# О продукте

**Audiogram** – это сервис на базе нейронных сетей и методов машинного обучения для распознавания и синтеза речи. Audiogram позволяет выполнять:

- **Синхронное распознавание речи**. В этом случае сервис получает запрос с аудиофайлом, который необходимо расшифровать, и возвращает распознанный текст. Данный способ выполняется последовательно и считается наиболее точным. Подходит, например, для расшифровки телефонных разговоров.
- **Потоковое распознавание речи**. В случае потокового распознавания устанавливается соединение с Audiogram, по которому речь говорящего отправляется на распознавание частями в режиме online. Сервис возвращает результаты расшифровки по мере обработки. Данный способ подходит, например, для создания голосовых помощников или субтитров к видео.
- **Синхронный синтез речи**. При синхронном синтезе запрос к Audiogram содержит текст, который необходимо озвучить, и дополнительную информацию по голосу, частоте дискретизации и кодировке. В ответ возвращается аудиофайл с озвученным текстом. Этот способ может использоваться, например, для озвучивания книг.
- **Потоковый синтез речи**. При потоковом синтезе текст отправляется в Audiogram и озвучивается по частям. Потоковый синтез подходит, например, для создания ответных реплик голосовых помощников, так как позволяет достичь эффекта живого общения без неестественных пауз.
- **Сбор аудиоартефактов**. Аудиофайлы, поступающие в Audiogram на распознавание речи, сохраняются в отдельном хранилище и могут быть использованы для обучения и усовершенствования ML-моделей, отвечающих за расшифровку речи.
- **Управление клиентами и просмотр статистики**. Это можно сделать с помощью удобного веб-клиента в любом браузере.

# Термины в документе

- **Audiogram** (также **Продукт**, **Система**) – сервис, выполняющий услуги по распознаванию речи (превращению аудиозаписей с речью в текст) и синтезированию речи (озвучиванию текстов).
- **ASR** (**Automatic Speech Recognition**) – запрос на распознавание речи.
- **TTS** (**Text-to-Speech**) – запрос на синтезирование речи.
- **ML-модель** (также **нейросеть**, **искусственный интеллект**) – программа, обученная распознаванию определенных типов закономерностей, которая используется в Audiogram для автоматизации и ускорения выполнения запросов на синтез и распознавание речи. В Audiogram используются различные ML-модели (в зависимости от типа запроса и деталей запросов).

- **Бот** (также **чат-бот**, **электронный помощник**) – программа, отвечающая на запросы пользователей и имитирующая живое общение между людьми.

# Демонстрация Audiogram

Вы можете бесплатно попробовать отправить какой-нибудь текст на озвучку или аудиофайл на распознавание, используя демонстрационную форму Audiogram, которая доступна НА ОСНОВНОЙ СТРАНИЦЕ ПРОДУКТА.

# Варианты поставки Audiogram

Возможны 2 варианта поставки Audiogram:

- **SaaS**: Audiogram установлен в облаке MTS AI. Доступ к сервису осуществляется через подключение по API (gRPC).
- **On-premise**: Audiogram развернут и функционирует в инфраструктуре заказчика.

# Справочник API

## Распознавание речи

Для взаимодействия с API распознавания речи используется протокол gRPC.

**Примечание**: Подробности об этом протоколе можно прочитать на HTTPS://GRPC.IO/

Чтобы пользоваться сервисом **Audiogram** для распознавания речи нужно создать клиентское приложение. Можно использовать любой язык программирования, который есть в библиотеке для работы с gRPC.

При написании приложения используйте PROTO-ФАЙЛ **STT.PROTO**.

Максимальная длина сообщения, принимаемого от клиентов по gRPC (в байтах): 31457280

### Методы

При обращении по gRPC-протоколу клиентское приложение использует нужный метод сервиса.

## Recognize

Распознавание аудио целиком.

| Имя метода | Тип запроса | Тип ответа | Описание |
| --- | --- | --- | --- |
| Recognize | RECOGNIZEREQUEST | RECOGNIZERESPONSE | Метод распознавания аудиофайла целиком. Ожидает аудиофайл, в этом же соединении возвращает результат и закрывает соединение. |

## StreamingRecognize

Потоковое распознавание речи.

| Имя метода | Тип запроса | Тип ответа | Описание |
| --- | --- | --- | --- |
| StreamingRecognize | stream STREAMINGRECOGNIZEREQUEST | stream STREAMINGRECOGNIZERESPONSE | Метод поточного распознавания. Принимает аудиоданные по мере их доступности. Распознавание заканчивается, когда поток закрывается клиентом. |

## GetModelsInfo

Запрос моделей для распознавания речи.

| Имя метода | Тип запроса | Тип ответа | Описание |
| --- | --- | --- | --- |
| GetModelsInfo | google.protobuf.Empty | MODELSINFO | Метод запроса списка моделей для распознавания речи. Ничего не принимает в качестве аргументов, возвращает список доступных моделей. |

# Сообщения PROTOBUF для распознавания речи

## VoiceActivityMark

Определяет разметку голосовой активности во входном акустическом сигнале. Сообщение включает в себя метку времени и тип метки.

| Поле | Тип | Описание |
| --- | --- | --- |
| mark_type | VOICEACTIVITYMARKTYPE | Тип разметки. |
| offset_ms | uint64 | Метка времени с точкой отсчета начала входного акустического сигнала, единицы измерения - миллисекунды. |

## VoiceActivityMarkType

Определяет тип метки голосовой активности.

| Имя | Значение | Описание |
| --- | --- | --- |
| VA_MARK_NONE | 0 | Тип метки отсутствия изменения голосовой активности. |
| VA_MARK_BEGIN | 1 | Тип метки начала голосовой активности. |
| VA_MARK_END | 2 | Тип метки конца голосовой активности. |

## VoiceActivityDetectionAlgorithmUsage

Тип используемого алгоритма VoiceActivity.

| Имя | Значение | Описание |
| --- | --- | --- |

| | | |
|---|---|---|
| USE_VAD | 0 | Включает использование VAD-алгоритма (Voice Activity Detection) для разбиения на фразы. Этот алгоритм определяет окончание фразы по паузам в речи и тишине. |
| DO_NOT_PERFORM_VOICE_ACTIVITY | 1 | Отключает разбиение на фразы по Voice Activity. То есть весь распознанный текст будет получен в виде одной фразы. |
| USE_DEP | 2 | Включает использование DEP-алгоритма для разбиения на фразы. Этот алгоритм реализует более сложную логику определения окончания фразы по законченности мысли. |

## VoiceActivityMarkEventsMode

Режим отправки VoiceActivity разметки клиенту.

