DATA COLLECTION AND PREPROCESSING

Vishnuvardhan(20mcme13)

Data Collection: It involves collection of diverse speech samples in languages like

- Hindi
- Bengali
- Telugu
- Tamil
- English(Indian accent)

to train an accurate and robust speech recognition model.Capturing various accents and speaking styles in those languages to enhance the model's comprehension of natural speech and taking care of the data should be well generalised over different speakers and contexts.

We can get this type of data from many open source datasets providers like common voice (kaggle) ,openslr.org,hugging face,etc. I have collected the data from the dataset "google/fleurs" available on Hugging Face Datasets.We have chosen the data from the Indian languages which are part of South-Asia geographical areas of the fleurs.And kaggle has a dataset for english where there is a accent column and many data samples belong to the indian accent.so we can get the indian accent from there.

Now we get datasets of Hindi,Bengali,Tamil and Telugu in which we have the audio files and tsv files with 9 attributes.And total audio samples size and number of audio samples for each language as follow,

Hindi :total audio samples size -132 MB ,no.of audio samples -239

Bengali : total audio samples size -279 MB ,no.of audio samples -402

Tamil : total audio samples size -239 MB ,no.of audio samples -377

Telugu : total audio samples size -167 MB ,no.of audio samples -311

We can download the audio sample from the links :

HINDI:

https://huggingface.co/datasets/google/fleurs/tree/main/data/hi_in/audio

BENGALI:

https://huggingface.co/datasets/google/fleurs/tree/main/data/bn_in/audio

TELUGU:

https://huggingface.co/datasets/google/fleurs/tree/main/data/te_in/audio

TAMIL:

https://huggingface.co/datasets/google/fleurs/tree/main/data/ta_in/audio

Attributes of the tsv file:

- id (int): ID of audio sample
- num_samples (int): Number of float values
- path (str): Path to the audio file
- audio (dict): Audio object including loaded audio array, sampling rate and path ot audio
- raw_transcription (str): The non-normalized transcription of the audio file
- transcription (str): Transcription of the audio file
- gender (int): Class id of gender
- lang_id (int): Class id of language
- lang_group_id (int): Class id of language group

And indian accent english has 7 attributes:

- filename - relative path of the audio file
- text - supposed transcription of the audio
- up_votes - number of people who said audio matches the text
- down_votes - number of people who said audio does not match text
- age - age of the speaker, if the speaker reported it
- gender - gender of the speaker, if the speaker reported it
- accent - accent of the speaker, if the speaker reported it

In Hindi,Telugu,Bengali and Tamil we need only some attributes and the remaining should be removed and the file should be in csv format.with some requirements like the transcripted text with no comma,full stops,etc.,and we need to divide our data collected into training and test data as per the required percentage of splits.so to achieve that the preprocessing of data is to be done.But in English we need to sort the file according to the accent Indian first and the we need remove the unwanted columns and the we need to move the audio sample according to the file name in the csv file and verify for the tools that convert audio to text so that the process is done properly.

And we can download the english data sets from the https://www.kaggle.com/datasets/mozillaorg/common-voice?resource=download website.

Preprocessing of data:

After the process of data collection the data samples and the transcription to be preprocessed so the the it is perfectly given to the model for training.Here the requirements for the preprocessing are

1. The data should be in the csv file format
2. The data should not contain commas,fullstops,hyphens,etc.,
3. The file should have two columns only . That is the audio file name/path and its transcription
4. Now the data needs to be divided into training and testing data based on the requirements.

Here for the preprocessing of the english datasets we need to convert the folder with large no.of audio sample file to the audio samples with the indian accent only based on the names in the csv file with audio sample name and its transcription.

We have hindi_transcription.tsv file to see the columns we use file_details.py

```
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/file_details.py"
Enter the input TSV file name: hindi_transcription.tsv
Number of rows: 238
Number of columns: 3

Column Attribute Names:
14584887621258891555.wav
इतालवी भाषा में उच्चारण करना तुलनात्मक रूप से आसान है क्योंकि ज़्यादातर शब्दों का उच्चारण वैसे ही किया जाता है जैसे उन्हें लिखा जा
ता है
इतालवी भाषा में उच्चारण करना तुलनात्मक रूप से आसान है क्योंकि ज़्यादातर शब्दों का उच्चारण वैसे ही किया जाता है जैसे उन्हें लिखा जा
ता है.1
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

In hindi_trascription there is an extra column so we need to remove the extra columns by using the file_editing.py code where file is asked as input and ask for the filename to which the required output of desired columns.

