факультет компьютерных наук

Анализ локальных моделей распознавания русскоязычной речи и оценка влияния дообучения на качество их работы

Analysis of local models of Russian speech recognition and assessment of the impact of additional training on their performance quality

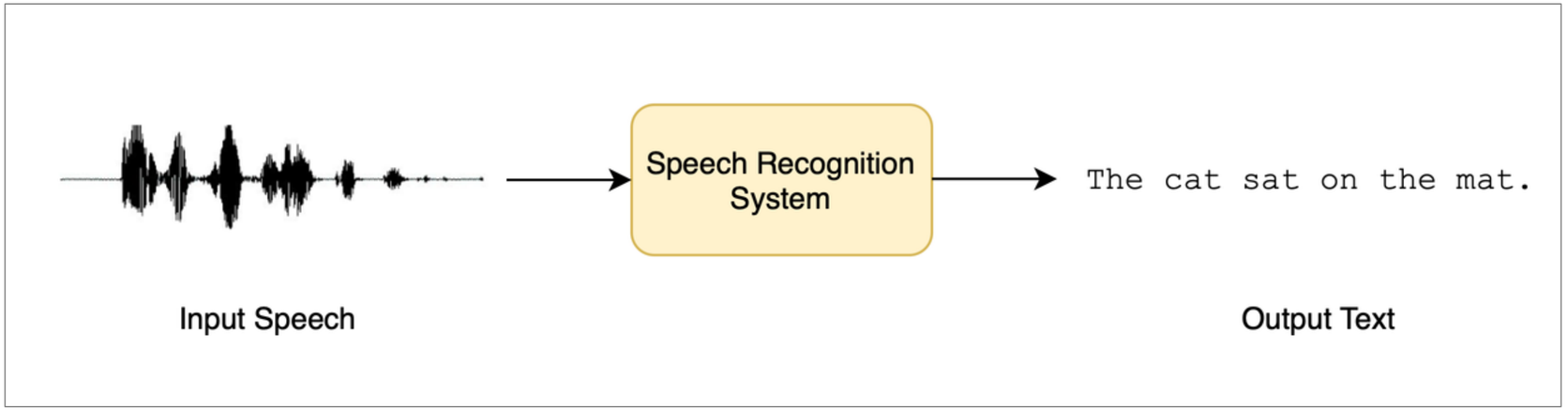
Индивидуальный исследовательский проект

Калинин Владислав Дмитриевич, БПИ 2310

Научный руководитель: Никулов Сергей Александрович, эксперт НИУ ВШЭ (БК Т-Банка)

факультет компьютерных наук

Automatic Speech Recognition

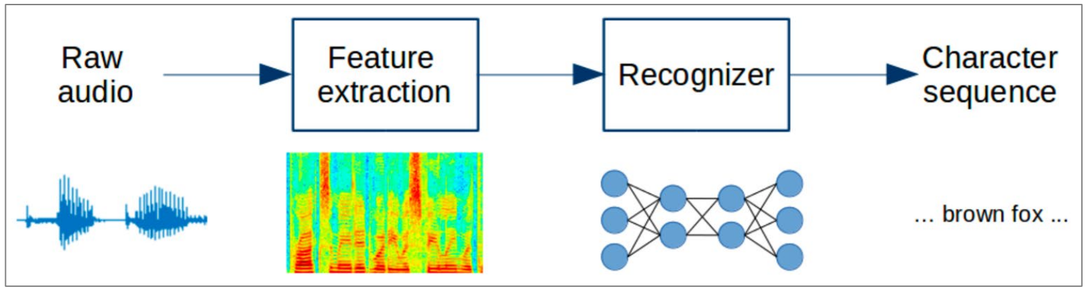


- Системы распознавания речи: от Hidden Markov Models до глубоких нейронных сетей
- Использование локальных моделей для сохранения конфиденциальности
- Применимость SOTA-моделей для конспектирования лекций ВШЭ
- Возможность дообучения на специализированных данных

факультет компьютерных наук

Цель и задачи исследования

Цель: Выработать рекомендации для практического использования локальных моделей для ASR при автоматизированном конспектировании русскоязычных лекций студентами ВШЭ



Задачи:

01 Анализ
Изучение архитектур существующих открытых решений для распознавания речи, поддерживающих работу с русским языком

