



Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Голосовой AI-агент 24/7 для записи пациентов в клинике

Как мы собираем телефонию, STT/LLM/TTS и МИС в
защищённый контур приёма звонков

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

Суть проблемы

Администратор клиники не отвечает на звонки ночью, в обед и в выходные — пациент слышит автоответчик и уходит к конкуренту. Каждый непринятый звонок — упущенная запись, а утренний обзвон восстанавливает меньше половины потока. Мы закрываем это голосовым AI-агентом: он принимает звонок, проверяет слоты в МИС и фиксирует запись без человека, соблюдая 152-ФЗ и врачебную тайну.

Почему это важно бизнесу

- Ночные и обеденные звонки — до трети входящего потока небольшой клиники, и без агента они просто теряются
- Один пропущенный звонок новой пациентки — это не просто минус визит, а минус вся её LTV-цепочка приёмов
- Администратор на ресепшене занят очно с пациентом и физически не берёт трубку — телефон уходит в фон
- Напоминания за 24 и 2 часа до приёма голосом или в мессенджере снижают неявки, которые съедают слот врача
- Штраф по 152-ФЗ за утечку данных о здоровье выше, чем стоимость грамотно спроектированного голосового контура

Ключевые параметры реализации

≤ 600 мс

Латентность «тишина→ответ» p95:
бюджет STT+LLM+TTS+сеть, за
которым диалог теряет живость
наш бюджет латентности

20 мс

Размер аудиофрейма в потоковом
канале ATC↔STT — телефонный
PCM без буферизации крупными
блоками
наш телефонный контур

8 кГц/16 бит

Формат PCM моно для канала
AudioSocket/SIP-транка —
узкополосный телефонный формат
по умолчанию
наш телефонный контур

150 мс

Порог тишины VAD для завершения
реплики (EOU), после которого
закрываем окно барж-ина
наш стандарт

24/7/365

Режим работы агента — закрываем
именно те окна, где сейчас звонок
падает на автоответчик
наш стандарт

3 попытки

Лимит переспросов при
нераспознанной дате/времени —
дальше перевод на
администратора
наш стандарт



Телефонный шлюз и захват аудио

Что настраиваем

Точка входа звонка: SIP-транк РФ-оператора или локальный Asterisk/FreePBX на площадке клиники

Как мы это делаем

- 1 Поднимаем SIP-транк у РФ-оператора телефонии (Mango Office, Zadarma, Voximplant) либо Asterisk 20 LTS на сервере клиники с модулем `app_audiosocket` (`res_audiosocket`)
- 2 В `dialplan` вызываем `Answer()` → `AudioSocket(uuid,host:port)` — двусторонний TCP-канал 8 кГц/16 бит signed-linear PCM, кадры 320 байт = 20 мс
- 3 Между шлюзом и оркестратором держим WebSocket-мост на `asyncio` с очередью `audioQueue` ≤50 кадров для мгновенного сброса при барж-ине
- 4 Пишем звонок в G.711 в локальное хранилище клиники, а не во внешний общий cloud STT/TTS-провайдер — требование локализации 242-ФЗ
- 5 Резервируем второй SIP-транк другого оператора: при падении номера основного клиники не теряет входящий поток целиком

РЕЗУЛЬТАТ

Клиника получает единый номер, который принимает звонок в любое время суток без физического телефонного аппарата на ресепшене, а обрыв связи с одним оператором не кладёт линию целиком.

КЛЮЧЕВОЙ НЮАНС

AudioSocket выигрывает у AGI/EAGI тем, что не блокирует диалплан и не пишет временные файлы — строим неблокирующий стриминг кадров по TCP, а не файловые интеграции.