```
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/file_editing.py"
Enter the input TSV file name: hindi_transcription.tsv
Enter the output TSV file name: hindi_modified.tsv
Modified data saved to hindi_modified.tsv
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

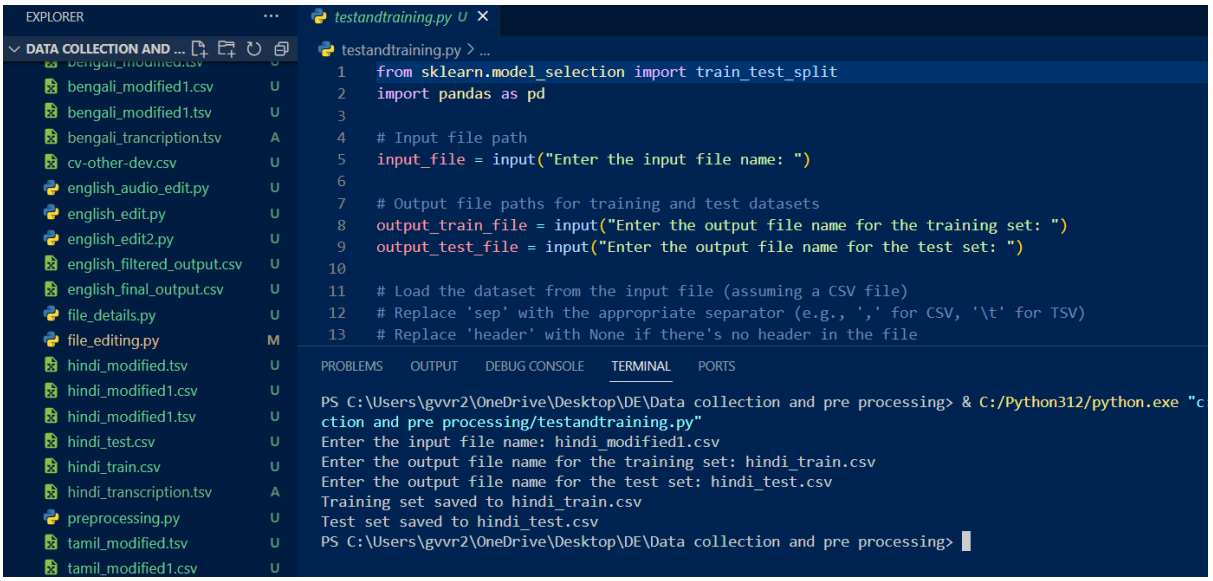
Then we need to remove all the commas,hyphens,full stops using the code preprocessing.py

```
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/preprocessing.py"
Enter the input TSV file name: hindi_modified.tsv
Enter the output file name: hindi_modified1.tsv
Data cleaning completed. Check the output file: hindi_modified1.tsv
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

Then we need to convert the file into csv format as per the model requirement.

```
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/tsv_to_csv.py"
Enter the input TSV file name: hindi_modified1.tsv
Enter the output CSV file name: hindi_modified1.csv
TSV to CSV conversion completed. Check the output file: hindi_modified1.csv
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

Then we need to divide the dataset into test and train according to the desired percentage and conditions.



Similarly the process is done for the Bengali,Telugu and Tamil datasets.and we can also separate the audio file according to the test and training files.

But for the **English** dataset the process is a little bit different.we get the english dataset with english accent of different countries.

First we need to separate the indian accents data by using the column accent where the value is indian.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/english_edit.py"
Filtered rows written to english_filtered_output.csv
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

english_filtered_output.csv

```
filename,text,up_votes,down_votes,age,gender,accent,duration
cv-other-dev/sample-000012.mp3,it was there that the wise man lived,0,0,twenties,male,indian,
cv-other-dev/sample-000013.mp3,our influence on their monopoly is tiny,0,0,twenties,male,indian,
cv-other-dev/sample-000015.mp3,holy moly you were fast on the zip line,0,0,twenties,male,indian,
cv-other-dev/sample-000046.mp3,i like cyndi wayne,0,0,fifties,female,indian,
cv-other-dev/sample-000052.mp3,remember the night we broke the windows in this old house,0,0,fifties,male,indian,
cv-other-dev/sample-000061.mp3,after all the spinning had stopped megan felt better,0,0,twenties,male,indian,
cv-other-dev/sample-000184.mp3,picture's in the paper,0,0,twenties,female,indian,
cv-other-dev/sample-000225.mp3,can i see william comes to town at kb theatres,0,0,twenties,male,indian,
cv-other-dev/sample-000253.mp3,we'll stay right here and celebrate,0,0,fifties,female,indian,
cv-other-dev/sample-000261.mp3,how can you tell,0,0,fifties,female,indian,
cv-other-dev/sample-000282.mp3,he remembered the sword,0,0,twenties,male,indian,
```

