УДК 004.8
DOI: 10.22213/2410-9304-2026-1-52-63

Разработка автоматической системы по расшифровке голосовых записей встреч в компании с помощью нейронных сетей

М. А. Полозов, Воронежский государственный университет инженерных технологий, Воронеж, Россия
Л. А. Коробова, кандидат технических наук, доцент, компания «Технопарк-В», Воронеж, Россия

В условиях увеличения доли удаленной работы и цифровой трансформации бизнеса все большее значение приобретает автоматизация рутинных процессов, включая обработку аудиозаписей совещаний и встреч. Современные системы видео-конференц-связи предлагают функции автоматической транскрибации, однако они зачастую ограничены платными тарифами, требуют подключения к интернету и не обеспечивают достаточного уровня конфиденциальности. В связи с этим актуальной становится разработка локального, экономически доступного и безопасного решения для расшифровки речи. В данной работе представлена автоматическая система по расшифровке голосовых записей встреч в проектной компании. Особенностью разработанной системы является интеграция открытых нейросетевых моделей: Whisper (для распознавания речи), pyannote.audio (для диаризации – идентификации спикеров) и GPT-oss (для постобработки и форматирования текста). Система реализована в гибридной архитектуре с использованием двух языков программирования Python и C#, что позволило совместить высокую производительность обработки аудио с удобным графическим интерфейсом. Ключевые преимущества решения – полная автономность (без подключения к облаку), поддержка русского языка, масштабируемость и соответствие требованиям информационной безопасности. Тестирование на контрольном аудиофрагменте показало значение метрик WER (Word Error Rate – коэффициент ошибок на уровне слов) и CER (Character Error Rate – коэффициент ошибок на уровне символов) на уровне, приемлемом для делового использования. Для оценки точности работы спроектированной системы были проведены дополнительные тесты в различных акустических ситуациях, которые показали, что система обеспечивает хорошее качество транскрибации в типичных условиях эксплуатации, а также при наличии фоновых шумов. Реализованный программный продукт позволит компаниям экономить на платном доступе к системам видео-конференц-связи и корпоративных подписках, одновременно повышая прозрачность и эффективность документирования встреч и совещаний. Представленная работа имеет как теоретическую, так и практическую значимость для развития отечественных ИТ-решений в сфере корпоративной автоматизации.

Ключевые слова: нейронные сети, автоматическая расшифровка аудиозаписей, нейросетевые модели, распознавание речи, диаризация, транскрибация, гибридная архитектура, локальная обработка, корпоративная безопасность, эффективное документирование, цифровая трансформация.

.

Введение

Современные системы видео-конференц-связи (TrueConf, Яндекс Телемост, VINTEO, eXpress, Битрикс24, Контур.Ток, Zoom, Google Meet и др.), с помощью которых сегодня можно организовывать встречи, совещания и рабочие сессии сотрудников компаний удаленно, предлагают широкий функционал, в том числе и возможность автоматической расшифровки аудиозаписей встреч с преобразованием устной речи участников в текстовый формат. «Автоматическая транскрибация значительно ускоряет и упрощает обработку больших объемов данных, позволяя компаниям оперативно создавать текстовые версии аудиоконтента» [1, с. 71]. Функция транскрибации аудиозаписей значительно облегчает последующий анализ обсуждений, подготовку отчетности и контроль выполнения решений.

Однако, несмотря на высокую востребованность, такая важная функция отсутствует во многих бесплатных версиях сервисов, а там, где есть такой функционал, он зачастую является платным. Более того, даже в платных тарифах расшифровка часто имеет ограничения по времени, количеству участников или качеству распознавания речи. В результате компаниям приходится нести дополнительные расходы: либо оплачивать индивидуальный доступ для каждого сотрудника, регулярно участвующего во встречах, либо приобретать корпоративные подписки, затраты на которые могут быть весьма ощутимыми, особенно для малого и среднего бизнеса.

Актуальность данной работы определяется необходимостью повышения экономической эффективности деятельности российских компаний в условиях увеличения доли удаленного формата работы и цифровой трансформации и поиском актуальных технологических решений для ведения бизнеса.

Целью данной работы является проектирование и разработка автоматической системы распознавания голосовых записей, не требующей больших финансовых затрат от конечного пользователя и имеющей возможность работать без подключения к интернету.

Для достижения цели были поставлены следующие *задачи*:

1) провести анализ предметной области;

2) выявить требования к разрабатываемой системе;

3) разработать объектно ориентированные модели системы;

4) спроектировать и разработать пользовательский интерфейс.

Теоретическая значимость работы заключается в обосновании архитектурного подхода к построению гибридных систем обработки речи на основе открытых нейросетевых моделей, сочетающих распознавание, диаризацию и постобработку текста.

Практическая значимость определяется разработкой готового программного продукта, который может быть внедрен в корпоративную инфраструктуру без зависимости от внешних поставщиков. Реализация данного проекта может стать основой для создания технически и экономически эффективного программного решения, включающего такие важные параметры, как функциональность, безопасность и адаптация к локальным условиям.

Новизна разработки заключается в использовании гибридного подхода к программированию системы, а также выборе в пользу открытых нейросетевых моделей с современной архитектурой и малым количеством зависимостей, что позволяет развертывать систему локально и на различных машинах.

Материалы и методы

Теоретическая база разработки опиралась на научные публикации последних лет, посвященные различным аспектам автоматического распознавания аудиоконтента. В работах [1–4] проведен сравнительный анализ архитектур и особенностей нейросетевых моделей для автоматического распознавания речи. Проблеме транскрибации зашумленной речи посвящены исследования [5–7]. Разработка и исследование алгоритмов для задач распознавания речи представлены в работах [8–11].

Практическая часть работы включала реализацию автоматической системы по расшифровке голосовых записей встреч с помощью новейших нейросетевых технологий и инструментов.

На этапе проектирования были определены основные требования к разрабатываемой системе: полная локальность, открытая архитектура, переносимость и масштабируемость.

Принцип работы спроектированной системы отражен на рис. 1 и 2.