| Имя | Значение | Описание |
|---|---|---|
| VA_DISABLE | 0 | Отключает отправку отметок VOICEACTIVITYMARK. |
| VA_ENABLE | 1 | Включает отправку отметок VOICEACTIVITYMARK синхронно вместе с транскрипцией. |
| VA_ENABLE_ASYNC | 2 | Включает отправку отметок VOICEACTIVITYMARK асинхронно (как только будет получена разметка, не дожидаясь работы asr). Для файлового режима работает идентично ENABLE. |

## VADOptions

Настройки работы алгоритма VAD.

| Поле | Тип | Описание |
|---|---|---|
| threshold | float | Порог срабатывания VAD. Если вероятность речи выше порога, значит обработанный чанк содержит речь. Возможные значения: (0, 1.0]. Значение по умолчанию - 0.2 |

| | | |
|---|---|---|
| speech_pad_ms | int32 | Отступ, добавляемый к границам найденных фрагментов (если speech_pad_ms < 0, отступ будет "внутрь" фрагмента). Опция применима только для offline режима (для Recognize() запросов). Единицы измерения - миллисекунды. Значение по умолчанию - 300 |
| min_silence_ms | uint32 | Если между двумя фрагментами речи встречается пауза короче min_silence_ms, то такая пауза не учитывается и фрагменты объединяются в один. Единицы измерения - миллисекунды. Возможные значения: min_silence_ms >= 0. Значение по умолчанию - 100 мс |
| min_speech_ms | uint32 | Минимальная продолжительность речи. Фрагменты короче min_speech_ms не учитываются. Опция применима только для offline режима (для Recognize() запросов). Единицы измерения - миллисекунды. Возможные значения: min_speech_ms >= 0. Значение по умолчанию - 250 |
| mode | VOICEACTIVITYDETECTIONMODE | Выбор типа разметки VAD-ом аудио файла для файлового запроса. |

## VoiceActivityDetectionMode

Выбор типа разметки аудио с помощью VAD.

| Имя | Значение | Описание |
|---|---|---|
| VAD_MODE_DEFAULT | 0 | Значение по умолчанию для файлового режима - ONLY_SPEECH, для стримингового (потокового) режима - SPLIT_BY_PAUSES. |
| SPLIT_BY_PAUSES | 1 | Аудио разделяется по паузам (ничего не вырезается). |
| ONLY_SPEECH | 2 | Вырезаются только сегменты с речью. |

## DEPOptions

Настройки работы алгоритма DEP.

| Поле | Тип | Описание |
|---|---|---|

| | | |
|---|---|---|
| smoothed_window_threshold | float | Порог срабатывания алгоритма DEP. На заданном окне сглаживания считается среднее значение вероятности завершения фразы. Если это значение больше порога то алгоритм срабатывает. Возможные значения: (0, 1.0). Значение по умолчанию - 0.754 |
| smoothed_window_ms | int32 | Окно, на котором происходит сглаживание при принятии решения о конце фразы. Единицы измерения - миллисекунды. Возможные значения: smoothed_window_ms >= 10. Значение по умолчанию - 970 мс. Значение должно быть кратно 10 мс |

## VoiceActivityConfig

Структура данных для хранения всех настроек VoiceActivity.

| Поле | Тип | Описание |
|---|---|---|
| dep_options | DEPOPTIONS | Опции алгоритма DEP. Используется при VoiceActivityDetectionAlgorithmUsage = USE_DEP |
| usage | VOICEACTIVITYDETECTIONALGORITHMUSAGE | Выбор алгоритма VoiceActivity. При DO_NOT_PERFORM_VOICE_ACTIVITY разметка аудио выключена. Значение по умолчанию - USE_VAD |
| vad_options | VADOPTIONS | Опции алгоритма VAD. Используется при VoiceActivityDetectionAlgorithmUsage = USE_VAD |

## AudioEncoding

Поддерживаемые форматы аудиоданных.

| Имя | Значение | Описание |
|---|---|---|
| ENCODING_UNSPECIFIED | 0 | На текущий момент не поддерживается. |
| LINEAR_PCM | 1 | PCM без заголовков с целыми знаковыми 16-битными сэмплами в линейном распределении (PCM 16bit). |

| | | |
|---|---|---|
| FLAC | 2 | На текущий момент не поддерживается. |
| MULAW | 3 | PCM без заголовков с 8-битными сэмплами в формате mu-law. |
| ALAW | 20 | PCM без заголовков с 8-битными сэмплами в формате a-law. |

## AttackType

Тип атаки.

| Имя | Значение | Описание |
|---|---|---|
| LOGICAL | 0 | Логическая атака. |
| PHYSICAL | 1 | Физическая атака (пока не поддерживается). |
| ALL_TYPES | 2 | Оба вида атак (логический и физический) (пока не поддерживается). |

## AntiSpoofingConfig

Конфигурация антиспуфинга.

| Поле | Тип | Описание |
|---|---|---|
| type | ATTACKTYPE | Тип атаки. |
| FAR | float | Допустимый процент принятия ботов за людей. |
| FRR | float | Допустимый процент отклонения людей (принятия их за ботов). |
| max_duration_for_analysis_ms | uint32 | Максимальная длительность анализа (в миллисекундах). Значение по умолчанию - 5000 миллисекунд. |

# RecognitionConfig

Конфигурация распознавания при вызове метода Recognize.

| Поле | Тип | Описание |
| --- | --- | --- |
| encoding | enum AUDIOENCODING | Формат аудио данных (кодировка). |
| sample_rate_hertz | int32 | Частота дискретизации аудио данных в герцах. |
| language_code | string | Язык, используемый в аудио файле. |
| max_alternatives | int32 | Максимальное количество возвращаемых гипотез.<br>Сервер может вернуть меньше, чем указано в этом параметре.<br>По умолчанию количество возвращаемых гипотез распознавания равно 1. |
| audio_channel_count | int32 | Количество каналов во входных аудиоданных. |
| enable_word_time_offsets | bool | Флаг, включающий вывод временных меток слов.<br>• при значении «true» лучший результат включает временные метки для этих слов;<br>• при значении «false» информация о временных метках не возвращается. |
| enable_automatic_punctuation | bool | Флаг, включающий модуль автоматической расстановки знаков препинания. |
| model | string | Модель распознавания. |
| va_config | VOICEACTIVITYCONFIG | Конфигурация Voice Activity. |
| va_response_mode | VOICEACTIVITYMARKEVENTSMODE | Режим отправки разметки клиенту. По умолчанию – VA_DISABLE. |
| enable_genderage | bool | Флаг, включающий модуль определения пола и возраста говорящего.<br>Примечание: пол ребенка в настоящий момент не определяется. |
| split_by_channel | bool | Этот флаг работает только для распознавания в файловом режиме. При его включении каждый канал будет распознаваться отдельно. Может применяться, |

| | | |
|---|---|---|
| | | например, для аудио из колл-центров, где в одном канале голос клиента, а в другом - оператора. |
| enable_antispoofing | bool | Флаг, включающий определение спуфинг-атак. |
| antispoofing_config | ANTISPOOFINGCONFIG | Конфигурация антиспуфинга. |