02 Датасет
Составление датасета из аудиозаписей лекций ВШЭ и их транскрипций

03 Оценка
Написание и запуск скрипта для оценки WER выбранных моделей на собранном датасете

04 Дообучение
Выборочный fine-tuning исследованных моделей, сравнение WER до и после дообучения

05 Выводы
Формулирование выводов об использовании разных моделей ASR для конспектирования лекций ВШЭ и о целесообразности дообучения

факультет компьютерных наук

Основные используемые архитектуры

0

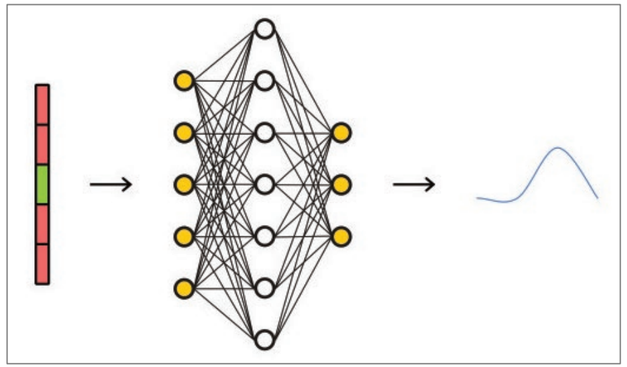
Скрытые марковские модели

Предположение об условной независимости каждой фонемы
Использование внешних языковых моделей для получения транскрипции

1

Простые нейронные сети

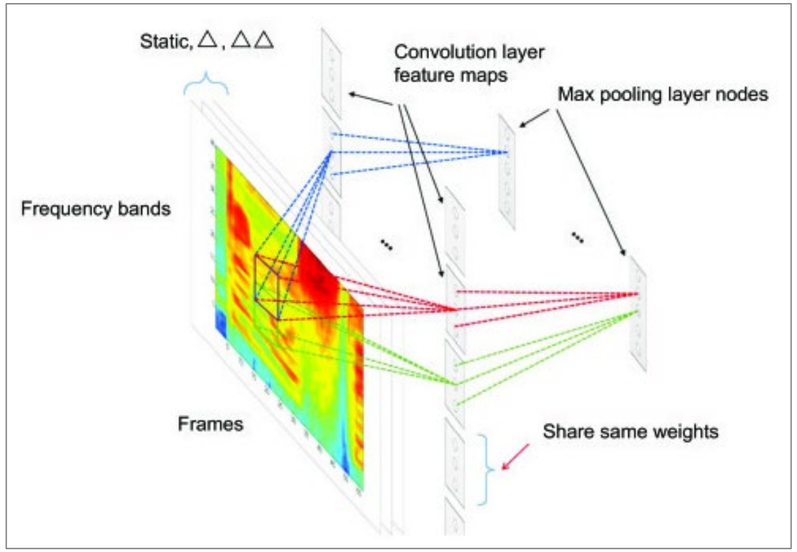
Простая реализация, хорошая точность
Плохо масштабируется



