Открытые речевые библиотеки - решения, тесты и рекомендации

Воск

- Работа с речью — распознавание, синтез
- Asterisk, Freeswitch, Unimrcp, Jitsi
- Потоковое распознавание
- 25+ языков: китайский, арабский, испанский

Современные речевые технологии

- Многозадачность
- Интегрированная или многокомпонентная структура
- Разнородные модули / компоненты

Пример

- Распознавание
- Идентификация голоса
- Отслеживание состояния диалога
- Определение качества звука (NISQA-s)
- Семантический анализ с помощью LLM

Разнородные модули / компоненты

- Открытые библиотеки
- Коммерческие API
- БЯМОДЫ

Тестирование речевых компонент

- Для каждой задачи существует много решений
- Необходимо проводить сложные тесты
- Тесты должны учитывать внутренности модели (белый ящик)
- Много дополнительных условий

Распознавание речи. WER

Мама МЫЛА ** раму у мамыры
Мама МИЛА У раму *** мамыры

WER (WORD ERROR RATE) =

(число подстановок + число удалений + число вставок) / общее число слов

Распознавание речи. Ещё WER

- WER в различных условиях (по различным датасетам)
- WER по сложным словам
- Семантический WER
- Оценка точности распознавания через LLM (LLM WER)
- WER по смеси языков

Распознавание речи

Чем точнее распознавание, тем больше нужно тестов:

WER 20%	1-2 часа
WER 2%	20 часов
WER по сложным словам	100 часов

Распознавание речи. Что кроме WER

- Задержка первого ответа
- Задержка финального ответа
- Скорость обработки
- Баланс между скоростью и пропускной способностью

Распознавание речи

04/2024

Sberdevices GigaAM RNNT на звонках — 15% WER

Синтез речи

- Чем выше качество речи, тем больше нужно данных для тестов
- Чем больше голосов поддерживается, тем тяжелее тестировать
- Современные тесты — автоматические, около 2000 примеров

Синтез речи. Метрики

- WER — чёткость речи
- UTMOS — качество звучания
- Похожесть голоса
- FAD — похожесть интонаций
- Скорость синтеза

