



AC
'22

Архитектура сервиса синтеза речи: gRPC, пайплайны, нейронные сети



Немного обо мне



- Больше 3.5 лет в синтезе речи
- Занимался NLP, чем-то типа MLOps и немного ресерчем
- Сейчас тимлид в команде синтеза Tinkoff VoiceKit
- тг: @sOMEMETH



О чем доклад



- История развития голосовых технологий в Tinkoff
- Какие элементы есть в пайплайне синтеза речи
- Как это все работает в продакшене
- Что еще мы делаем в Tinkoff VoiceKit



Как развивался Tinkoff VoiceKit

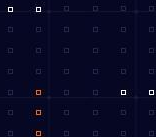


Аналитика
работы колл
центра

2016 год



АС
'22



Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год



AC
'22

Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год

Разработка
внутреннего TTS

2018 год



AC
'22

Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год

Разработка
внутреннего TTS

2018 год

Релиз Олега

2019 год



AC
'22

Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год

Разработка
внутреннего TTS

2018 год

Релиз Олега

2019 год

Выступление
Гриши О. на
Asterconf

2019 год



АС
'22

Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год

Разработка
внутреннего TTS

2018 год

Релиз Олега

2019 год

Выступление
Гриши О. на
Asterconf

2019 год

voice conversion

2021 год



АС
'22

Как развивался Tinkoff VoiceKit

Аналитика
работы колл
центра

2016 год

Разработка
внутреннего STT

2017 год

Разработка
внутреннего TTS

2018 год

Релиз Олега

2019 год

Выступление
Гриши О. на
Asterconf

2019 год

voice conversion

2021 год

Нейросетевая
биометрия

2021 год



AC
'22

Синтез, зачем?



