Министерство цифрового развития

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра прикладной математики и кибернетики

Отчёт

по лабораторной работе № 7 «Распознавание голосовых сообщений с использованием
готовой модели Whisper от OpenAI.»

Выполнил:

студент группы ИП-312

Прозоренко К.В

Работу проверил: старший преподаватель кафедры ПМиК
Дементьева К.И.

Новосибирск 2025 г.

Задание

Выберите один набор данных из предложенных:

1. LibriSpeech.

2. Common Voice.

3. VoxCeleb.

4. Nexdata.

5. Свой собственный набор: 5–10 голосовых сообщений из Telegram/WhatsApp (формат .mp3, .wav или .ogg).

Задание 1. Подготовка данных

1. Установите необходимые библиотеки

2. Загрузите 5-10 аудиофайлов, при необходимости преобразуйте в формат .waw.

Задание 2. Транскрибация аудио с помощью модели Whisper

Для каждого аудиофайла выполните транскрибацию и сохраните результаты в файл в формате: файл: транскрипция.

Задание 3. Визуализация и анализ результатов

Для одного аудиофайла визуализируйте спектрограмму и проанализируйте, какие слова модель распознала неверно и почему.

Ход выполнения:

Установил необходимые библеотеки
pip install -q git+https://github.com/openai/whisper.git

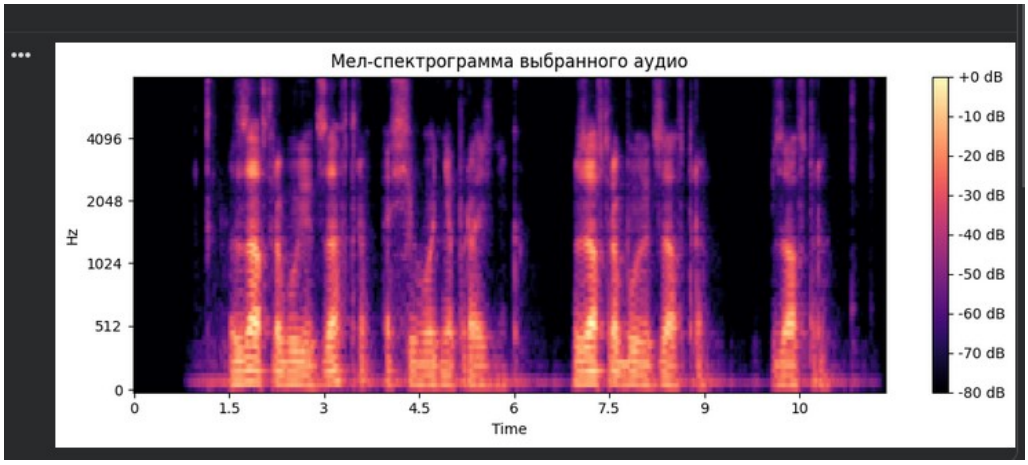
pip install -q torch librosa matplotlib

Загрузил 6 голосовых сообщений из Telegram. Каждое сообщение длится 5~7 секунд.

Выполнил транскрибацию для каждого файла и сохранил результат в отдельный файл

```
Обрабатываю: audio/audio_2025-12-22_02-03-31.ogg
audio_2025-12-22_02-03-31.ogg -> Завтра будет завтра, а сегодня праздник.
Обрабатываю: audio/audio_2025-12-22_02-03-44.ogg
audio_2025-12-22_02-03-44.ogg -> Sun sunnich. Sun sunnich. Sun sunnich.
Обрабатываю: audio/audio_2025-12-22_02-03-48.ogg
audio_2025-12-22_02-03-48.ogg -> Доварищ сухом говорит.
Обрабатываю: audio/audio_2025-12-22_02-03-51.ogg
audio_2025-12-22_02-03-51.ogg -> Это манба фантастика.
Обрабатываю: audio/audio_2025-12-22_02-03-54.ogg
audio_2025-12-22_02-03-54.ogg -> Отсюда она тут свет. В керзовых сопогах я не хотел.
Обрабатываю: audio/audio_2025-12-22_02-04-01.ogg
audio_2025-12-22_02-04-01.ogg -> Where is my mind?
Сохранено в transcriptions.txt
```

Построил спектрограмму для 1 файла.



Можно четко увидеть в каких местах были паузы.

Для анализа ошибок я подсчитал WER

```python
from jiwer import wer

refs = [
    "Завтра будет завтра, а сегодня праздник.",
    "Сан Саныч. Сан Саныч. Сан Саныч.",
    "Товарищ Сухов говорит.",
    "Это мамба фантастика.",
    "Отсюда на тот свет. В керзовых сапогах я не хотел.",
    "Where is my mind? Where is my mind?"
]

hyps = [
    "Завтра будет завтра, а сегодня праздник.",
    "Sun sunnich. Sun sunnich. Sun sunnich.",
    "Доварищ сухом говорит.",
    "Это манба фантастика.",
    "Отсюда она тут свет. В керзовых сопогах я не хотел.",
    "Where is my mind?"
]

print("Средний WER по всем сообщениям:", wer(refs, hyps))
```

```
Средний WER по всем сообщениям: 0.4444444444444444
```

Как видно модель распознала сообщения с долей ошибок 44%. Что в целом неплохой результат, учитывая что в одном из сообщений она не угадала русский язык.

Ссылка на Google Collab:
https://colab.research.google.com/drive/1cnkMIbSqbU2vUsyIt1ayvHgv8sUdf2IO?usp=sharing