

Санкт-Петербургский государственный университет

Программная инженерия

Группа 24.М71-мм

Разработка и оптимизация системы голосового управления для автоматизированного заполнения медицинской документации

Альшаеб Басель

Отчёт по производственной практике
в форме «Решение»

Научный руководитель:
ст. преподаватель, к.ф.-м.н. С. С. Сысоев.

Санкт-Петербург
2026

Оглавление

Введение	4
1. Постановка задачи	6
2. Обзор	7
2.1. Обзор существующих решений	7
2.2. Обзор используемых технологий	8
2.3. Выводы	9
3. Описание решения	10
3.1. Архитектура и поток данных	10
3.2. Клиентский модуль: захват аудио и VAD	10
3.3. Клиентский модуль: активация и выделение команды . .	11
3.4. Серверный модуль: приём аудио, ASR и интерпретация .	12
3.5. Интеграция со STOMMIS и особенности предметной об- ласти	13
3.6. Нефункциональные требования: ресурсы и приватность .	13
4. Алгоритмы обработки речевого сигнала	15
4.1. Захват и предварительная обработка аудиосигнала . . .	15
4.2. Детекция речевой активности	15
4.3. Выделение участков тишины	16
4.4. Активация системы по ключевому слову	16
4.5. Оптимизация вычислительных ресурсов	17
4.6. Выводы по главе	17
5. Алгоритмы обработки команд и автоматического запол- нения медицинской карты	18
5.1. Разделение речевых фрагментов и управляющих команд	18
5.2. Формирование структурного представления команды . .	19
5.3. Использование языковой модели для интерпретации команд	19
5.4. Ограничение пространства решений и фильтрация команд	19
5.5. Особенности заполнения стоматологической карты . . .	20

5.6. Интеграция с медицинской информационной системой .	20
5.7. Сравнение с традиционным подходом	21
5.8. Выводы по главе	21
6. Эксперимент	22
6.1. Условия эксперимента	22
6.2. Исследовательские вопросы	23
6.3. Метрики	23
6.4. Результаты	24
6.5. Обсуждение результатов	25
6.6. Угрозы нарушения корректности	25
6.7. Воспроизводимость	25
7. Заключение	27
Список литературы	28

Введение

Цифровизация медицинских учреждений приводит к тому, что часть клинической работы переносится в информационные системы: врач фиксирует жалобы, анамнез, результаты осмотра и назначений в электронных формах. В стоматологии это особенно заметно: итогом приёма является структурированная «карта пациента» (диагнозы, манипуляции, зубная формула, рекомендации), которая должна быть заполнена быстро и без потери качества. На практике в системе *STOMMIS* (используемой в партнёрской клинике) обнаружилась типовая проблема: во время осмотра врач работает в перчатках и часто не может удобно взаимодействовать с клавиатурой/мышью и деревьями выбора в интерфейсе, из-за чего растёт время приёма и увеличивается количество пропусков и неточностей при заполнении карты.

Существующие решения на рынке нередко сводятся к диктовке «сплошного текста» в аудио/текстовый протокол или к полной записи консультации с последующей расшифровкой. Подходы такого типа уменьшают объём ручного ввода, но не решают задачу *структурирования* данных (заполнение конкретных полей карты) и создают дополнительную нагрузку: постоянное распознавание речи требует вычислительных ресурсов и, при использовании облачных сервисов, приводит к росту стоимости обработки и сетевого трафика. Кроме того, при непрерывной обработке возрастают риски ложных срабатываний на разговоры пациента и фоновые звуки.

Целью работы является разработка и прототипирование инструмента голосового управления для стоматологической ИС, который позволяет выделять фрагменты речи и команды врача с *минимальным потреблением ресурсов* и интегрируется с существующим серверным контуром распознавания и интерпретации команд. Ключевое противоречие проекта: система должна потреблять минимум ресурсов и не выполнять дорогостоящее распознавание «всегда», но при этом надёжно фиксировать команды врача и не пропускать их. В рамках работы рассматриваются две ключевые подзадачи: (1) автоматическое разделение звукового по-

тока на участки с речью и «тишиной» (voice activity detection, VAD), (2) выделение в речи команд из ограниченного списка, запускаемых ключевым словом (wake word), и последующая интерпретация команд для заполнения полей карты пациента.

