Яндекс Облако

Распознавание и синтез речи как сервис - введение для разработчиков

Вадим Челышков
Архитектор облачных решений - ML

План выступления

01 | Yandex SpeechKit

02 | Синтез речи

03 | Распознавание речи

04 | Встраивание SpeechKit в телефонию

05 | Полезные материалы

01

Yandex SpeechKit

Yandex SpeechKit. Главные особенности.

- Распознавание и синтез речи в режиме реального времени, транскрибация длинных записей
- Ключевой компонент голосового помощника Алиса (8M ежедневно) средняя загрузка 800 RPS
- Доступен через GRPC/REST API
- Не храним данные, соответствуем 152-ФЗ

Сценарии

- Виртуальный оператор колл-центра, обзвон
- Аналитика звонков и расшифровка аудиозаписей
- Озвучивание роликов, новостей, текстов
- Голосовое управление в приложениях

02

Синтез речи

Познакомьтесь с Алёной и Филиппом

Поддержка SSML-разметки

<speak> Вот несколько примеров использования SSML.

Вы можете добавить в текст паузу любой длины:
<break time="2s"/> та-дааам!

Или разметить текст на параграфы и предложения.
Паузы между параграфами длиннее.
<s>Первое предложение</s><s>Второе предложение</s> .

А ещё вы можете подменять фразы. Например, чтобы произносить аббревиатуры и _{т.п.}</speak>

Методы авторизации

Токены .

В Яндекс.Облаке *OAuth-токен* используется в процедуре аутентификации для получения IAM-токена.
Срок жизни OAuth-токена 1 год.

IAM-токен — уникальная последовательность символов, которая выдается пользователю после прохождения аутентификации. С помощью этого токена пользователь авторизуется в API Яндекс.Облака и выполняет операции с ресурсами.
IAM-токен действует не больше 12 часов

Api Key

API-ключ — секретный ключ, используемый для упрощенной авторизации в API Яндекс.Облака. API-ключи используются только для сервисного аккаунта.

API-ключи не имеют срока действия, поэтому этот способ аутентификации проще, но менее безопасный.

Синтез речи — пример запроса

```
In [10]:  %%bash
          export SSML_REQUEST='''
          <speak>
            <p>Вот несколько примеров использования <phoneme alphabet="ipa" ph="esasemel  ʲ">SSML</phoneme>.</p>
            <p>Вы можете добавить в текст паузу любой длины: <break time="2s"/> та-даааам! </p>
            <p>Или разметить текст на параграфы и предложения. Паузы между параграфами длиннее. </p>
            <p><s>Первое предложение</s><s>Второе предложение</s></p>
            <p>А еще вы можете подменять фразы. Например, чтобы произносить аббревиатуры и <sub alias="тому
          подобное">т.п.</sub>
          </speak>

          '''
          export APIKEY=AQVNwGu1aM8v0VfLzMJhVwi2EpNqJYn15vTdpinB
          curl -X POST \
              -s \
              -H "Authorization: Api-Key ${APIKEY}" \
              --data-urlencode "ssml=$SSML_REQUEST" \
               -o speech-test.wav \
               -d "lang=ru-RU&voice=alena&format=lpcm&sampleRateHertz=8000"
          "https://tts.api.cloud.yandex.net/speech/v1/tts:synthesize"
```

cloud.yandex.ru/docs/speechkit/tts/request

03

Распознавание речи

Распознавание речи — 3 варианта

Потоковый режим

- позволяет одновременно отправлять аудио на распознавание и получать результаты распознавания в рамках одного соединения (сессии) в реальном времени
- возможность получать промежуточные результаты распознавания, пока фраза еще незакончена
- после паузы сервис вернет финальные результаты и начнет распознавание новой фразы

Распознавание коротких аудио

- отличается быстрой скоростью ответа и подходит для одноканального аудио небольшого размера
- ответ на запрос возвращается синхронно

Распознавание длинных аудио

- подходит для многоканальных аудиофайлов до 1 ГБ
- дешевле других способов распознавания
- не подходит в сценариях распознавания речи онлайн — время ответа больше

04

Встраивание SpeechKit в телефонию

HTTPv2 + GRPC

Бинарный вместо текстового

Используются frames и streams. Protobuf в качестве инструмента описания типов данных и сериализации позволяет повысить производительность.

Мультиплексирование

HTTP/2 в качестве транспорта позволяет переиспользовать один сокет для нескольких параллельных запросов.