2

Свёрточные нейронные сети

Простая реализация,
высокая скорость работы



3

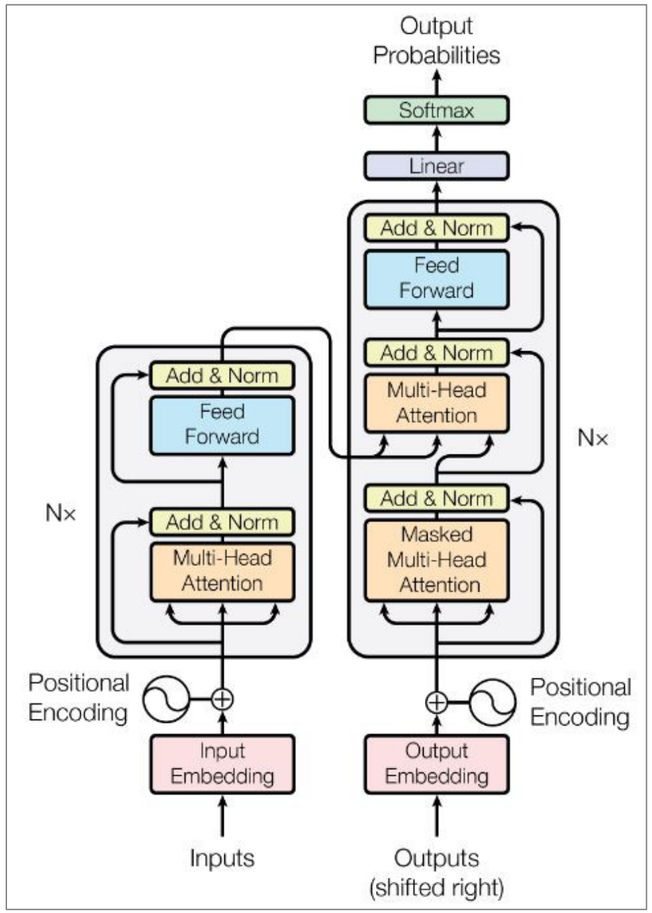
Рекуррентные нейронные сети

Использование «памяти» для
сохранения контекста

4

Трансформеры

Механизм внимания
Энкодер+декодер

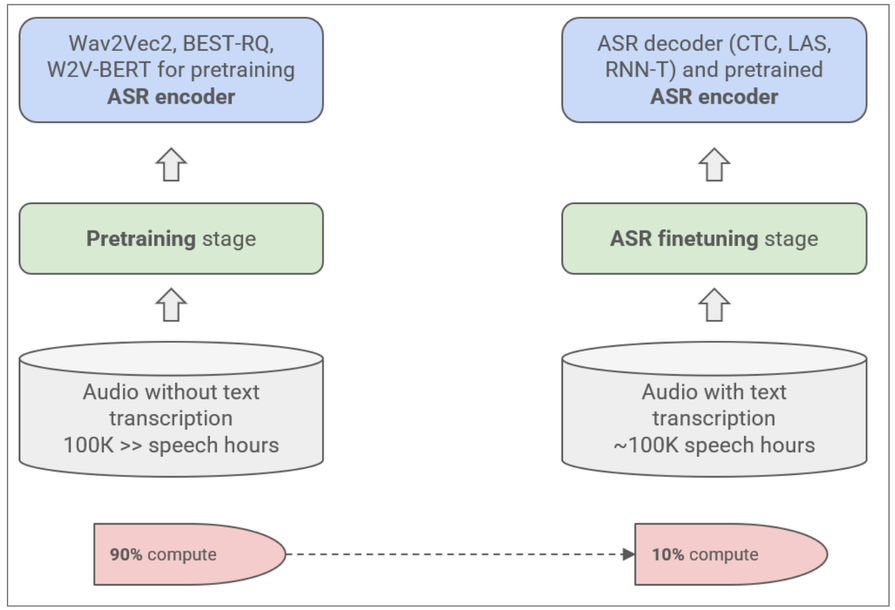


факультет компьютерных наук

Современные тренды

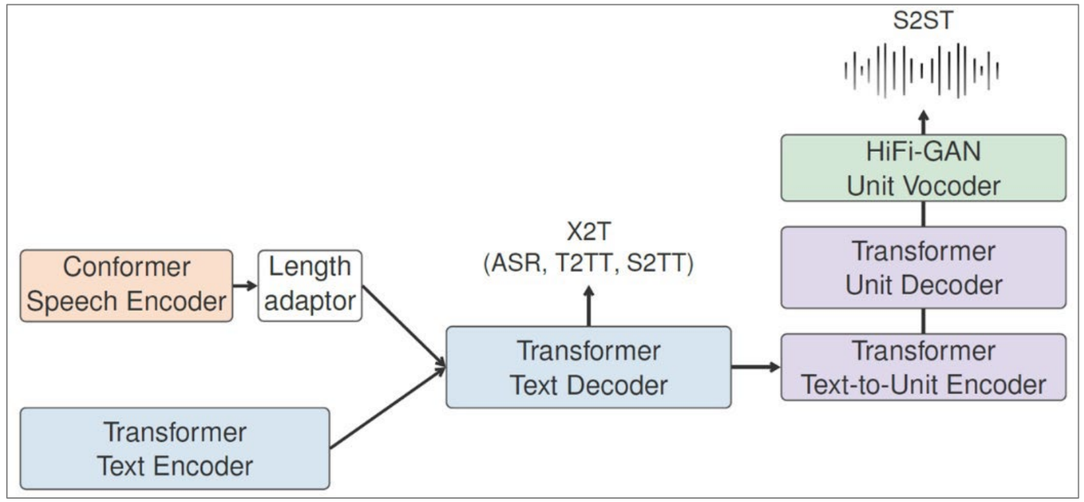
Масштабный pre-training

Обучение модели извлечения аудиопризнаков на неразмеченных данных



Мультиязычность, многозадачность

Обучение моделей работе с аудио и текстами, переводом речи и текста



факультет компьютерных наук

Выбор конкретных моделей для сравнения

Семейство моделей	SeamlessM4T	MMS	Whisper	FastConformer	GigaAM
Авторы-разработчики, год	Facebook AI Research, 2023	Facebook AI Research, 2024	OpenAI, 2023	NVIDIA, 2023	Sber Devices, 2024
Базовая архитектура	w2v-BERT 2.0	wav2vec 2.0	Transformer	Conformer	Conformer
Модальности входных и выходных данных	S/T -> S/T	S/T -> S/T	S -> T	S -> T	S -> T
Количество параметров	От 1.2B до 2.3B	От 300M до 1B	От 38M до 1.55B	От 114M до 120M	От 242M до 243M
Количество поддерживаемых языков	101 -> 96	102/1107	99	1/2/10	1
Количество исследованных версий моделей	3	3	8	4	4

*S=Speech, T=Text

https://huggingface.co/models?pipeline_tag=automatic-speech-recognition&language=ru&sort=downloads