Предлагаемый подход основан на переносе части обработки на клиентскую сторону: аудиопоток захватывается в браузере средствами Web Audio API, далее лёгкий VAD-модуль выделяет фрагменты речи и завершает сегмент после паузы заданной длительности. На сервер передаются только релевантные сегменты (после активации ключевым словом), что уменьшает сетевую нагрузку и стоимость распознавания. На сервере сегменты транскрибируются с помощью Yandex SpeechKit, после чего текстовая команда сопоставляется со структурированным действием (выбор диагноза/процедуры/поля карты) с использованием правил и/или LLM-модуля. Эффективность решения оценивается экспериментально по метрикам ресурсоёмкости, задержки и качества выделения команд.

1. Постановка задачи

Целью работы является разработка и прототипирование инструмента голосового управления для стоматологической информационной системы, обеспечивающего выделение команд врача из аудиопотока при минимальном потреблении вычислительных и сетевых ресурсов.

Для достижения цели были поставлены следующие задачи:

1. разработать клиентский модуль захвата аудио в браузере и определения участков с речью/без речи (VAD) в реальном времени (раздел 3.2);
2. реализовать механизм активации и выделения команд (ключевое слово + завершение команды по паузе), передающий на сервер только релевантные аудиосегменты (раздел 3.3);
3. интегрировать полученные аудиосегменты с серверным контуром распознавания речи и интерпретации команд для заполнения полей карты пациента в *STOMMIS*, а также провести экспериментальную оценку решения (разделы 3.4 и 6).

Ограничения и допущения работы:

- решение должно исполняться либо в браузере (JavaScript), либо на сервере в поликлинике (Python), без необходимости в специализированном оборудовании;
- обработка должна быть устойчивой к типовым помехам кабинета (разговоры, шум оборудования), а также не должна существенно увеличивать задержку взаимодействия врача с системой.

2. Обзор

В разделе приведён обзор подходов и открытых решений, необходимых для понимания разработки голосового управления в медицинской ИС. Цель обзора — выбрать способ локальной предварительной обработки аудиопотока, который позволит уменьшить стоимость и ресурсоёмкость распознавания речи за счёт передачи на сервер только информативных фрагментов.

2.1. Обзор существующих решений

Выделение участков с речью (VAD). Наивный способ отделения речи от «тишины» основан на пороговой обработке энергии сигнала (RMS/амплитуда) и может быть реализован прямо в браузере. Преимущество такого подхода — минимальная вычислительная сложность, однако на практике он плохо работает в условиях кабинета из-за фонового шума и непостоянной громкости голоса. Более устойчивым решением является классический VAD из проекта WebRTC [9], использующий статистические признаки спектра; он широко применяется в VoIP и видеоконференциях и рассчитан на работу в реальном времени.

В последние годы получили распространение нейросетевые VAD-модели, обученные на больших корпусах аудио. К популярным открытым решениям относится *Silero VAD* [8], обеспечивающий высокую точность и работающий на CPU. Недостатком является необходимость серверного выполнения (Python) или интеграции через ONNX/WebAssembly, что усложняет развёртывание на клиентской стороне.

Для браузерных приложений существует специализированный класс библиотек, которые поставляют готовую ONNX-модель и обвязку для запуска в WebAssembly. В данной работе рассматривается библиотека *@ricky0123/vad-web* [3], которая позволяет запускать VAD непосредственно в браузере, используя ONNX Runtime Web [7], и предоставляет события начала/окончания речевого сегмента.

Ключевое слово (wake word) и выделение команды. Для ак-

тивации голосового режима распространены подходы keyword spotting (KWS): система постоянно анализирует поток и при обнаружении ключевого слова переводит интерфейс в режим приёма команды. Реализации могут быть (а) правилами и DTW-сопоставлением MFCC-признаков, (б) компактными нейросетями. Открытым решением является *openWakeWord* [1], ориентированный на запуск на CPU. Однако внедрение KWS в браузер усложняется требованиями к аудио-фреймам (фиксированный размер буфера, частота дискретизации) и необходимостью стабильного источника признаков.

Альтернативный практический подход, хорошо сочетающийся с VAD, состоит в двухэтапной схеме: (1) на клиенте выделяются короткие речевые сегменты, (2) на сервер отправляются только сегменты, потенциально содержащие ключевое слово или команду. Ключевое слово может проверяться по текстовой транскрипции (ASR), что повышает требования к задержке, но упрощает клиент и позволяет использовать уже существующий сервер распознавания.