## StreamingRecognitionConfig

Конфигурация распознавания при вызове метода StreamingRecognize.

| Поле | Тип | Описание |
|---|---|---|
| config | RECOGNITIONCONFIG | Конфигурация распознавания. |
| single_utterance | bool | Флаг для включения режима распознавания одной фразы. В этом режиме (при выставленном значении «true») распознавание завершается сервисом сразу после распознавания первой фразы и соединение разрывается. |
| interim_results | bool | Конфигурация для промежуточных результатов:<br>• при значении «true» возвращаются промежуточные результаты (промежуточные гипотезы) и конечные результаты;<br>• при значении «false» возвращаются только конечные результаты (у которых is_final = true). |

## StreamingRecognizeRequest

Запрос на распознавание аудио.

Первое сообщение типа StreamingRecognizeRequest должно содержать данные в поле «streaming_config» и не должно содержать данные в поле «audio». Все последующие сообщения StreamingRecognizeRequest наоборот не должны иметь данных в поле «streaming_config», а в поле «audio» передаются аудиоданные. Аудиобайты должны быть закодированы, как указано в RECOGNITIONCONFIG.

| Поле | Тип | Описание |
| --- | --- | --- |
| streaming_config | STREAMINGRECOGNITIONCONFIG | Конфигурация распознания. |
| audio | bytes | Аудиофрагменты для распознания. |

## RecognitionAudio

Аудио, отправляемое на распознавание.

| Поле | Тип | Описание |
| --- | --- | --- |
| audio | bytes | Аудиофрагменты для распознания. |
| uri | string | Ссылка на скачивание аудио.<br>В настоящий момент не поддерживается. |

## RecognizeRequest

Запрос на распознавание аудио.

| Поле | Тип | Описание |
| --- | --- | --- |
| config | RECOGNITIONCONFIG | Конфигурация распознавания. |
| audio_content | RECOGNITIONAUDIO | Структура для передачи аудиоданных в метод распознавания. |

## RecognizeResponse

Ответ с результатами распознавания для метода RECOGNIZE.

| Поле | Тип | Описание |
| --- | --- | --- |

| | | |
|---|---|---|
| results | repeated SPEECHRECOGNITIONRESULT | Результат распознавания. |

## StreamingRecognizeResponse

Ответ с результатами распознавания для метода STREAMINGRECOGNIZE.

| Поле | Тип | Описание |
|---|---|---|
| results | repeated STREAMINGRECOGNITIONRESULT | Этот повторяющийся список содержит последнюю расшифровку, соответствующую аудио, которое в настоящее время обрабатывается. В настоящее время возвращается один результат, причем каждый результат может иметь несколько значений (альтернативы). |

## EmotionsRecognition

Эмоциональная окраска голоса спикера.

| Поле | Тип | Описание |
|---|---|---|
| positive | float | Положительный тон. |
| neutral | float | Нейтральный тон. |
| negative_angry | float | Сердитый тон. |
| negative_sad | float | Печальный тон (в данный момент этот параметр не поддерживается). |

## SpeakerGenderAgePrediction

Пол, возраст и эмоциональная окраска голоса спикера.

| Поле | Тип | Описание |
| --- | --- | --- |
| gender | GENDERCLASS | Значение, определяющее пол. |
| age | AGECLASS | Значение, определяющее возраст. |
| emotion | EMOTIONSRECOGNITION | Значение, определяющее эмоциональную окраску голоса спикера. |

## GenderClass

Пол говорящего.

| Имя | Значение | Описание |
| --- | --- | --- |
| GENDER_UNDEF | 0 | Пол не определен. |
| GENDER_MALE | 1 | Мужчина. |
| GENDER_FEMALE | 2 | Женщина. |

## AgeClass

Возраст говорящего.

| Имя | Значение | Описание |
| --- | --- | --- |
| AGE_UNDEF | 0 | Возраст не определен. |
| AGE_ADULT | 1 | Взрослый. |
| AGE_CHILD | 2 | Ребенок. |

## SpoofingResult

| Поле | Тип | Описание |
| --- | --- | --- |
| type | ATTACKTYPE | Тип спуфинг-атаки. |
| result | ATTACKRESULT | Результат определения является ли звонок спуфинг-атакой. |
| confidence | float | Уверенность в принятом решении в поле result. |
| start_time_ms | uint64 | Начальная метка временного отрезка, который анализировался на предмет спуфинг-атаки. |
| end_time_ms | uint64 | Конечная метка временного отрезка, который анализировался на предмет спуфинг-атаки. |

## AttackResult

| Имя | Значение | Описание |
| --- | --- | --- |
| ATTACK_DETECTED | 0 | Зафиксирована спуфинг-атака (бот пытается выдать себя за человека). |
| GENUINE | 1 | Голос принадлежит человеку. |

## SpeechRecognitionAlternative

Альтернативные гипотезы.

| Поле | Тип | Описание |
| --- | --- | --- |
| transcript | string | Распознанный текст. |
| confidence | float | Коэффициент достоверности (степень уверенности) распознанной фразы. |
| words | repeated WORDINFO | Список информации об объектах распознанных слов. |

| | | |
|---|---|---|
| start_time | google.protobuf.Duration | Временная метка начала фразы относительно начала аудиопотока. |
| end_time | google.protobuf.Duration | Временная метка конца фразы относительно начала аудиопотока. |

## WordInfo

Объект, содержащий информацию, относящуюся к распознанному слову.

| Поле | Тип | Описание |
|---|---|---|
| start_time | google.protobuf.Duration | Временная метка начала слова относительно начала аудиопотока. |
| end_time | google.protobuf.Duration | Временная метка конца слова относительно начала аудиопотока. |
| word | string | Распознанное слово. |
| confidence | float | Коэффициент достоверности (степень уверенности) распознанного слова. |

## StreamingRecognitionResult

Результат вызова метода распознавания потокового аудио STREAMINGRECOGNIZE.

| Поле | Тип | Описание |
|---|---|---|
| result | SPEECHRECOGNITIONRESULT | Результат распознавания. |
| is_final | bool | Флаг, указывающий, что сформирована окончательная гипотеза и меняться она больше не будет:<br>• «true», если пришла финальная гипотеза;<br>• «false» для промежуточных гипотез. |

