

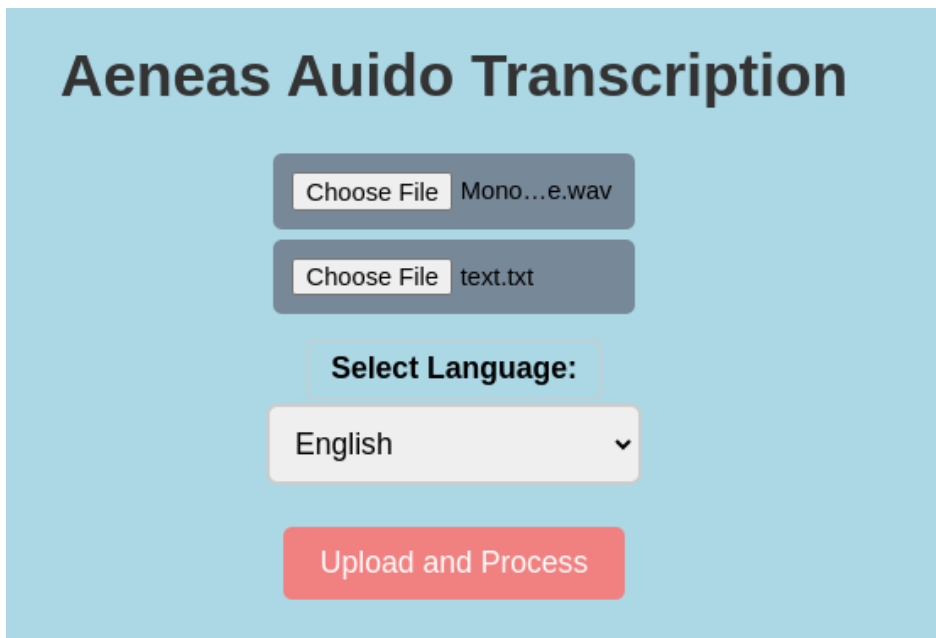
BTP - 2

Mitul Garg 2020102026
Atharv Sujlegaonkar 2020102025

Overview

- Our problem statement is to develop/Improvise current **Open-source** Solutions for forced aligners to our use case - **Indian Languages**.
- So we worked on an Aligner called **Aeneas** and improved its working.
- The Overall project is deployed on a local web server which can be used to obtain a transcription for a given text and audio sample.

The Website UI :



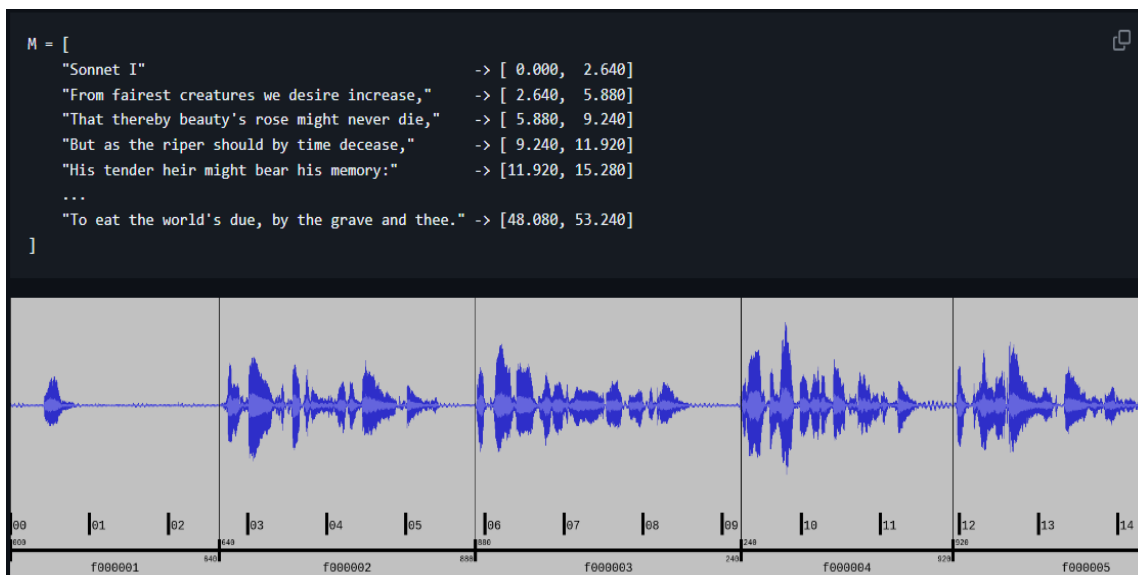
The screenshot shows a web interface titled "Aeneas Audio Transcription" on a light blue background. It features two file upload sections, each with a "Choose File" button and a text input field. The first section has "Mono...e.wav" in the input field, and the second has "text.txt". Below these is a "Select Language:" label above a dropdown menu currently showing "English" with a downward arrow. At the bottom is a red "Upload and Process" button.

What is Aligning ?