Распознавание и интерпретация команд. Для распознавания речи в русскоязычных приложениях применяются как локальные движки (например, Vosk), так и облачные сервисы. В проекте используется Yandex SpeechKit [10], поскольку он обеспечивает высокое качество ASR на русском языке и имеет готовый API, уже интегрированный в серверный контур системы. После получения текста команды возможны два подхода: (1) строгое сопоставление с шаблонами/регулярными выражениями, (2) интерпретация через LLM с последующим приведением к фиксированному формату (intent + slots). В работе используется гибридный подход: ограниченный список команд задаёт допустимые действия, а LLM применяется для устойчивого сопоставления формулировок врача с интентами и параметрами.

2.2. Обзор используемых технологий

Захват аудио в браузере. Для доступа к микрофону применяется API `MediaDevices.getUserMedia` [4], а для потоковой обработки и/или

записи — Web Audio API [6] и MediaRecorder API [5]. Указанные технологии доступны во всех современных браузерах и не требуют установки дополнительного ПО на рабочее место врача.

Браузерный VAD. В качестве VAD-ядра выбран `@ricky0123/vad-web` [3]. Данный вариант обеспечивает баланс между простотой интеграции (JS, события `onSpeechStart/onSpeechEnd`) и ресурсной эффективностью (WebAssembly + оптимизированные вычисления ONNX Runtime Web [7]). В работе используется завершение сегмента после паузы фиксированной длительности (например, 2 секунды), что соответствует требованиям к выделению команд.

Серверная обработка и интеграция. Серверный контур реализован на Python (Flask) и принимает аудио сегменты как файл `multipart/form-data`. При необходимости входной формат приводится к Ogg Opus с помощью `ffmpeg` [2], после чего аудио подаётся в модуль распознавания речи (Yandex SpeechKit [10]). Далее результат передаётся в модуль интерпретации команд (правила/LLM), который возвращает структурированное действие для заполнения полей карты пациента.

2.3. Выводы

По результатам обзора для прототипа выбрана архитектура *VAD в браузере + ASR и интерпретация на сервере*. Она позволяет:

- снизить стоимость и сетевую нагрузку за счёт отправки на сервер только речевых сегментов, релевантных командам;
- сохранить качество распознавания русской речи благодаря использованию готового ASR-сервиса;
- минимизировать требования к рабочему месту врача, ограничившись браузером и микрофоном.

Браузер (рабочее место врача)

(1) getUserMedia → (2) VAD → (3) сегменты речи
(4) детектор активации (wake word) → (5) сегмент команды

Сервер (Flask, в контуре клиники)

(6) приём файла file → (7) транскодирование (при необходимости)
(8) ASR (Yandex SpeechKit) → (9) интерпретация команды (rules/LLM)
(10) запись результата в STOMMIS / возврат ответа клиенту

Рис. 1: Укрупнённая схема обработки аудиопотока и команд.

3. Описание решения

В разделе описывается реализованный прототип, который захватывает аудиопоток в браузере, выделяет речевые сегменты и передаёт на сервер только те фрагменты, которые относятся к голосовым командам врача. Далее сервер выполняет распознавание речи и преобразование текста команды в структурированное действие для заполнения полей карты пациента в *STOMMIS*.

3.1. Архитектура и поток данных

Поток обработки построен как конвейер из клиентской и серверной частей (рис. 1).

Ключевой принцип архитектуры — минимизировать «дорогие» операции распознавания речи, выполняя на клиенте лёгкую предварительную фильтрацию (VAD и логика активации). Это снижает стоимость и сетевую нагрузку, а также уменьшает объём данных, передаваемых за пределы браузера.

3.2. Клиентский модуль: захват аудио и VAD

Клиентская часть реализована как веб-страница (HTML + JavaScript + jQuery), которая получает доступ к микрофону и запускает VAD-движок. Доступ к микрофону осуществляется через

`getUserMedia` [4]; обработка аудио выполняется в браузере с использованием Web Audio API [6].

Для выделения речи используется библиотека `@cricky0123/vad-web` [3], предоставляющая событийную модель:

- `onSpeechStart` — фиксируется начало речевого сегмента;
- `onSpeechEnd` — сегмент считается завершённым после паузы (периода «тишины») заданной длительности;
- `onUpdate` — обновление вероятности речи (для визуализации и отладки).

