

SPEAKNODE

Описание функциональных характеристик

Версия 1.0



Настоящая документация может быть использована только для поддержки работоспособности продуктов, поставленных платформой Voxagent.

Все примеры конфигураций, реквизитов и других данных в документе вымышлены и не относятся к реальным сущностям. Любое совпадение случайно.

Все встречающиеся в тексте товарные знаки и зарегистрированные товарные знаки являются собственностью их владельцев и используются исключительно для идентификации программного обеспечения или компаний.

Данная документация может не отражать некоторых модификаций программного обеспечения. Если вы заметили ошибки или опечатки, сообщите об этом по контактными данным ниже.

Все имущественные авторские права сохраняются за Voxagent в соответствии с действующим законодательством.

© **Voxagent, 2026**

Россия

help@mail.voxagent.app · voxagent.app

Содержание

Обзор функционала	6
Основные возможности	6
Как это работает	6
AI-агенты	6
Настройка агента	7
Основные параметры	7
Языки и локализация	7
LLM-модель	7
Распознавание речи (STT)	9
Детекция голосовой активности (VAD)	9
Голос (TTS)	9
Динамические переменные	10
Инструменты агента	10
Публикация	10
Дублирование агента	10
Аналитика	10
История разговоров	11
Список сессий	11
Детали сессии	11
Телеметрия и трейсинг	11
Пайплайн телеметрии	12
Хранение данных	12
Биллинг	12
Модель биллинга	12
Подписки и тарифы	12
Отслеживание использования	12
Обработка платежей	13
Управление биллингом	13
Исходящие кампании	13
Обзор	13
Настройка кампании	13
Элементы кампании	13

<u>Как это работает</u>	14
<u>Мониторинг</u>	14
<u>Команды и пространства</u>	14
<u>Пространства</u>	14
<u>Что принадлежит пространству</u>	14
<u>Управление пространством</u>	14
<u>Роли пользователей</u>	15
<u>Приглашения</u>	15
<u>Аутентификация</u>	15
<u>Телефония</u>	15
<u>Поддерживаемые провайдеры</u>	15
<u>Входящие звонки</u>	16
<u>Управление номерами</u>	16
<u>Исходящие звонки</u>	16
<u>Свой SIP (Custom)</u>	16
<u>Интеграция с VoxImplant</u>	17
<u>SIP Trunk (LiveKit)</u>	17
<u>Инструменты и интеграции</u>	17
<u>Обзор</u>	17
<u>Типы инструментов</u>	17
<u>Webhook-инструменты</u>	17
<u>Системные инструменты</u>	18
<u>Настройка инструмента</u>	18
<u>Речь во время выполнения инструмента</u>	18
<u>Правила переключения</u>	18
<u>Как это работает</u>	18
<u>Управление вебхуками</u>	18
<u>Примеры</u>	18
<u>Голос и коммуникация</u>	19
<u>Как это работает</u>	19
<u>Голосовой пайплайн</u>	19
<u>Распознавание речи (STT)</u>	19
<u>Языковая модель (LLM)</u>	19
<u>Синтез речи (TTS)</u>	20
<u>Детекция голосовой активности (VAD)</u>	20

Запись аудио	20
Уведомления в реальном времени	20
Жизненный цикл сессии	20
Виджет для сайта	20
Обзор	20
Установка	21
Настройка	21
Атрибуты	21
Передача переменных через URL	21
Настройка виджета в админ-панели	21
Как это работает	22
Требования	22

Описание функциональных характеристик

Обзор функционала

SpeakNode — платформа для создания и управления голосовыми AI-агентами, которые общаются с пользователями по телефону, через виджет на сайте или напрямую в браузере.