- Роботы для обзвонів
- Роботы для входящих
- Аудиокниги
- Чтение статей



Верхнеуровневая схема синтеза

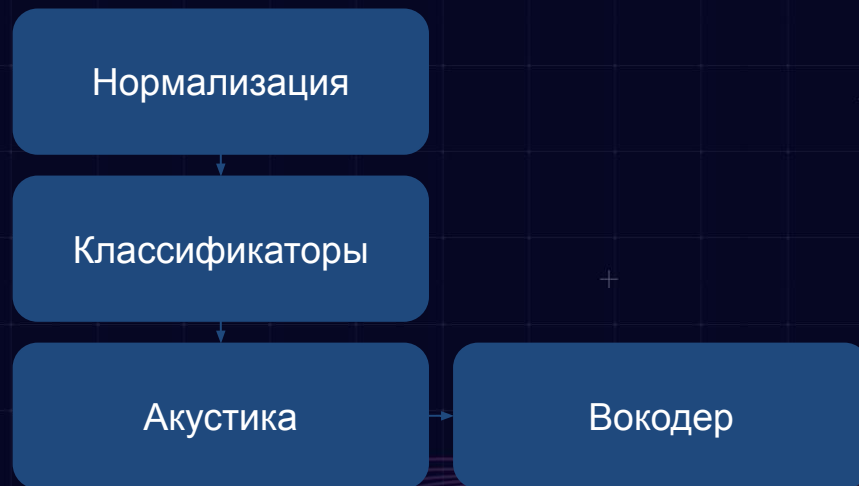


Нормализация

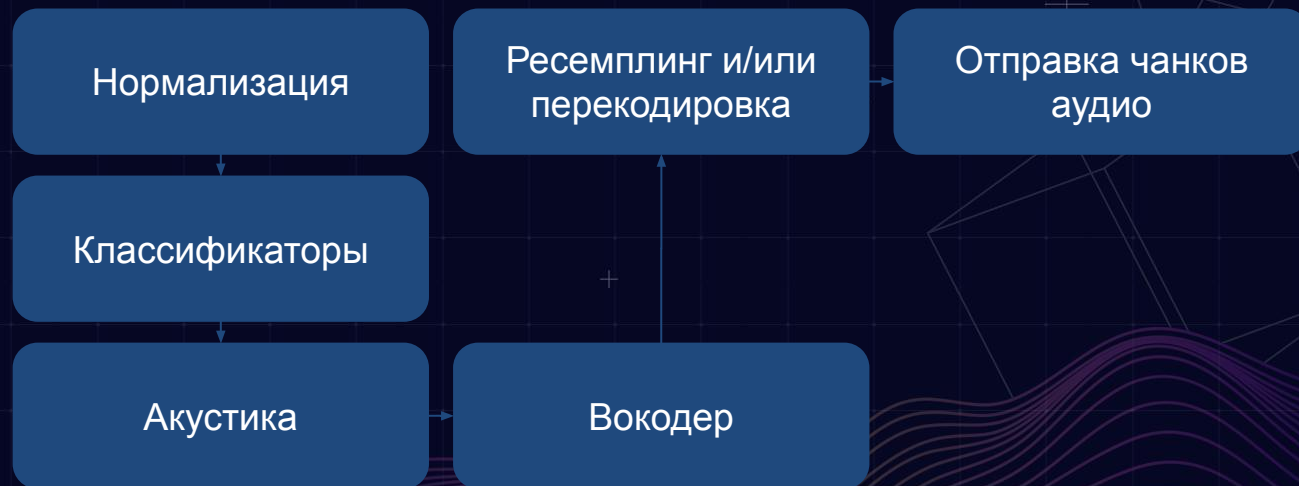


Классификаторы

Верхнеуровневая схема синтеза



Верхнеуровневая схема синтеза





Верхнеуровневая схема синтеза

Вызывается отдельно для каждого предложения



Разбиение текста на предложения

- Уменьшает latency
- Позволяет синтезировать большие тексты
- Позволяет брать из кэша отдельные предложения
- Работает на правил

И.С. Петров не
пришел т.к. он
заболел. Зайдите к
нам 05.06.07

И.С. Петров не
пришел т.к. он
заболел.

Зайдите к нам
05.06.07



AC
'22

Кэш



- Существенно ускоряет синтез
- Работает на уровне отдельных предложений
- Хранение по SSML-ключам
- Нужно обновлять при каждом релизе
- Нуждается в бэкапах



AC
'22



Нормализация

- Трансформирует текст из письменной записи в разговорную
- Делает это с помощью правил и нейронной сети
- Одна из наиболее критических в плане качества компонент синтеза и одна из самых сложных

До 05.07
переведите 20 руб.

До пятого июля
переведите
двадцать рублей.



AC
'22

Нормализация

- Трансформирует текст из письменной записи в разговорную
- Делает это с помощью правил и нейронной сети
- Одна из наиболее критических в плане качества компонент синтеза и одна из самых сложных

До 05.07
переведите 20 руб.

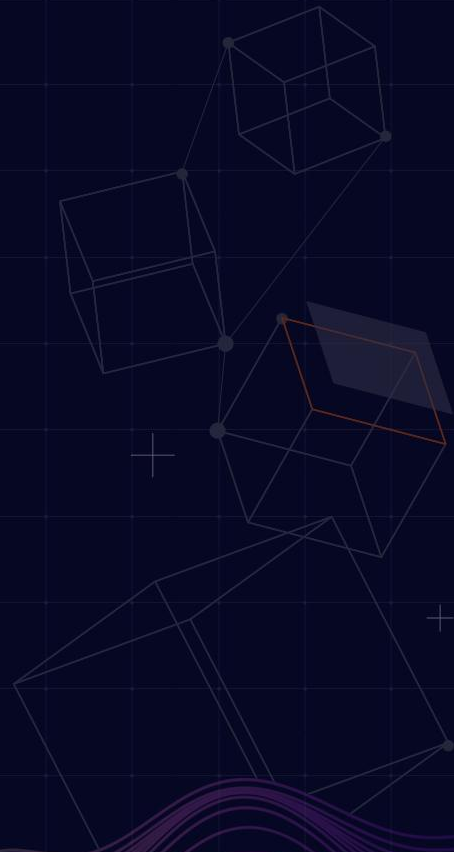
До пятого июля
переведите
двадцать рублей.