# SpeechRecognitionResult

Результат вызова метода файлового распознавания аудио RECOGNIZE.

| Поле | Тип | Описание |
| --- | --- | --- |
| alternatives | repeated SPEECHRECOGNITIONALTERNATIVE | Список альтернативных результатов распознавания.<br>Может содержать одну или несколько гипотез распознавания (максимальное значение указано в параметре RecognitionConfig.max_alternatives).<br>Альтернативы упорядочены по точности, причем первая альтернатива является наиболее вероятной в соответствии с SpeechRecognitionAlternative.confidence. |
| channel | int32 | Идентификатор канала (в настоящее время данный параметр не поддерживается). |
| va_marks | repeated VOICEACTIVITYMARK | Voice Activity разметка. Массив меток отправляется только если VOICEACTIVITYMARKEVENTSMODE = VA_ENABLE / VA_ENABLE_ASYNC. При VoiceActivityMarkEventsMode = VA_ENABLE_ASYNC все остальные поля структуры SpeechRecognitionResult могут быть пустые. |
| genderage | SPEAKERGENDERAGEPREDICTION | Результат работы модели классификации мужчина/женщина/ребенок. Включается флагом enable_genderage в RECOGNITIONCONFIG. |
| spoofing_result | repeated SPOOFINGRESULT | Результат работы модели антиспуфинга. Включается флагом enable_antispoofing в RECOGNITIONCONFIG. |

# ModelsInfo

Доступные модели распознавания речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| models | repeated MODELINFO | Список доступных моделей. |

## ModelInfo

Информация о модели распознавания речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| name | string | Имя модели распознавания речи. |
| sample_rate_hertz | uint32 | Частота дискретизации аудио данных в герцах. |
| language_code | string | Язык, используемый в аудио файле, по умолчанию ru. |

# Синтез речи

Взаимодействие с API синтеза речи осуществляется по протоколу gRPC.

**Примечание**: Подробности об этом протоколе можно прочитать на HTTPS://GRPC.IO/

Чтобы пользоваться сервисом **Audiogram** для синтеза речи нужно создать клиентское приложение. Можно использовать любой язык программирования, который есть в библиотеке для работы с gRPC.

При написании приложения используйте PROTO-ФАЙЛ **TTS.PROTO**.

Максимальная длина сообщения, принимаемого от клиентов по gRPC (в байтах): 31457280

# Методы

## StreamingSynthesize

Потоковый синтез речи.

| Имя метода | Тип запроса | Тип ответа | Описание |
| --- | --- | --- | --- |

| | | | |
|---|---|---|---|
| StreamingSynthesize | SYNTHESIZESPEECHREQUEST | stream STREAMINGSYNTHESIZESPEECHRESPONSE | Метод потокового синтеза речи. Разбивает текст на короткие фразы, и возвращает результат по мере их синтеза. |

## Synthesize

Синхронный файловый синтез речи.

| Имя метода | Тип запроса | Тип ответа | Описание |
|---|---|---|---|
| Synthesize | SYNTHESIZESPEECHREQUEST | SYNTHESIZESPEECHRESPONSE | Метод синхронного файлового синтеза речи. Возвращает целый аудиофайл в формате, заданном в encoding с заголовками выбранного контейнера, пригодный для сохранения на диск. |

## GetModelsInfo

Запрос моделей для синтеза речи.

| Имя метода | Тип запроса | Тип ответа | Описание |
|---|---|---|---|
| GetModelsInfo | google.protobuf.Empty | MODELSINFO | Метод запроса списка моделей для синтеза речи. Ничего не принимает в качестве аргументов, возвращает список доступных моделей. |

# Сообщения PROTOBUF для синтеза речи

## AudioEncoding

Поддерживаемые форматы аудиоданных.

| Имя | Значение | Описание |
|---|---|---|

| | | |
|---|---|---|
| ENCODING_UNSPECIFIED | 0 | На текущий момент не поддерживается. |
| LINEAR_PCM | 1 | 1. Если данное поле выбрано при использовании метода SYNTHESIZE то в поле SynthesizeSpeechResponse.audio вернётся WAV linear PCM аудиофайл с заголовком, содержащий целые знаковые 16-битные сэмплы в линейном распределении (PCM 16bit) и заданной частотой дискретизации в соответствии с полем sample_rate_hertz.<br>2. При использовании метода STREAMINGSYNTHESIZE в поле StreamingSynthesizeSpeechResponse.audio по мере синтеза отправляются чанки linear PCM без заголовка WAV с целыми знаковыми 16-битными сэмплами в линейном распределении (PCM 16bit). |
| FLAC | 2 | На текущий момент не поддерживается. |
| MULAW | 3 | 1. Если данное поле выбрано при использовании метода SYNTHESIZE, то в поле SynthesizeSpeechResponse.audio вернётся WAV PCM аудиофайл с заголовком, содержащий 8-битные сэмплы в формате mu-law и заданной частотой дискретизации в соответствии с полем sample_rate_hertz.<br>2. При использовании метода STREAMINGSYNTHESIZE в поле StreamingSynthesizeSpeechResponse.audio по мере синтеза отправляются чанки PCM без заголовка WAV с 8-битными сэмплами в формате mu-law. |
| ALAW | 20 | 1. Если данное поле выбрано при использовании метода SYNTHESIZE, то в поле SynthesizeSpeechResponse.audio вернётся WAV PCM аудиофайл с заголовком, содержащий 8-битные сэмплы в формате a-law и заданной частотой дискретизации в соответствии с полем sample_rate_hertz.<br>2. При использовании метода STREAMINGSYNTHESIZE в поле StreamingSynthesizeSpeechResponse.audio по мере синтеза отправляются чанки PCM без заголовка WAV с 8-битными сэмплами в формате a-law. |

## VoiceStyle

Эмоциональная окраска голоса.

| Имя | Значение | Описание |
|---|---|---|
| VOICE_STYLE_NEUTRAL | 0 | Спокойное состояние. |

| | | |
|---|---|---|
| VOICE_STYLE_HAPPY | 1 | Радость. |
| VOICE_STYLE_ANGRY | 2 | Злость. |
| VOICE_STYLE_SAD | 3 | Грусть. |
| VOICE_STYLE_SURPRISED | 4 | Удивление. |

## SynthesizeOptions

Опции синтеза.