The mapping of audio and text in a time interval, one can relate them to **captions/subtitles**. Ex -

```
"Sonnet I" -> [ 0.000, 2.640]
"From fairest creatures we desire increase," -> [ 2.640, 5.880]
"That thereby beauty's rose might never die," -> [ 5.880, 9.240]
"But as the ripper should by time decease," -> [ 9.240, 11.920]
"His tender heir might bear his memory:" -> [11.920, 15.280]
...
"To eat the world's due, by the grave and thee." -> [48.080, 53.240]
```

Forced Aligner - Aeneas



How Aeneas works :

- Converts audio to mono channel and marks the intervals of **voiced, unvoiced and silence regions**.
- Converts text to plain format to allocate fragments timestamps after aligning.
- Using a **TTS model**, it obtains a Synthesized audio.
- Both the initial audio (R) and synthesized (S), are then used to obtain **MFCC** matrices for both.
- Afterwards, it performs a **DTW** (Dynamic time warping) technique to obtain a mapping between R and S.
- This approach is **$O(N*M)$** , N and M being the sizes of R and S.

As the approach is of exponential complexity, (**M and N being directly proportional to Audio lengths**), we need to optimize the way we feed the text data to aeneas and our audio files. It has been found out that Aeneas works best when the audio files are less than **30 seconds** and delivers good speed even on **smaller GPUs**.

Improvements done :

- As noted that our audio should be in the range of 30 seconds, we needed to work on fragmenting the text.
- We noted that there is a general trend that a person stops speaking for around **0.5-3 seconds** during a **Full-Stop (.), and a Comma (,)**.
- So, we used this to **track silence regions** in the audio and split the audio up.
- Another thing to note is that sentences which are spoken for longer than **30 seconds**, generally have multiple commas in between, so that case is also taken care of .

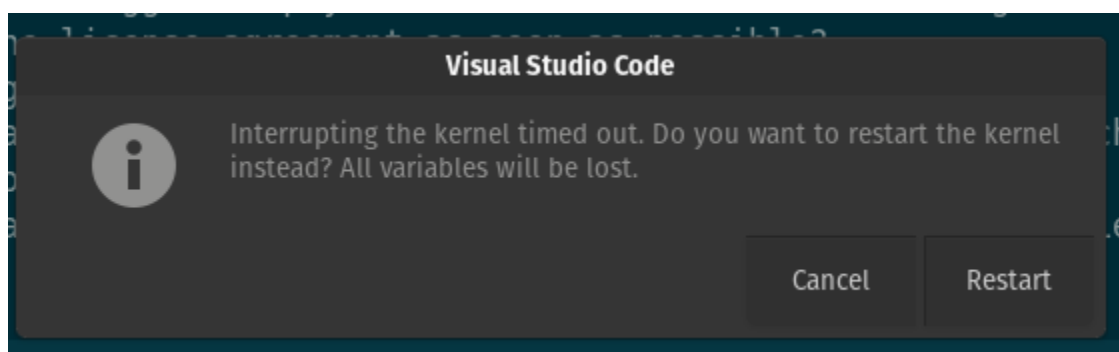
Datasets Used :

English and Hindi audiotext samples ranging in
10 seconds to >5 minutes.

```
[
  {
    "audioFilename": "train_hindifullmale_00001.wav",
    "text": "परसिद् कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, उस रामानंद की खोज करता है",
    "speaker": "iitm_ph2_hindi_male_spk1",
    "duration": 6.960770833333333,
    "gender": "male"
  },
  {
    "audioFilename": "train_hindifullmale_00002.wav",
    "text": "किन्तु आधुनिक पांडित्य, न सिर्फ़ एक ब्राह्मण रामानंद के, एक जुलाहे कबीर का गुरु होने से, बल्कि दोनों के समकालीन",
    "speaker": "iitm_ph2_hindi_male_spk1",
    "duration": 11.2083125,
    "gender": "male"
  },
  {
    "audioFilename": "train_hindifullmale_00003.wav",
    "text": "उस पर, इन चार कवियों का गहरा असर है",
    "speaker": "iitm_ph2_hindi_male_spk1",
    "duration": 3.435583333333333,
    "gender": "male"
  },
  {
    "audioFilename": "train_hindifullmale_00004.wav",
    "text": "इसे कई बार, मंचित भी किया गया है",
    "speaker": "iitm_ph2_hindi_male_spk1",
    "duration": 2.843125,
    "gender": "male"
  },
  {
    "audioFilename": "train_hindifullmale_00005.wav",
    "text": "यहाँ परस्तुत है, हिन्दी कवि कथाकार, तेजी ग़रेवर के अंग्रेज़ी के मार्फ़त किए गए अनुवाद के कुछ अंश",
    "speaker": "iitm_ph2_hindi_male_spk1",
    "duration": 7.661895833333333,
    "gender": "male"
  }
]
```

Issues in Normal Aeneas :

Fails on longer audios and Longer samples in
transcriptions gives poor resolution.



