**Задача 1:**

**Для распознавания текста из аудиодорожки я рассмотрел несколько open-source нейросетей на платформе huggingface, обученные на русскоязычных текстах, где для дальнейшей работы выделил whisper и wav2vec модели, как sota-модели с обилием документации и приемлимым качеством распознавания. Учитывая бенчмарк (https://deepgram.com/learn/benchmarking-top-open-source-speech-models) и условие задачи, я решил остановиться на wav2vec модели, поскольку она требует меньше вычислительных ресурсов, что влияет на скорость распознавания, по качеству на сгенерированном аудио обе модели равны (**<https://arxiv.org/pdf/2212.04356>**). После на huggingface я наткнулся на модель от Nvidia Nemo (**<https://huggingface.co/nvidia/stt_ru_conformer_transducer_large>**) и бенчмарк (**<https://alphacephei.com/nsh/2023/01/22/russian-models.html>**), Nvidia RNNT показала более низкую ошибку на тестах и высокую скорость, решил использовать именно ее. В результате личного тестирования скорость у wav2vec для первой фразы 0.19 сек, для Nvidia RNNT 0.04 сек. на видеокарте 2080ti. Для генерации текста использовал pretrained на русском языке модель Silero, которая доступна в pytorch.**

#### **Задача 2:**

#### Для разделения аудио на отдельные персоны использовал sota метод pyannote (<https://huggingface.co/pyannote>), который показывает хорошее качество сегментации на большинстве тестах.

1. https://github.com/moksyasha/TestASR

2. Python 3.11.9

3. CUDA 12.5

4. Linux Ubuntu 24.04

5. Получил 19.07.2024 – Отправил 20.07.2024

6. Корнеич Никита Алексеевич