А. Д. Чупеев, Н. В. Силин, И. Н. Сергеев
Тюменский государственный университет, г. Тюмень
УДК 004.852, 004.855.5, 004.942

РАЗРАБОТКА ПРОГРАММНОГО РЕШЕНИЯ ДЛЯ ТРАНСКРИБАЦИИ И СУММАРИЗАЦИИ АУДИОЗАПИСЕЙ СОВЕЩАНИЙ ДЛЯ ПРЕДПРИЯТИЯ «СТУДИЯ 42»

Аннотация. Разработан веб-сервис «Референта» для автоматизации транскрипции и суммаризации аудиозаписей совещаний. Система на базе цепочки нейросетевых моделей обеспечивает транскрибацию с точностью 87,5%, диаризацию и генерацию рефератов, сокращая время обработки в 3,8 раза. Проведено тестирование на 74 записях, подтверждающая эффективность решения

Ключевые слова: транскрибация, реферирование, автоматизация документооборота, веб-сервис, обработка аудио, NLP.

Введение. В последние годы развитие информационных технологий и усложнение организационных структур приводят к росту объемов аудиоданных, требующих оперативной обработки и анализа. Проведение регулярных совещаний и конференций является неотъемлемой частью бизнес-процессов, а качественное протоколирование встреч способствует прозрачности и оперативности управленческих решений. Вместе с тем ручной подход к оформлению протоколов характеризуется значительной трудоемкостью, временными затратами и высокой вероятностью ошибок, что подтверждается в исследованиях Е. Г. Ходжаяна [1] и Д. А. Елизарова [2].

Современные методы машинного обучения и обработки естественного языка, в частности система Whisper, позволяют автоматизировать процесс транскрибирования аудиозаписей в текстовый формат и значительно оптимизировать подготовку аналитических материалов [3]. Методические основы оценки эффективности моделей распознавания речи и суммаризации текста, предложенные А. В. Носкиной [4], обеспечивают научно обоснованный выбор алгоритмов для автоматического распознавания и реферирования.

Особую актуальность данная проблематика приобретает для предприятий, обрабатывающих большие объемы аудиоматериалов. Для партнеров «Студии 42» разработка веб-сервиса, способного интегрировать функции транскрипции и автоматической генерации рефератов обсуждений, представляет собой существенный шаг к повышению эффективности документооборота. Современные подходы к суммаризации, рассмотренные Т. В. Любимовой и И. Н. Евсеенко [5], а также сравнительный анализ методов суммаризации, выполненный М. Дорошем, Д. И. Райковским и К. В. Пугиным [6], подтверждают потенциал интеграции подобных технологий в корпоративные информационные системы.

Таким образом, внедрение автоматизированных средств транскрипции и суммаризации аудиозаписей позволяет снижать временные и материальные затраты, минимизировать количество ошибок и совершенствовать информационные системы предприятия «Студия 42».

Проблема исследования. Предприятие «Студия 42» испытывает трудности в автоматизации обработки аудиозаписей совещаний: текущий ручной или полуавтоматический процесс требует значительных ресурсов и чреват искажениями. Существующие решения не подходят из-за слабой поддержки русского языка и невозможности локального развертывания, что нарушает требования безопасности.

Для преодоления этих ограничений поставлены следующие исследовательские задачи:

1. Предобработка аудио. Анализ спектра шумов, сравнение фильтрации, вейвлетов и адаптивного шумоподавления, оценка влияния на WER.

2. Автоматическая транскрипция. Исследование ASR-архитектур (End-to-End, гибридных), дообучение под русский язык м.

3. Диаризация. Сравнение подходов (x/i-vector, нейросети), настройка кластеризации, оценка по DER и JER.

4. Автоматическое саммари. Изучение методов суммаризации (экстрактивных и абстрактных), сравнение моделей (BART, T5, TextRank), разработка гибрида, оценка по ROUGE и экспертным критериям.

Результат — веб-сервис, преобразующий аудиозаписи в структурированный текст с временными метками, диаризацией и саммари, снижая трудозатраты на документирование более чем на 50 %.

Материалы и методы. В настоящем исследовании в качестве основного корпуса материалов использовались аудиозаписи совещаний, проведенных на предприятии в период с 20 января по 15 апреля 2025 года. Общая выборка включала 74 аудиофайла в форматах MP3 и WAV общей длительностью от 10 до 120 минут каждый. Все записи были выполнены с частотой дискретизации не менее 16 кГц и отражали обсуждение профессиональных и технических вопросов, в том числе с использованием отраслевой терминологии. Исходные материалы получены посредством записи экрана участником видеоконференции без дополнительной предварительной обработки аудиосигнала, поскольку качество исходных записей оказалось достаточным для корректного функционирования алгоритмов распознавания речи. С этической точки зрения все участники совещаний были заранее проинформированы о целях и способах использования записей, дали свое добровольное согласие на участие и имели возможность отказаться от съемки в любой момент. Полученные аудиофайлы хранились в локальном файловом хранилище сервера, доступ к которым имели исключительно авторизованные пользователи.