факультет компьютерных наук

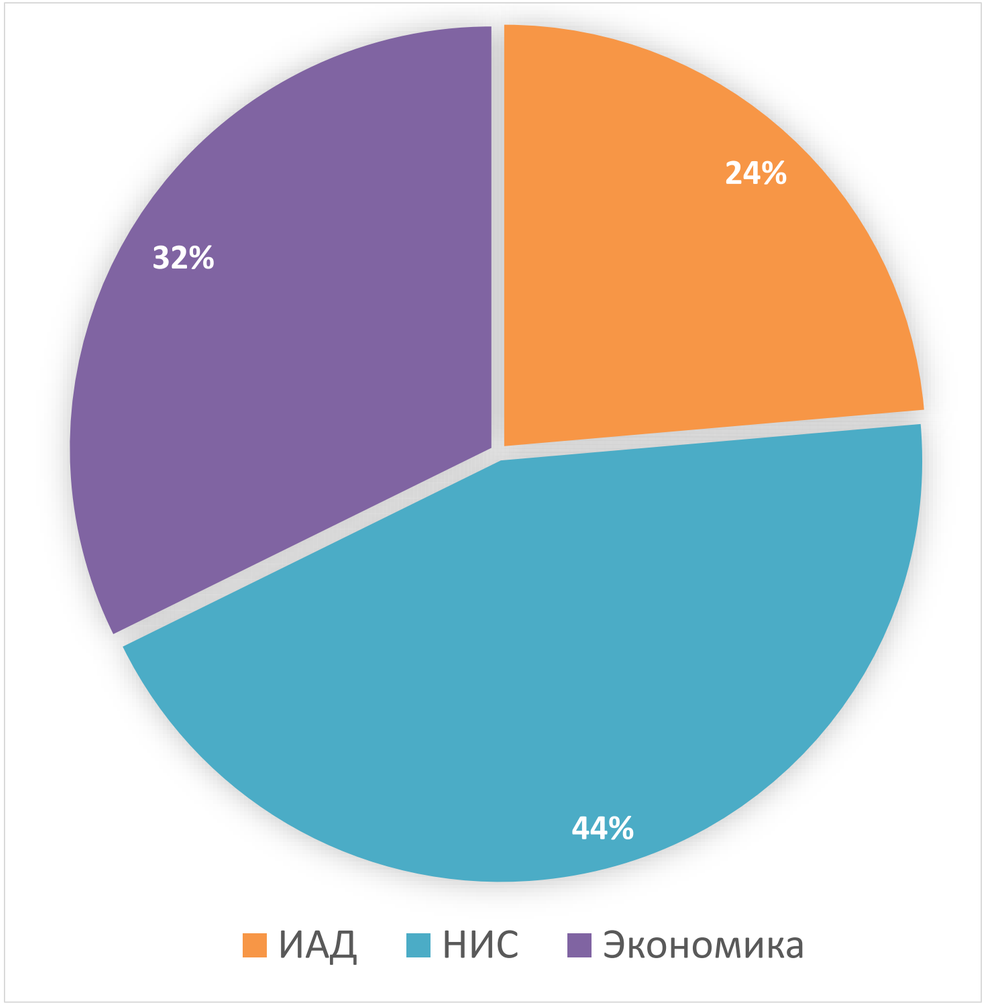
Экспериментальная установка

Данные – датасет из лекций 1 курса

- Майнор «Интеллектуальный анализ данных», поток «Базы данных» – 15 часов
- Научно-исследовательский семинар «Нейросетевые технологии» – 28 часов
- Курс «Экономика для неэкономистов» – 20.5 часов

- Разбиение датасета на train и test:
