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:

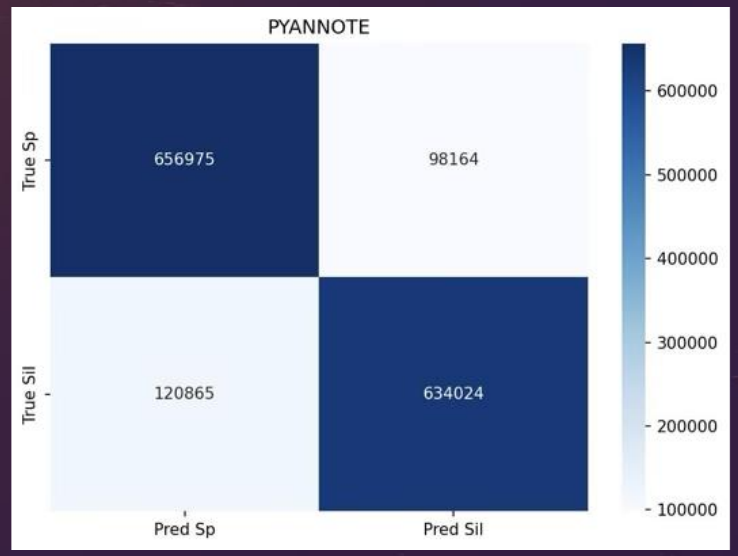
- Облачные API решают задачу диаризации «из коробки», но когда речь заходит о специфичных сценариях, требованиях к данным и стоимости — компании упираются в ограничения.
  - Где находится баланс между удобством готовых сервисов и возможностями собственной разработки?
  - Как развитие экспертизы внутри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

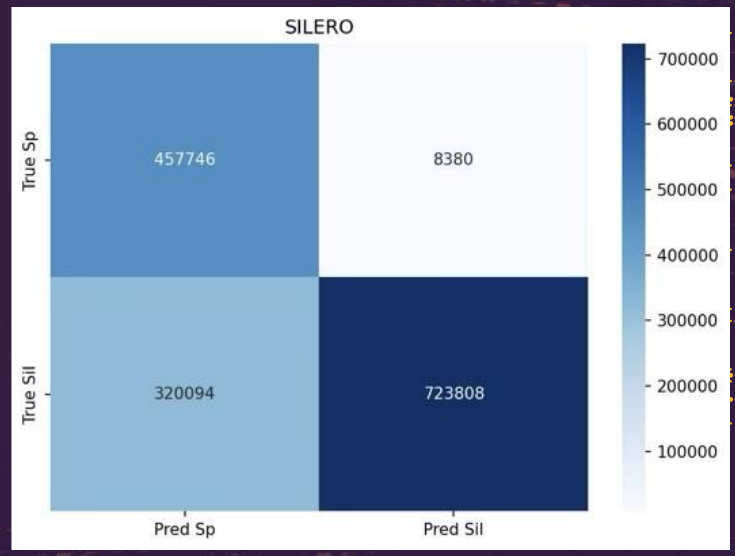
Конкретный кейс: call-центр

- 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

Как мы выбирали VAD?