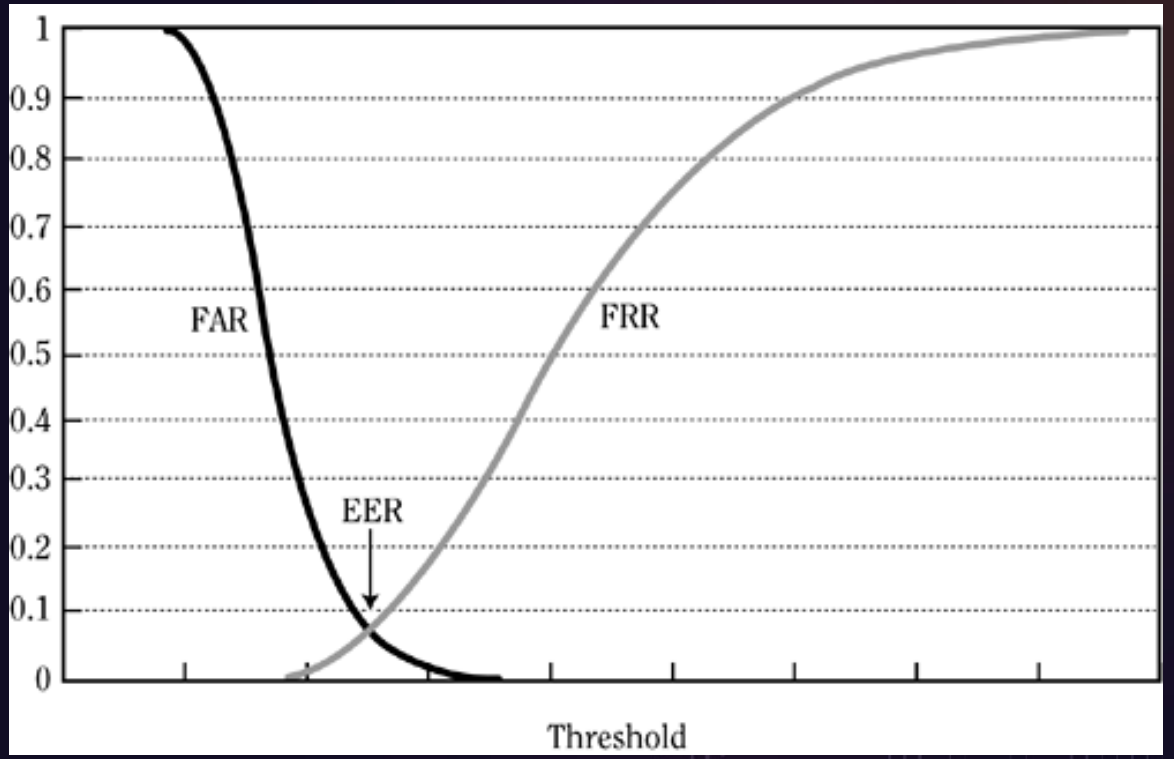
Синтез речи

Движок	Голос	CER	xRT GPU	xRT CPU	UTMOS	Похожесть Avg/Min	Encodec FAD
Silero v3_1	Aidar	0.7	0.0177	0.1256	2.544	-	97.36
Silero v3_1	Baya	0.7	0.0177	0.1256	2.978	-	170.53
Silero 4	Aidar	1.0	0.0149	0.0544	1.755	-	79.33
Silero 4	Baya	0.9	0.0149	0.0544	2.144	-	118.63
Vosk-TTS 0.6	Multi	2.3	-	0.0605	3.283	0.869/0.571	9.99
TeraTTS	Natasha	1.6	-	0.1945	3.281	-	70.10
UtrobinTTS	Female	2.1	0.0265	0.1323	2.851	-	73.34
UtrobinTTS	Male	2.1	0.0265	0.1323	3.186	-	46.14
XTTS2	Multi	2.7	0.3458	-	3.035	0.762/0.468	97.05
Vosk-TTS GPT	Multi	2.1	0.2690	-	3.381	0.814/0.544	10.08
Piper	Denis	3.7	-	0.045	3.056		142.91
Piper	Dmitry	3.6	-	0.045	2.864		130.9
Piper	Irina	1.4	-	0.045	3.672		74.98
Piper	Ruslan	3	-	0.045	2.975		72.22
BeneGes	Ruslan	2.4	-	0.321	2.537		63.02
EdgeTTS	Dmitry	0.7	-	0.076 (cloud)	3.565		32.69
EdgeTTS	Svetlana	0.7	-	0.076 (cloud)	3.513		30.60
Yandex	Alexander	0.6	-	0.028 (cloud)	3.413		54.10
Yandex	Marina	0.6	-	0.028 (cloud)	3.482		49.40
Tortoise Ruslan	Multi	6.2	25.0300	-	2.893	0.660/0.483	14.21
Bark Small	Ru_4	10.3	1.201	-	2.554		61.71