  2 к 1 с сохранением пропорций по предметам

Структура датасета



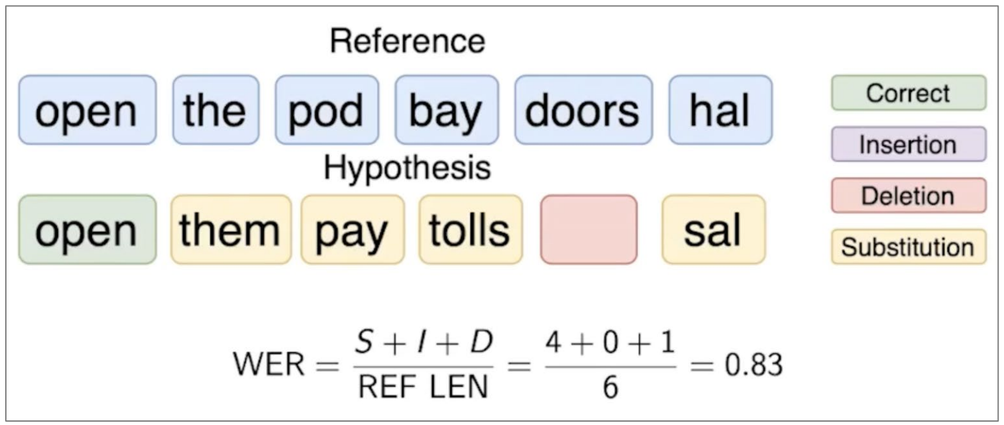
факультет компьютерных наук

Экспериментальная установка

Метрика WER

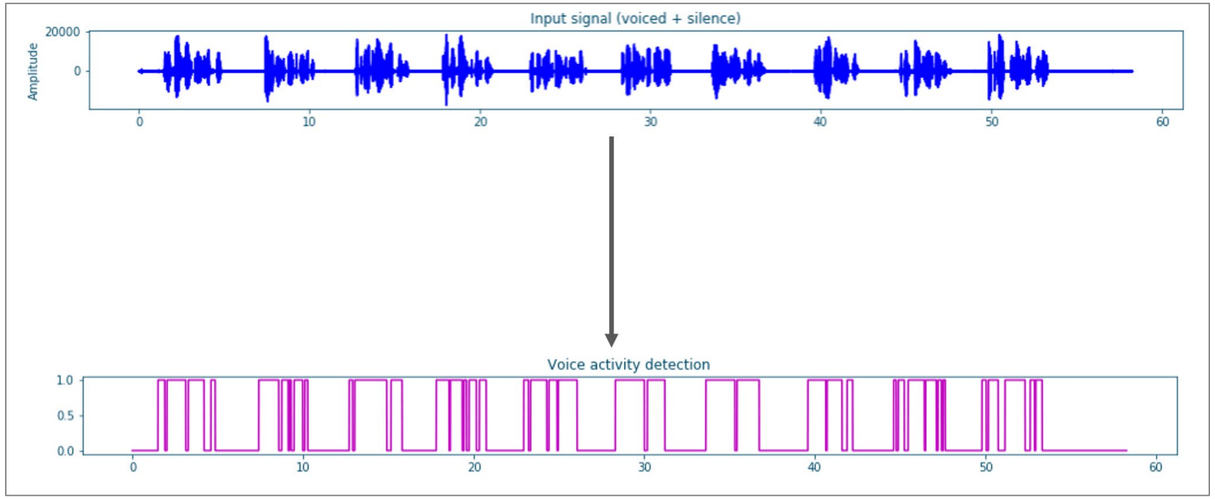
$$\mathrm{WER} = \frac{S + D + I}{N}$$

S — количество замен
D — количество удалений
I — количество вставок
N — количество слов



