sanatel

ЛОКАЛЬНАЯ РЕЧЕВАЯ АНАЛИТИКА ДЛЯ ЗВОНКОВ СО СМЕШАННОЙ РЕЧЬЮ

Распознавание и анализ телефонных разговоров на русском, казахском и узбекском языках — в контуре заказчика, без передачи данных внешним сервисам

Sanatel Consulting · Казахстан, Узбекистан · 2026

КОНТЕКСТ

Зачем компании речевая аналитика

Речевая аналитика превращает телефонные разговоры в структурированные данные. Она востребована везде, где телефон остаётся основным каналом продаж и обслуживания: в банках и страховых компаниях, в интернет-магазинах и службах доставки, в автосалонах и медицинских центрах.

Контроль качества

Проверка соблюдения скриптов и стандартов обслуживания по каждому звонку, а не по выборке из нескольких записей в месяц.

Потребности и возражения

Систематическое выявление причин отказов, жалоб и повторяющихся запросов клиентов на всём массиве разговоров.

Оценка менеджеров

Сопоставимая оценка работы сотрудников по единым критериям и выявление конкретных зон для обучения.

Корпус данных компании

Накопление размеченной базы реальных коммуникаций — основа для будущей аналитики и голосовых ассистентов.

ПОСТАНОВКА ЗАДАЧИ

Почему это не типовая задача распознавания

Готовые решения распознавания рассчитаны на один язык и на качественную запись. Телефонные разговоры в Казахстане и Узбекистане не соответствуют ни одному из этих допущений.

Смешанная речь

В одном разговоре встречаются русский и казахский или русский и узбекский языки. Чаще всего переключение происходит в начале разговора, сразу после приветствия.

Разговорная лексика

Клиенты говорят на живом разговорном языке с заимствованиями. Модели, обученные на литературной речи и новостных корпусах, теряют на этом заметную часть точности.

Качество записи

Запись поступает с телефонной станции: узкая полоса 8 кГц, сжатие, разговор через трубку. Это принципиально хуже, чем студийная или гарнитурная запись.

Два собеседника

Запись стерео: менеджер и клиент разведены по каналам. Это преимущество — роли определяются точно, без разделения дикторов, но обрабатывать нужно оба канала.

Итог: на русском языке готовые модели дают приемлемый результат, на казахском и узбекском — заметно слабее. Именно этот разрыв решение и закрывает.

РЕГУЛИРОВАНИЕ

Правовая рамка: Казахстан и Узбекистан

Требования к размещению персональных данных в двух странах различаются и в последние годы менялись. Локальное развёртывание закрывает вопрос независимо от редакции нормы.

КАЗАХСТАН

- Закон «О персональных данных и их защите» № 94-V: хранение персональных данных ведётся в базе, находящейся на территории Республики Казахстан (п. 2 ст. 12).
- Трансграничная передача регулируется отдельно и допускается при соблюдении установленных условий — то есть облако запрещено не как класс, но требует отдельного обоснования.
- Для банков поверх этого действует режим банковской тайны: сведения о клиенте раскрываются третьим лицам только с его письменного согласия.

УЗБЕКИСТАН

- Закон «О персональных данных» № ЗРУ-547, ст. 27^{1} в редакции закона № ЗРУ-1125 от 26 марта 2026 года: перечень данных, подлежащих обязательному хранению в стране, сужен.
- В перечень входят биометрические и генетические данные, а также данные пользователей услуг операторов телекоммуникаций.
- Для остальных категорий обработка за пределами страны допускается при выполнении условий закона; подзаконные акты по ним на момент подготовки материала не приняты.

Размещение системы в контуре заказчика снимает и вопрос локализации, и вопрос передачи данных третьей стороне — независимо от того, как будет развиваться регулирование.

ПРАКТИКА

Что требуют заказчики помимо закона

По итогам переговоров с потенциальными заказчиками требование локального размещения формулируется не только юристами. Три причины повторяются чаще других.

