Точечная консультация: Whisper + Claude

Кейс: сеть «Хлеб & Соль» — расшифровка и анализ всех звонков на бронь

С чем пришёл клиент

Управляющий слышит, что хостес «иногда грубят» и теряют брони. Видит дыры в выручке — около 12% звонков не доходят до подтверждённой брони. Прослушивать 120 звонков в день руками — нереально. Старший администратор слушает выборочно ~5%, и этого хватает только чтобы «уволить очевидных».

Что нужно на выходе

Полнота	Анализ 100% звонков, а не 5% выборочно
Метрики	Конверсия в бронь, причины отказа, проблемные хостес, частые вопросы гостей
Стоимость	Уложиться в 15 000 ₽/мес на всю сеть (3 точки)
Безопасность	ПДн гостей не уходят за рубеж. По возможности — без VPN.
Скорость старта	Запустить за 2 недели силами одного сотрудника + я на созвонах

Что обсудили на созвоне (60 мин)

- ✓ Архитектура: телефония → Whisper (расшифровка) → Claude (анализ) → Google Sheets / 1C
- ✓ Сравнение Whisper vs SaluteSpeech vs YandexSpeechKit на ваших аудио — где лучше распознаёт
- ✓ Чем разбирать смысл диалога: Claude Sonnet (точнее) vs GigaChat Pro (без VPN, дешевле)
- ✓ Где деплоить: своя VM (10 USD/мес) vs Replicate / Yandex Cloud по факту использования
- ✓ Как закрыть ПДн: обезличивание ФИО и телефонов до отправки в модель
- ✓ Какие промпты использовать для классификации (тип брони, причина отказа, тон диалога)

Вердикт: задача решается. Стек — Whisper (open-source, self-host) + Claude Sonnet через API-посредника. Бюджет уложится в 6–9 тыс. ₽/мес при текущем потоке ~3 600 звонков/мес. Запуск — 10–14 рабочих дней.

План внедрения

Конкретные шаги, инструменты, ссылки. Делает один человек уровня «уверенный пользователь + минимум IT». Я на созвоне сопровождаю первые 2 шага.

Архитектура (одной строкой)

```
Телефония (МангоТелеком / UIS) → запись .mp3 → cron-скрипт → Whisper Large v3 (self-host VM) → текст диалога → обезличиватель (regex) → Claude Sonnet через API-посредник → JSON с метриками → Google Sheets + 1С
```

Шаги по неделям

Дни 1–2	Доступ к записям звонков В кабинете телефонии (МангоТелеком / UIS) включаем хранение записей и автоматический экспорт по API. У UIS — webhook на каждую новую запись, у Манго — pull по REST.
Дни 3–4	Поднимаем Whisper VM на Yandex.Cloud (4 vCPU + 16 ГБ RAM, ≈ 3 500 ₽/мес) или Selectel. Ставим Whisper Large v3 через faster-whisper. 1 минута аудио = ~7 сек обработки. Скрипт-демон забирает новые mp3, расшифровывает, складывает .txt.
День 5	Обезличивание Regex-скрипт на Python: ФИО заменяем на [GUEST], телефоны на [PHONE], адреса на [ADDRESS]. До отправки в Claude. 50 строк кода, отдам шаблон.
Дни 6–7	Промпт для Claude Один системный промпт на все звонки. На входе — обезличенная расшифровка. На выходе — JSON: {результат, тип_брони, причина_отказа, тон_оператора, проблемные_фразы, рекомендация_по_обучению}. Промпт-шаблон в приложении.
Дни 8–10	Сборка пайплайна Cron каждые 30 мин: новые звонки → Whisper → обезличить → Claude API → запись JSON в Google Sheets через service account. Дашборд в Looker Studio — конверсия, топ причин отказа, рейтинг операторов.
Дни 11–14	Калибровка Прогоняем 200 исторических звонков. Сравниваем с ручной оценкой администратора. Подкручиваем промпт. Целевая точность классификации — ≥ 85%. Дальше пайплайн работает в проде.

Сервисы и ссылки

Компонент	Что использовать	Тариф / условия
Распознавание речи	Whisper Large v3 (open-source, self-host)	Бесплатно. VM 3 500 ₽/мес.
Альтернатива RU	Yandex SpeechKit (без VPN, картой РФ)	≈ 60 коп. за минуту аудио
LLM-анализ	Claude Sonnet 4.7 через API-посредник	≈ $3 за 1M входных токенов
Альтернатива RU	GigaChat Pro API (точность ниже на 8–12%)	≈ 1 500 ₽/мес, без VPN
Хранение и отчёт	Google Sheets + Looker Studio	Бесплатно
Оркестрация	Python + cron, либо n8n (low-code)	Бесплатно

Как поймёте, что заработало

Через 2 недели после запуска. Если по 5 из 6 пунктов «да» — система живёт и приносит.

1. Покрытие звонков	В таблице — ≥ 95% звонков за прошлую неделю. Если меньше — сломалась интеграция с телефонией.
2. Точность классификации	На 50 случайных звонках совпадение с ручной оценкой — ≥ 85%. Особенно по полю «результат: бронь / отказ / переоформление».
3. Утечки ПДн	В тексте, ушедшем в Claude, нет реальных ФИО и телефонов. Только токены [GUEST_N], [PHONE_N]. Проверяется выборочно админом.
4. Скорость	Звонок появляется в дашборде не позже чем через 1 час после окончания разговора. Если сильно дольше — переходим на webhook вместо cron.
5. Деньги	Затраты на API + VM за неделю — не выше 2 500 ₽. Если выше — режем длину входного промпта или переключаемся на GigaChat для простых звонков.
6. Действие	По итогам недели управляющий запустил хотя бы одно изменение: разговор с конкретной хостес, скрипт ответа, новое поле в анкете брони.

Бюджет в месяц (после запуска)

Статья	Стоимость	Комментарий
VM для Whisper	≈ 3 500 ₽	Yandex.Cloud или Selectel, без VPN
Claude Sonnet API	≈ 2 500 ₽	При ~3 600 звонков/мес и средней длине 3 мин
Посредник для API	≈ 500 ₽	Комиссия за оплату OpenAI/Anthropic
Google Workspace	0 ₽	Используете уже существующий
ИТОГО	**≈ 6 500 ₽/мес**	Укладываемся в бюджет 15 000 ₽ с двукратным запасом

Что прилагается к консультации

✓	**Запись созвона**	60 мин видео + автоматическая расшифровка
✓	**Шаблон промпта для Claude**	JSON-схема классификации звонков, готов к копированию
✓	**Скрипт-обезличиватель**	Python, ~50 строк, под ваш формат записи звонков
✓	**Список проверенных посредников**	Конкретные сервисы оплаты Anthropic API
✓	**Поддержка 2 недели в Telegram**	Любые вопросы по этому внедрению

Цена точечной консультации — 12 000 ₽. 60–90 мин + 3-страничный план + 2 недели поддержки.

Максим Чирков — AI-аудитор. chirkov.tech · Telegram @on_max · канал «Раздеваю AI»