Синтез речи. Ещё метрики

- **Вариативность синтеза**
- Пропускная способность
- Управляемость

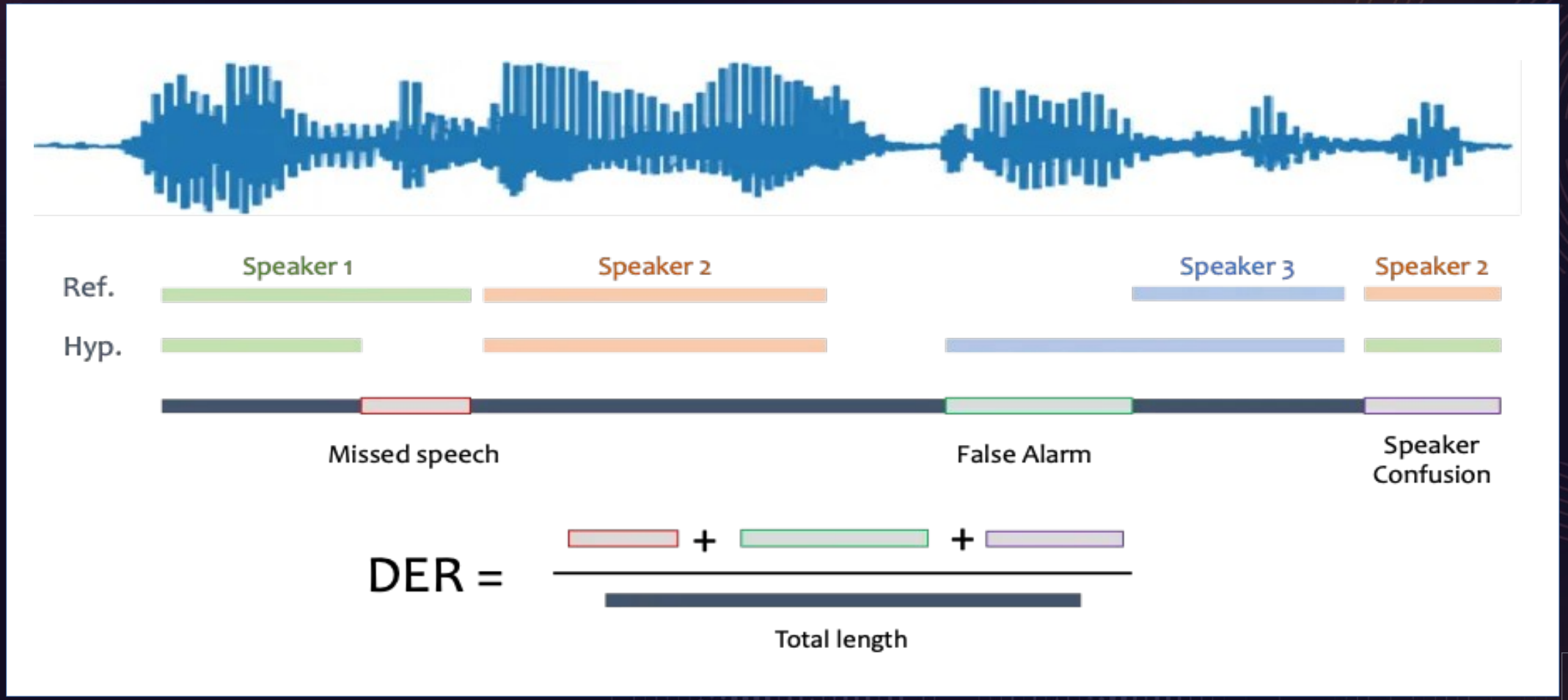
Идентификация голосов. EER



Идентификация голосов. EER

VoxBlink2 Resnet293	1.106
VoxBlink2 Resnet34	1.635
RedimNetB6	2.864
Titanet Large	3.669
ECAPA TDNN	4.402

Разделение голосов. DER



Разделение голосов. DER

Wespeaker VoxBlink2	20.10
Nemo Telephonic Cluster	22.08
Pyannote 3.1	24.4
Английский Callhome	10.4

Мультимодальные LLM

- TTFT (Time to first token) — отлично, 150ms
- Перевод — отлично, метрика BLUE 40+
- WER — обычно плохо (2 раза хуже простого распознавания)

Заключение

- Тестировать речевые системы сложно
- Вместо одной метрики — много метрик, отражающих разные аспекты системы
- Ручные тесты уходят, тестирование автоматизируется

Спасибо за внимание!

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

Шмырев Николай

nshmyrev@alphacephei.com

+7 926 106 60 08

tg://speech_recognition