Дополнительная обработка

- Сегментация длинных аудио с помощью SileroVAD
- Нормализация транскрипций перед подсчётом WER: приведение к нижнему регистру, замена Ё на Е



Fine-tuning

- Использование официальных скриптов для дообучения моделей на размеченных данных

факультет компьютерных наук

Результаты экспериментов

Итоговая сравнительная таблица

Полное название модели	*Число параметров*	*WER* на test до fine-tuning*	*WER* на test после fine-tuning на train*	*Длительность fine-tuning*
SeamlessM4T Medium	1.2B	20,2%	19,49%	5 часов
SeamlessM4T Large-v1	2.3B	19,23%	19,03%	5 часов
SeamlessM4T Large-v2	2.31B	21,92%	20,75%	5 часов
MMS-1B FL102	965M	36,38%	34,88%	3 часа
MMS-1B L1107	965M	43,92%	39,44%	3 часа
MMS-1B All	965M	32,44%	31,63%	3 часа
Whisper Tiny	37.8M	38,77%	35,36%	3 часа
Whisper Base	72.6M	29,02%	27,56%	2 часа
Whisper Small	242M	19,82%	17,17%	2 часа
Whisper Medium	764M	17,36%	16,48%	2 часа
Whisper Large-v1	1.54B	16,22%	15,95%	2 часа
Whisper Large-v2	1.54B	18,49%	16,36%	2 часа
Whisper Large-v3	1.54B	12,45%	11,25%	3 часа
Whisper Turbo	809M	12,51%	11,58%	3 часа
NVIDIA STT Ru Conformer-CTC Large	120M	15,16%	14,38%	4 часа
NVIDIA STT Multilingual FastConformer Hybrid Transducer-CTC Large P&C	114M	16,39%	15,72%	4 часа
NVIDIA FastConformer-Hybrid Large ru	115M	14,06%	13,58%	4 часа
NVIDIA FastConformer-Hybrid Large kk-ru	115M	17,82%	17,04%	4 часа
GigaAM CTC-v1	242M	9,42%	-	-
GigaAM RNNT-v1	243M	8,74%	-	-
GigaAM CTC-v2	242M	8,35%	-	-
GigaAM RNNT-v2	243M	7,99%	-	-

* WER ниже – лучше

факультет компьютерных наук

Выводы и дальнейшая работа

Выводы

- Использовать Whisper или GigaAM
- Fine-tuning слишком затратный
- Нужна пост-обработка для полноценных конспектов