Конвейер STT → LLM → TTS с барж-ином

Что настраиваем

Распознавание речи, диалоговая логика на Claude с `tool_use`, синтез голоса — весь речевой контур

Как мы это делаем

- 1 STT: Yandex SpeechKit Recognizer API v3 (gRPC RecognizeStreaming, `stt.api.cloud.yandex.net:443`) в режиме `REAL_TIME` с промежуточными результатами и словарём терминов...
- 2 LLM-оркестратор — Claude через Messages API с `tool_use`: схемы `check_slot` / `book_appointment` / `reschedule` / `cancel_appointment`, `temperature 0.2`
- 3 TTS: SpeechKit Synthesizer API v3, нейросетевой голос ru-RU в OGG_OPUS 48 кГц, ресемплинг в 8 кГц телефонного канала перед отдачей в AudioSocket
- 4 Барж-ин: флаг `isBotSpeaking` + семантический VAD — при перебивании пациента сбрасываем `audioQueue` и отменяем TTS-стрим быстрее 150 мс
- 5 Системный промпт жёстко ограничивает `score` (запись/перенос/справка о врачах) — на жалобы и симптомы агент не отвечает, а эскалирует

РЕЗУЛЬТАТ

Пациент говорит с агентом как с живым администратором, может перебить и уточнить дату, а модель никогда не строит диалог о симптомах — это остаётся за живым медработником.

КЛЮЧЕВОЙ НЮАНС

`tool_use` обязателен: без реального вызова API МИС в момент диалога LLM может «придумать» свободный слот — сверяем актуальность на каждом подтверждении, а не по памяти модели.



Интеграция с МИС, ПДн и напоминания

Что настраиваем

Коннектор к медицинской информационной системе клиники + контур 152-ФЗ/242-ФЗ + исходящие напоминания

Как мы это делаем

- 1 REST/HL7-коннектор к МИС (1С:Медицина, МедЭлемент, Ident): GetFreeSlots/CreateAppointment/CancelAppointment с idempotency-key против дублей
- 2 В начале звонка голосовое уведомление о записи разговора и обработке ПДн, факт согласия фиксируем отдельным полем в БД приёма
- 3 Аудиозаписи и транскрипты храним на сервере в юрисдикции РФ (ч.5 ст.18 152-ФЗ, требование 242-ФЗ), шифрование at rest AES-256
- 4 Retention: аудио 90 дней, запись о факте обращения — согласно внутреннему регламенту клиники по хранению медицинской документации
- 5 Напоминания за 24 ч и 2 ч до приёма тем же речевым конвейером либо WhatsApp Business API/Telegram, при отсутствии ответа — SMS-фолбэк

РЕЗУЛЬТАТ

Клиника получает запись приёма без ручного переноса из тетради или мессенджера в МИС, юридически корректное согласие на голосовые ПДн и снижение неявок за счёт двух точек напоминания.

КЛЮЧЕВОЙ НЮАНС

Здоровье — специальная категория ПДн: смотрим не только 152-ФЗ, но и требование локализации 242-ФЗ к самой инфраструктуре STT/TTS, а не только к записи разговора.



Подводные камни

✗ Сбор анамнеза голосовым агентом

Агент уточняет жалобы и симптомы — это врачебная тайна без должной защиты канала. Ограничиваем intent-classifier темами записи/переноса/справки.

✗ STT/TTS на зарубежном облаке

Аудио с голосом пациента уходит за периметр РФ — нарушение локализации 242-ФЗ. Держим STT/TTS и хранилище внутри РФ-юрисдикции.

✗ Нет уведомления о записи разговора

Без голосового уведомления в начале звонка обработка ПДн идёт без согласия. Фиксируем факт согласия отдельной записью в БД на каждый звонок.

✗ VAD не учитывает эхо АТС

Без эхоподавления VAD ловит речь агента как барж-ин и обрывает его же фразу. Настраиваем echo cancellation на шлюзе до передачи в STT.

✗ LLM без tool_use выдумывает слоты

Безынструментный диалог даёт модели галлюцинировать свободное время. Каждое подтверждение слота — реальный вызов API МИС, а не память модели.

✗ Нет idempotency-key на запись

Обрыв связи и повторный дозвон без ключа идемпотентности создают дубль приёма в МИС. Вызов book_appointment несёт уникальный order_id сессии звонка.

✗ Единая точка отказа — один SIP-транк

При сбое у оператора клиника теряет весь входящий поток разом. Резервируем второй транк другого оператора с автоматическим failover.