| Поле | Тип | Описание |
|---|---|---|
| model_type | string | Тип модели. Доступные варианты:<br>1) **light** - это менее ресурсоемкая модель, рекомендуется для синтеза в файловом и потоковом режимах при промышленном использовании, когда важен быстрый ответ системы;<br>2) **high_quality** - относится к поколению моделей, которые отличаются улучшенными характеристиками синтеза речи, но потребляет больше системных ресурсов. Рекомендуется к использованию при небольшой нагрузке.<br><br>В запросе на синтез укажите название модели, которую необходимо использовать. Если не передать название модели, то при синтезе в потоковом режиме по умолчанию будет использоваться **light**, а при синтезе в файловом режиме - **high_quality**. |
| model_sample_rate_hertz | uint32 | Частота дискретизации модели (в герцах). Если поле не указано, то будет подобрана наиболее близкая модель к указанной частоте дискретизации аудио. |
| voice_style | VOICESTYLE | Стиль речи. Значение по умолчанию - VOICE_STYLE_NEUTRAL |
| postprocessing_mode | POSTPROCESSINGMODE | Постобработка аудио. |
| custom_options | map<string, CUSTOMSYNTHESIZEOPTIONVALUE> | Дополнительный набор опций по настройке синтеза. В custom_options выносятся экспериментальные настройки, которые еще не прошли полную проверку. На текущий момент список дополнительных настроек пустой, так как этот функционал проходит тестирование. |

## CustomSynthesizeOptionValue

Индивидуальные настройки синтеза.

| Поле | Тип | Описание |
| --- | --- | --- |
| int32_value | int32 | Значение типа int32. |
| int64_value | int64 | Значение типа int64. |
| number_value | double | Значение типа double. |
| string_value | string | Значение типа string. |
| bool_value | bool | Значение типа bool. |

## PostprocessingMode

Постобработка аудио (удаление фоновых шумов, выравнивание громкости, эквализация и другие улучшения).

| Имя | Значение | Описание |
| --- | --- | --- |
| POST_PROCESSING_DISABLE | 0 | Постобработка выключена. |
| POST_PROCESSING_PHONE_CHANNEL | 1 | Рекомендуется использовать для телефонного канала. Допускается использовать только с SynthesizeSpeechRequest.sample_rate_hertz = 8000 Гц. Наилучший результат дает при использовании модели синтеза default в сочетании с SynthesizeOptions.model_sample_rate_hertz = 22050 Гц. **Важно!** - данные настройки не рекомендуется применять к модели синтеза high_quality. |
| POST_PROCESSING_PRETTIFY | 2 | Не рекомендуется использовать этот параметр. |

# SynthesizeSpeechRequest

Настройки синтеза для файлового метода синтеза речи.

| Поле | Тип | Значение | Описание |
| --- | --- | --- | --- |
| text | string | | Текст для синтеза без SSML разметки<br>(необходимо задавать только одно из полей – text или ssml).<br>Если в поле text отправить на озвучку текст с SSML-тегами,<br>Audiogram озвучит не только текст, но и теги. |
| ssml | string | | Текст для синтеза в формате SSML<br>(необходимо задавать только одно из полей – text или ssml).<br>Если в поле ssml отправить на озвучку текст без SSML-разметки,<br>вернется ошибка синтеза. |
| language_code | string | | Язык, который используется для синтеза. В настоящее время поддерживается только русский язык. |
| encoding | AUDIOENCODING | | Формат аудио данных (кодировка). |
| sample_rate_hertz | int32 | | Частота дискретизации синтеза (в герцах). |
| voice_name | string | | Имя голоса. Список доступных моделей:<br>женские голоса:<br>• borisova<br>• kishchik<br>мужские голоса:<br>• gandzhaev<br>• gavrilov |
| synthesize_options | SYNTHESIZEOPTIONS | | Опции синтеза аудио. |

# StreamingSynthesizeSpeechResponse

Результат работы потокового синтеза речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| audio | bytes | Байты аудиоданных без заголовка, закодированные, как указано в encoding и заданной в sample_rate_hertz частотой дискретизации.<br>Результат синтеза может прийти в нескольких ответах, по мере их синтезирования. |

## SynthesizeSpeechResponse

Результат работы файлового синтеза речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| audio | bytes | В данном поле передаётся целый синтезированный аудиофайл с заголовками в формате, заданном в encoding и заданной в sample_rate_hertz частотой дискретизации. |

## ModelsInfo

Доступные голоса для синтеза речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| models | repeated MODELINFO | Список доступных голосов для синтеза речи. |

## ModelInfo

Информация о модели синтеза речи.

| Поле | Тип | Описание |
| --- | --- | --- |
| name | string | Название модели синтеза речи. |
| sample_rate_hertz | uint32 | Частота дискретизации аудио данных в герцах. |
| language_code | string | Язык, которым озвучивается текст, отправленный на синтез. По умолчанию ru. |
| type | string | Тип модели. Возможные значения:<br>• **light** - это менее ресурсоемкая модель, рекомендуется для синтеза в файловом и потоковом режимах при промышленном использовании;<br>• **high_quality** - относится к новейшему поколению моделей, которые отличаются улучшенными характеристиками синтеза речи, но потребляет больше системных ресурсов. Рекомендуется к использованию при небольшой нагрузке.<br>В запросе на синтез укажите название модели, которую необходимо использовать. Если не передать название модели, то при синтезе в потоковом режиме по умолчанию будет использоваться **light**, а при синтезе в файловом режиме - **high_quality**. |

# Использование SSML-разметки

SSML (Speech Synthesis Markup Language) – это язык разметки с фиксированным набором тегов и атрибутов, основанный на XML (но без тега xml в начале) и применяемый для синтеза речи. Его можно использовать, чтобы настроить скорость и звучание голоса.

Настройка проводится с помощью SSML-тегов, которые нужно указать в тексте, отправляемом на синтез.

**Примечание:** текст без тегов необходимо писать в поле SynthesizeSpeechRequest.text

На данный момент в **Audiogram** поддерживаются следующие SSML-теги:

| Тег | Описание | Параметры тега |
| --- | --- | --- |
| speak | Обязательный тег для работы с SSML. В него должен быть обернут весь текст, отправляемый на синтез. | • speed (скорость) – положительное число, рекомендуемый интервал [0.1 – 2.0] (значения меньше единицы – медленнее, больше – быстрее)<br>Если параметр не указан, его значение по умолчанию = 1. |

| | | |
|---|---|---|
| | При помощи дополнительных параметров может управлять скоростью и высотой тона (питчем) всего текста.<br>Должен сопровождаться закрывающим тегом \ ... \. | • pitch (высота тона) – рекомендуемый интервал [-1 – 3], но допустимы и значения вне данного диапазона. К примеру, при значении 5 получится металлический голос.<br>Отрицательные значения – низкий тон, положительные – высокий.<br>Если параметр не указан, его значение по умолчанию = 0. |
| prosody | Управляет скоростью и высотой тона (питчем) произвольного количества предложений и слов. Должен сопровождаться закрывающим тегом (\ ... \). | • speed (скорость) – положительное число, рекомендуемый интервал [0.1 – 2.0] (значения меньше единицы – медленнее, больше – быстрее)<br>Если параметр не указан, его значение по умолчанию = 1.<br>• pitch (высота тона) – рекомендуемый интервал [-1 – 3], но допустимы и значения вне данного диапазона. К примеру, при значении 5 получится металлический голос.<br>Отрицательные значения – низкий тон, положительные – высокий.<br>Если параметр не указан, его значение по умолчанию = 0. |
| break | Добавляет паузу произвольной длины в секундах в любое место. Этот тег является самозакрывающимся (<break .../>). | • time (время продолжительности паузы в секундах) - данный параметр является опциональным. В случае его отсутствия длительность паузы по умолчанию - 1 секунда.<br>(единица измерения s (секунды) после количества секунд является обязательным параметром - \ |