Модели-классификаторы



Большая предобученная на
огромном корпусе текстов модель,
которая используется для
векторизации текста



Модели-классификаторы



Паузы

Делаем ли мы паузы между словами и если да, то какой длины?



AC
'22



Модели-классификаторы



Паузы

Акцентные слова

Какие слова мы выделяем
интонацией?



AC
'22



Модели-классификаторы



Паузы

Акцентные слова

Омографы

Адреса или адресА, всЕ или всЁ



AC
'22



Модели-классификаторы



Паузы

Акцентные слова

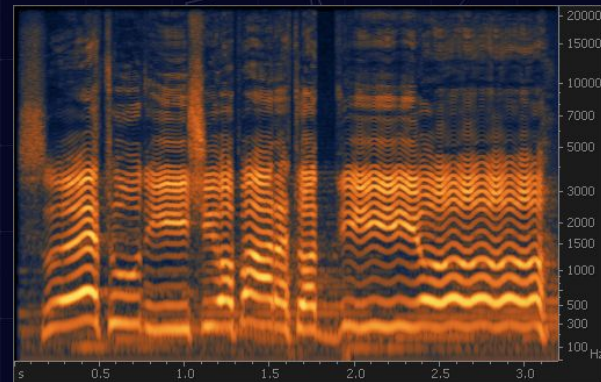
Омографы

Ударения

Куда ставим ударение в
незнакомых словах?

Акустическая модель

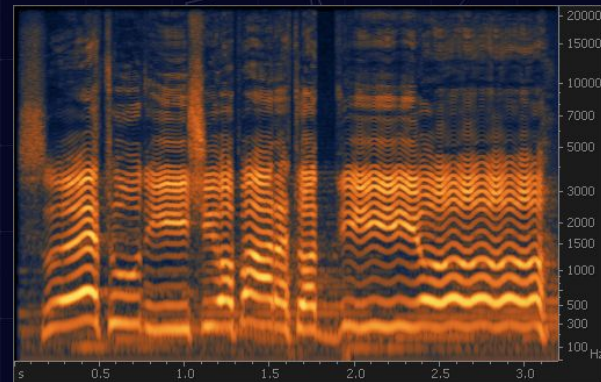
ПРИВЕОТ <p1> ка0к
дела0?



- Отвечает за моделирование глобальных зависимостей в речи, интонацию
- Самая тяжеловесная модель синтеза
- Позволяет управлять скоростью, громкостью и высотой речи

Акустическая модель

ПРИВЕОТ <p1> ка0к
дела0?



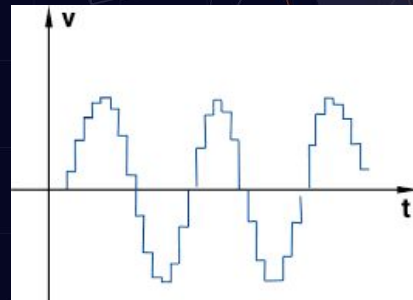
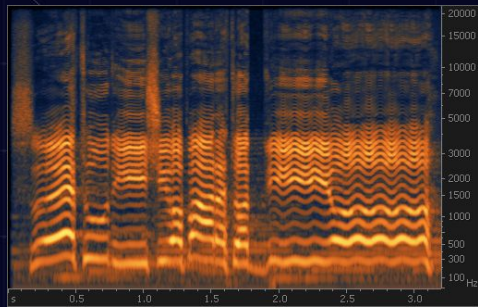
- Отвечает за моделирование глобальных зависимостей в речи, интонацию
- Самая тяжеловесная модель синтеза
- Позволяет управлять скоростью, громкостью и высотой речи