Экспериментальная платформа представляла собой сервер с операционной системой Ubuntu 22.04 LTS, оснащенный процессором Intel(R) Core(TM) i5-3550 CPU 3.30GHz, 16 ГБ оперативной памяти и графическим ускорителем NVIDIA Quadro P4000 с 8 ГБ видеопамяти. Для организации веб-сервиса использован стек Python 3.10: фреймворк Django реализует MVC-приложение, а Django REST Framework — API-шлюз. Фронтенд выполнен как Single Page Application на JavaScript с использованием сборщика Vite, что обеспечивает динамическое обновление интерфейса без полной перезагрузки страниц. Сессии и пользовательские данные сохраняются в реляционной базе SQLite, упрощает развертывание в локальной инфраструктуре без сложных СУБД.

Архитектура приложения состоит из четырех слоев (см. рис. 1):

Web Application (MVC) — обрабатывает HTTP(S)-запросы, выполняет аутентификацию/авторизацию, загрузку аудиофайлов и выдачу документов, используя Django и защищенный обмен по HTTPS.

SPA-фронтенд — взаимодействует с API через AJAX: отображает прогресс, опрашивает статус задач, уведомляет о завершении обработки.

API-шлюз (Django REST Framework) — управляет задачами, отдает результаты и метаданные в формате JSON, реализует пагинацию и токенизированный доступ.

Фоновые воркеры — отдельные Python-процессы под systemd/supervisor, обрабатывающие задания из файловой очереди.

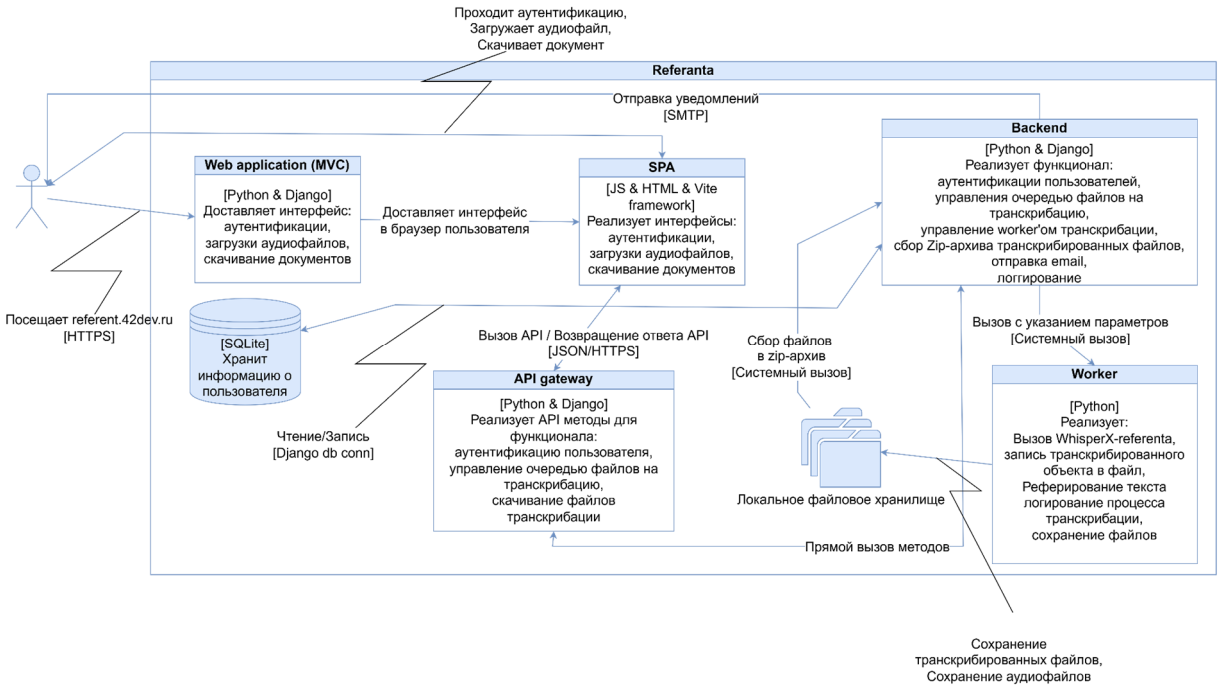

Рис. 1. Архитектура приложения

Основной вычислительный конвейер (pipeline) обработки аудиоданных включает четыре последовательно выполняемых этапа: 1. шумоподавление, 2. автоматическая транскрипция, 3. диаризация участников и 4. суммаризация текста. На этапе шумоподавления использована модель DEMUCS V4, основанная на U-Net-архитектуре с добавленными рекуррентными связями для воспроизведения спектрально-временных характеристик речевого сигнала (рис. 2).