Основные возможности

Функция	Описание
AI-агенты	Создание и настройка голосовых агентов с кастомным поведением, личностью и знаниями
Голос и коммуникация	Голосовые разговоры в реальном времени на базе WebRTC
Виджет для сайта	Встраиваемый виджет для добавления голосового агента на любой сайт
Телефония	Входящие и исходящие звонки через Twilio и VoxImplant
Исходящие кампании	Массовые исходящие обзвоны с динамическими переменными
Инструменты и интеграции	Подключение агентов к внешним API и вебхукам
Аналитика	История разговоров, записи и телеметрия
Команды и пространства	Рабочие пространства для совместной работы
Биллинг	Тарификация по использованию

Как это работает

Пользователь настраивает агента → Агент запускается в LiveKit-комнате → AI-агент подключается и начинает разговор → Сессия записывается и анализируется

1. **Настройте** — задайте поведение агента: системный промпт, голос, язык, инструменты
2. **Разверните** — встройте виджет на сайт, подключите номер телефона или используйте API
3. **Общайтесь** — агент ведёт разговоры в реальном времени
4. **Анализируйте** — просматривайте записи, транскрипции и данные телеметрии

AI-агенты

AI-агенты — основа SpeakNode. Каждый агент — это настраиваемый голосовой ассистент, способный вести разговоры с пользователями в реальном времени.

Настройка агента

Основные параметры

- **Название** — имя агента (до 50 символов)
- **Системный промпт** — инструкции, определяющие поведение, личность и знания агента. Поддерживает динамические переменные в формате `{{имя_переменной}}`
- **Temperature** — контролирует случайность ответов LLM (0 = детерминированные, 1 = креативные)
- **Лимит токенов** — максимальное потребление токенов за разговор

Языки и локализация

Каждый агент поддерживает несколько языков. Для каждого языка настраивается:

- **Первое сообщение** — приветствие, которое агент произносит в начале разговора
- **TTS-голос** — голос и провайдер синтеза речи
- **Скорость речи** — множитель скорости воспроизведения
- **Фоновый звук** — звуковой фон во время разговора (тишина, офис, кофейня и т.д.)

Агент может автоматически определять язык пользователя и переключаться на соответствующую локаль.

LLM-модель

Выбор языковой модели, которая обеспечивает логику агента:

- LLM-провайдер и версия модели

- Temperature и дополнительные параметры модели

Рекомендованные модели для Voice Sales

Рейтинг лучших LLM для сценариев исходящих **голосовых продаж** — по критериям latency, качества русского языка и надёжности вызова инструментов.

Ключевые факторы для voice:

- **Latency (TTFT)** — время до первого токена напрямую влияет на воспринимаемую скорость ответа; модели с reasoning (chain-of-thought) добавляют 300–800 мс
- **Качество русского** — Qwen3 и GLM-4.7 обучены на больших кириллических корпусах
- **Управляемость** — насколько точно модель следует персоне, скриптам отработки возражений и схемам tool-call, не добавляя лишних шагов

Подробный разбор моделей

▢ **Qwen3-30B-A3B (DeepInfra)** — MoE-флагман среднего класса от Alibaba: 30.5B параметров с 3.3B активных за токен (8 из 128 экспертов), unified thinking/non-thinking режим (для voice используем non-thinking). В коротких репликах ведёт себя сдержанно, не сбивается в маркетинговый тон, чётко исполняет инструкции из system prompt (на independent IFEval $\approx 84.7\%$), уверенно вызывает tool'ы по схеме (BFCL v3 ≈ 65.1). Сильный русский — модель обучена на больших кириллических корпусах, не «соскальзывает» в английский даже когда клиент задаёт смешанные фразы. Хорошо держит диалог в роли менеджера: короткие фразы, один вопрос за раз, мало отступлений. Для Voice Sales — оптимальный baseline по соотношению качество/предсказуемость поведения. Источники: [Qwen3 technical report \(arXiv 2505.09388\)](#), [HuggingFace model card](#), [BFCL v3 leaderboard](#).

▢ **GLM-4.7 non-reasoning (Cerebras)** — agent-флагман от Zhipu AI / Z.ai, хостится на чип-инфре Cerebras. По t-bench (multi-turn agentic tasks: бронирование, retail-сценарии) даёт $\approx 87.4\%$ — один из лучших результатов среди open-weight моделей такого класса. В разговоре «инженерно-вежливый»: чётко закрывает шаги, аккуратно подтверждает услышанное, охотно вызывает tool'ы (parallel + single tool-calling из коробки). Важно: **строго в non-reasoning режиме** — thinking-блоки нелогично распухают на коротких репликах и режут latency. По русскому слабее Qwen3 (заметнее калька с английских конструкций), но если разговор скриптованный — справляется. Источники: [Cerebras GLM-4.7 announcement](#), [Cerebras inference docs](#), [t-bench paper \(arXiv 2406.12045\)](#).

