Подсчёт ведётся по количеству вхождений каждой фразы в каждой категории (рис. 5).

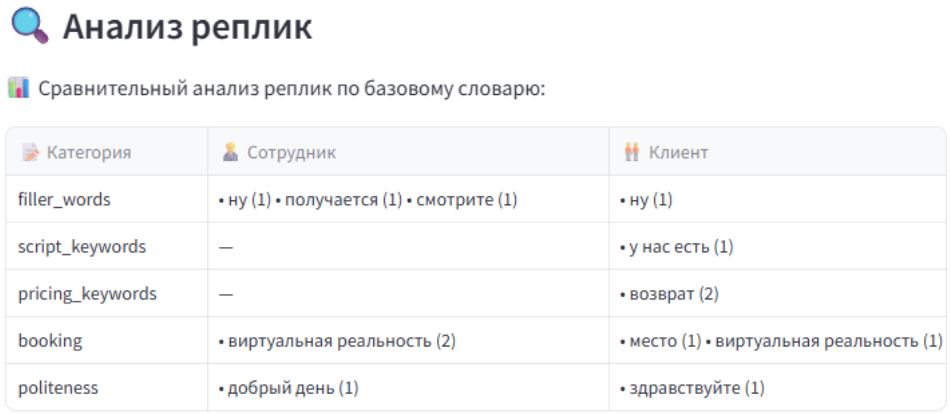

Рис. 5 – Анализ реплик по базовому JSON-словарю

Результаты исследования

Для оценки эффективности предлагаемого подхода анализа аудиозаписей была проведена серия экспериментов с использованием реальных аудиофайлов различной длины, качества и содержания. В тестировании участвовали 150 записей длительностью от 30 секунд до 5 минут, содержащие разговоры между клиентами и сотрудниками клуба «Арена виртуальной реальности». Результаты представлены в табл. 1.

Таблица №1

Результаты серии экспериментов

Показатель	**Значение**
Количество протестированных записей	150
Средняя длина аудио	2 мин 14 сек
Среднее время обработки файла	~100.6 сек
Средняя точность транскрибации (Whisper Large-v2)	**86%**
WER	**4%**
CER	**9%**
Точность диаризации (pyannote.audio)	**96%**
DER	**6%**
Найдено ключевых фраз (всего)	812
Ложных срабатываний в поиске фраз (rapidfuzz)	~7.3%
Аудио с серьёзными ошибками	4 из 150 (3%)

После анализа результатов экспериментов были сделаны следующие выводы:

1) Предложенный авторами подход для автоматической транскрибации и анализа аудиозаписей телефонных разговоров показал устойчивую и надёжную работу в реальных условиях. Полученные данные подтверждают его практическую применимость для задач контроля качества в отделах продаж или call-центрах.

2) Использование современных языковых моделей и алгоритмов позволяет добиться высокой степени автоматизации обработки аудиозаписей, а fuzzy-поиск обеспечивает адекватную точность при использовании качественных словарей.

– Whisper продемонстрировал высокую устойчивость к неидеальным условиям записи, однако стоит отметить, что качество транскрибации падало при наличии фоновых шумов, перекрывающихся реплик, неразборчивой речи или технических искажений.

– pyannote.audio в большинстве случаев корректно разделял роли собеседников, однако иногда происходило объединение реплик одного участника с другим в случаях, когда один из собеседников говорит значительно дольше другого, между фразами отсутствуют чёткие паузы, а также в диалогах с быстрым чередованием коротких фраз.

– Fuzzy-поиск (rapidfuzz) с порогом 85% оказался эффективным: он находил варианты фраз с незначительными искажениями, что важно при анализе естественной речи.

Обобщая вышесказанное, можно утверждать, что использование модели Whisper для автоматической транскрибации в сочетании с инструментом pyannote.audio для диаризации обеспечило приемлемое качество распознавания речи и разделения реплик между участниками диалога. Такая связка позволяет эффективно проводить последующий анализ, включая выявление ключевых фраз, проверку соответствия речевых конструкций заданным скриптам и фиксацию отклонений от установленных стандартов общения. Также на ее основе можно реализовать систему оценки качества коммуникаций, которая будет актуальной в тех организациях, где требуется контроль за соблюдением стандартов обслуживания клиентов.