AC
'22



Вокодер



SSML



- Язык разметки для управления синтезом речи
- Голос
- Сэмпл рейт
- Скорость речи
- Паузы
- Высота речи

Сервинги

- Превращают обученные модели в полноценные grpc сервисы
- Позволяют batch-обработку запросов
- Очереди запросов
- Отвечают за менеджмент ресурсов, благодаря ним можно класть несколько моделей на одну карт
- Метрики и работа с3 из коробк
- Triton serving/tfserving

Обучаем модель
на python

Сохраняем ее на
s3 в .proto
формате

Поднимаем
контейнер

Модель доступна
как сервис



AC
'22



- Синтез почти всегда стриминговая технология: мы получаем кусочки итогового результата постепенно
- gRPC поддерживает стриминг
- Он очень быстрый

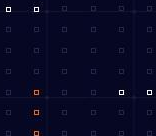
Golang



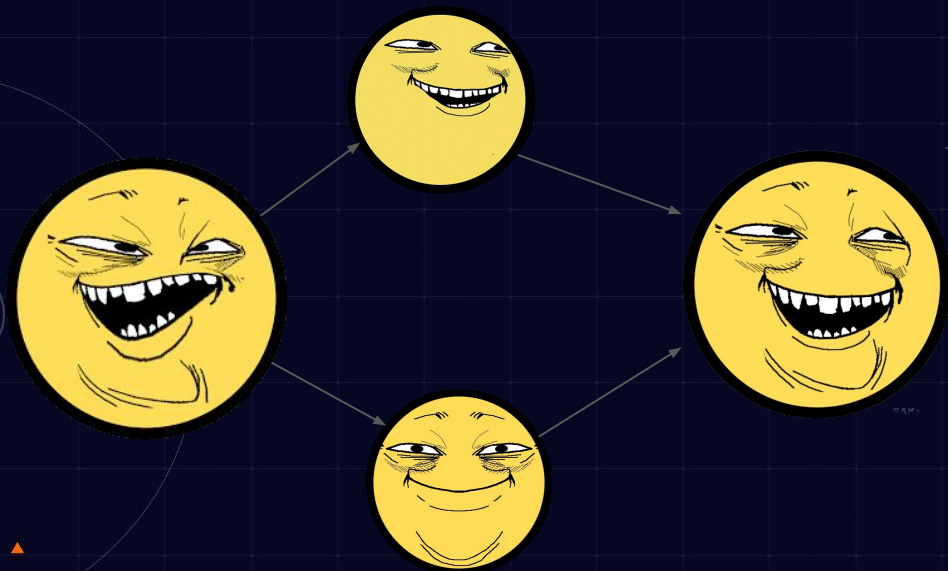
- У нас много логики, требующей параллельного исполнения
- Большое количество обработки текста
- Golang прекрасно сочетается с gRPC
- Go крайне простой в изучении язык
- Поэтому в проде у нас Go



AC
'22



Пайплайны



- Собирается на каждый запрос
- Каждая нода обрабатывает одновременно только одну фразу
- В DAG легко делать параллельную обработку запросов
- <https://go.dev/blog/pip>



AC
'22



Что еще есть в VoiceKit

- Одно из лучших, если не лучшее STT для русского и английского языков.
- Голосовая биометрия
- Активно развиваем on-prem решения
- На сайте можно нас протестировать



Voice conversion



Target
Speaker



Source
audio



VC



Result



AC
'22



Материалы



- [Наша статья про voice conversion](#)
- [Сайт VoiceKit](#)
- [Мое выступление на датафесте о процессах разработки](#)
- [Про пайплайны](#)
- [Triton serving](#)

Спасибо за внимание!

У кого-нибудь есть вопросы?

s.shishkin@tinkoff.ru

tg: @somemeth

<https://voicekit.tinkoff.ru/>

Пожалуйста, дайте оценку
моему докладу прямо сейчас