Контроль над данными

Записи разговоров содержат персональные данные и коммерческую информацию. Заказчики формулируют требование прямо: данные не должны покидать периметр компании, а доступ к ним разграничивается их собственными средствами.

Собственная инфраструктура

Типовой запрос — выделенный сервер в помещении компании либо собственная стойка в центре обработки данных. Разделяемая облачная среда, как правило, не рассматривается.

Предсказуемая стоимость

Облачное распознавание тарифицируется за минуту. На потоке в тысячи разговоров в сутки стоимость владения становится основным аргументом в пользу собственного оборудования.

Локальное развёртывание превращает эти требования из ограничения в характеристику решения: сервер, модели и данные находятся под контролем заказчика.

РЕШЕНИЕ

Архитектура: один сервер в контуре заказчика

Все компоненты работают на одном сервере внутри периметра компании. Наружу не уходит ни аудио, ни текст расшифровки.

КОНТУР ЗАКАЗЧИКА

Телефония

ATC Asterisk: стерео-запись разговора, канал на каждого собеседника

Потребители

CRM, речевая аналитика, отчётность. Обращаются к общему API и получают готовый диалог и результат анализа

ASR Service

Собственная разработка: Node.js, TypeScript, Fastify

- Единый API для всех систем компании
- Очередь заданий и управление состояниями
- Подготовка аудио, выбор результата, журналы

Распознавание

Python, FastAPI, faster-whisper. Модель Whisper large-v3 и специализированная модель для отдельного языка

Смысловой анализ

Ollama, языковая модель Qwen2.5. Суммаризация, проверка по чек-листу, выявление потребностей

Компоненты обмениваются данными только внутри сервера. Внешние сетевые обращения в работе решения не используются.

СОБСТВЕННАЯ РАЗРАБОТКА

ASR Service: слой, который делает движки сервисом

faster-whisper и Ollama — это библиотека и локальный сервер моделей. Они умеют обработать один файл по запросу, но не умеют принимать поток заданий от нескольких систем, удерживать очередь, восстанавливаться после сбоя и отчитываться о состоянии. Этот слой и есть наша разработка: Node.js, TypeScript, Fastify.

Единый API

Постановка задания на распознавание и на анализ текста, получение результата. Авторизация по токену, проверка запроса по схеме.

Очередь и состояния

Задание проходит фиксированную последовательность состояний. На каждом шаге результат сохраняется, ошибка переводит задание в отдельное состояние с описанием причины.

Три независимых обработчика

Подготовка аудио, распознавание и анализ работают своими циклами: пока распознаётся один разговор, для следующего уже готовятся каналы.

Журналы и статистика

Страница состояния: задания по состояниям, время распознавания, температура процессора, оценки вариантов, покрытие речи.

Управление хранением

Временные файлы каналов удаляются сразу после обработки, завершённые задания — по истечении суток. Сервис не накапливает записи разговоров.

Управление моделями

Сервис указывает, каким движком выполнять конкретный проход, и переключает модель на стороне распознавания между запросами.

ЖИЗНЕННЫЙ ЦИКЛ ЗАДАНИЯ

РАСПОЗНАВАНИЕ

Почему одного прохода недостаточно

В режиме автоматического определения языка модель выбирает один язык на весь фрагмент и на смешанной речи ошибается. Принудительное указание языка эту ошибку снимает, но только там, где фрагмент действительно одноязычный. Поэтому каждый канал распознаётся трижды, а лучший результат выбирается по формальным метрикам.

Узбекистан

языковая пара проекта: uz, ru
каждый проход — по двум каналам

ПРОХОД 1	ПРОХОД 2	ПРОХОД 3
авто: uz + ru	**только uz**	**только ru**

Казахстан

языковая пара проекта: kk, ru
каждый проход — по двум каналам

ПРОХОД 1	ПРОХОД 2	ПРОХОД 3
авто: kk + ru	**только kk**	**только ru**

Архитектура конвейера от языковой пары не зависит: один и тот же сервис обслуживает и узбекских, и казахстанских заказчиков — меняется только параметр задания.