Параметры тегов указываются внутри треугольных скобок в виде

<название_тега название_параметра1="величина_параметра1" название_параметра2="величина_параметра2" закрытие тега>

**Как поставить ударение в слове**

Некоторые слова могут читаться по-разному. Например, «жАркое» или «жаркОе». При помощи SSML-разметки можно указать где надо делать ударение. Для этого после ударной гласной необходимо вставить {'} :

- «жа{'}ркое», чтобы получилось «жАркое»; и
- «жарко{'}е», чтобы получилось «жаркОе».

**Дополнительно**

1. Если отправить какой-то текст без разметки на озвучку в поле text, а потом этот же текст обернуть только в тег и отправить на озвучку в поле ssml, то текст будет озвучен одинаково.

2. Если текст с SSML-тегами отправить на озвучку в поле text, то Audiogram озвучит не только текст, но и теги. Например, если отправить "Привет мир", то озвучка будет - "спик привет мир спик".

3. Знак ударения {'} синтез не считает ssml-тегом и он будет обработан как ударение и в text, и в ssml.

4. Синтез текста без SSML-разметки и синтез этого же текста с разметкой не будут различаться по нагрузке на систему.

# Примеры использования SSML-тегов

**Пример 1** (весь текст с SSML-разметкой, отправляемый на синтез, должен быть обернут в тег speak):

```
<speak>Глава 1.
Три девицы под окном
Пряли поздно вечерком.
«Кабы я была царица, —
Говорит одна девица, —
То на весь крещеный мир
Приготовила б я пир».</speak>
```

**Пример 2** (добавляем паузу после «Глава 1.»):

```
<speak>Глава 1. <break time="2s"/>
Три девицы под окном
Пряли поздно вечерком.
«Кабы я была царица, —
Говорит одна девица, —
То на весь крещеный мир
Приготовила б я пир».</speak>
```

**Пример 3** (поставим ударение, чтобы сервис правильно произнес «девИца», а не «дЕвица»):

```
<speak>Глава 1. <break time="2s"/>
Три девицы под окном
Пряли поздно вечерком.
«Кабы я была царица, —
Говорит одна деви{'}ца, —
То на весь крещеный мир
Приготовила б я пир».</speak>
```

**Пример 4** (укажем высоту и скорость произнесения всего текста):

```
<speak speed="0.8" pitch="-0.4">Глава 1. <break time="2s"/>
Три девицы под окном
Пряли поздно вечерком.
«Кабы я была царица, —
Говорит одна деви{'}ца, —
То на весь крещеный мир
Приготовила б я пир».</speak>
```

**Пример 5** (Изменим высоту и скорость произнесения прямой женской речи. Если выбрана мужская голосовая модель, это сделает звучание более аутентичным.):

```
<speak speed="0.8" pitch="-0.4">Глава 1. <break time="2s"/>
Три девицы под окном
Пряли поздно вечерком.
```

```
<prosody speed="1.1" pitch="0.7">«Кабы я была царица, —
Говорит одна деви{'}ца, —
То на весь крещеный мир
Приготовила б я пир».</prosody></speak>
```

# Метаданные gRPC-запросов

- Каждый запрос к Audiogram API должен содержать токен доступа. Передавайте токен следующим способом:

```
authorization: Bearer <access_token>
```

- В запросе можно передать уникальный идентификатор, который позволит детально проследить за историей выполнения запроса. Для этого используйте ключ trace-id:

```
external_trace_id: <id>
```

# Список ML-моделей и голосов в Audiogram

Модель для распознавания речи (**ASR**) или голос для синтеза речи (**TTS**) следует указывать, используя псевдоним (**alias**). Это позволяет обновлять модели без необходимости проводить повторную интеграцию клиентов.

**Внимание!** Если в запросе указана частота дискретизации (sample rate) отличная от значений, поддерживаемых моделью, то:

- в случае распознавания речи (**ASR**) произойдет перекодирование частоты дискретизации на значение, поддерживаемое моделью (16000 Гц).
- в случае синтеза речи (**TTS**) будет использована модель с ближайшей частотой дискретизации в большую сторону.

# ASR e2e

| **Alias – это имя модели, которое указывается в конфиге распознавания RecognitionConfig.model** | **Alias** | **Sample rate (Hz)** | **Описание** |
|---|---|---|---|
| e2e-v1 | 16000 | Модель конформер онлайн. | |
| e2e-v1 | 16000 | Модель конформер оффлайн. | |

# TTS

| **Alias – это имя голоса, которое указывается в конфиге распознавания SynthesizeSpeechRequest.voice_name** | **Alias** | **Sample rate (Hz)** | **Описание** |
|---|---|---|---|
| **женские голоса**: | | | |
| borisova | 8000 | Голос Борисовой 8000 Гц | |
| borisova | 22050 | Голос Борисовой 22050 Гц | |
| borisova | 44100 | Голос Борисовой 44100 Гц | |
| kishchik | 8000 | Голос Кищик 8000 Гц | |
| kishchik | 22050 | Голос Кищик 22050 Гц | |
| kishchik | 44100 | Голос Кищик 44100 Гц | |
| **мужские голоса**: | | | |
| gandzhaev | 8000 | Голос Ганджаева 8000 Гц | |
| gandzhaev | 22050 | Голос Ганджаева 22050 Гц | |
| gandzhaev | 44100 | Голос Ганджаева 44100 Гц | |

| | | |
|---|---|---|
| gavrilov | 8000 | Голос Гаврилова 8000 Гц |
| gavrilov | 22050 | Голос Гаврилова 22050 Гц |
| gavrilov | 44100 | Голос Гаврилова 44100 Гц |

# Сообщения об ошибках

| Код | Описание |
|---|---|
| PERMISSION_DENIED | Данная ошибка может вернуться в следующих ситуациях:<br>• срок действия токена доступа истек;<br>• токен недействителен;<br>• у клиента нет прав использовать Audiogram (например, клиент заблокирован через консоль администратора);<br>• ошибка при попытке авторизации (например, из-за сбоя внутренних сервисов). |
| INTERNAL | Не работают внутренние сервисы |
| UNKNOWN | Ошибки, которые пока не обрабатываются на стороне сервиса |

# Примеры кода клиентских приложений Audiogram

Для работы с Audiogram необходимо создать клиентское приложение. В этом разделе можно посмотреть примеры кода клиентских приложений, написанных на Python. Однако, вы можете использовать и другие языки программирования, для которых существуют библиотеки gRPC.