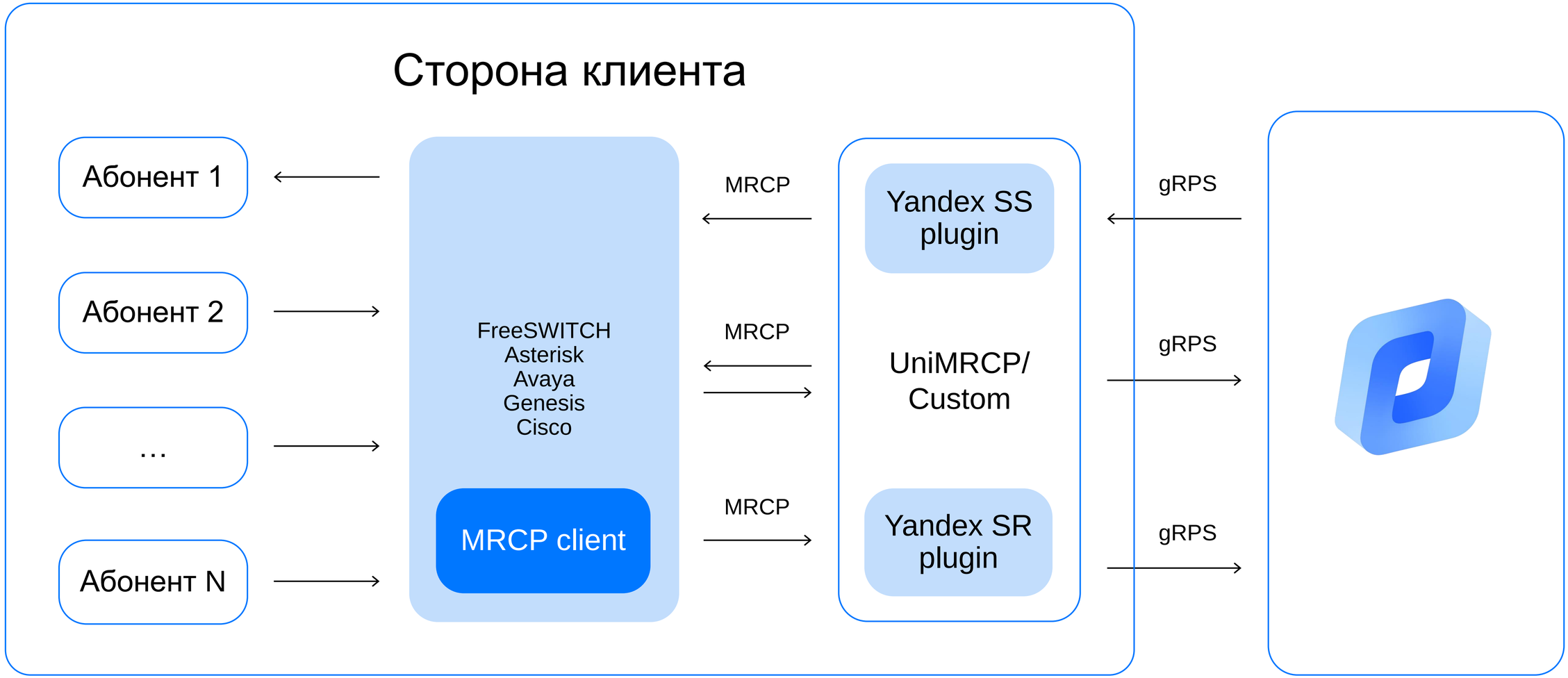
Сжатие заголовков

Для сжатия литеральные значения сжимаются по алгоритму Хаффмана, а клиент и сервер поддерживают единую таблицу заголовков

Приоритизация запросов

Можно назначить приоритет потоку, добавив информацию о приоритете во фрейм HEADERS, которым открывается поток или отправить фрейм PRIORITY, меняющий приоритет потока.

Типичная роль SpeechKit в телефонии



Tips & Tricks

- ✓ Отправляйте аудио максимально приближенное к реальному времени, например лучше каждые 100мс отправлять 100мс готового аудио. Время отправки между двумя чанками не должно превышать 5 секунд, иначе сессия будет разорвана.
- ✓ Оптимальный размер одного чанка – не более 400мс, но не менее 100мс.
- ✓ Проще использовать для авторизации API Key, т.к. время жизни OAUTH-токена ограничено.
- ✓ Заведите контрольный датасет для оценки WER и других метрик.
- ✓ Для Go, C#, C++ и Java производительность будет выше за счёт более удачных реализаций протокола GRPC.

Опрос по поводу SDK

05

Полезные материалы

Полезные материалы

- Документация
  cloud.yandex.ru/docs/speechkit/
- Посмотреть демонстрацию сервиса
  cloud.yandex.ru/services/speechkit
- Грант на знакомство с платформой
  cloud.yandex.ru/prices

Запросите расширенный грант
на тестирование SpeechKit

- Видеоматериалы
  - youtu.be/Cn3hbsrRRSw
  - youtube.com/playlist?list=PL1x4ET76A10YtMqNF2gFqOUOFWV0YKzD4

Яндекс Облако

Спасибо!

Вадим Челышков
Архитектор облачных решений - ML

vachel@yandex-team.ru