Пороговые параметры VAD настраиваются экспериментально под условия кабинета: `positiveSpeechThreshold` (чувствительность старта), `negativeSpeechThreshold` (чувствительность окончания), а также `redemptionFrames`, задающий длительность «тишины» до завершения сегмента. В прототипе используется правило «конец команды — 2 секунды тишины», реализуемое через `redemptionFrames`.

3.3. Клиентский модуль: активация и выделение команды

Основная проблема непрерывного распознавания речи состоит в том, что на сервер уходит *весь* аудиопоток (включая фоновые разговоры и паузы), что приводит к росту стоимости и нагрузки. Для решения используется двухэтапная логика:

1. **Режим мониторинга:** VAD работает постоянно, но речевые сегменты не отправляются на сервер.
2. **Режим команды:** после обнаружения ключевого слова (например, «старт») ближайший речевой сегмент считается командой и отправляется на сервер.

В прототипе предусмотрены два варианта активации:

- **Ручная активация** (кнопка / горячая клавиша) — используется как контрольный сценарий и позволяет сравнить качество выделения сегмента команды;
- **Голосовая активация** (wake word) — ключевое слово определяется по коротким речевым сегментам, полученным от VAD. Практический вариант для прототипа состоит в проверке ключевого слова по текстовой транскрипции короткого сегмента на сервере (ASR), что позволяет избежать сложной KWS-модели на клиенте.

После активации команда считается завершённой, когда VAD фиксирует паузу заданной длительности. Получившийся сегмент команды кодируется в аудиофайл и отправляется на сервер как `multipart/form-data` с полем `file` (см. раздел 3.4).

3.4. Серверный модуль: приём аудио, ASR и интерпретация

Серверная часть реализована на Python (Flask) и предоставляет HTTP-endpoint для приёма командных аудиосегментов. Клиент отправляет запрос вида:

- метод `POST`;
- тело `FormData` с полем `file`;
- дополнительные поля метаданных (идентификатор врача, пациента, режим команд), если требуется контекст.

На сервере выполняются шаги:

1. проверка наличия поля `file` и чтение байтов аудио в память;
2. приведение формата аудио к поддерживаемому входу ASR: если MIME-тип отличается от ожидаемого, выполняется транскодирование в Ogg Opus через `ffmpeg` по пайпам (без записи на диск) [2];

3. распознавание речи посредством Yandex SpeechKit [10];
4. интерпретация результата: преобразование текста в структурированную команду (например, `intent` + набор параметров), пригодную для заполнения полей карты пациента.

Интерпретация реализуется гибридно: для короткого списка команд задаётся допустимый набор интенгов, а для сопоставления свободных формулировок используется LLM-модуль. Результат возвращается клиенту в JSON и/или применяется на стороне STOMMIS.

3.5. Интеграция со STOMMIS и особенности предметной области

В отличие от решений, ориентированных на «протоколирование» консультации, в рассматриваемом сценарии цель — заполнение структурированной карты. Это требует:

- приведения распознанного текста к заранее определённым категориям (диагноз, манипуляция, зуб, рекомендация);
- сопоставления со справочниками (например, список процедур и заболеваний), используемыми в STOMMIS;
- минимизации количества действий врача в интерфейсе (особенно при работе в перчатках).

Командный режим позволяет врачу произносить короткие инструкции, которые конвертируются в конкретные изменения в карте. Например, команда «старт, кариес на шестом верхнем справа» может быть интерпретирована как выбор диагноза «кариес» и установка соответствующего зуба в формуле.

3.6. Нефункциональные требования: ресурсы и приватность

Решение спроектировано с учётом ограничений клиники:

- **Ресурсоэффективность:** постоянная работа VAD в браузере должна быть существенно дешевле постоянного ASR;
- **Минимизация трафика:** на сервер отправляются только короткие сегменты команд;
- **Приватность:** исключается постоянная передача полного аудио приёма на сервер; объём передаваемых данных ограничивается командными сегментами.

Перечисленные свойства проверяются экспериментально в разделе 6.

4. Алгоритмы обработки речевого сигнала

Обработка речевого сигнала в разрабатываемой системе осуществляется в несколько этапов, каждый из которых направлен на снижение вычислительной нагрузки, уменьшение количества ошибок распознавания и повышение удобства использования системы в реальных условиях медицинского приёма.