▢ **Gemini 3.1 Flash Lite (Google AI Studio)** — managed Google-модель с контекстом 1M токенов и поддержкой мультимодальности (text/image/audio/video). Подкупает **предсказуемостью**: меньше сюрпризов open-weight моделей, стабильная грамматика на русском, ровная tone-of-voice. Tool-calling работает через structured outputs (без отдельной

BFCL-сертификации). Используется как managed-fallback: когда нужно гарантированно работающее SLO без своей инфры — отдаёт ровный результат, но в сравнении с Qwen3/GLM-4.7 проигрывает по богатству тонкого следования персоне и по сложности агентного поведения.

Источники: [Google model card](#), [Google Cloud blog](#).

4. Nemotron Super 49B (DeepInfra) — dense 49B-модель (NAS-сжатие с Llama-3.3-70B через NVIDIA «Puzzle»-алгоритм), 128K контекст, post-trained для reasoning, tool-calling и RAG. В разговоре звучит «академично-развёрнуто»: любит подробно проговаривать ход мысли, делает явные reasoning-шаги, что **противоречит инструкции "одна короткая реплика"** Voice Sales-сценария. Хорошо подходит для агентов где качество вывода важнее скорости (анализ длинных диалогов, итоговая саммаризация звонка, сложная фильтрация tool-результатов). Для короткой реплики менеджера — избыточна.

Источники: [NVIDIA Nemotron model card](#), [NVIDIA build page](#).

Резервные (fallback) LLM-модели

Под основным LLM-пикером в настройках агента можно добавить до **двух резервных моделей**. Каждая строка — отдельная модель + версия, и список **упорядоченный**: runtime идёт сверху вниз и использует первую модель, которая успешно ответила.

Когда срабатывает резерв. Fallback запускается только если предыдущая модель упала — недоступность провайдера, rate-limit / квота, network timeout, ошибка ключа. Успешный ответ (даже медленный) fallback **не** триггерит — отдаём то, что вернула активная модель.

Почему порядок важен.

- **Основная модель** работает на каждом ходу. Ставьте сюда ту, чьё звучание и tool-calling вы реально хотите слышать на звонках.
- **Fallback #1** — на самый вероятный сбой основной. Если primary хостится на Cerebras — fallback #1 разумно взять с DeepInfra (другой инференс-провайдер, похожее поведение): тональность диалога не уплывёт.
- **Fallback #2** — «во что бы то ни стало». Сюда хорошо ложится managed-модель Google / OpenAI: совсем другой вендор, предсказуемый русский, минимум риска что мульти-провайдерный outage положит агента целиком.

Промпт / персону между попытками мы не меняем — та же история диалога просто проигрывается следующей модели. Выбирайте резервы так, чтобы стиль ответа был достаточно близок к основному и абонент не заметил подмены посреди реплики.

Поведение в UI.

- В пикере каждой строки скрыты модели, уже выбранные на этом агенте (primary или другая fallback-строка) — задублировать цепочку случайно нельзя.
- Перетаскивайте ручку слева от строки (), чтобы сменить приоритет. Порядок в UI = порядок попыток.

- Кнопка «Select this version» внутри тестера модели прокидывает версию прямо в ту строку, из которой вы открыли тестер.
- Иконка корзины справа удаляет резервную строку. Можно удалить все — тогда у агента просто нет резервов и при падении primary ход проваливается.

Нюансы.

- Каждый fallback добавляет latency на пути сбоя (мы ждём пока primary упадёт прежде чем дёргать следующий). Не пихайте сюда много моделей в надежде замаскировать нестабильность — лучше починить primary.
- Токены и биллинг считаются по той модели, что фактически ответила, а не по primary.
- Если fallback-версия снята с производства провайдером, строка остаётся с оранжевым ⚠ — замените её на актуальную.

Распознавание речи (STT)

Настройка транскрипции речи пользователя:

- Модель и STT-провайдер
- Параметры, специфичные для модели

Детекция голосовой активности (VAD)

Управление детекцией речи пользователя:

- Выбор модели VAD
- Параметры чувствительности и таймингов
- Возможность отключить прерывания во время первого сообщения

Голос (TTS)

Настройка голоса агента для каждого языка:

- TTS-провайдер
- Выбор голоса и его параметры
- Регулировка скорости
- Тип и громкость фонового звука

Динамические переменные

Агенты поддерживают динамические переменные в системных промптах и первых сообщениях. Переменные задаются через OpenAPI-схему и подставляются при запуске.