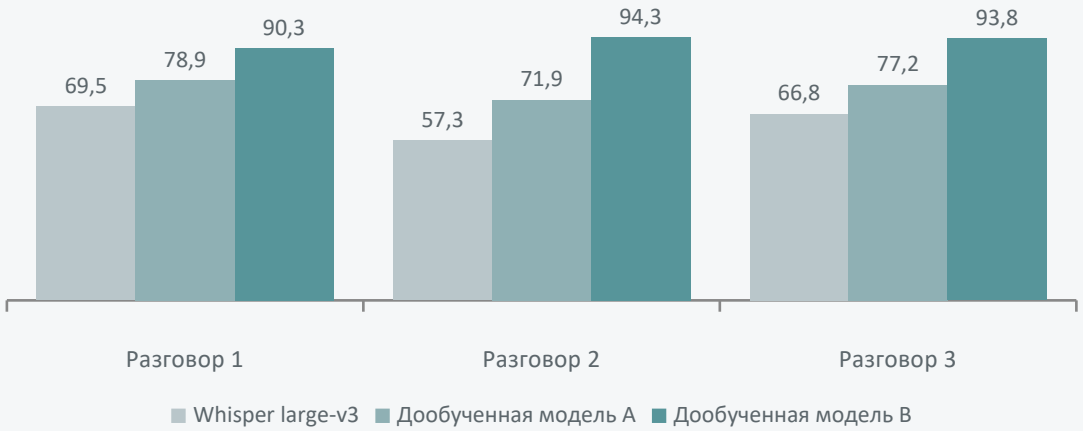
ВЫБОР МОДЕЛЕЙ

Базовая модель и специализированная

Whisper large-v3 остаётся базовой моделью: она отвечает за автоматический проход и за русский язык. На форсированном проходе по второму языку её заменяет специализированная модель, дообученная под этот язык.

Сравнение на реальных разговорах

Взвешенная оценка уверенности, узбекский проход, три разговора



0,095 → 0,814

Доля разговора, распознанная как речь, на записи, где базовая модель практически не справилась: было 9,5 %, стало 81,4 %.

Критерии отбора модели

- Совместимость с рабочим движком: архитектура Whisper либо возможность конвертации в формат CTranslate2.
- Качество на собственных записях, а не на публичных тестовых наборах.
- Устойчивость к артефактам: у выбранной модели фильтры не отбраковали ни одного сегмента на всех трёх разговорах.

Обе модели одновременно в памяти графического процессора не помещаются, поэтому движок распознавания подменяется между запросами. Измеренное время подмены — около 0,9 секунды.

КАЧЕСТВО ДАННЫХ

Что приходится вычищать из результата распознавания

Модели семейства Whisper на тишине, шуме и обрывках речи склонны выдавать правдоподобный, но несуществующий текст. Без отдельного слоя очистки такой текст попадает в диалог и искажает последующий анализ.

Повтор подсказки

Модель дословно воспроизводит служебную подсказку, переданную ей вместе с аудио, как реплику собеседника.

Известные артефакты

Следы обучающих данных: фразы из субтитров видеороликов, призывы подписаться на канал, благодарности зрителям.

Чужая письменность

На шуме декодирование уходит в посторонний алфавит. Допустимы кириллица и латиница; превышение доли прочих знаков отбраковывает сегмент.

Зацикливание

Одна и та же фраза повторяется подряд десятки раз. Пороги подобраны так, чтобы не задеть естественные повторы коротких реплик.

Повтор внутри сегмента

Бессмысленные повторы слогов внутри одной реплики выявляются по степени сжимаемости текста.

Недостоверная длительность

Короткой реплике приписан интервал в десятки секунд. Текст сохраняется, но для метрик берётся правдоподобное время произнесения.

Исходный результат распознавания при этом не изменяется: очистка влияет только на собираемый диалог и на метрики, поэтому разбор любого спорного случая возможен по исходным данным.