Пример – лучшие модели

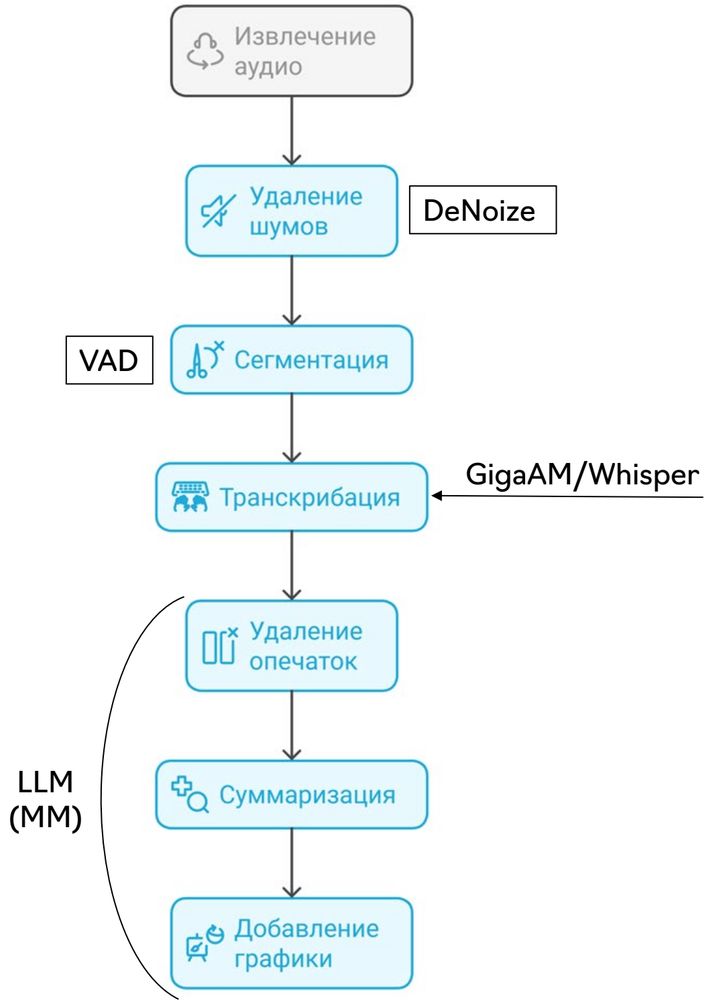
- GigaAM RNNT-v2

  вот предположим есть у нас преподаватель который работает там репетитором допустим за десять долларов в час и при зарплате в десять долларов он отрабатывает в неделю энное количество часов что будет если его зарплата повысится до ста долларов в час ну мечтать не вредно предположим тогда вероятно этот работник будет работать еще больше часов потому что теперь для него отдых станет слишком дорогим удовольствием представьте себе не поработать один час это означает потерять сто долларов слишком дорогим удовольствием становится отдых

- Whisper large-v3

  Вот, предположим, есть у нас преподаватель, который работает репетитором, допустим, за 10 долларов в час. И при зарплате в 10 долларов он отрабатывает в неделю энное количество часов. Что будет, если его зарплата повысится до 100 долларов в час? Ну, мечтать не вредно, предположим. Тогда, вероятно, этот работник будет работать еще больше часов. Потому что теперь для него отдых станет слишком дорогим удовольствием. Представьте себе, не поработать один час, это означает потерять 100 долларов. Слишком дорогим удовольствием становится отдых.

Полный pipeline системы конспектирования



факультет компьютерных наук

Список источников

1) Pratap V. et al. Scaling speech technology to 1,000+ languages //Journal of Machine Learning Research. – 2024. – Т. 25. – №. 97. – С. 1-52.
2) Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – С. 28492-28518.
3) Господинов, Г. GigaAM: класс открытых моделей для обработки звучащей речи [Электронный ресурс] / Habr. Режим доступа: https://habr.com/ru/companies/sberdevices/articles/805569, свободный. (дата обращения: 21.04.2025).
4) Gales, M.J.F. & Young, Steve. (2007). The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing. 1. 195-304. 10.1561/2000000004.
5) Graves, Alex & Fernández, Santiago & Gomez, Faustino & Schmidhuber, Jürgen. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning. 2006. 369-376. 10.1145/1143844.1143891.
6) Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.
7) Schneider S. et al. wav2vec: Unsupervised pre-training for speech recognition //arXiv preprint arXiv:1904.05862. – 2019.
8) Gulati A. et al. Conformer: Convolution-augmented transformer for speech recognition //arXiv preprint arXiv:2005.08100. – 2020.
9) Graves A. Sequence transduction with recurrent neural networks //arXiv preprint arXiv:1211.3711. – 2012.
10) Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-12460.
11) Barrault L. et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation //arXiv preprint arXiv:2308.11596. – 2023.
12) Chung Y. A. et al. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training //2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). – IEEE, 2021. – С. 244-250.
13) salute-developers. GigaAM: Foundational Model for Speech Recognition Tasks [Электронный ресурс] / GitHub. Режим доступа: https://github.com/salute-developers/GigaAM, свободный. (дата обращения: 21.04.2025).
14) OpenAI. Whisper: Robust Speech Recognition via Large-Scale Weak Supervision [Электронный ресурс] / GitHub. Режим доступа: https://github.com/openai/whisper, свободный. (дата обращения: 21.04.2025).
15) NVIDIA. NeMo: A scalable generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (Automatic Speech Recognition and Text-to-Speech) [Электронный ресурс] / GitHub. Режим доступа: https://github.com/NVIDIA/NeMo, свободный. (дата обращения: 21.04.2025).
16) snakers4. Silero VAD: pre-trained enterprise-grade Voice Activity Detector [Электронный ресурс] / GitHub. Режим доступа: https://github.com/snakers4/silero-vad, свободный. (дата обращения: 21.04.2025).

ФКН
факультет компьютерных наук

Спасибо за внимание!