Пример системного промпта:

Ты агент поддержки компании {{company_name}}.
Имя клиента — {{customer_name}}, номер заказа — {{order_id}}.

Переменные передаются через:

- API-вызов при диспатче
- URL-параметры (для виджета)
- Данные элемента кампании (для исходящих звонков)

Каждая переменная имеет:

- Имя
- Тип (string, number, boolean, integer)
- Описание
- Значение по умолчанию
- Флаг nullable

Инструменты агента

Агенты могут вызывать внешние API во время разговора. Подробнее в разделе [Инструменты и интеграции](#).

Публикация

Агент может быть опубликован для предоставления доступа:

- **Опубликованный агент** — доступен по публичной ссылке для тестирования
- **Виджет** — встраивается на любой сайт
- **Номер телефона** — привязывается к входящей/исходящей телефонии
- **API** — запускается программно

Дублирование агента

Любой агент можно продублировать, создав копию со всеми настройками. Полезно для создания вариаций или шаблонов.

Аналитика

SpeakNode собирает детальные данные о каждом разговоре для анализа и мониторинга.

История разговоров

Каждая сессия агента записывается с:

- **Метаданные сессии** — агент, звонящий, время начала/окончания, длительность, статус
- **Аудиозапись** — полное аудио сессии в объектном хранилище
- **Транскрипция** — текстовая расшифровка разговора
- **Телеметрия** — детальные OpenTelemetry-трейсы каждого шага

Список сессий

Просмотр всех разговоров с фильтрами:

- По агенту
- По статусу (завершён, ошибка, активен)
- По диапазону дат
- По информации о звонящем

Детали сессии

Для каждой сессии можно просмотреть:

- Полную транскрипцию с разбивкой по репликам
- Воспроизведение аудио с визуализацией волны
- Метрики по каждой реплике (задержка, длительность)
- Вызовы инструментов во время сессии
- Детали ошибок (если произошёл сбой)

Телеметрия и трейсинг

Платформа использует OpenTelemetry для сквозной трассировки:

Пользователь говорит → STT span → LLM span → Tool call span → TTS span →
Пользователь слышит

Каждый разговор — это трейс, содержащий спаны для каждой операции:

- **Имя спана** — тип операции (STT, LLM inference, tool call, TTS)
- **Длительность** — сколько времени занял каждый шаг
- **Статус** — успех или ошибка
- **Атрибуты** — теги и метаданные
- **События** — детальные события внутри спана

Пайплайн телеметрии

```
Python Worker → OTEL Collector → Kafka → ASP.NET Consumer → База данных
                                     → Langfuse (внутренний мониторинг)
```

Хранение данных

Данные	Хранилище
Метаданные сессий	PostgreSQL
Аудиозаписи	S3-совместимое объектное хранилище
Спаны телеметрии	PostgreSQL (через Kafka)
Агрегации аналитики	ClickHouse
LLM-трейсы	Langfuse

Биллинг

SpeakNode использует тарификацию по использованию на базе GetLago.

Модель биллинга

Тарификация основана на **голосовых минутах** — фактическом времени разговоров с AI-агентами.

Подписки и тарифы

Каждое пространство имеет тарифный план, определяющий:

- Включённые голосовые минуты за период
- Стоимость превышения
- Уровни доступа к функциям
- Лимиты по агентам

Отслеживание использования

Платформа автоматически отслеживает использование:

1. Каждая сессия агента фиксирует время начала и окончания

2. Длительность рассчитывается и передаётся в GetLago
3. GetLago агрегирует использование относительно тарифа
4. Счета формируются в конце биллингового периода

Обработка платежей

Платежи проходят через биллинговый движок платформы с подключённым защищённым платёжным провайдером:

- Автоматическая синхронизация платёжных методов
- Генерация и доставка счетов
- Биллинг за превышение

Управление биллингом

В панели управления:

- Просмотр текущего использования и оставшихся минут
- Управление тарифным планом
- Просмотр истории счетов
- Обновление платёжного метода

Исходящие кампании

SpeakNode позволяет запускать массовые исходящие обзвоны с помощью AI-агентов.

