AC'25
ASTERCONF

У нас есть речевая аналитика дома: как обогнать коммерческие API и не разориться

Дмитрий Шатнёв

ML Engineer

Инвентос

Веду Speech AI направление, координирую небольшую команду инженеров.

Проблематика доклада:

- Облачные API решают задачу диаризации «из коробки», но когда речь заходит о специфичных сценариях, требованиях к данным и стоимости — компании упираются в ограничения.
  - Где находится баланс между удобством готовых сервисов и возможностями собственной разработки?
  - Как развитие экспертизы внутри компании позволяет адаптировать систему под данные и инфраструктуру?
  - Что даёт такой подход бизнесу?

Конкретный кейс: call-центр

- Диалог с конфликтным клиентом → нецензурная лексика
- Система ошибочно приписала реплику оператору
- Ошибка пошла с уровня диаризации, дальше автоматически исправить не получилось
- Потенциальное последствие: увольнение оператора
- Собственный пайплайн позволил быстро найти и устранить проблему

Архитектура решения

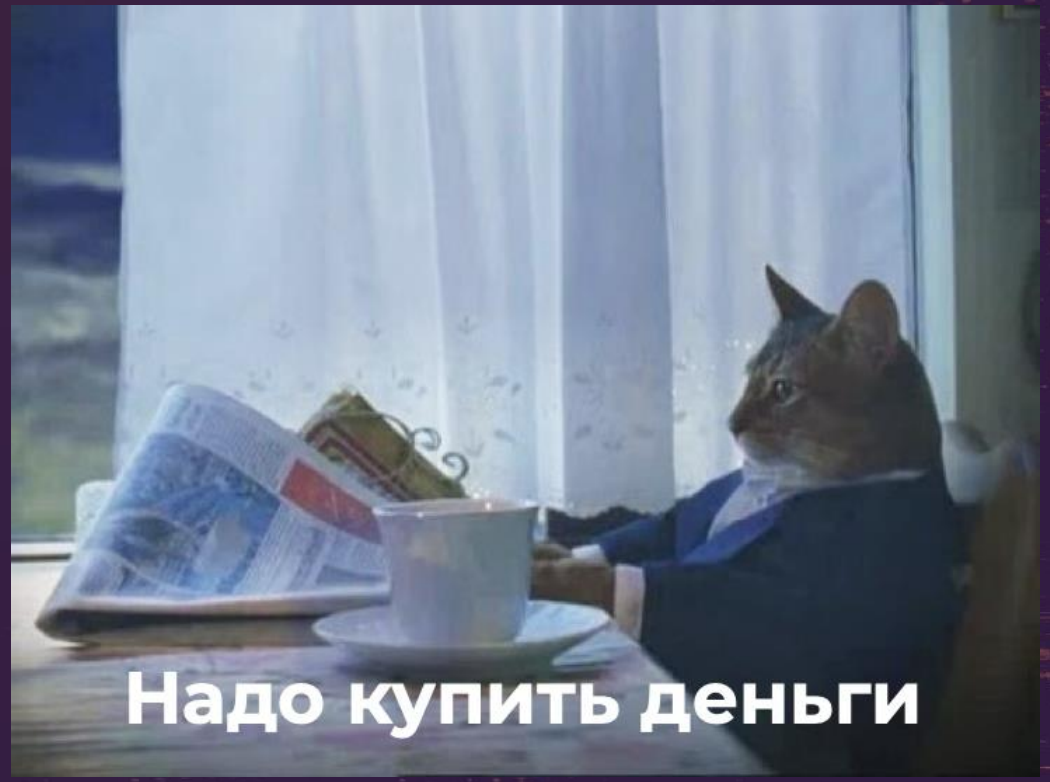
Live Demo

Практические результаты

Метрика	Наш сервис	Облачные API
Качество (DER)	VoxConverse: **11.3%**	Google ~16%, AWS ~17.4%
	OpenSTT (RU): **13.2%**	Яндекс ~18.7%, Сбер ~19.8%
Производительность	GPU: **0.025 RTF** → ~40× быстрее реального времени CPU: **0.34 RTF** → быстрее реального времени	—
Устойчивость	При SNR=5 дБ рост DER всего **+0.8%**	—

Что это дает бизнесу?

- **Кастомизация** — адаптация под конкретный сценарий (контакт-центры, медицина, банки)
- **Контроль данных** — записи не покидают инфраструктуру, соответствие ФЗ-152
- **Новые возможности** — аналитика разговоров: кто говорил, сколько, перебивания, эмоции
- **Экономика** — меньше регулярных затрат, инвестиции в собственную команду



Итоги

- **Собрали модульный пайплайн:** VAD → Эмбеддинги → Кластеризация → Постобработка
- **Достигли DER ~11–14%** (лучше API на 5–8%)
- **Скорость: 0.025 RTF на GPU** (в 40× быстрее реального времени)
- **Сервис полностью работает внутри компании, без утечки данных**
- **Главный вывод:** при достаточной экспертности open source решение становится конкурентным вариантом

AC'25 ASTERCONF

Спасибо за внимание!

Буду рад ответить на все ваши вопросы
сейчас или свяжитесь со мной в будущем:

INVENTOS

Дмитрий Шатнёв

dmitry.shatnev@mail.ru

+7 (961) 624-03-17

tg://dmitrii_shatnev