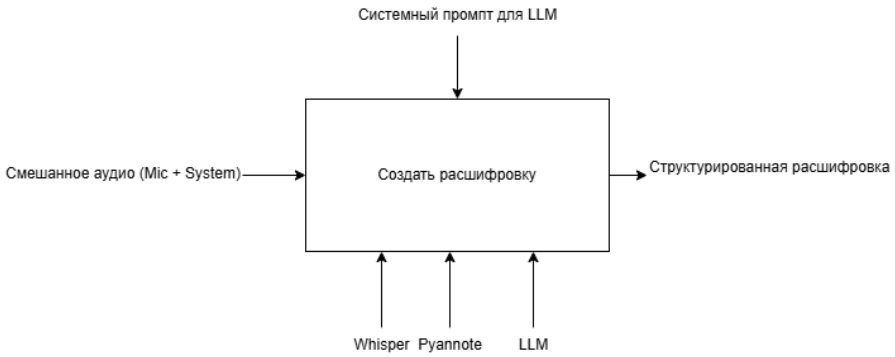

Рис. 1. Общий принцип работы IDEF0

Fig. 1. General operating principle of IDEF0

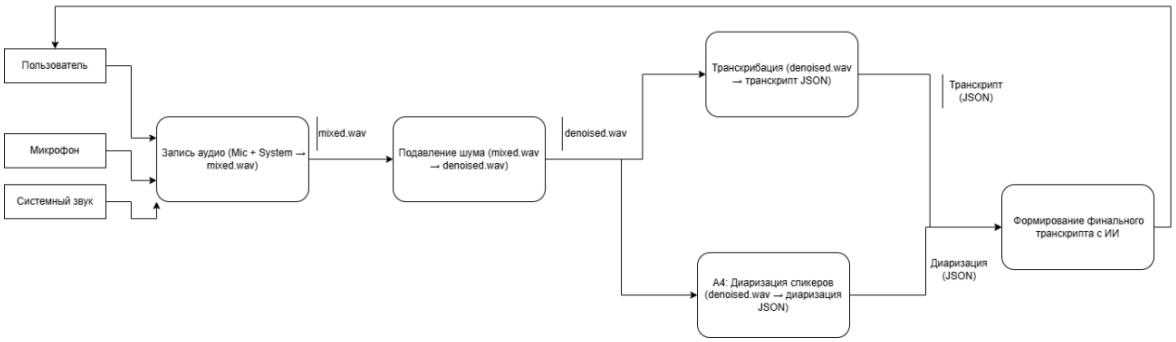

Рис. 2. Декомпозиция в DFD

Fig. 2. Decomposition in DFD

На диаграммах показаны инструменты системы и взаимодействия между ее модулями. Архитектура разработанной системы отражает многоэтапный пайплайн обработки аудио, включающий прием файла, подавление шумов, диаризацию, распознавание речи и структурирование результата в виде расшифровки с указанием говорящих. В дальнейшем представленные инструменты и модули будут рассмотрены подробнее.

Для проектирования и реализации системы были использованы следующие материалы и методы.

Языки программирования

Для создания программного кода были выбраны такие языки, как C# и Python. Такой гибридный подход обусловлен тем, что каждый из языков обладает уникальными преимуществами в соответствующей сфере применения.

Язык C# был использован для разработки пользовательского интерфейса и основной логики системы. Сильной стороной этого языка является встроенная поддержка объектно ориентированного программирования, а также строгая типизация, которая позволяет избегать проблем с типами и выявлять такие ошибки на ранних этапах разработки без написания большого количества Unit-тестов.

Язык Python, в свою очередь, был выбран как основной язык для реализации модулей, связанных с обработкой аудио и нейросетевым анализом. Выбор определен его богатым набором библиотек, таких как PyTorch, transformers, librosa, noisereduce и других. Для нашего проекта особенно важна поддержка CUDA через библиотеку Torch, которая позволяет задействовать графические процессоры NVIDIA для ускорения обработки данных в десятки раз, благодаря чему во много раз сокращается время обработки аудиофайлов. Также в Python удобнее работать с моделями, так как большинство библиотек для работы с ними написаны под этот язык

Нейросетевые модели

В проекте задействованы три ключевые нейросетевые модели Whisper, pyannote.audio и GPT-oss, каждая из них решает свою задачу в конвейере обработки.

Для реализации проекта были рассмотрены такие современные системы распознавания речи, как Kaldi, MozillaDeepSpeech, Whisper иWav2Vec2.0. Классический фреймворк Kaldi остается гибким инструментом для экспертов, но требует глубокой настройки и уступает по точности современным нейросетевым моделям. MozillaDeepSpeech хотя и проста в развертывании, демонстрирует устаревшую производительность и практически не развивается. Wav2Vec2.0 при высоких показателях точности и скорости оказывается непригодной для обработки зашумленных данных. Таким образом, опираясь на проведенный в работах [1–4] обзор и анализ современных систем транскрибации, а также на собственные исследования, мы сделали выбор в пользу Whisper.

Whisper – это нейросеть, созданная компанией OpenAI для распознавания речи. Она реализована на основе архитектуры трансформера с кодировщиком и декодировщиком и способна преобразовывать аудио в текст с высокой точностью даже в условиях шумного фона или неидеального произношения. Принцип преобразований следующий: входной аудиосигнал разбивается на фрагменты длиной 30 секунд, которые конвертируются в спектрограммы log-Mel и подаются на вход кодировщику [12]. Кодировщик преобразует эти спектрограммы в представления высокой размерности, которые затем обрабатываются декодировщиком с генерацией текстовых токенов, которые содержат не только распознанную речь, но и определение языка и расстановку временных меток [13, 14].

Whisper поддерживает более 100 языков, включая русский, и распространяется под открытой лицензией MIT, что позволяет свободно интегрировать ее в коммерческие и некоммерческие проекты. Нейросеть работает с аудиофайлами в формате WAV в моноканале на частоте 16 кГц. Whisper полностью совместима с PyTorch и CUDA, что обеспечивает ее эффективное выполнение на видеокартах.