Обзор

Кампания — это набор исходящих звонков по списку контактов. Каждый звонок обрабатывается AI-агентом с динамически подставленными переменными для каждого контакта.

Настройка кампании

1. **Выберите агента** — укажите AI-агента, который будет обрабатывать звонки
2. **Добавьте контакты** — загрузите список с номерами телефонов и данными переменных
3. **Настройте переменные** — сопоставьте данные контакта с переменными агента (например, `{{customer_name}}`, `{{order_id}}`)
4. **Запустите** — платформа последовательно обзванивает каждый контакт

Элементы кампании

Каждый контакт в кампании — это **элемент кампании** с:

- Номером телефона для звонка
- Значениями переменных для конкретного звонка
- Отслеживанием статуса (ожидание, в процессе, завершён, ошибка)
- Ссылкой на сессию разговора после звонка

Как это работает

1. Создаётся кампания со списком элементов
2. Для каждого элемента платформа:
 - Создает LiveKit-комнату
 - Запускает агента с переменными элемента
 - Иницирует исходящий звонок через телефонного провайдера
 - Отслеживает статус сессии
3. Результаты доступны по каждому элементу (запись, транскрипция, статус)

Мониторинг

Прогресс кампании отображается в панели управления:

- Общий статус кампании
- Результаты звонков по каждому элементу
- Ссылки на отдельные сессии разговоров

Команды и пространства

SpeakNode поддерживает командную работу через рабочие пространства — **Spaces**.

Пространства

Спэс — это общее рабочее пространство, где участники команды совместно работают над агентами, кампаниями и настройками.

Что принадлежит пространству

- Агенты и их конфигурации
- Номера телефонов
- Инструменты и интеграции

- Кампании
- Вебхуки
- Биллинг и подписка

Управление пространством

- **Создание** нового пространства для команды
- **Приглашение** участников по email
- **Управление** ролями и правами участников

Роли пользователей

Роль	Возможности
Owner	Полный доступ, управление биллингом, управление участниками
Member	Создание и управление агентами, кампаниями, инструментами
Admin (платформа)	Настройки платформы, управление моделями, контроль пользователей

Приглашения

Участники приглашаются по email. Процесс приглашения:

1. Владелец пространства отправляет приглашение
2. Приглашённый получает email со ссылкой активации
3. Пользователь регистрируется или входит через Keycloak SSO
4. Пользователь добавляется в пространство

Аутентификация

Все пользователи авторизуются через Keycloak (OpenID Connect):

- Вход по email/паролю
- Интеграция с Google OAuth
- Поддержка SSO
- Управление сессиями на основе токенов

Телефония

SpeakNode интегрируется с телефонными провайдерами, позволяя AI-агентам принимать и совершать телефонные звонки.

Поддерживаемые провайдеры

Провайдер	Входящие	Исходящие	Примечания
Twilio	Да	Да	Прямая интеграция через SIP trunk
VoxImplant	Да	Да	Через WebSocket Media Bridge (нет прямого доступа к SIP)
Свой SIP (Custom)	Да	Да	Подключение собственной АТС / SIP-trunk по стандартному SIP-протоколу

Входящие звонки

При поступлении звонка на подключённый номер:

1. Телефонный провайдер отправляет вебхук на платформу
2. Вебхук принимается через Kafka
3. Бэкенд определяет агента, привязанного к номеру
4. Создаётся LiveKit-комната с SIP-участником
5. AI-агент запускается в комнате
6. Разговор начинается

Управление номерами

- Импорт существующих номеров из Twilio или VoxImplant
- Привязка номера к конкретному агенту
- Настройка параметров для каждого номера

Исходящие звонки

Исходящие звонки можно инициировать:

- Индивидуально через API
- Массово через [Исходящие кампании](#)

Платформа создаёт LiveKit-комнату, запускает агента и инициирует исходящий звонок через настроенного телефонного провайдера.

Свой SIP (Custom)

Помимо готовых интеграций с Twilio и VoxImplant, SpeakNode позволяет подключить **любую SIP-телефонию** — собственную АТС (Mango Office, Asterisk, FreeSWITCH, PBX-as-a-Service и т.д.) — по стандартному SIP-протоколу. LiveKit-сервер выступает SIP-узлом по обе стороны:

- **Входящие** (PBX → LiveKit): ваш SIP-провайдер отправляет `INVITE` на публичный SIP-домен LiveKit. Проверка по списку разрешённых IP (allowlist) и опциональной Digest-аутентификации.
- **Исходящие** (LiveKit → PBX): LiveKit инициирует `INVITE` к SIP-серверу вашей АТС. Аутентификация — опциональная Digest-пара логин/пароль либо IP-trust (PBX доверяет нашим IP).