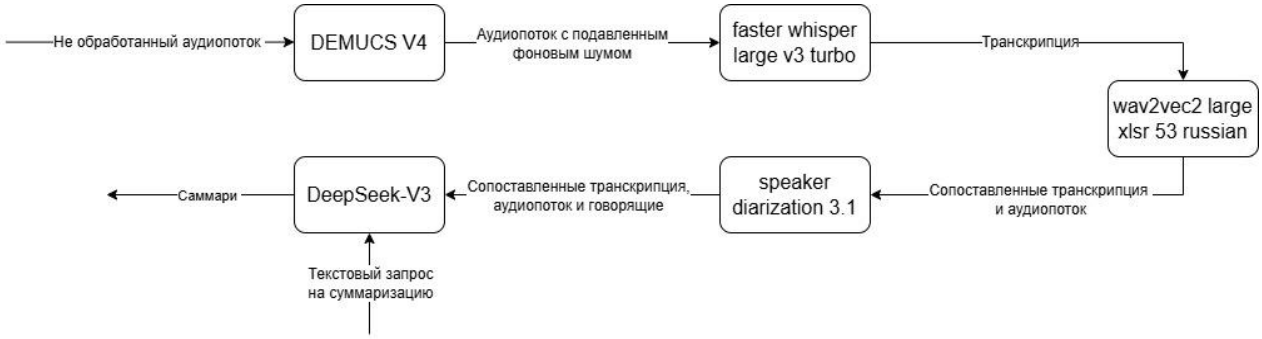

Рис. 2. Конвейер нейросетевых моделей

Входной аудиопоток без предварительной обработки поступает в DEMUCS, которая снижает уровень фоновых шумов при сохранении качества речи. Затем очищенный сигнал проходит двухэтапную транскрипцию:

- Faster Whisper large v3 turbo быстро сегментирует речь и формирует черновой текст.
- Wav2Vec2 large XLSR-53 Russian, дообученная на русскоязычном корпусе, уточняет транскрипт с точностью ≥85 % (WER).

Для временной привязки применяется forced alignment, формирующий тайм-коды на уровне слов и фраз. На этапе диаризации используется Speaker Diarization 3.1, извлекающая x-vector-эмбеддинги и кластеризующая их с PLDA. Алгоритм определяет число говорящих, группирует сегменты и оценивает точность с помощью DER и JER. Финальный этап — генерация саммари моделью DeepSeek-V3, которая на основе размеченного текста выделяет смысловые кластеры, ключевые фразы и формирует связный абстрактно-экстрактивный реферат обсуждения.

Для оценки эффективности использовались: точность транскрибации (WER), время обработки аудиофайла и генерации реферата, время ручного редактирования и удовлетворенность пользователя (в %, на основе анкетирования). Время фиксировалось автоматически и агрегировалось по каждому файлу.

Все промежуточные и итоговые файлы хранились локально до 30 дней или до удаления пользователем. Логи этапов сохранялись в формате JSON для аудита и анализа ошибок.

Результаты. В ходе валидации веб-сервиса «Референта» на 74 аудиозаписях совещаний (средняя длительность — 58 мин, ≥16 кГц, MP3/WAV) получены количественные показатели производительности и качества. Средняя точность транскрипции составила 87,5 % (WER = 12,5 %). Обработка одного файла занимала 4,7 ± 0,9 мин (транскрипция) и 3,4 ± 0,7 мин (саммари), в то время как ручной процесс требовал в среднем 113 ± 15 мин. Информация приведена в табл. 1.

Таблица 1

Сравнительный анализ сервиса «Референта» и ручной транскрибации

Показатель	«Референта»	Ручной метод
Среднее время транскрипции, мин	4,7	67,5
Среднее время генерации саммари, мин	3,4	25,0
Полное время документооборота, мин	21,4	113
Точность транскрипции, %	87,5 ± 3,2	100 (эталон)

Дополнительно была проведена оценка распределения основных источников ошибок распознавания речи. Результаты анализа логов свидетельствуют, что 65 % неточностей приходится на узкоспециализированную лексику, отсутствующую в первоначальном корпусе моделей ASR; 23 % ошибок связаны с остаточными шумовыми артефактами, несмотря на применение DEMUCS V4; 7 % — обусловлены высокой скоростью речи; 3 % — наложением говорящих (перебивания); 2 % приходится на низкий битрейт и искажения записи (см. рис. 3).