Pyannote.audio – библиотека, созданная французским исследователем Herve Bredin и его командой. В нашем проекте она используется для диаризации – разделения аудиопотока на однородные сегменты в соответствии с принадлежностью тому или иному говорящему. Иными словами, она решает задачу, заключающуюся в определении того, кто и в какой момент говорит в аудиозаписи. Это особенно важно для анализа совещаний или встреч с участием нескольких человек: без диаризации все слова сливаются в единый поток текста, что затрудняет последующую интерпретацию. Аудиопоток в pyannote.audio разделяется на сегменты с речью каждого спикера с помощью конвейера SpeakerDiarization. Конвейер проходит несколько этапов: загрузка аудиофайла, сегментация с помощью нейронной модели, обнаружение активности речи внутри сегментов, извлечение векторов, которые представляют характеристики спикеров, кластеризация (группировка вложений от одного спикера) и окончательная диаризация. Выбор данной нейросети обусловлен ее относительной простотой в установке за счет малого количества зависимостей по сравнению, например, с NVIDIANeMo или Kaldi, что упрощает развертывание системы на различных машинах.

GPT-oss – открытая реализация языковой модели от OpenAI, выпущенная в августе 2025 года. В проекте используется ее версия с 20 миллиардами параметров (GPT-oss-20b), благодаря чему достигается компромисс между скоростью обработки и нагрузкой на систему. Основные преимущества модели – открытость, что позволяет запускать ее локально, гибкость и современная архитектура. Эта модель применяется на финальном этапе обработки: она получает «сырую» расшифровку от Whisper и структурированные метки от pyannote.audio, после чего форматирует результат в читаемый текст. Для цели обработки текста после диаризации и первичной расшифровки она подходит идеально. Архитектура GPT-oss дает возможность использовать ее локально, без подключения к облачным API, что критически важно для компаний, работающих с конфиденциальной информацией.

Кроме нейросетевых решений в системе также применяется библиотека noisereduce, которая не использует нейросети в своей работе, а полагается на метод «спектральное подавление шума». Ее задача – предварительное удаление фоновых шумов перед основной обработкой аудио.

Результаты исследования и их обсуждение

После завершения разработки ключевых компонентов системы была составлена диаграмма последовательности. Она детально отображает, в какой последовательности происходят процессы в программе и как с ними взаимодействует пользователь (рис. 3).

На диаграмме видно, что для работы с нейросетями используются внешние программы, из C# вызывается лишь их API либо же запускаются нужные процессы с аргументами из C#.

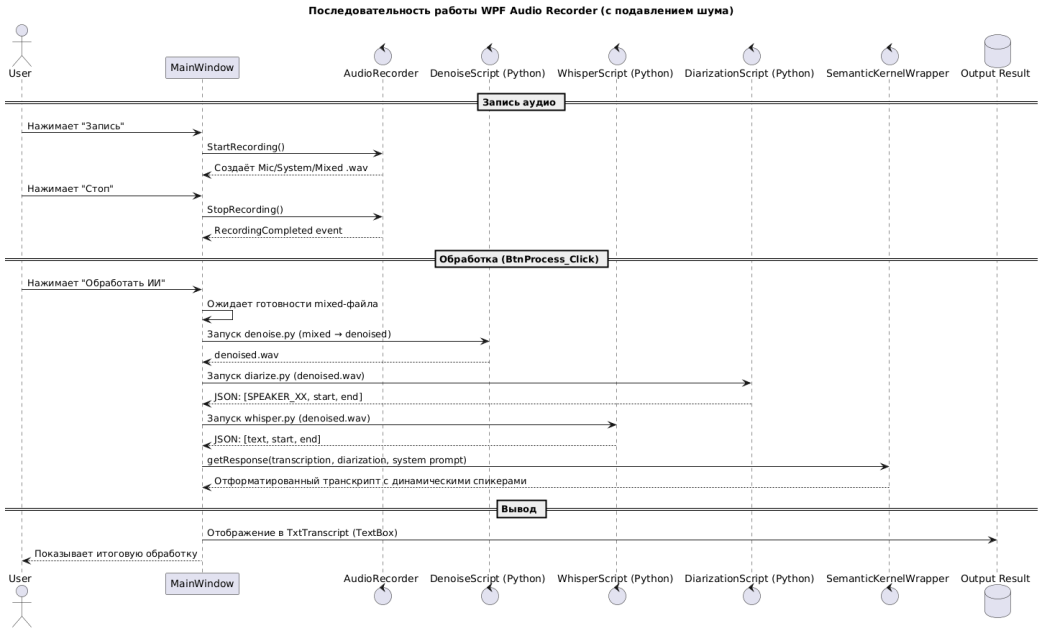

Рис. 3. Диаграмма последовательности работы программы

Fig. 3. Programflowchart

На рис. 4 представлена диаграмма развертывания, на которой показано, как именно взаимодействуют все части программы в системе.