Для каждого аккаунта администратор задаёт SIP-сервер, транспорт, страну назначения, список разрешённых IP и пары кренделов. После создания аккаунта платформа отдаёт **домен публичного SIP-шлюза** и **список наших исходящих IP** — эти данные администратор передаёт владельцу АТС, чтобы тот настроил маршрутизацию.

Пошаговая настройка обеих сторон — в руководстве [«Свой SIP \(Custom\)»](#).

Интеграция с VoxImplant

VoxImplant не предоставляет прямой доступ к SIP trunk. Платформа использует **WebSocket Media Bridge** — сервис на Node.js, который:

1. Получает аудио от VoxImplant через WebSocket
2. Стримит его в LiveKit-комнату
3. Отправляет аудио из LiveKit обратно в VoxImplant

Это обеспечивает полную функциональность агента с VoxImplant в качестве телефонного бэкенда.

SIP Trunk (LiveKit)

LiveKit Server включает встроенный SIP-компонент, который обеспечивает:

- Регистрацию SIP trunk
- Правила диспатча для маршрутизации звонков в комнаты
- Прямой мост SIP-to-WebRTC (для провайдеров с поддержкой SIP)

Инструменты и интеграции

Агенты SpeakNode могут вызывать внешние API во время разговоров, получая данные, выполняя действия и интегрируясь с вашими системами.

Обзор

Инструменты — это HTTP-интеграции, которые LLM может вызывать во время разговора. Когда агент решает, что ему нужны внешние данные или нужно выполнить действие, он вызывает настроенный инструмент и использует результат в своём ответе.

Типы инструментов

Webhook-инструменты

Пользовательские HTTP-эндпоинты, которые агент может вызывать:

- **Метод** — GET, POST, PUT, DELETE
- **URL** — адрес эндпоинта (поддерживает подстановку переменных)
- **Заголовки** — кастомные заголовки, определённые через OpenAPI-схему
- **Тело** — тело запроса, определённое через OpenAPI-схему
- **Описание** — объясняет LLM, когда и как использовать инструмент

Системные инструменты

Встроенные инструменты платформы (например, определение языка, перевод разговора).

Настройка инструмента

Каждый инструмент определяется OpenAPI-схемой для его параметров:

- **Схема тела** — определяет структуру JSON-тела, которое LLM должен сформировать
- **Схема заголовков** — определяет кастомные заголовки для отправки
- **Код** — уникальный идентификатор, по которому LLM ссылается на инструмент

Речь во время выполнения инструмента

Инструменты поддерживают настраиваемую речь при выполнении:

- **Pre-tool speech** — что агент говорит перед вызовом инструмента (например, «Сейчас проверю...»)
- **Progress speech** — что агент говорит в ожидании ответа

Правила переключения

Инструменты могут включать правила переключения, перенаправляющие разговор:

- **Переключение на другого агента** — передача разговора другому AI-агенту

- **Условие** — когда должно произойти переключение

Это позволяет создавать мультиагентные сценарии, где специализированные агенты обрабатывают конкретные темы.

Как это работает

1. LLM решает вызвать инструмент на основе контекста разговора
2. Python worker формирует HTTP-запрос по OpenAPI-схеме
3. Запрос отправляется на настроенный эндпоинт
4. Ответ передаётся обратно в LLM
5. LLM включает результат в свой ответ

Управление вебхуками

Платформа также предоставляет отдельную систему вебхуков для приёма событий:

- Настройка эндпоинтов для получения событий сессий
- Просмотр истории запросов и полезных нагрузок
- Управление разрешёнными путями и маршрутизацией через Webhook Receiver Client

Примеры

Поиск в CRM:

```
Инструмент: get_customer_info
Метод: GET
URL: https://crm.example.com/api/customers/{{customer_id}}
Описание: Поиск информации о клиенте по ID
```

Статус заказа:

```
Инструмент: check_order_status
Метод: POST
URL: https://api.example.com/orders/status
Тело: { "order_id": "string" }
Описание: Проверить текущий статус заказа клиента
```

Голос и коммуникация

SpeakNode использует WebRTC для голосовой связи в реальном времени между пользователями и AI-агентами.