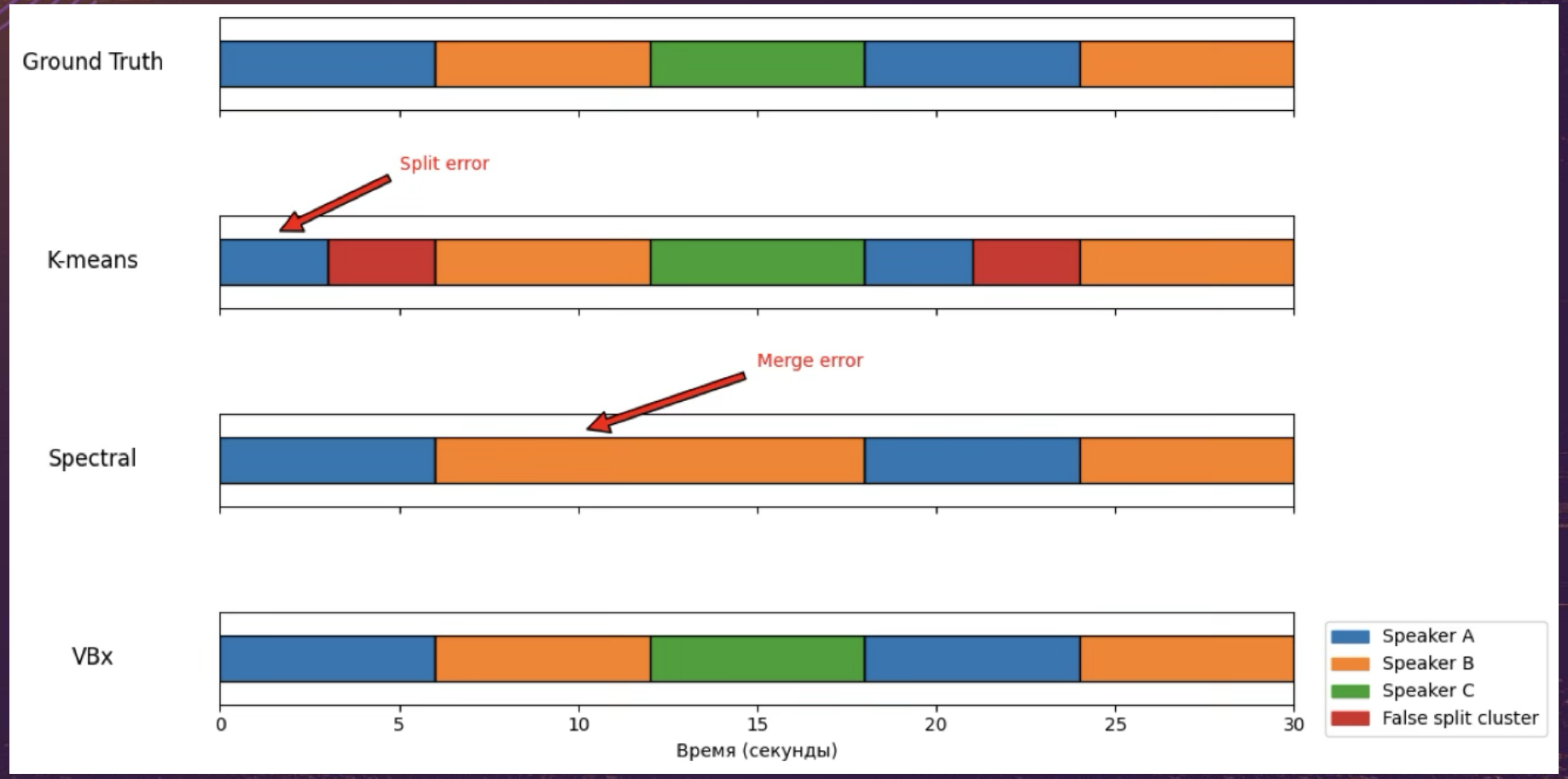


Для оценки использовали размеченные людьми субтитры

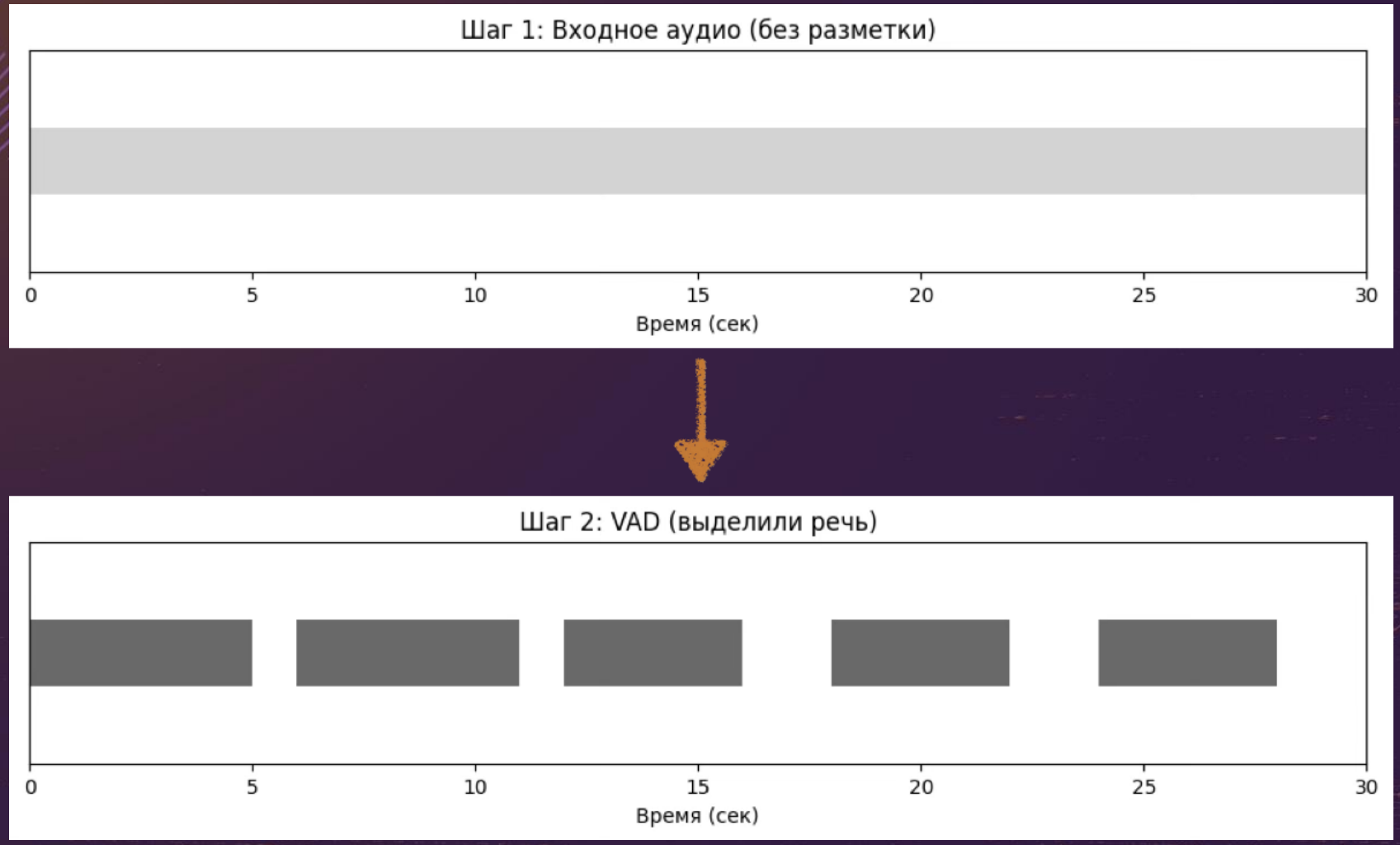
Эмбеддинги: какой моделью извлекать «отпечаток голоса»?