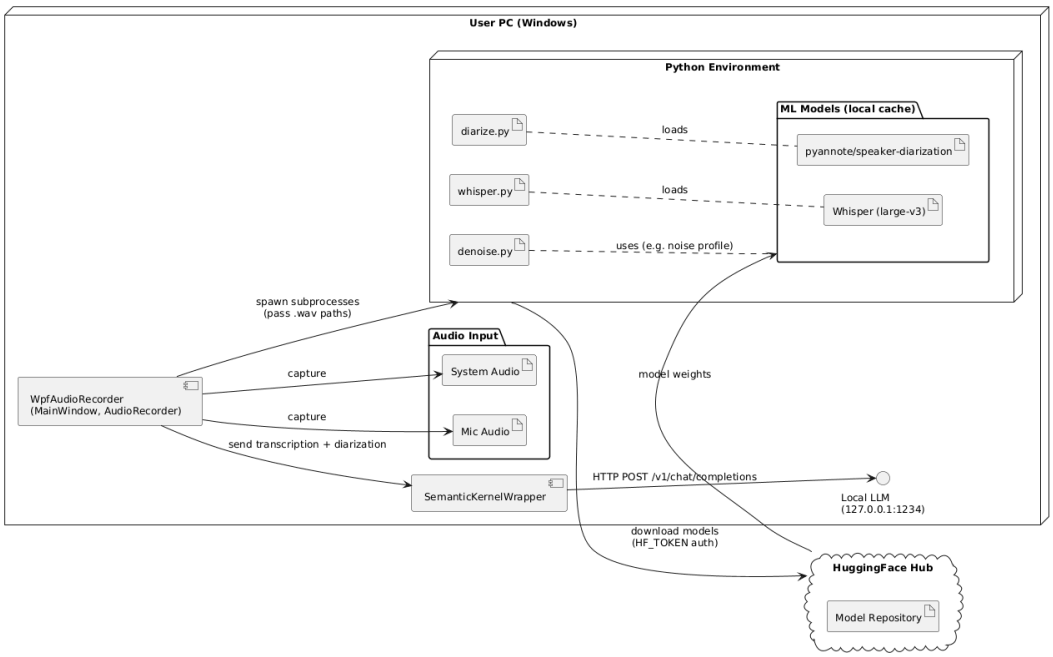

Рис. 4. Диаграмма развертывания проекта

Fig. 4. Project deployment diagram

По этой диаграмме можно увидеть, что все необходимые нейросетевые модели загружаются с платформы HuggingFace непосредственно перед использованием, а также то, что локальная LLMGPT-oss находится на localhost и хостится отдельно от нашей программы. Также наглядно показано, что каждая обработка аудиофайла в Python реализована в отдельном файле, что помогает делать код более чистым и простым для редактирования и чтения [8, 15]. Такой подход соответствует принципам чистой архитектуры и SOLID: модули слабо связаны, легко тестируемы и могут быть заменены или улучшены независимо друг от друга. Это значительно упрощает сопровождение кода.

Интерфейс программы выполнен в минималистичном стиле (рис. 5).

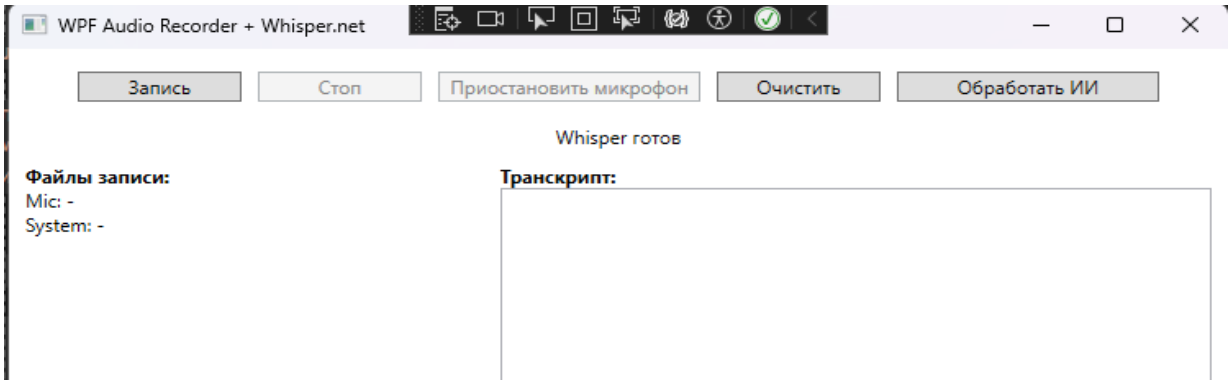

Рис. 5. Окно программы
Fig. 5. Program window

В пользовательском интерфейсе реализован набор интуитивно понятных элементов управления, включающий следующие кнопки.

1. «Запись» - запускает одновременный захват аудиосигнала с микрофона пользователя и системного аудио (например, звук из браузера или видеоконференции).

2. «Стоп» – завершает процесс записи и автоматически микширует два потока (микрофон и системный звук) в единый монофонический WAV-файл с частотой дискретизации 16 кГц – формат, требуемый для Whisper.

3. «Приостановить микрофон» – временно отключает микрофон без остановки общей записи. Эта функция может быть использована, если пользователь не хочет, чтобы сказанное им попало в финальную запись.

4. «Очистить» – удаляет текущее содержимое окна транскрипт.

5. «Обработать ИИ» – запускает пайплайн обработки с помощью нейросетей. После завершения в интерфейсе отображается структурированная, читаемая расшифровка с разбивкой по спикерам и корректной пунктуацией.

Принцип работы программы следующий: пользователю достаточно нажать кнопку «Запись», провести встречу, остановить запись и запустить обработку. Далее программа замикширует звук с микрофона и из системы в один аудиофайл, которые будут подвергаться дальнейшей обработке в следующем порядке:

1. Подавление шумов – с помощью библиотеки noisereduce удаляются фоновые шумы.

2. Диаризация – pyannote.audio формирует JSON-файл с временными метками и идентификаторами спикеров.

3. Расшифровка – Whisper генерирует текстовую транскрипцию также в формате JSON.

4. Обработка LLM – GPT-oss объединяет данные диаризации и транскрипции, исправляет грамматику, добавляет структуру и выводит финальный результат (рис. 6)

Результаты работы нейросетей показаны на рис. 6.

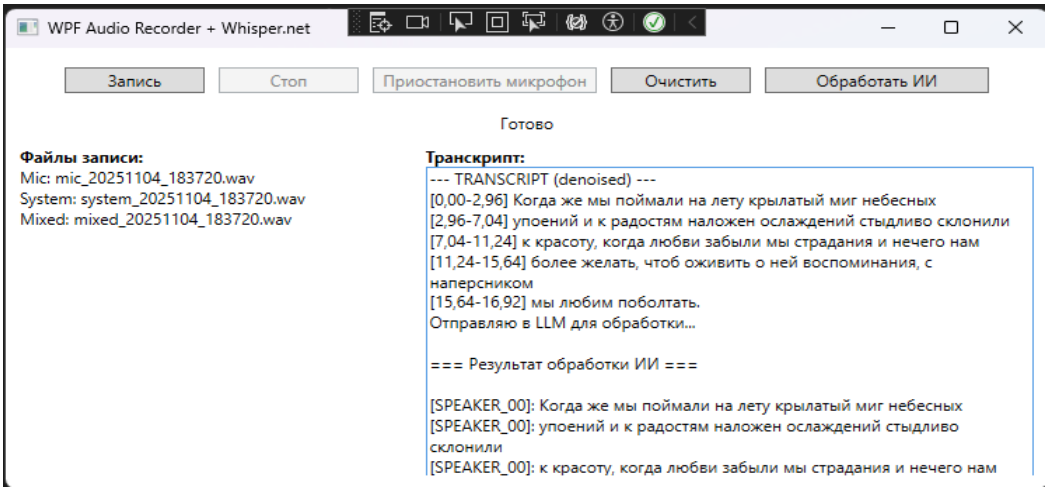

Рис. 6. Результат работы

Fig. 6. Result of work

Для оценки точности работы системы был проведен контрольный тест. В качестве тестового аудио использовался отрывок из стихотворения А. С. Пушкина:

«Когда же мы поймали на лету
Крылатый миг небесных упоений
И к радостям на ложе наслаждений
Стыдливую склонили красоту,
Когда любви забыли мы страданье
И нечего нам более желать,
Чтоб оживить о ней воспоминанье,
С наперсником мы любим поболтать.
И ты, господь!»