При написании клиентских приложений используйте proto-файлы.

# Распознавание аудио в файловом режиме

ПОСМОТРЕТЬ ПРИМЕР...

# Распознавание аудио в потоковом режиме

ПОСМОТРЕТЬ ПРИМЕР...

# Синтез аудио в файловом режиме

ПОСМОТРЕТЬ ПРИМЕР...

# Синтез аудио в потоковом режиме

ПОСМОТРЕТЬ ПРИМЕР...

# История релизов Audiogram

## Текущая версия - Audiogram 3.17.0

Обновление Audiogram 3.17.0 содержит следующие улучшения, обновления и исправления:

**Новый функционал**

- Добавлена возможность передавать новый ключ session-id в метаданных запросов на распознавание речи. По этому ключу можно агрегировать из аудиоархива все запросы, относящиеся к одному диалогу.

**Улучшения**

- Обновлена модель high_quality: произведено ускорение тритон-сервера, улучшен интонационный рисунок синтеза.

**Исправление ошибок**

- Исправлена ошибка, которая приводила к погрешностям при сборе статистики по синтезу и распознаванию речи.

- Исправлена ошибка, из-за которой количество каналов в многоканальных аудио не изменялось на 1 при значении параметра split_by_channels = False, что приводило к ошибкам и сбоям в работе транскодера.

# Предыдущие версии

## Audiogram 3.16.0

Audiogram 3.16.0 содержит следующие улучшения, обновления и исправления:

**Новый функционал**

- Добавлен новый сервис Антиспуфинг, позволяющий отличить реального человека от бота при входящем звонке.

**Улучшения**

- Произведен рефакторинг компонента asr-e2e-agent.

**Исправление ошибок**

- Исправлена ошибка, из-за которой для некоторых аудио, отправленных на распознавание в потоковом режиме с выключенным VAD, приходили пустые результаты распознавания.

## Audiogram 3.15.0

Audiogram 3.15.0 содержит следующие улучшения и исправления:

**Оптимизация**

- Ускорена ML-нормализация чисел более, чем в 2 раза, за счёт изменения алгоритмов работы.

**Исправление ошибок**

- Исправлена ошибка, из-за которой иногда происходила некорректная озвучка дат и сумм в некоторых падежах.

## Audiogram 3.14.0

Audiogram 3.14.0 содержит следующие улучшения:

**Новый функционал**

- Добавлена поддержка определения тональности аудио при распознавании речи. По тону высказывание может быть позитивным, нейтральным, грустным или сердитым.
  Определение тональности работает по сегментам VAD (Voice Activity Detection). Если VAD выключен, определяется тональность всего аудио. Если VAD включен, то определяется тональность каждого сегмента, отмеченного VAD.

**Оптимизация**

- Проведены работы по общей оптимизации Audiogram, направленные на повышение быстродействия сервиса и уменьшение потребления системных ресурсов:
  - Библиотека log-kit была дополнительно переработана и обновлена в следующих сервисах:
  - **api**
  - **asr-e2e-agent**
  - **tts-agent**
  - **vad-agent**
  - **genderage-agent**
  - **statistics-api**
  - **statistics-ingester**
  - **grpc-service-template**
  - Сервисы Audiogram, работающие по gRPC, переведены на формат пропагации трассировки ABNF.
  - Оптимизирована упаковка чанков аудио в InferInput.
  - Удалена метка **grpc-code** для метрики **grpc_requests_processing_time_seconds**.

  Результаты оптимизации по сравнению с предыдущей версией Audiogram:

  Распознавание речи (ASR):

  - 100 одновременных запросов в потоковом режиме - Latency p95 сократился на 8%; Latency mean сократился на 11%
  - 150 одновременных запросов в потоковом режиме - Latency p95 сократился на 17.4%; Latency mean сократился на 29.8%
  - 100 одновременных запросов в файловом режиме - RTFx увеличился на 7%

- 150 одновременных запросов в файловом режиме - RTFx увеличился на 6.1%

Синтез речи (TTS) (проверялся только потоковый режим; замеры по файловому режиму отдельно не проводились, так как он основан на потоковом):

- 100 одновременных запросов на синтез в потоковом режиме моделью light - Latency p95 сократился на 4.5%; Latency mean сократился на 3%; RTFx увеличился на 7.5%
- 100 одновременных запросов на синтез в потоковом режиме моделью high_quality - Latency p95 сократился на 4.2%; Latency mean сократился на 6%; RTFx увеличился на 5.8%

- Проведены доработки сервиса **transcriber**:
  - Triton Inference Server обновлен до версии 23.07.
  - С целью улучшения качества синтеза речи пополнены следующие словари:
  - эфикация
  - ёфикация
  - однозначные ударения
  - морфологические омонимы
  - контекстуальные омонимы
  - слова с дефисом
- Проведены доработки и интеграция обновленного сервиса **transcoder**:
  - Изменен способ задания формата кодеков. Теперь приходящий в api запрос имеет следующий маппинг по полю encoding сообщения AudioEncoding:

| AudioEncoding | Значение поля format | Значение поля codec |
|---|---|---|
| ENCODING_UNSPECIFIED | не поддерживается | не поддерживается |
| LINEAR_PCM | s16le | s16le |
| FLAC | не поддерживается | не поддерживается |
| MULAW | s16le | pcm_mulaw |

| | | |
|---|---|---|
| ALAW | s16le | pcm_alaw |

- Передача трассировочной информации теперь осуществляется через метаданные в формате ABNF.
- Добавлена поддержка версионирования api и обратная совместимость между версиями сервиса **transcoder**.
- Проведена прямая интеграция библиотек FFmpeg в **transcoder**, что позволило увеличить пропускную способность сервиса в 10-15 раз.

- Повышена точность работы сервиса **genderage** за счет изменения значения CHUNK_ALIGNMENT_SIZE по умолчанию с 25600 до 32000 секунд.