МАТЕМАТИКА ВЫБОРА

Как определяется лучший из трёх вариантов

После очистки каждый вариант получает набор метрик, и они сводятся в одну оценку. Формулу подбирали на реальных разговорах: ранние варианты, основанные только на средней уверенности модели, выбирали неправильный результат.

оценка = взвешенная уверенность × относительное покрытие речи × относительное число сегментов

1

Взвешенная уверенность

Собственная оценка модели по каждому сегменту, приведённая к шкале от 0 до 100 и усреднённая с весом по длительности речи, а не по числу сегментов.

2

Относительное покрытие речи

Сколько секунд речи вариант распознал — в сравнении с лучшим вариантом того же разговора. Отсекает варианты, потерявшие часть разговора.

3

Относительное число сегментов

Дополнительная поправка на дробность: вариант, собравший разговор в несколько длинных блоков, не получает преимущества перед детальным.

Почему одной уверенности мало

У выдуманного текста собственная уверенность модели регулярно выше, чем у настоящей речи. Вариант, потерявший половину разговора, по средней уверенности выглядит лучшим.

Границы применения формулы

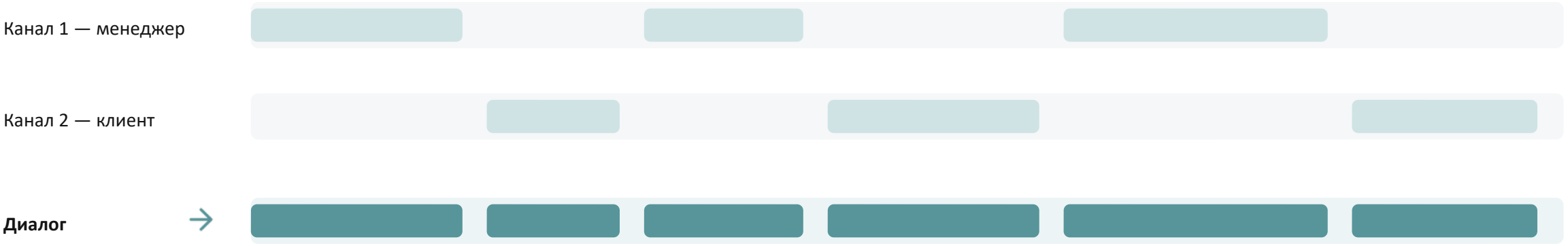
Это функция ранжирования вариантов одного разговора, а не абсолютная оценка качества. Для отбраковки используется отдельный показатель — доля распознанной речи.

РЕЗУЛЬТАТ

Сборка диалога и отбраковка

Победивший вариант превращается в диалог: сегменты обоих каналов сводятся в общую временную шкалу, каждому присваивается роль согласно настройке задания.

- Временные метки сохраняются исходными: ограничение длительности применяется только к метрикам и не искажает разметку диалога.
- Роли определяются каналом записи, разделение дикторов не требуется.
- Метрики проигравших вариантов сохраняются вместе с результатом — это материал для разбора спорных случаев.



0,28

Порог отбраковки по доле распознанной речи

Если победивший вариант покрыл речью менее 28 % длительности записи, задание помечается как нераспознанное. Такой разговор не уходит в смысловой анализ и не портит отчётность.

СМЫСЛОВОЙ АНАЛИЗ

Зачем в решении языковая модель

Расшифровка сама по себе не отвечает на вопросы бизнеса. Содержательные выводы по разговору делает языковая модель, развёрнутая на том же сервере.

Что модель делает с диалогом

- Краткое содержание разговора и его результат
- Проверка по чек-листу: приветствие, выявление потребности, работа с возражением, договорённость о следующем шаге
- Потребности, возражения и причины отказа клиента
- Оценка работы менеджера по единым критериям

Qwen2.5

Выбрана за качество работы с русским, казахским и узбекским языками. В эксплуатации версия на 7 млрд параметров; при большем объёме видеопамяти — версия на 14 млрд.