Полученный от Whisper текст был сопоставлен с оригиналом для расчета метрики WER (WordErrorRate) – коэффициента ошибок на уровне слов. WER вычисляется по формуле:

$$WER = \frac{S + I + D}{N}, \qquad (1)$$

где S – это количество замененных слов, I – количество вставленных (лишних) слов, D – количество удаленных (пропущенных) слов, N – общее количество слов в эталонном тексте.

В расшифрованном тексте следует отметить следующие расхождения:

1. Фраза «на ложе наслаждений» была распознана как «наложен ослаждений» – две ошибки замены, что привело к искажению смысла.

2. «Стыдливую склонили красоту» → «стыдливо склонили к красоту», одна вставка и одна замена;

3. «Страданье» → «страдания», одна незначительная замена;

4. Фраза «И ты, господь!» – в конце отрывка полностью отсутствовала в расшифровке, это два удаления.

Подставив значения в формулу (1), получаем следующий результат:

$$WER = \frac{2+1+1+3}{45} = 0{,}13.$$

Расчет метрики CER (Character Error Rate) – коэффициента ошибок на уровне символов. CER вычисляется по формуле:

$$CER = \frac{S+I+D}{N}, \qquad (2)$$

где S – это количество замененных символов, I – количество вставленных (лишних) символов, D – количество удаленных (пропущенных) символов, N – общее количество символов в эталонном тексте.

Для расчета этой метрики была использована Python-библиотека jiWER, так как подсчет вручнуюдовольно сложен. В итоге былополучено значение CER, равное 0,09, что является удовлетворительным результатом.

В итоге, учитывая полученные значения WER 0,13 и CER 0,09, имеем хорошее качество текста. Расчеты подтверждают, что потери смысла в тексте незначительны, то есть смысловая целостность сохранена в большинстве фрагментов.

Более того, важно отметить, что локальная LLM (GPT-oss) не внесла дополнительных ошибок. Напротив, она улучшила читаемость текста, добавила пунктуацию и логически сгруппировала реплики. Это под-

утверждает целесообразность включения этапа постобработки в пайплайн.

Также в рамках эксперимента были проведены 2 теста с фоновыми шумами, моделирующими реальные условия эксплуатации системы.

Первый тестовый сценарий имитировал офисную среду с характерными акустическими помехами (голоса людей, звуки работающей техники, телефонные звонки и т. п.). Второй сценарий воспроизводил домашнюю обстановку во время уборки с включенным пылесосом, представляющим собой источник интенсивного низкочастотного шума.

Результаты теста 1 показаны на рис. 7.

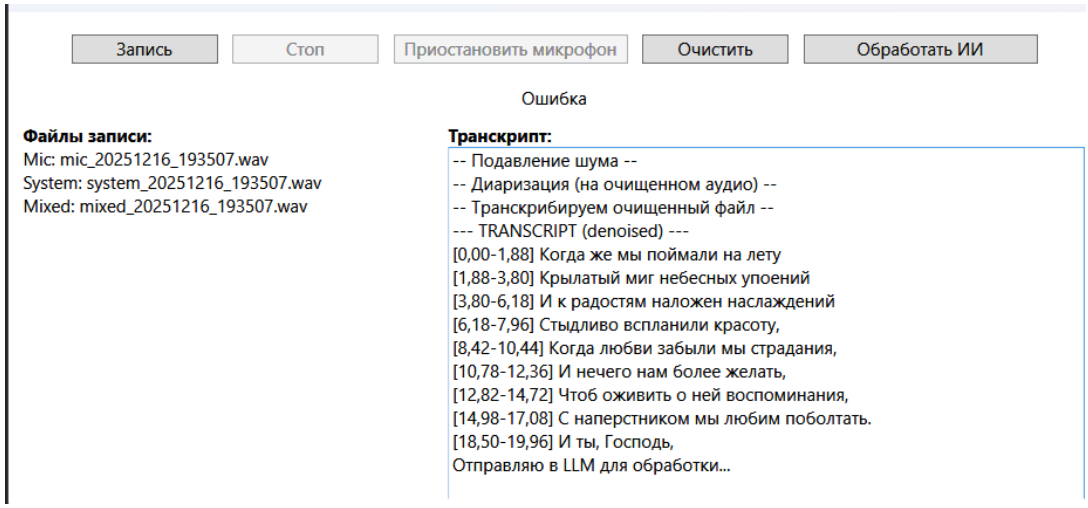

Рис. 7. Результат теста 1

Fig. 7. Result of the test 1

В тесте 1 значение WER составило 0,21, что указывает на ухудшение качества распознавания по сравнению с контрольным тестом без фоновых шумов. Однако, в целом, потеря качества остается умеренной и не достигает критического уровня. Значение CER составило 0,10, то есть этот показатель почти не ухудшился по сравнению с вариантом без акустических помех.

Полученный в результате распознавания текст показывает, что большинство ошибок, учтенных в WER, представляют собой замены слов на семантически или фонетически близкие аналоги, что в совокупности не приводит к существенной потере смысловой целостности распознанного текста.

Результаты теста 2 показаны на рис. 8.