✗ Агент даёт медицинские советы

На провокационный вопрос LLM без ограничений может дать рекомендацию по лечению — запрещено юридически. Guardrail жёстко исключает эту тему.

Как правильно

МИНИМУМ

- SIP-транк + Asterisk/FreePBX, STT/TTS SpeechKit, Claude с tool-функциями записи
- Аудиозаписи хранятся локально в РФ-контуре клиники, retention 90 дней
- Голосовое уведомление о записи разговора и обработке ПДн в начале звонка
- Эскалация на живого администратора по DTMF «0» в рабочее время

НОРМАЛЬНО

- Резервный SIP-транк второго оператора с автоматическим failover
- Семантический барж-ин, латентность «тишина→ответ» p95 ≤ 600 мс
- Напоминания за 24 ч и 2 ч через WhatsApp Business API с SMS-фолбэком
- Idempotency-key на все вызовы записи/переноса/отмены + аудит-лог tool-вызовов

ХОРОШО

- Intent-classifier исключает медицинские вопросы до вызова LLM, а не после
- On-prem fallback STT (Whisper large-v3) при недоступности облачного API
- Шифрование транскриптов AES-256 at rest + RBAC-доступ к записям
- Мониторинг качества распознавания медтерминов с дообучением словаря-подсказок



Чек-лист самопроверки

- | | |
|--|--|
| ☐ Первая фраза агента уведомляет о записи разговора и обработке ПДн? | ☐ Есть idempotency-key, исключающий дубль записи при обрыве и повторном звонке? |
| ☐ Аудиозаписи и транскрипты хранятся на серверах в юрисдикции РФ по 242-ФЗ? | ☐ Барж-ин работает с latency p95 ≤ 600 мс, пациент может перебить агента? |
| ☐ Есть резервный SIP-транк или номер на случай падения основного оператора? | ☐ Есть эскалация на живого администратора по ключевому слову или DTMF в любой момент? |
| ☐ Настроен intent-classifier, исключающий сбор жалоб/симптомов голосовым каналом? | ☐ Напоминания за 24 ч и 2 ч настроены с фолбэк-каналом при недоступности основного? |
| ☐ Слот записи проверяется через API МИС в момент подтверждения, а не по памяти LLM? | ☐ Определён retention аудио и разграничен доступ к транскриптам по ролям? |

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.



Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Разворачиваем телефонный шлюз (Asterisk/FreePBX или облачная АТС РФ-оператора) с резервированием номера
- Собираем конвейер STT→LLM(Claude tool_use)→TTS под медицинскую терминологию и интегрируем с МИС клиники по REST/HL7
- Закладываем контур 152-ФЗ/242-ФЗ: локализация записей в РФ, шифрование, уведомление о записи, retention-политика
- Настраиваем напоминания пациентам (WhatsApp Business API/Telegram/SMS) и эскалацию медвопросов на администратора
- Берём конвейер на сопровождение: латентность, деградацию STT, мониторинг ошибок tool-вызовов записи

15+

лет в ИТ-поддержке

50

рабочих мест — наш профиль

МТС

дата-центр, Москва



КОНТАКТЫ

Обсудить вашу задачу

Сайт **itfresh.ru**

Телефон **+7 903 729-62-41**

Telegram **@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.



itfresh.ru



Техническая база

- 01** AudioSocket — Channel Driver / Dialplan Application (docs.asterisk.org — 20 LTS)
- 02** Claude Messages API — Tool use (function calling) (docs.anthropic.com — 2026)
- 03** SpeechKit — Recognizer API v3 (RecognizeStreaming) (yandex.cloud — 2026)
- 04** SpeechKit — Synthesizer API v3 (синтез речи) (yandex.cloud — 2026)
- 05** Voximplant VoxEngine — серверная логика телефонии (voximplant.com — 2026)
- 06** ФЗ №152 «О персональных данных» (pravo.gov.ru — 2006)
- 07** ФЗ №242 (локализация ПДн в РФ) (pravo.gov.ru — 2014)
- 08** Наш шаблон голосового контура приёма звонков (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.