Основная идея алгоритмического подхода заключается в том, что непрерывный звуковой поток не передаётся целиком на сервер и не подвергается постоянному распознаванию. Вместо этого используется последовательная фильтрация аудиоданных, позволяющая выделять только информативные участки речи и игнорировать фоновые шумы и нерелевантные звуки.

4.1. Захват и предварительная обработка аудиосигнала

Захват аудиосигнала осуществляется на стороне клиента с использованием стандартных средств Web Audio API. Входной сигнал представляет собой непрерывный поток аудиоданных, поступающих с микрофона пользователя.

На этапе предварительной обработки к аудиосигналу применяются базовые методы улучшения качества записи, включая подавление шума, автоматическую регулировку усиления и компенсацию эха. Данные методы позволяют стабилизировать амплитуду сигнала и повысить устойчивость последующих этапов обработки к изменениям громкости речи и внешним акустическим условиям.

4.2. Детекция речевой активности

Для разделения непрерывного аудиопотока на участки речи и условной тишины в системе используется алгоритм детекции речевой активности (Voice Activity Detection, VAD). Задачей данного алгоритма является определение моментов начала и окончания речевых сегментов в

звуковом сигнале.

Использование VAD позволяет отказаться от постоянной записи и передачи аудиоданных на сервер, что существенно снижает объём обрабатываемой информации и нагрузку на вычислительные ресурсы. Алгоритм VAD работает в режиме реального времени и принимает решение о наличии речи на основе анализа кратковременных характеристик сигнала.

4.3. Выделение участков тишины

Завершение речевой команды определяется на основе анализа продолжительности интервалов тишины между фрагментами речи. После обнаружения речевого сегмента система продолжает наблюдение за входным сигналом и фиксирует момент завершения команды при достижении заданной длительности условной тишины.

Использование интервалов тишины в качестве критерия завершения команды позволяет отказаться от явных маркеров окончания речи и делает взаимодействие с системой более естественным для пользователя. Такой подход особенно удобен в медицинской практике, где речь врача может сопровождаться паузами, не являющимися признаком завершения команды.

4.4. Активация системы по ключевому слову

Для предотвращения случайной активации системы и снижения количества ложных срабатываний используется механизм активации по ключевому слову. В неактивном состоянии система выполняет только детекцию речевой активности и анализ коротких речевых фрагментов.

При обнаружении короткого речевого сегмента он передаётся на сервер для проверки на соответствие ключевой фразе активации. Только после подтверждения ключевого слова система переходит в активный режим и начинает обработку основной команды врача.

Данный подход позволяет существенно сократить количество обращений к сервису распознавания речи и, как следствие, снизить эксплу-

атационные затраты и задержки обработки.

4.5. Оптимизация вычислительных ресурсов

Одним из ключевых требований к разрабатываемой системе является минимизация вычислительных затрат при сохранении приемлемой точности распознавания. Для достижения данной цели применяется многоуровневая фильтрация аудиосигнала.

На клиентской стороне выполняется первичная фильтрация и сегментация речи с использованием VAD. На сервер передаются только короткие аудиофрагменты, потенциально содержащие команды или ключевые слова. Полноценное распознавание речи и семантическая интерпретация выполняются исключительно для подтверждённых сегментов.

Таким образом, большая часть нерелевантных данных отбрасывается на ранних этапах обработки, что обеспечивает эффективное использование вычислительных ресурсов и повышает масштабируемость системы.

4.6. Выводы по главе

В данной главе был рассмотрен алгоритмический подход к обработке речевого сигнала, основанный на последовательной фильтрации аудиоданных и активации системы по ключевому слову. Предложенные алгоритмы позволяют обеспечить устойчивую работу системы в реальных условиях медицинского приёма при ограниченных вычислительных ресурсах.

Использование детекции речевой активности и анализа интервалов тишины делает взаимодействие с системой естественным и не требует от врача дополнительного обучения или изменения привычного рабочего процесса.

5. Алгоритмы обработки команд и автоматического заполнения медицинской карты

После завершения этапа распознавания речи система получает текстовую транскрипцию голосовой команды врача. Однако прямое использование данного текста для заполнения медицинской документации является неэффективным и может приводить к ошибкам, поскольку речь врача часто содержит уточнения, комментарии и фразы, не предназначенные для прямого сохранения в медицинской карте.