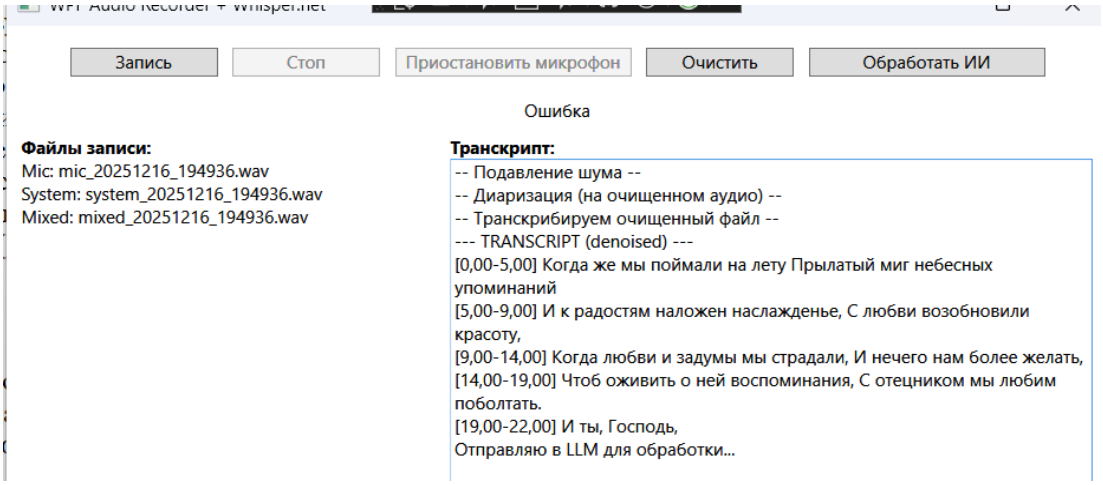

Рис. 8. Результат теста 2

Fig. 8. Result of the test 2

Результаты теста 2 показали значения WER= 0,31 и CER= 0,16. Полученные показатели демонстрируют заметное ухудшение качества транскрибации. Хотя примененный алгоритм денойзинга эффективно подавил типичные офисные акустические помехи, присутствие интенсивного низкочастотного шума (имитирующего работу пылесоса) практически полностью перекрыло целевой аудиофайл. Тем не менее, несмотря на столь неблагоприятные условия, смысловая целостность распознанного текста в целом сохранилась. В то же время метрики качества свидетельствуют о значительной деградации: значение WER достигло критического уровня, а значение CER также существенно ухудшилось. Однако, следует отметить, что на практике условия максимальной зашумленности встречаются крайне редко, поскольку пользователи, как правило, избегают выполнения задач транскрибации в столь шумной среде.

Проведенное тестирование позволило оценить точность работы спроектированной автоматической системы по расшифровке голосовых записей встреч в различных акустических ситуациях. В контрольном тесте (без фоновых шумов) достигнуты высокие показатели качества: WER = 0,13 и CER = 0,09, что свидетельствует о минимальной потере смысловой целостности распознанного текста.

В условиях, имитирующих реальные сценарии использования, система продемонстрировала устойчивость к умеренным помехам: в офисной среде (тест 1) значения WER и CER составили 0,21 и 0,10 соответственно, при этом обнаруженные ошибки искажали содержание некритично. В условиях интенсивного низкочастотного шума, моделирующего работу пылесоса (тест 2), наблюдалось значительное ухудшение показателей (WER =0,31, CER= 0,16), что указывает на пределы применимости системы в крайне зашумленных средах. Тем не менее даже в этих условиях сохранялась общая смысловая связность текста.

Таким образом, разработанная система обеспечивает хорошее качество транскрибации в типичных условиях эксплуатации, включая умеренный фоновый шум.

Заключение

В заключение отметим, что поставленные в работе задачи выполнены и цель достигнута.

Разработанная система представляет собой экономически и технологически обоснованное решение для современных российских компаний. Она позволяет сократить расходы на облачные сервисы транскрибации. По предварительным расчетам, экономия может составлять до 1500 рублей на одного сотрудника в месяц при регулярном использовании. Годовая экономия может составить 18 000 рублей на человека, что особенно существенно для средних и крупных организаций.

Ключевые преимущества разработанной системы:

1. Полная локальность – все данные обрабатываются внутри корпоративной сети, что снижает риски утечки конфиденциальной информации.

2. Открытая архитектура – использование моделей с открытым исходным кодом (Whisper, pyannote.audio, GPT-loss) и стандартных протоколов (JSON, RESTAPI) обеспечивает гибкость и независимость от вендоров.

3. Интеграция с существующими инструментами – систему можно адаптировать для работы с любыми платформами, предоставляющими доступ к аудиопотоку через API.

4. Масштабируемость – при необходимости модули можно развернуть на выделенном сервере, обеспечив обработку записей от сотен пользователей параллельно.

5. Точность – проведенные тесты экспериментально подтвердили достаточную точность работы системы в различных акустических ситуациях.

Таким образом, разработанная система по расшифровке голосовых записей встреч не только решает задачу экономии, но и формирует основу для построения безопасного и эффективного цифрового рабочего пространства.

Библиографические ссылки

1. *Баруздин М. М., Раскатова М. В., Щеголев П.* Развитие современных систем транскрибации аудио- и видеоконтента // Вестник Российского нового университета. Серия:

Сложные системы: модели, анализ и управление. 2024. № 4. С. 71–78. DOI 10.18137/RNU.V9187.24.04.P.71. EDN BYNYRY.

2. *Долженко А. И., Школина А. В.* Обзор существующих систем распознавания речи с открытым исходным кодом // Проблемы проектирования, применения и безопасности информационных систем в условиях цифровой экономики : материалы XXII Международной научно-практической конференции, Ростов-на-Дону, 21–22 ноября 2022 года. Ростов-на-Дону: Ростовский государственный экономический университет «РИНХ», 2022. С. 341–345. EDN XSGZTA.

3. *Липскеров М. А., Финк Г. Д., Корецкий В. П.* Повышение эффективности проведения корпоративных совещаний за счет автоматического анализа речевых данных и создания протоколов встреч с использованием технологий искусственного интеллекта // Цифровизация в социально-экономических системах : сборник статей II кафедральной научно-практической конференции, Москва, 21 апреля 2025 года. М. : ЗАО «Университетская книга», 2025. С. 72–76. EDN CDOZXA.

4. *Мещанинов В. Е., Поляк М. Д.* Нейросетевая модель транскрибации русской речи // Обработка, передача и защита информации в компьютерных системах: Первая Всероссийская научная конференция, Санкт-Петербург, 14–22 апреля 2020 года. СПб. : Санкт-Петербургский государственный университет аэрокосмического приборостроения, 2020. С. 75–79. DOI 10.31799/978-5-8088-1452-3-2020-1-75-79. EDN XSWIZV.