Строгий формат ответа

Ответ запрашивается по заданной схеме данных. Результат возвращается пригодным для загрузки в систему аналитики без ручного разбора текста.

Анализируется не каждый разговор

В смысловой анализ передаются только диалоги длиннее двадцати реплик. Короткие разговоры — ошибочные звонки, переводы на другого сотрудника — содержательных выводов не дают и вычислительный ресурс не занимают.

РАЗВЁРТЫВАНИЕ

Требования к оборудованию

Решение рассчитано на одну серверную видеокарту. Объём видеопамяти — главный ограничитель: на ней одновременно находятся модель распознавания и языковая модель.

16 ГБ

видеопамяти — практический минимум

Текущий стенд: NVIDIA RTX 5060 Ti 16 ГБ, Ubuntu 24.04. Этого достаточно, чтобы держать распознавание и языковую модель в памяти одновременно.

Две модели распознавания одновременно не помещаются, поэтому движок подменяется между запросами — около 0,9 секунды.

Что занимает видеопамять

Модель распознавания	Whisper large-v3 в рабочем режиме вычислений
Языковая модель	Qwen2.5 на 7 млрд параметров
Служебные буферы	Промежуточные данные обработки аудио и контекст языковой модели

Если видеопамяти меньше

- Языковая модель меньшего размера — с потерей качества смысловых выводов.
- Либо выгрузка одной модели перед загрузкой другой — с дополнительной задержкой на каждом разговоре.

ПРОИЗВОДИТЕЛЬНОСТЬ

Сколько разговоров обрабатывает один сервер

Видеокарта выполняет задания последовательно: распознавание и смысловой анализ делят один вычислительный ресурс, а не складывают его. Поэтому пропускная способность считается по сумме времени обработки, а не по наибольшему из двух.

≈ 30 с

на распознавание одного разговора длительностью 2–3 минуты — включая все три прохода по двум каналам

2500–3000

разговоров в сутки при круглосуточной работе — ориентир для текущей конфигурации

> 20 реплик

порог передачи диалога в смысловой анализ: короткие разговоры ресурс не занимают

пропускная способность = 86 400 секунд в сутках ÷ (время распознавания + время смыслового анализа)

Очередь заданий сглаживает пики: разговоры поступают неравномерно, а обрабатываются равномерно, в том числе ночью.

Для большего объёма конфигурация масштабируется второй видеокартой или вторым сервером — очередь распределяется между ними.

РАЗВИТИЕ

Направления дальнейшей работы

Перечень отражает фактическое состояние работ, включая то, что сознательно отложено.

БЛИЖАЙШАЯ ЗАДАЧА

Специализированная модель для казахского языка

Подход воспроизводится без изменений архитектуры: меняется модель на форсированном проходе. Отбор кандидатов ведётся по тем же критериям.

БЛИЖАЙШАЯ ЗАДАЧА

Работа с исходным качеством записи

Сейчас на вход поступает сжатая запись с узкой полосой 8 кГц. Получение исходных файлов с телефонной станции — самый дешёвый способ поднять точность.

БЛИЖАЙШАЯ ЗАДАЧА

Отраслевой словарь заказчика

Передача модели списка терминов, названий продуктов и типовых имён повышает точность на специфичной лексике без какого-либо обучения.

ОПЦИЯ ВТОРОГО ЭТАПА

Дообучение модели под домен

Требует порядка 10–30 часов размеченных записей заказчика. Обучение выполняется на его данных и в его контуре; целесообразно при большом постоянном объёме.

Спасибо за внимание

Готовы обсудить пилотный проект на ваших записях

АДРЕС
Алматы, пр. Достык, 180, офис 200

ЭЛЕКТРОННАЯ ПОЧТА
info@sanatel.kz

ТЕЛЕФОН
+7 701 705 5947

TELEGRAM
@sanatel_kz

WHATSAPP
+7 701 705 5947

sanatel