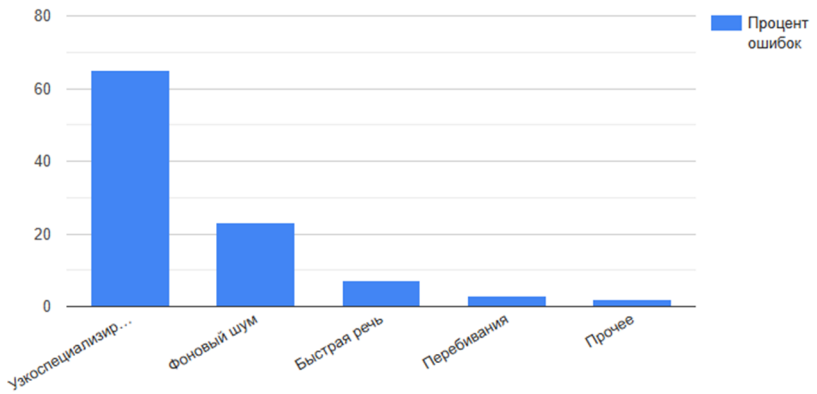

Рис. 3. Распределение типов ошибок транскрибации в сервисе "Референта".

Критический анализ полученных данных позволяет сделать следующие **выводы**.

Во-первых, заявленная эффективность шумоподавления оказалась недостаточной для условий реального офиса: подавление артефактов на 23 % не снизило общую WER до обозначенного порога 10 %.

Во-вторых, основное снижение точности (65 % ошибок) связано с недостаточной поддержкой профессиональной терминологии. Предварительное дообучение Wav2Vec2 на общекорпусных данных не обеспечило требуемой полноты словарного запаса. Дальнейшее развитие должно включать сбор специализированных текстовых данных и использование методов активного обучения для расширения лексической базы моделей.

В-третьих, при попытке параллельной обработки более пяти задач производительность сервера начинает снижаться из-за роста очередей во воркере, что приводит к увеличению латентности свыше нормативных 15 минут для файлов длиной более 120 минут. Это свидетельствует о необходимости внедрения системы управления ресурсами (resource scheduler) и масштабирования вычислительных мощностей, а также оптимизации этапа forced alignment для больших объемов данных.

Наконец, в интерфейсной части пользователи отметили удобство визуального контроля за состоянием задач и легкость скачивания результатов. Пример пользовательского экрана с перечнем заданий и ссылками на скачивание транскриптов, диаризованных разметок и саммари приведен на рис. 4.

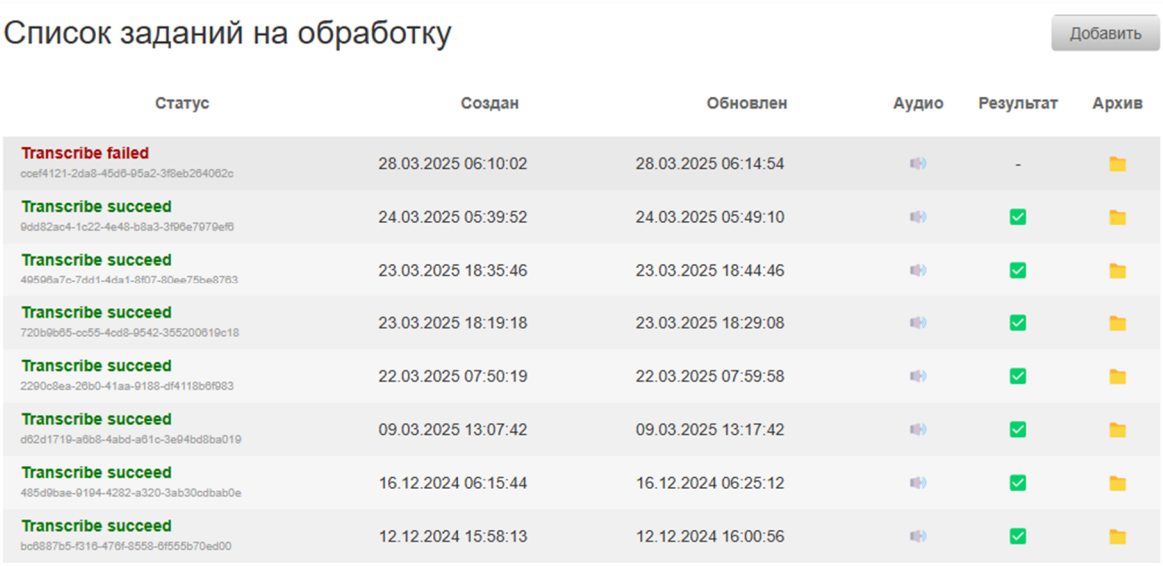

Рис. 4. Скриншот интерфейса «Референта»: список задач на обработку, статусы и кнопки скачивания.