Модель	Плюсы	Минусы	DER (%)	RTF
ECAPA-TDNN	Очень быстрая	Падает качество на коротких фразах	~14.5	0.05
WavLM Large	Лучшая робастность к шумам	Очень тяжёлая, медленная	~13.1	0.285
SimAM-ResNet100	Баланс качество/скорость, работает на русском	Крупнее ECAPA	12.6	0.017

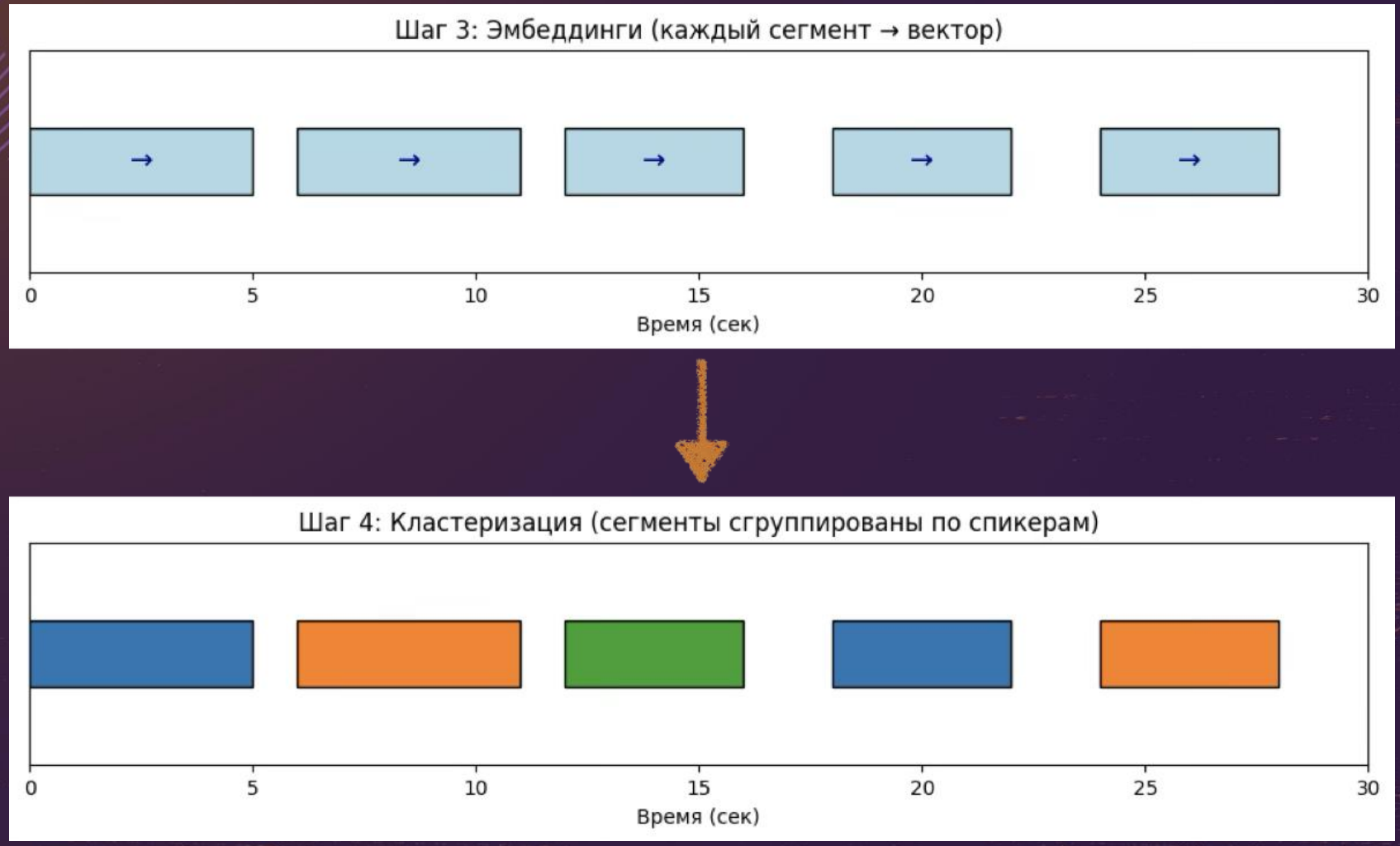
Кластеризация: как объединять сегменты одного говорящего?