В связи с этим в разрабатываемой системе реализован отдельный этап интерпретации команд, целью которого является преобразование неструктурированной текстовой информации в формализованное представление, пригодное для автоматизированного внесения данных в электронную медицинскую карту пациента.

5.1. Разделение речевых фрагментов и управляющих команд

Речь врача в процессе приёма пациента может содержать как управляющие команды, адресованные системе, так и пояснительные комментарии, не имеющие прямого отношения к заполнению медицинской карты. Для корректной обработки голосового ввода необходимо различать данные типы речевых фрагментов.

В предлагаемом подходе речевая команда рассматривается как управляющее высказывание, содержащее указание на конкретное действие системы, например выбор диагноза, добавление симптома или заполнение определённого поля медицинской карты. Все остальные речевые фрагменты рассматриваются как нерелевантные для автоматической обработки и отбрасываются на данном этапе.

5.2. Формирование структурного представления команды

Для обеспечения однозначной интерпретации команд врача используется структурное представление команды, включающее тип действия и набор параметров. Каждая команда преобразуется в формализованную структуру следующего вида:

$$(тип_команды, объект, параметры)$$

Тип команды определяет общее направление действия, например добавление записи, выбор значения или подтверждение состояния пациента. Объект команды указывает на элемент медицинской карты, к которому относится действие, а параметры содержат дополнительные данные, необходимые для выполнения команды.

5.3. Использование языковой модели для интерпретации команд

Для преобразования текстовой транскрипции в структурированное представление используется языковая модель, способная выполнять семантический анализ естественного языка. Языковая модель получает на вход текст команды и контекст текущего состояния медицинской карты пациента.

На основе анализа входных данных модель определяет наиболее вероятный тип команды и соответствующие параметры. Использование языковой модели позволяет учитывать вариативность формулировок команд, характерную для естественной речи, и снижает зависимость системы от строго фиксированного набора шаблонов.

5.4. Ограничение пространства решений и фильтрация команд

Для повышения надёжности работы системы и предотвращения некорректных действий языковая модель не имеет доступа к произ-

вольному пространству решений. Вместо этого интерпретация команд выполняется в рамках заранее определённого множества допустимых действий и объектов.

В частности, выбор диагнозов, симптомов и процедур осуществляется исключительно из предопределённых медицинских справочников и классификаторов. Такой подход позволяет снизить вероятность логических ошибок и обеспечивает соответствие результатов обработки требованиям медицинских информационных систем.

5.5. Особенности заполнения стоматологической карты

Стоматологическая медицинская карта обладает высокой степенью структурированности и включает множество взаимосвязанных параметров, таких как состояние отдельных зубов, наличие патологий и выполненные процедуры. Ручной выбор соответствующих значений из иерархических списков является трудоёмкой задачей и отвлекает врача от лечебного процесса.

В разрабатываемой системе голосовые команды используются для навигации по структуре стоматологической карты и выбора необходимых элементов без прямого взаимодействия с пользовательским интерфейсом. Это позволяет врачу сосредоточиться на пациенте и существенно ускоряет процесс заполнения документации.

5.6. Интеграция с медицинской информационной системой

После формирования структурированного представления команды выполняется её интеграция с медицинской информационной системой. На данном этапе проверяется корректность параметров команды и соответствие текущему состоянию медицинской карты пациента.

При успешной проверке система автоматически вносит изменения в соответствующие поля медицинской карты. В случае возникновения

неоднозначностей или ошибок команда отклоняется, а врач получает уведомление о необходимости уточнения запроса.

5.7. Сравнение с традиционным подходом

В традиционных системах голосового ввода распознанная речь сохраняется в медицинской карте в виде текстового комментария без дальнейшей обработки. Такой подход не позволяет автоматизировать работу с данными и требует последующего ручного анализа.

В отличие от этого, предлагаемый в работе метод ориентирован на интерпретацию речи как источника управляющих команд. Результатом обработки является не текст, а структурированные изменения в медицинской карте, что обеспечивает более высокий уровень автоматизации и снижает нагрузку на медицинский персонал.

5.8. Выводы по главе

В данной главе был рассмотрен алгоритмический подход к интерпретации голосовых команд врача и автоматизированному заполнению медицинской карты пациента. Использование языковой модели в сочетании с ограничением пространства допустимых решений позволяет обеспечить надёжную и контролируемую обработку команд в медицинской информационной системе.