Тем не менее наблюдались единичные жалобы на задержки при генерации реферата для длинных записей, что связано с линейным ростом времени инференса модели DeepSeek-V3. Это указывает на целесообразность разработки облегченных версий абстрактивных моделей или применения двухуровневой схемы: экстрактивная предварительная фильтрация ключевых фрагментов и последующая абстрактивная генерация на значительно сокращенном тексте.

Заключение. Проведенное исследование продемонстрировало, что разработанный веб-сервис «Референта» обеспечивает улучшение показателей автоматизированного протоколирования по сравнению с ручным методом. Основные достижения заключаются в следующем:

1. Внедрение конвейера обработки позволило уменьшить полное время документооборота в 3,84 раза — до 21,4 минуты на файл против 113 минут при ручной обработке.

2. Средняя точность транскрипции — 87,5 %

3. Развертывание компонентов в локальной инфраструктуре предприятия обеспечивает соблюдение политики конфиденциальности и гибкость масштабирования.

Однако выявленные ограничения предъявляют следующие требования к дальнейшей разработке:

- Необходимо внедрить адаптивные методы подавления фонового шума, учитывающие конкретные акустические параметры помещения, либо добавить этап обучения на реальных шумовых записях.
- Требуется собрать и интегрировать специализированные текстовые корпуса для дообучения ASR-модели и реализации механизма онлайн обновления словаря с учетом доменной терминологии «Студии 42».
- Следует ввести динамический менеджмент ресурсов, позволяющий автоматически увеличивать число воркеров при возрастании нагрузки и оптимизировать фазу forced alignment для длинных аудиозаписей.
- Актуальна разработка гибридного подхода с применением экстрактивных методов для предварительной фильтрации и сокращения объема входного текста для DeepSeek-V3, что позволит снизить время инференса без потери качества саммари.

Таким образом, «Референта» представлена как основа для автоматизации документооборота совещаний, однако для достижения промышленного уровня качества необходимо провести доразработку критических компонентов, в особенности шумоподавления и лексической адаптации моделей.

СПИСОК ЛИТЕРАТУРЫ

1. Ходжаян, Е. Г. Проведение и протоколирование коллегиального обсуждения вопросов и принятия решения / Е. Г. Ходжаян // Симбирский научный вестник. — 2020. — № 3-4(41-42). — С. 104-119. — EDN MCTUOL. (https://www.elibrary.ru/item.asp?id=44419663)
2. Елизаров, Д. А. Разработка системы транскрибации аудио-и видеоконтента / Д. А. Елизаров // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. — 2023. — № 4. — С. 87-95. — DOI 10.18137/RNU.V9187.23.04.P.87. — EDN JNAYUA (https://www.elibrary.ru/item.asp?id=59759054)
3. Майбородина, И. А. Автоматическое распознавание речи из видеолекций с использованием открытой модели Whisper / И. А. Майбородина // Наука настоящего и будущего. — 2024. — Т. 1. — С. 16-19. — EDN ULQNAB. (http://elibrary.ru/item.asp?id=68923530)
4. Носкина, А. В. Сравнение NLP-моделей на задаче суммаризации технических текстов на русском языке / А. В. Носкина // Статистические методы анализа экономики и общества : 14-я Международная научно-практическая конференция студентов и аспирантов: Труды конференции, Москва, 16–19 мая 2023 года. — Москва: Национальный исследовательский университет «Высшая школа экономики», 2023. — С. 220-221. — EDN JCDVOS. (https://www.elibrary.ru/item.asp?id=59430677)
5. Любимова, Т. В. Суммаризация текста: подходы и алгоритмы / Т. В. Любимова, И. Н. Евсеенко // Университетская наука. — 2024. — № 1 (17). — С. 182-185. — EDN MUEZZP (https://www.elibrary.ru/item.asp?id=67862378)
6. Дорош, М. Задача суммаризации текста / М. Дорош, Д. И. Райковский, К. В. Пугин // Инновации. Наука. Образование. — 2022. — № 49. — С. 2036-2044. — EDN ZNZFHC (https://www.elibrary.ru/item.asp?id=47949377)