Пример работы



Пример работы



Пример работы

[Speaker A | 0–5s]: Добрый день, коллеги.

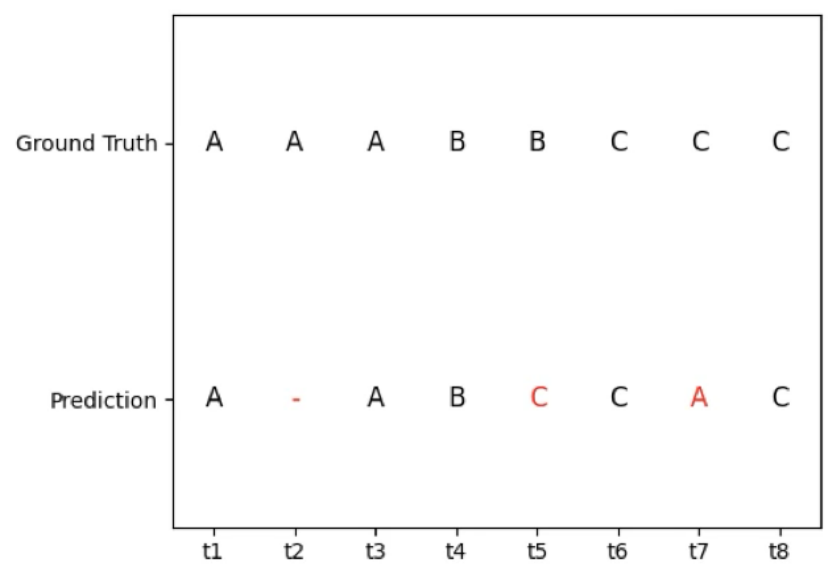
[Speaker B | 6–11s]: Да, здравствуйте.

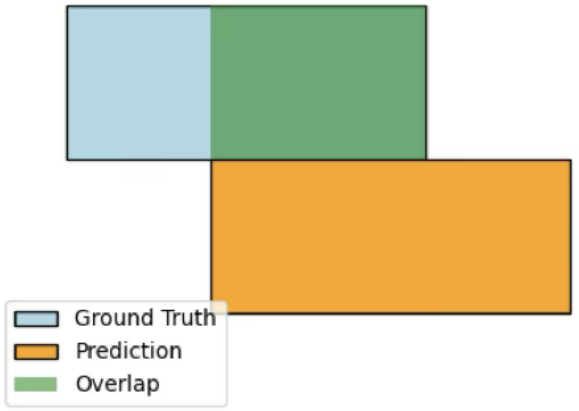
[Speaker C | 12–16s]: Всем привет.