6. Эксперимент

Цель эксперимента — проверить, что предложенная архитектура (VAD в браузере + отправка только командных сегментов) действительно уменьшает ресурсоёмкость и стоимость обработки по сравнению с наивным подходом непрерывного распознавания речи, сохраняя приемлемое качество выделения и распознавания команд.

6.1. Условия эксперимента

Эксперимент проводился в условиях, приближенных к рабочему месту врача:

- **Клиент:** современный браузер (Chrome/Chromium), запуск веб-страницы с VAD; доступ к микрофону через `getUserMedia`.
- **Сервер:** Python 3.x, Flask, модуль транскодирования через `ffmpeg` [2], модуль распознавания речи Yandex SpeechKit [10].
- **Сеть:** локальная сеть клиники / тестовая сеть; фиксируются объём отправленных данных и задержки.

Данные. Для оценки использовались записи (или воспроизведение через микрофон) типовых сценариев приёма: команды врача, обычная речь (объяснения пациенту), паузы, а также фоновые шумы кабинета. Разметка включает:

- временные границы команд (начало/конец);
- текстовую расшифровку команды;
- ожидаемое структурированное действие в системе (intent + параметры).

При сборе данных предполагается соблюдение требований приватности: данные деперсонифицированы, не содержат ФИО и иной чувствительной информации.

6.2. Исследовательские вопросы

RQ1: Насколько уменьшается объём аудио, отправляемого на сервер, и связанная с этим стоимость распознавания по сравнению с непрерывным ASR?

RQ2: Каково качество выделения и распознавания команд (precision/recall) при использовании VAD + активации по ключевому слову?

RQ3: Как влияет настройка параметров VAD (пороги, длительность «тишины» до завершения) на ложные срабатывания и задержку?

6.3. Метрики

Для ответа на исследовательские вопросы используются следующие метрики:

- **Доля переданного аудио** $S = \frac{T_{\text{sent}}}{T_{\text{total}}}$, где T_{sent} — суммарная длительность сегментов, отправленных на сервер, T_{total} — длительность исходного аудиопотока (RQ1).
- **Сетевой трафик** (байты/МБ), отправленный клиентом на сервер (RQ1).
- **Задержка** L — время от окончания произнесения команды до получения серверного ответа (медиана и 95-й перцентиль) (RQ2, RQ3).
- **Качество выделения команд:** precision, recall и F_1 по факту обнаружения команды и её корректной интерпретации (RQ2).
- **Ресурсоёмкость клиента:** средняя загрузка CPU и потребление памяти браузером в режиме мониторинга и в режиме команды (RQ1).

Таблица 1: Сравнение базового подхода и предложенного решения.

Метрика	Baseline	Proposed
Доля переданного аудио S	—	—
Трафик, МБ	—	—
Задержка L , мс (медиана)	—	—
Precision команд	—	—
Recall команд	—	—
CPU в мониторинге, %	—	—

6.4. Результаты

Результаты эксперимента представляются в виде таблиц и графиков, сопоставляющих базовый подход и предложенное решение.

Сравниваемые варианты.

1. **Baseline:** непрерывная запись и отправка аудио на сервер с последующим ASR.
2. **Proposed:** VAD в браузере + активация + отправка только сегментов команд.

Шаблон таблицы результатов. В таблице 1 приводится рекомендуемый формат представления ключевых метрик (значения подставляются по результатам замеров).

6.4.1. RQ1: ресурсоёмкость и стоимость

Ожидается, что предложенное решение значительно уменьшит T_{sent} и сетевой трафик за счёт исключения пауз и нерелевантной речи. Дополнительно фиксируется загрузка CPU в браузере: в режиме мониторинга должна наблюдаться существенно меньшая нагрузка, чем при постоянной записи/кодировании и передаче аудио.

6.4.2. RQ2: качество команд

Для качества выделения команд важно учитывать два типа ошибок: ложные срабатывания (команда выделена, но её не было) и пропуски

(команда произнесена, но не выделена/не распознана). Оценка проводится на размеченных сценариях и отражается через $\text{precision}/\text{recall}/F_1$.

6.4.3. RQ3: влияние параметров

Проводится серия запусков при различных значениях порогов VAD и длительности паузы завершения. Анализируется компромисс между задержкой (слишком длинная пауза увеличивает L) и устойчивостью (слишком короткая пауза может преждевременно обрывать команду).