Results

The obtained transcription look like this :

```
{
  "fragments": [
    {
      "begin": "0.000",
      "children": [],
      "end": "12.120",
      "id": "f000001",
      "language": "eng",
      "lines": [
        "Nigel: Glad to see things are going well and business is starting to pick up."
      ]
    },
    {
      "begin": "12.120",
      "children": [],
      "end": "18.760",
      "id": "f000002",
      "language": "eng",
      "lines": [
        "Andrea told me about your outstanding numbers on Tuesday."
      ]
    },
    {
      "begin": "18.760",
      "children": [],
      "end": "22.840",
      "id": "f000003",
      "language": "eng",
      "lines": [
        "Keep up the good work."
      ]
    },
    {
      "begin": "22.840",
      "children": [],

```

```

{
  "fragments": [
    {
      "begin": "0.000",
      "children": [],
      "end": "2.120",
      "id": "f000001",
      "language": "hin",
      "lines": [
        "किन्तु आधुनिक पांडित्य,"
      ]
    },
    {
      "begin": "2.120",
      "children": [],
      "end": "4.480",
      "id": "f000002",
      "language": "hin",
      "lines": [
        "न सिर्फ़ एक ब्राह्मण रामानंद के,"
      ]
    },
    {
      "begin": "4.480",
      "children": [],
      "end": "7.360",
      "id": "f000003",
      "language": "hin",
      "lines": [
        "एक जुलाहे कबीर का गुरु होने से,"
      ]
    },
    {
      "begin": "7.360",
      "children": [],
      "end": "10.040"
    }
  ]
}

```

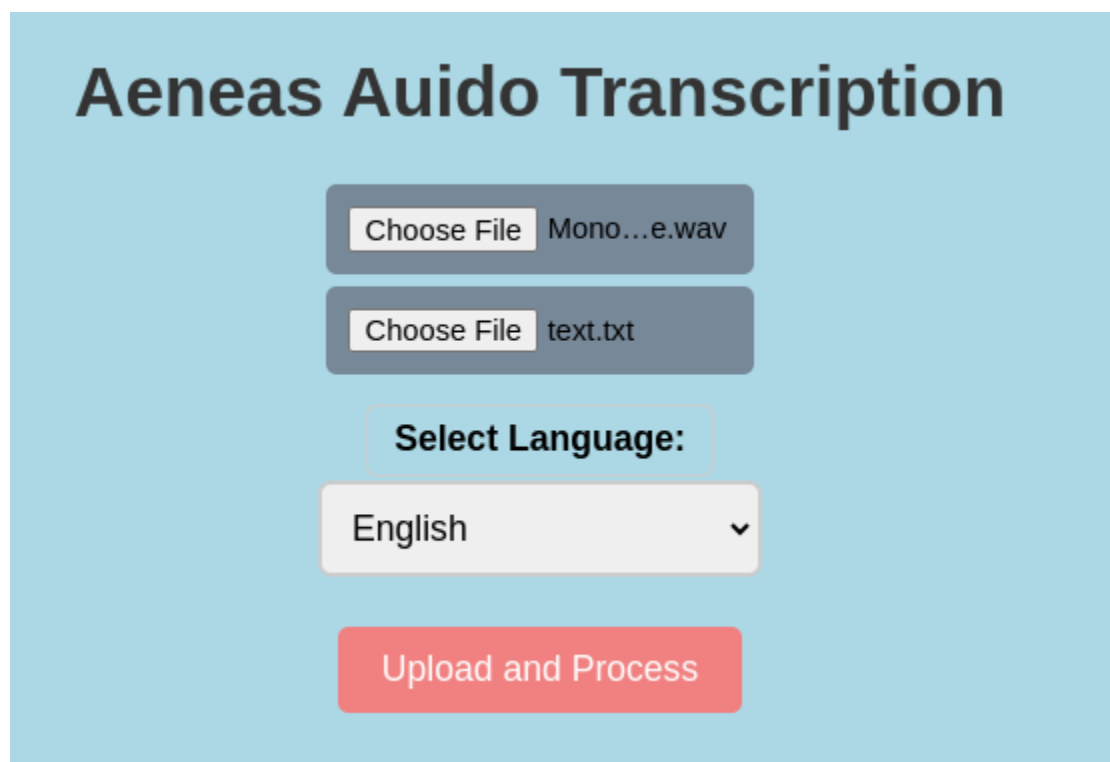
The Website:

We deployed the **Python backend using Flask**(an open source Python framework) and the Frontend using **React JS and Axios**.

Our Results:

- The **overall per segment run-time has been brought down** so that it can be run on smaller GPUs, as the chunks now are at <30 seconds.
- Better Alignment results than normal Aeneas, as there is a tracking error in longer audios.
- Better Resolution on sentence level, as Aeneas merges 2-3 sentences in case of longer audios and confuses it with the synthesized audio from TTS.

The website UI :



The screenshot shows the user interface of the 'Aeneas Audio Transcription' website. The title 'Aeneas Audio Transcription' is displayed in a large, bold, dark font at the top. Below the title, there are two file upload sections. The first section has a 'Choose File' button and a text input field containing 'Mono...e.wav'. The second section also has a 'Choose File' button and a text input field containing 'text.txt'. Below these, there is a 'Select Language:' label above a dropdown menu that currently shows 'English' with a downward arrow. At the bottom of the form is a large red button labeled 'Upload and Process'.

Thank you