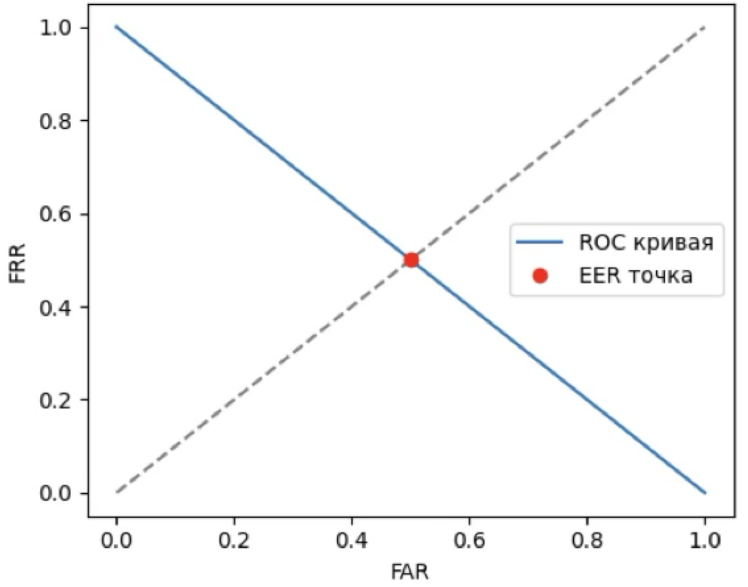
[Speaker A | 18–22s]: У меня есть вопрос.

[Speaker B | 24–28s]: Давайте обсудим.

Как оцениваем качество?







DER = (Ложные срабатывания + Пропуски + Ошибки спикера) ÷ Общее время речи

JER = 1 - (Пересечение ÷ Объединение)

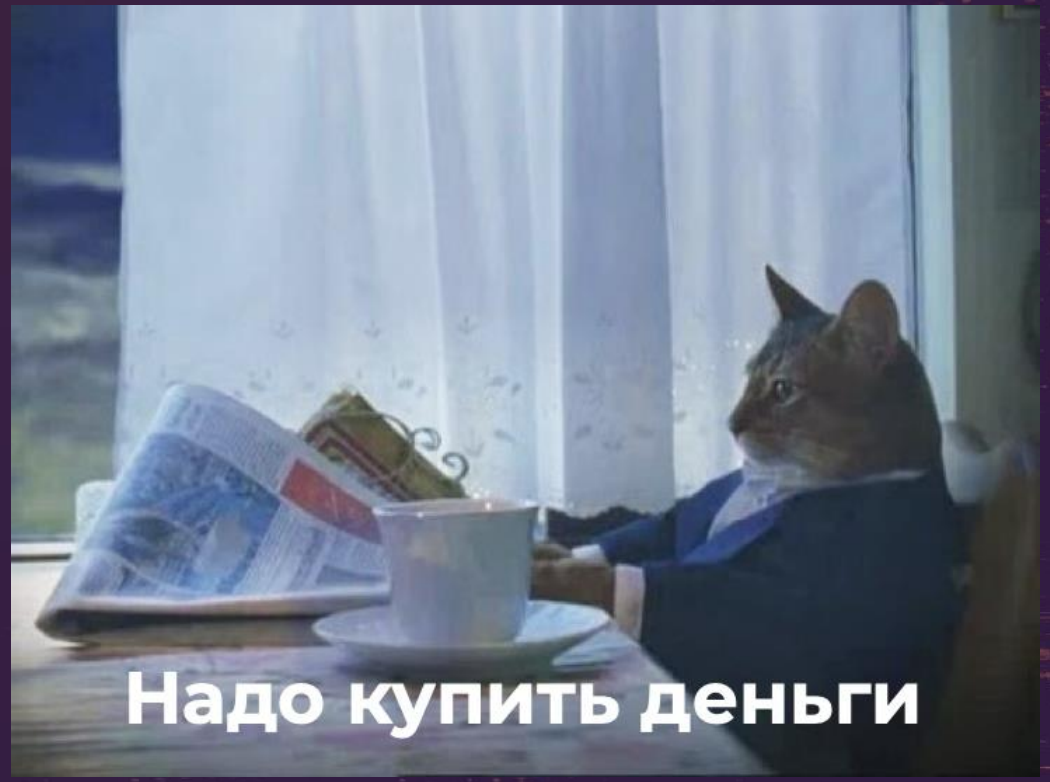
EER (точка, где процент ложных принятий = процент ложных отказов)

Практические результаты

Метрика	Наш сервис	Облачные API
Качество (DER)	VoxConverse: **12.8%**	Google ~16%, AWS ~17.4%
	OpenSTT (RU): **14.6%**	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 **+1.4%**	—

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 ~12–14%** (лучше API на 2–5%)
- **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