6.5. Обсуждение результатов

При интерпретации результатов важно учитывать особенности стоматологического кабинета: наличие фоновой речи, инструмента и оборудования, а также различия в манере диктовки у разных врачей. Планируется отдельно анализировать типовые причины ошибок (например, смешение речи врача и пациента, тихая речь, высокая шумовая нагрузка) и влияние этих факторов на качество системы.

6.6. Угрозы нарушения корректности

Основные угрозы валидности эксперимента:

- **Разметка:** неточность временных границ команд и субъективность в определении «команды».
- **Сеть и инфраструктура:** задержки и потери в сети могут влиять на измеряемую латентность.

6.7. Воспроизводимость

Для воспроизводимости эксперимента в репозитории проекта фиксируются:

- версии зависимостей клиентского и серверного модулей;
- параметры VAD, использованные в замерах;

- скрипты для запуска и сбора метрик (тайминг, объём отправленных данных);
- (при возможности) набор обезличенных тестовых аудиофайлов или инструкции по его получению.

7. Заключение

В работе разработан и реализован прототип голосового управления для стоматологической ИС, ориентированный на минимизацию ресурсоёмкости и стоимости распознавания за счёт предварительной фильтрации аудиопотока на клиенте.

Получены следующие результаты:

- Реализован клиентский модуль захвата аудио в браузере и определения участков с речью/без речи в реальном времени на основе браузерного VAD (@ricky0123/vad-web) (результат к задаче №1, раздел 3.2).
- Разработана логика активации и выделения команд (wake word + завершение по паузе), позволяющая отправлять на сервер только командные сегменты, тем самым снижая сетевую нагрузку и стоимость обработки (результат к задаче №2, раздел 3.3).
- Выполнена интеграция с серверным контуром на Flask: приём сегмента как поля `file`, приведение формата аудио через `ffmpeg`, распознавание речи Yandex SpeechKit и преобразование текста команды в структурированное действие для заполнения карты пациента в STOMMIS (результат к задаче №3, раздел 3.4).
- Предложен дизайн эксперимента и набор метрик для оценки ресурсоёмкости, задержки и качества выделения команд, позволяющий валидировать работоспособность решения в условиях клиники (раздел 6).

Дальнейшее развитие работы включает: расширение словаря команд и справочников предметной области, улучшение механизма ключевого слова (в том числе за счёт специализированного KWS-модуля), а также проведение полномасштабной апробации на данных нескольких врачей с анализом устойчивости к шумам и индивидуальным особенностям речи.

Список литературы

- [1] AI Mycroft, contributors. openWakeWord: open-source wake word detection. — 2025. — URL: <https://github.com/dscripka/openWakeWord> (дата обращения: 2026-01-04).
- [2] FFmpeg Developers. FFmpeg Documentation. — 2025. — URL: <https://ffmpeg.org/documentation.html> (дата обращения: 2026-01-04).
- [3] Но Ricky. @ricky0123/vad-web: User Guide and API. — 2025. — URL: <https://docs.vad.ricky0123.com/user-guide/api/> (дата обращения: 2026-01-04).
- [4] MDN Web Docs. MediaDevices.getUserMedia(). — 2025. — URL: <https://developer.mozilla.org/en-US/docs/Web/API/MediaDevices/getUserMedia> (дата обращения: 2026-01-04).
- [5] MDN Web Docs. MediaRecorder API. — 2025. — URL: <https://developer.mozilla.org/en-US/docs/Web/API/MediaRecorder> (дата обращения: 2026-01-04).
- [6] MDN Web Docs. Web Audio API. — 2025. — URL: https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API (дата обращения: 2026-01-04).
- [7] Microsoft. ONNX Runtime Web. — 2025. — URL: <https://onnxruntime.ai/docs/get-started/with-javascript.html> (дата обращения: 2026-01-04).
- [8] Silero Team. Silero VAD: pre-trained Voice Activity Detection models. — 2025. — URL: <https://github.com/snakers4/silero-vad> (дата обращения: 2026-01-04).
- [9] WebRTC Project. Voice Activity Detector (VAD) in WebRTC. — 2025. — URL: https://webrtc.googlesource.com/src/+refs/heads/main/common_audio/vad/ (дата обращения: 2026-01-04).

- [10] Yandex Cloud. SpeechKit: распознавание и синтез речи. — 2025. — URL: <https://cloud.yandex.ru/services/speechkit> (дата обращения: 2026-01-04).