**Исправление ошибок**

- Исправлена ошибка, из-за которой распознавание в файловом режиме проходило успешно, но в логах сервиса **asr-e2e-agent** иногда записывалась ложная ошибка о незакрытой последовательности контекста.
- Исправлена ошибка, из-за которой при запросе доступных ML-моделей в список попадали более не поддерживаемые модели.
- Исправлена ошибка, из-за которой клиентское приложение зависало при отправке в сервис **genderage** нечетного количества байт.
- Исправлена ошибка, из-за которой в запросах от **api** к **transcoder** иногда выставлялось неверное значение аудиоканалов.

# Audiogram 3.10.0

Audiogram 3.10.0 содержит следующие улучшения:

**Новый функционал**

- Добавлена поддержка обработки сокращений и ML-нормализации чисел (с сохранением функционала нормализации чисел, работающей на правилах).
- Добавлена поддержка распознавания многоканального аудио (каждый канал распознаётся по отдельности).

**Оптимизация**

- Улучшена работа библиотеки log-kit, что позволит сервисам Audiogram тратить меньше времени и ресурсов на запись логов.

**Исправление ошибок**

- Исправлена ошибка в работе сервиса DEP, из-за которой в некоторых случаях, когда голос звучал в самом начале первого чанка, некорректно идентифицировалось начало речи.
- Исправлена ошибка в работе сервиса VAD, из-за которой иногда некорректно выставлялись метки начала и конца речи.

# Audiogram 3.9.0

В обновлении Audiogram 3.9.0 вдвое увеличена скорость синтеза речи моделью high_quality.

# Audiogram 3.8.0

Audiogram 3.8.0 содержит следующие улучшения:

**Оптимизация**

- Переработана и оптимизирована трассировка для компонентов **asr-e2e-agent** и **vad-agent**, что позволило сократить задержку (latency) p99.
- Переработано и оптимизировано логирование для компонента **asr-e2e-agent**, что позволило сократить задержку и нагрузку на систему.
- Проведены исследования и увеличена длительность аудиофрагмента, отправляемого на расшифровку, с 0.8 до 2 секунд. Это позволило снизить нагрузку на систему и количество ошибок распознавания речи (WER - Word Error Rate).
- Упрощено извлечение информации из логов за счет введения структурированного логирования для компонентов **auth** и **audio-archive-back**.

**Исправление ошибок**

- Исправлена ошибка, из-за которой в логах сборки обфусцированных образов **asr-e2e-agent**, **vad-agent** и **audio-archive-back** появлялась строка о проблеме, но при этом сами образы были рабочими.
- Исправлена ошибка, из-за которой метрики производительности компонента **vad-agent** не отображались на обзорной панели.
- Исправлена ошибка, из-за которой в ключе tts-cache не учитывалось применение постобработки, вследствие чего один и тот же текст нельзя было озвучить с другим видом постобработки.
- Исправлена ошибка, из-за которой наблюдалось снижение производительности синтеза речи при включенной авторизации.
- Исправлена ошибка, из-за которой в ответе на запрос распознанных фраз из архива **audio-archive-back** возвращал только первую фразу.

**Дополнительно**

- Версия Triton в VAD обновлена до 23.07 для более стабильной работы сервиса.
- Для избежания проблем с уязвимостью обновлена версия утилиты grpc_health_probe до 0.7.0.
- Обновлён образ **audio-archive-back** (Python > v3.11).

# Audiogram 3.7.0

Audiogram 3.7.0 содержит следующие улучшения:

**Новый функционал:**

- Добавлены gRPC-метрики, позволяющие отслеживать взаимодействие между компонентами **asr-e2e-agent** и **genderage_agent** (количество запросов, количество запросов в прогрессе, время выполнения запросов, количество ошибок и т.д.) с помощью Prometheus.
- Реализована возможность работы сервиса **genderage** отдельно от встроенного **VAD** (Voice Activity Detection), что позволяет заказчикам при необходимости подключать свой собственный VAD (компонент, отвечающий за определение голосовой активности в аудио).

**Оптимизация**

- Фразы, полученные при распознавании в файловом режиме, больше не объединяются в один элемент, а перечисляются по отдельности. Это позволяет впоследствии корректно выполнить определение пола и возраста участников разговора, если в аудио несколько спикеров.
- Повышена скорость взаимодействия **asr-e2e-agent** и Triton-сервера благодаря оптимизации подготовки инпутов (объектов для входа).

**Исправление ошибок**

- Исправлена ошибка с отправкой некорректной метки, из-за которой в Prometheus не отображались алерты (alerts) для компонента **asr-e2e-agent**.
- Исправлена ошибка с отправкой некорректной метки, из-за которой в Prometheus не отображались алерты (alerts) для компонента **vad-agent**.
- Исправлена ошибка, из-за которой возникали разные результаты определения пола при отправке запроса к **genderage** через Audiogram API и напрямую.
- Исправлена ошибка, из-за которой распознавание речи не работало при выключенном **VAD**.

# Audiogram 3.6.0

Audiogram 3.6.0 содержит следующие улучшения:

1. В ходе тестирования были выявлены и исправлены несколько ошибок, что повысило общую стабильность работы сервиса.
2. Оптимизированы развертывание, работа и названия моделей для синтеза речи. Теперь обе модели доступны одновременно. Чтобы выбрать какой моделью необходимо произвести озвучку, передайте её название в запросе:

- **light** - эта модель является менее ресурсоемкой и рекомендуется к использованию в производственных целях.
- **high_quality** - данная модель потребляет больше системных ресурсов, так как относится к новому поколению моделей, отличающихся более хорошими показателями качества синтеза, На данный момент ее рекомендуется использовать при невысокой нагрузке.

Вы можете выбрать любую модель, но если в запросе не указать конкретное название, то для синтеза в потоковом режиме по умолчанию будет использоваться **light**, а для синтеза в файловом режиме - **high_quality**.

# Audiogram 3.5.1

Audiogram 3.5.1 содержит следующее улучшение:

1. Исправлена ошибка трассировки, из-за которой в один трейс попадали спаны нескольких запросов.

# Audiogram 3.5.0

Audiogram 3.5.0 содержит следующие улучшения:

**Для сервиса распознавания речи (ASR)**

1. Проведены работы по снижению задержки (latency). Например: убрано шифрование модуля sentence-piece (данное изменение не влияет на безопасность); внедрен новый предиктор ML-модели; и др.

**Для сервиса синтеза речи**

1. Внедрена новая модель, что привело к улучшению интонаций и уменьшению количества потребляемых ресурсов видеокарты.
2. Улучшено качество синтеза голоса Ганджаева благодаря дополнительному обучению моделей.