Как это работает

1. При запуске агента создаётся **LiveKit-комната**
2. **AI-агент** (Python worker) подключается к комнате и начинает слушать
3. **Пользователь** подключается через браузер (виджет или клиент) или телефон
4. Аудио передаётся в обе стороны в реальном времени
5. Сессия **записывается** автоматически (audio egress)

Голосовой пайплайн

Пользователь говорит → WebRTC аудио → STT (речь в текст) →
LLM обрабатывает текст → генерирует ответ →
TTS (текст в речь) → WebRTC аудио → Пользователь слышит

Распознавание речи (STT)

Преобразует речь пользователя в текст для LLM. STT-провайдер настраивается для каждого агента отдельно.

Языковая модель (LLM)

Обрабатывает разговор и генерирует ответы. LLM-провайдер и модель настраиваются для каждого агента отдельно (напрямую или через агрегатор).

Синтез речи (TTS)

Преобразует ответы агента в речь. TTS-провайдер настраивается для каждого агента отдельно.

Детекция голосовой активности (VAD)

Определяет, когда пользователь начинает и прекращает говорить. Управляет поведением очередности — когда агент должен начать или прекратить говорить.

Запись аудио

Каждый разговор автоматически записывается через LiveKit Egress. Записи включают:

- Полное аудио сессии (композитная запись)
- Треки отдельных участников
- Хранение в S3-совместимом объектном хранилище

Уведомления в реальном времени

Платформа использует SignalR для уведомления фронтенда о событиях сессии:

- **Agent Ready** — агент подключился к комнате
- **Session Completed** — разговор завершён, аудио доступно

Жизненный цикл сессии

Статус	Описание
Pending	Сессия создана, комната ещё не готова
Dispatching	LiveKit-комната создана, ожидание подключения агента
Active	Агент подключился, разговор идёт
Completed	Разговор завершён нормально
Failed	Произошла ошибка во время сессии

Виджет для сайта

SpeakNode предоставляет встраиваемый веб-компонент, который добавляет голосового агента на любой сайт.

Обзор

Виджет — это Angular веб-компонент, упакованный как `<speaknode-agent>`. Он подключается к опубликованному агенту и позволяет посетителям сайта вести голосовой разговор прямо в браузере.

Установка

Разместите кастомный элемент в нужном месте страницы, а скрипт-загрузчик добавьте либо сразу после элемента, либо перед `</body>`:

```
<speaknode-agent agent-id="your-agent-id"></speaknode-agent>
<script src="https://widget.dev.voxagent.ru/speaknode-widget-loader.js">
</script>
```

Скрипт-загрузчик регистрирует кастомный элемент `<speaknode-agent>`, подгружает конфигурацию опубликованного агента и инициализирует виджет, когда пользователь нажимает кнопку начала разговора.

Настройка

Атрибуты

Атрибут	Тип	Описание
<code>agent-id</code>	string	ID опубликованного агента
<code>agent-token</code>	string	Предгенерированный JWT-токен (опционально)
<code>language</code>	string	Переопределение языковой локали
<code>initialized</code>	boolean	Режим внешней конфигурации

Передача переменных через URL

Динамические переменные можно передавать как query-параметры URL. Виджет считывает их автоматически и передаёт агенту при запуске.

Пример:

```
https://yoursite.com/page?customer_name=Иван&order_id=12345
```

Настройка виджета в админ-панели

На странице редактирования агента вкладка **Widget** позволяет:

- Настроить расположение и внешний вид виджета
- Сгенерировать код для встраивания
- Задать кастомные цвета темы

Как это работает

1. Виджет загружается на странице и получает публичную конфигурацию агента
2. Пользователь нажимает, чтобы начать разговор
3. Виджет запрашивает токен диспатча у бэкенда
4. Создаётся LiveKit-комната, запускается агент
5. Виджет подключается к комнате через WebRTC
6. Голосовой разговор начинается в реальном времени
7. После завершения разговора сессия сохраняется

Требования

- Агент должен быть **опубликован** (`is_published: true`)
- Браузер пользователя должен поддерживать WebRTC
- Необходим доступ к микрофону