5. *Леохин Ю. Л., Фатхулин Т. Д., Ментус М. В.* Разработка и применение методов распознавания зашумленных аудиофайлов посредством нейросетевых технологий // Вестник Рязанского государственного радиотехнического университета. 2024. № 88. С. 65–73. DOI 10.21667/1995-4565-2024-88-65-73. EDN NMXASI.

6. *Мамаев И. Д., Риехакайнен Е. И.* Автоматическая расшифровка записей устной речи: тестирование программы Whisper // Социо- и психолингвистические исследования. 2023. № 11. С. 19–22. EDN ONBYJY.

7. *Мхаммад С., Молодяков С. А.* Разработка и исследование алгоритма для раздельной записи речи нескольких спикеров // International Journal of Open Information Technologies. 2025. Т. 13, № 5. С. 41–48. EDNDUBSXW.

8. Telemarketing automation based on the MIKO IP-telephony module / T. Gladkikh, L. Korobova, S. Chernyaeva [et al.] // Proceedings II International Scientific Conference on Advances in Science, Engineering and Digital Education (ASEDU-II-2021): Conference Proceedings, Krasnoyarsk, 28 октября 2021 года. Vol. 2647 A. Krasnoyarsk: AIPPUBLISHING, 2022. P. 30029. DOI 10.1063/5.0104592. EDNAHGXXE.

9. *Митюля Е. С., Голубович Ю. И., Марков А. Н.* Нейросетевые технологии обработки речи: преобразование звуков в фонемы и перспективы их применения // Сборник трудов международной молодежной школы «Инженерия-XXI», Новороссийск, 15–18 апреля 2025 года. Новороссийск : Белгородский государственный технологический университет им. В. Г. Шухова, 2025. С. 211–212. EDN EOULXJ.

10. *Морозов В. П.* Синхронизация речи и текста: ключевые инструменты // Образование России и актуальные вопросы современной науки : сборник статей VII Всероссийской научно-практической конференции, Пенза, 20–21 мая 2024 года. Пенза : Пензенский государственный аграрный университет, 2024. С. 299–302. EDN MESVYX.

11. *Тушев А. Н., Фещенко Д. Н., Деменко А. М.* Анализ необходимого инструментария для разработки программы преобразования человеческой речи в текст // Измерение, контроль, информатизация : материалы XIX Международной научно-технической конференции, Барнаул, 23 мая 2018 года / под ред. Л. И. Сучковой. Т. 1. Барнаул : Алтайский государственный технический университет им. И. И. Ползунова, 2018. С. 44–48. EDN YQMNQL.

12. *Тукаев В. Р., Беляева М. Б.* Оценка качества распознавания русской речи на чистых и зашумленных аудиоданных // Научное обозрение. Технические науки. 2025. № 3. С. 50–55. DOI 10.17513/srts.1514. EDN EPVBPD.

13. *Шилов Н. М.* Алгоритмы и подходы для решения задачи распознавания речи // Наукосфера. 2021. № 2-1. С. 89–95. EDN PALLXG.

14. Introducing Whisper // Open AI. 2022. September 21. URL: https://https://openai.com/index/whisper (дата обращения: 05.10.2025).

15. Prototype mobile application definitions fresh products based on neural network / L. A. Korobova, I. S. Tolstova, I. A. Matytsina, M. S. Mironova // Journal of Physics: Conference Series : Current Problems, Voronezh, 07–09 декабря 2020 года. Voronezh, 2021. P. 012118. DOI 10.1088/1742-6596/1902/1/012118. EDN XIEBCT.

References

1. Baruzdin M.M., Raskatova M.V. & Shchegolev P. [Development of modern systems for transcription of audio and video content]. *Vestnik Rossijskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie*. 2024. Pp. 71-78 (in Russ.). Available at: http://doi.org /10.18137/RNU.V9187.24.04.P.71.

2. Dolzhenko A.I., Shkolina A.V. *Obzor sushchestvuyushchih sistem raspoznavaniya rechi s otkrytym iskhodnym kodom* [Review of existing open source speech recognition systems]. *Problemy proektirovaniya, primeneniya i bezopasnosti informacionnyh sistem v usloviyah cifrovoj ekonomiki : materialy XXII Mezhdunarodnoj nauchno-prakticheskoj konferencii* [Problems of design, application and security of information systems in the digital economy: Proceedings of the XXII International Scientific and Practical Conference]. 2022. Rostov-on-Don: Rostov State University of Economics "RINH". Pp. 341-345 (in Russ.).

3. Lipskerov M.A., Fink G.D., Koretsky V.P. *Povyshenie effektivnosti provedeniya korporativnyh soveshchanij za schet avtomaticheskogo analiza rechevyh dannyh i sozdaniya protokolov vstrech s ispol'zovaniem tekhnologij iskusstvennogo intellekta* [Increasing the efficiency of corporate meetings through automatic analysis of speech data and creation of meeting minutes using artificial intelligence technologies]. *Cifrovizaciya v social'no-ekonomicheskih sistemah : sbornik statej II kafedral'noj nauchno-prakticheskoj konferencii* [Proc. Digitalization in socio-economic systems, Collection of articles from the II-nd departmental scientific and practical conference, April 21, 2025]. 2025. Moscow: ZAO "Universitetskayakniga", pp. 72-76 (in Russ.).

4. Meshchaninov V.E., Polyak M.D. *Nejrosetevaya model' transkribacii russkoj rechi* [Neural network model of Russian speech transcription]. *Obrabotka, peredacha i zashchita informacii v komp'yuternyh sistemah: Pervaya Vserossijskaya nauchnaya konferenciya* [Information processing, transmission and protection in computer systems, First All-Russian scientific conference, St. Petersburg, April 14-22, 2020]. St. Petersburg: St. Petersburg State University of Aerospace Instrumentation], 2020. Pp. 75-79 (in Russ.). Available at: http://doi.org/10.31799/978-5-8088-1452-3-2020-1-75-79.

5. Leokhin Yu. L., Fatkhulin T.D., Mentus M.V. [Development and application of methods for recognizing noisy audio files using neural network technologies]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2024, no. 88, pp. 65-73 (in Russ.). Available at: http://doi.org/10.21667/1995-4565-2024-88-65-7.