Литература

1. Коробкин Д. М., Фоменков С. А., Брызгалин Н. А., Васильев А. А. Автоматизация распознавания заявок радиослушателей // Инженерный вестник Дона, 2024, № 7 URL: ivdon.ru/ru/magazine/archive/n7y2024/9338.

2. Орлова Ю. А., Дмитриев А. С., Колчева Д. В. Адаптация модели распознавания речи для упрощения редактирования исходного кода

программ для ЭВМ с мобильных устройств // Инженерный вестник Дона, 2021, №2. URL: ivdon.ru/ru/magazine/archive/n2y2021/6822.

3. Kim J.Y., Liu C., Calvo R.A., McCabe K., Taylor S.C.R., Schuller B.W., Wu K. A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech // arXiv preprint arXiv: 1904.12403. 2019 URL: arxiv.org/abs/1904.12403.

4. Filippidou F., Moussiades L. A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems // IFIP Advances in Information and Communication Technology. 2020. Vol. 584. pp. 73–82.

5. Ангапов В. Д. Анализ методов распознавания голоса в голосовых помощниках // Проблемы современной науки и образования. 2023. № 1. С. 69-75. URL: elibrary.ru/download/elibrary_59904781_45364663.pdf.

6. Казакова М. А., Султанова А. П. Анализ технологии обработки естественного языка: современные проблемы и подходы // Advanced Engineering Research (Rostov-on-Don). 2022. Т. 22, № 2. С. 169-176.

7. Баруздин М.М., Раскатова М.В., Щёголев П. Развитие современных систем транскрибации аудио- и видеоконтента // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2024. № 4. С. 71-78.

8. Елизаров Д. А. Разработка системы транскрибации аудио-и видеоконтента // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2023. № 4. С. 87-95.

9. Чернышев К. А., Лысов Г. М. Транскрибация по методу Whisper и выделение паттернов в задаче анализа регламента служебных переговоров на железнодорожном транспорте Российской Федерации // Транспортное дело России. 2023. № 5. С. 305-307.

10. Нечаев В. А., Косяков В. А. Метод разработки моделей распознавания речи для использования в информационных системах

энергетики // Вестник Ивановского государственного энергетического университета. 2023. № 4. С. 94-100.

References

1. Korobkin D. M., Fomenkov S. A., Bry`zgalin N. A., Vasil`ev A. A. Inzhenernyj vestnik Dona, 2024, №7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9338.

2. Orlova Yu. A., Dmitriev A. S., Kolcheva D. V. Inzhenernyj vestnik Dona, 2021, №. 2. URL: ivdon.ru/ru/magazine/archive/n2y2021/6822.

3. Kim J.Y., Liu C., Calvo R.A., McCabe K., Taylor S.C.R., Schuller B.W., Wu K. A. arXiv preprint arXiv:1904.12403. 2019. URL: arxiv.org/abs/1904.12403.

4. Filippidou F., Moussiades L. IFIP Advances in Information and Communication Technology. 2020. Vol. 584. pp. 73–82.

5. Angapov V. D. Problemy` sovremennoj nauki i obrazovaniya. 2023. №. 1. pp. 69-75. URL: elibrary.ru/download/elibrary_59904781_45364663.pdf.

6. Kazakova M. A., Sultanova A. P. Advanced Engineering Research (Rostov-on-Don). 2022. V. 22, no. 2. pp. 169−176.

7. Baruzdin M.M., Raskatova M.V., Shhyogolev P. Vestnik Rossijskogo novogo universiteta. Seriya: Slozhny`e sistemy`: modeli, analiz i upravlenie. 2024. № 4. pp. 71-78.

8. Elizarov D.A. Vestnik Rossijskogo novogo universiteta. Seriya: Slozhny`e sistemy`: modeli, analiz i upravlenie. 2023. № 4. pp. 87-95.

9. Cherny`shev K. A., Ly`sov G. M. Transportnoe delo Rossii. 2023. № 5. pp. 305-307.

10. Nechaev V. A., Kosyakov V. A. Vestnik Ivanovskogo gosudarstvennogo e`nergeticheskogo universiteta. 2023. № 4. pp. 94-100.

Дата поступления: 7.06.2025 Дата публикации: 25.07.2025