**Task 1**

**Сравнение моделей распознавания речи: Google, OpenAI и Сбер**

При анализе и тестировании различных моделей распознавания речи были рассмотрены три основных модели: Google Speech Recognition, OpenAI Whisper Large v3 и Сбер (GigaAM или SaluteSpeech). Каждый из этих сервисов был протестирован по ряду параметров, таких как скорость работы, точность распознавания, поддержка русского языка и доступность. В этом отчете представлены результаты сравнительного анализа этих моделей.

**1. Google Speech Recognition**

**Преимущества:**

* **Поддержка русского языка:** Единственная модель из трех, которая может работать с русским языком в рамках модуля speech\_recognition.
* **Скорость работы:** Быстро обрабатывает запросы:
  + 0.58 секунд на распознавание трех слов.
  + 6.11 секунд на распознавание пары предложений.

**Недостатки:**

* **Пунктуация:** Не учитывает пунктуацию, что может усложнять последующую обработку текста.

**2. OpenAI Whisper Large v3**

**Преимущества:**

* **Точность распознавания:** Превосходит Google Speech Recognition в точности распознавания.
* **Пунктуация:** Учитывает пунктуацию, что значительно улучшает читаемость и качество выходного текста.

**Недостатки:**

* **Скорость работы:** Крайне медленная:
  + 30 секунд на распознавание трех слов.
* **Требования к оборудованию:** Работает локально без мощного оборудования, которое предоставляет сервис когда используешь API, что может быть затратно и неудобно для пользователей без доступа к необходимым вычислительным ресурсам.

**P.S**.: После некоторых манипуляций, которые указаны на странице модели на HuggingFace (догрузка некоторых библиотек), скорость уменьшилась в 6 раз, но качество распознавания упало в ноль. За 10 попыток ни разу не удалось определить ключевую фразу.

**3. Сбер (GigaAM/SaluteSpeech)**

**Преимущества:**

* **API и локальная установка:** Возможность использования API или локальной версии.
* **Скорость работы:**
  + На коротких текстах чуть медленнее Google (1.7 секунд).
  + На длинных текстах значительно опережает Google (3.1 секунд на двух предложениях).

**Недостатки:**

* **Ограничения бесплатной версии:** Ограничение на загрузку текстов более 30 секунд, что может быть неудобно для длительных записей.

**Заключение**

Из трех рассмотренных моделей, **Сбер (GigaAM или SaluteSpeech)** является наиболее рекомендованной для использования. Она предоставляет баланс между точностью и скоростью распознавания, а также удобство использования через API. В случае превышения лимитов бесплатной версии, **Google Speech Recognition** является хорошей альтернативой. Модель **OpenAI Whisper Large v3** из-за своей медлительности и высоких требований к оборудованию не рекомендуется для локального использования без мощных ресурсов.

Этот анализ был проведен на основе тестирования и сравнения различных параметров каждой модели, что позволило определить наиболее подходящий вариант для практического применения.

**Таблица сравнений:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Модель** | **Скорость (короткий текст)** | **Скорость (длинный текст)** | **Поддержка русского языка** | **Пунктуация** | **Ограничения** |
| Google Speech Recognition | 0.58 сек | 6.11 сек | Да | Нет | Нет |
| OpenAI Whisper Large v3 | 30 сек | 30 сек | Да | Да | Требует мощного оборудования |
| Сбер (GigaAM/SaluteSpeech) | 1.7 сек | 3.1 сек | Да | Да | Лимит 30 секунд |

**Рекомендация:**

* **Сбер (GigaAM или SaluteSpeech)** для большинства задач.
* **Google Speech Recognition** в случае необходимости обойти лимиты.
* **OpenAI Whisper Large v3** не рекомендуется для локального использования без мощного оборудования.

Для оценки качества распознавания речи была также рассмотрена метрика отклонения от изначального текста. Из-за различных ограничений (со стороны моделей или ресурсов) объективно проверить этот параметр самостоятельно не удалось. Тем не менее, на основе найденной таблицы (<https://alphacephei.com/nsh/2024/04/14/russian-models.html>) с метриками отклонений, лучшее качество распознавания показала модель от Сбера.

**Task 2.**

В ходе исследования платформы Hugging Face были рассмотрены несколько моделей от пользователя pyannote, в частности модели для диаризации речи под названием speaker-diarization. Также была изучена возможность использования диаризации, встроенной в модель Whisper от OpenAI, однако эта идея была отвергнута из-за многочисленных ограничений, выявленных в предыдущем анализе.

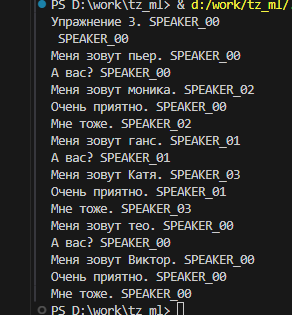
**Выбор моделей:**

Для анализа использовалась последняя версия модели speaker-diarization от pyannote, так как она показала лучшие результаты по метрикам, указанным на странице проекта. Возможно, использование Whisper было бы более удобным, так как эта модель способна выполнять обе задачи: разделение текста и перевод на русский язык. Однако, исходя из предыдущего анализа, было решено использовать лучшее решение из прошлого задания и интегрировать его с текущим.

**Алгоритм решения:**

* Загрузка модели диаризации и аудио
* Загрузка аудио в модель для получения разбитого списка
* Запуск цикла по полученному списку
  + Загружается копия аудио
  + Аудио разбивается по разнице конечной точки этого промежутка и прошлого
  + Левая часть аудио переводится в текст с помощью модели Сбера и выводится вместе с говорящим
  + Правая часть аудио становится копией аудио из первого шага цикла

По итогу работы алгоритма получается список

Сбер

Спикеров должно быть больше, но из-за того, что диктора всего 2 (мужчина и женщина), определяет он ошибочно.

Также изначально планировал работать с моделью гугла, из-за простоты её использования и более быстрой работе на коротких текстах, однако модель показала себя слабо. Текст, полученный с её помощью, был трудно читаем, содержал множество ошибок и были "съедены" все имена, что существенно ухудшило качество результата.

A screen shot of a computer

Description automatically generated  
гугл