6. Mamaev I.D., Riekhakainen E.I. [Automatic transcription of oral speech recordings]. *Socio- i psiholingvisticheskie issledovaniya*. 2023. No. 11, pp.19-22 (in Russ.).

7. Mhammad S., Molodyakov S.A. [Development and study of an algorithm for separate recording of speech of several speakers]. International Journal of Open Information Technologies. 2025. No. 13, pp. 41-48 (in Russ.).

8. Gladkikh T., Korobova L., Chernyaeva S., Tolstova I. & Pracheva E. (2021) Telemarketing automation based on the MIKO IP-telephony module *Proceedings II International Scientific Conference on Advances in Science, Engineering and Digital Education (ASEDU-II-2021): Conference Proceedings, Krasnoyarsk, October* 28, 2021, 2647 A., 30029 (in Russian).

9. Mityulya E.S., Golubovich Yu.I., Markov A.N. *Nejrosetevye tekhnologii obrabotki rechi: preobrazovanie zvukov v fonemy i perspektivy ih primeneniya* [Neural network technologies for speech processing: transformation of sounds into phonemes and prospects for their application]. *Sbornik trudov mezhdunarodnoj molodezhnoj shkoly «Inzheneriya-XXI»* [Collection of works of the international youth school "Engineering - XXI", Novorossiysk, April 15-18, 2025 g.]. 2025, pp. 211-212. Novorossiysk: Belgorod State Technological University named after V. G. Shukhov (in Russ.).

10. Morozov V.P. *Sinhronizaciya rechi i teksta: klyuchevye instrumenty* [Synchronization of speech and text: key tools]. *Obrazovanie Rossii i aktual'nye voprosy sovremennoj nauki : sbornik statej VII Vserossijskoj nauchno-prakticheskoj konferencii* [*E*ducation of Russia and topical issues of modern science: Collection of articles of the VII All-Russian scientific and practical conference, May 20-21, 2024 g.]. Penza: Penza State Agrarian University, 2024, pp. 299-302 (in Russ.).

11. Tushev A.N., Feshchenko D.N., Demenko A.M. *Analiz neobhodimogo instrumentariya dlya razrabotki programmy preobrazovaniya chelovecheskoj rechi v tekst* [Analysis of the tools required to develop a program for converting human speech into text]. *Izmerenie, kontrol', informatizaciya : materialy XIX Mezhdunarodnoj nauchno-tekhnicheskoj konferencii* [Proc. of the XIX international scientific and technical conference, Barnaul, May 23, 2018]. Barnaul: Altai State Technical University named after I. I. Polzunov. 2018, pp. 44-48 (in Russ.).

12. Tukaev V.R., Belyaeva M.B. [Evaluation of the quality of Russian speech recognition using

clean and noisy audio data]. *Nauchnoe obozrenie. Tekhnicheskie nauki*. 2025, no. 3, pp. 50-55 (in Russ.). Available at: http://doi.org/10.17513/srts.1514.

13. Shilov N.M. [Algorithms and approaches for solving the speech recognition problem]. *Naukosfera.* 2021, no. 2-1, pp. 89-95 (in Russ.).

14. Introducing Whisper Open AI. 2022. September 21. Retrieved from https:// https://openai.com/index/whisper.

15. Korobova, L. A., Tolstova, I. S., Matytsina, I. A. & Mironova, M. S.(2020) Prototype mobile application definitions fresh products based on neural network. *Journal of Physics: Conference Series: Current Problems, Voronezh, December 7–9, 2020 g*. [Journal of Physics: Conference Series: Current Problems, Voronezh, December 7–9, 2020],p. 012118. http://doi.org/10.1088/1742-6596/1902/1/012118.

***

Development of an Automatic System for Transcribing Voice Recordings of Company Meetings Using Neural Networks

M. A. Polozov, Voronezh State University of Engineering Technologies, Voronezh, Russia
L. A. Korobova, Candidate of Technical Sciences, Associate Professor, "Technopark-V" Company, Voronezh, Russia

Amid the growing prevalence of remote work and the digital transformation of business, the *automation of routine processes—including the processing of audio recordings of meetings and conferences—is becoming increasingly important. Modern video conferencing systems offer automatic transcription features; however, these are often restricted to paid subscription plans, require an internet connection, and do not provide a sufficient level of confidentiality. Consequently, the development of a local, economically accessible, and secure solution for speech transcription has become highly relevant. This paper presents an automatic system designed for transcribing voice recordings of meetings within a project company. A distinctive feature of the developed system is the integration of open-source neural network models: Whisper (for speech recognition), pyannote.audio (for diarization – speaker identification), and an open-source GPT model (for post-processing and text formatting). The system is implemented using a hybrid architecture employing two programming languages, Python and C#, which combines high-performance audio processing with a user-friendly graphical interface. The key advantages of the solution include complete autonomy (no cloud connection required), support for the Russian language, scalability, and compliance with information security requirements. Testing on a control audio fragment yielded Word Error Rate (WER) and Character Error Rate (CER) metrics at levels acceptable for business use. To assess the accuracy of the designed system, additional tests were conducted in various acoustic environments, demonstrating that the system ensures good transcription quality under typical operating conditions, as well as in the presence of background noise. The implemented software product will enable companies to save on paid access to video conferencing systems and corporate subscriptions, while simultaneously increasing the transparency and efficiency of documenting meetings and conferences. This work holds both theoretical and practical significance for the development of domestic IT solutions in the field of corporate automation.*

Keywords: neural networks, automatic audio transcription, neural network models, speech recognition, diarization, transcription, hybrid architecture, local processing, corporate security, efficient documentation, digital transformation.

Получено: 18.12.25

Образец цитирования

Полозов М. А., *Коробова Л. А.* Разработка автоматической системы по расшифровке голосовых записей встреч в компании с помощью нейронных сетей // Интеллектуальные системы в производстве. 2026. Т. 24, № 1. С. 52–63. DOI: 10.22213/2410-9304-2026-1-52-63.

For Citation

Polozov M.A., Korobova L.A. [Development of an Automatic System for Transcribing Voice Recordings of Company Meetings Using Neural Networks]. *Intellektual'nye sistemy v proizvodstve.* 2026, vol. 24, no. 1, pp. 52-63 (in Russ.). DOI: 10.22213/2410-9304-2026-1-52-63.