Then we need to remove the unwanted columns.so that we get filename/path and its transcription only.

english_edit2.py U | english_final_output.csv U

english_final_output.csv

```
cv-other-dev/sample-000616.mp3,earlier this year we found a really nice place near my office and we moved in together
cv-other-dev/sample-000618.mp3,just drop a module into your macrosystem directory turn your microphone off and on and it wi
cv-other-dev/sample-000619.mp3,and he gave you that five dollar raise
cv-other-dev/sample-000622.mp3,you can't do that to me
cv-other-dev/sample-000660.mp3,we don't have to give up our club
cv-other-dev/sample-000662.mp3,the wind screamed with delight and blew harder than ever
cv-other-dev/sample-000730.mp3,then she again took his hands and studied them carefully
cv-other-dev/sample-000793.mp3,you could run an interview that would prove it
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing> & C:/Python312/python.exe "c:/Users/gvvr2/OneDrive/Deskt
op/DE/Data collection and pre processing/english_edit2.py"
Columns 1 and 2 written to english_final_output.csv
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```

Then we can use that file and separate indian accents audio samples(110) from a vast number of samples(3022) to a separate folder named as the path mentioned in the csv file

EXPLORER

DATA COLLECTION AND ...

- cv-other-dev
- english_dataset
- screenshots
- source_folder
  - sample-000000.mp3 U
  - sample-000001.mp3 U
  - sample-000002.mp3 U
  - sample-000003.mp3 U
  - sample-000004.mp3 U
  - sample-000005.mp3 U
  - sample-000006.mp3 U
  - sample-000007.mp3 U
  - sample-000008.mp3 U
  - sample-000009.mp3 U
  - sample-000010.mp3 U
  - sample-000011.mp3 U
  - sample-000012.mp3 U

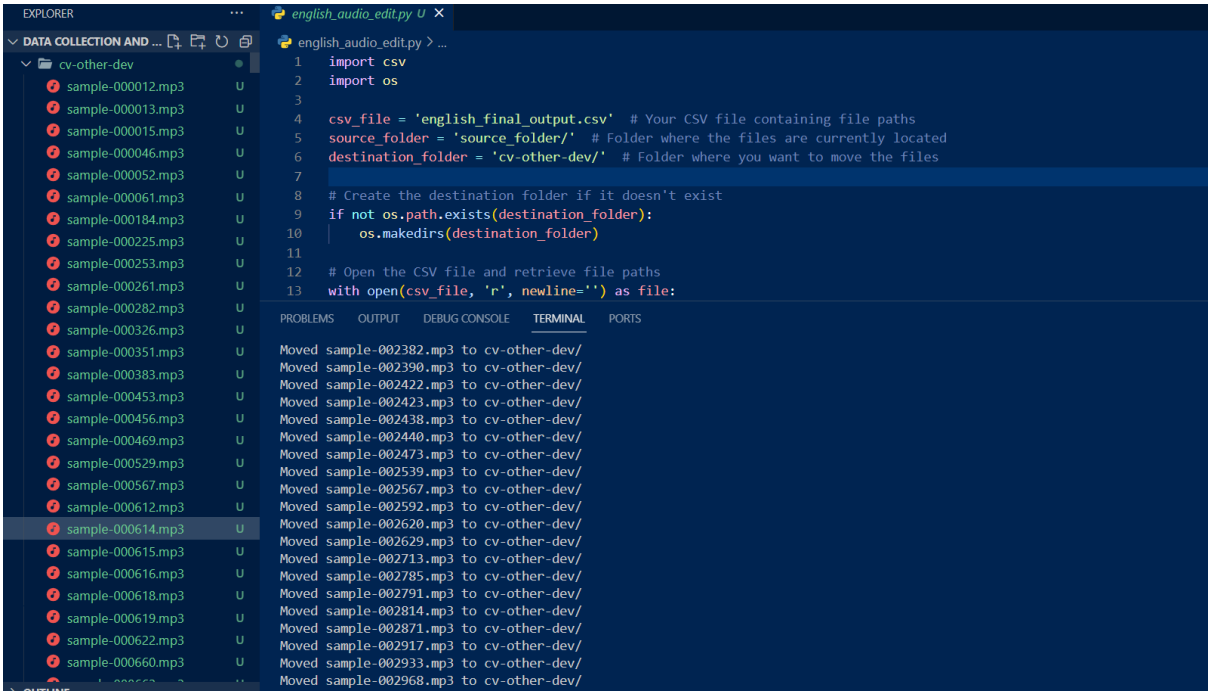
english_audio_edit.py U

english_audio_edit.py > ...

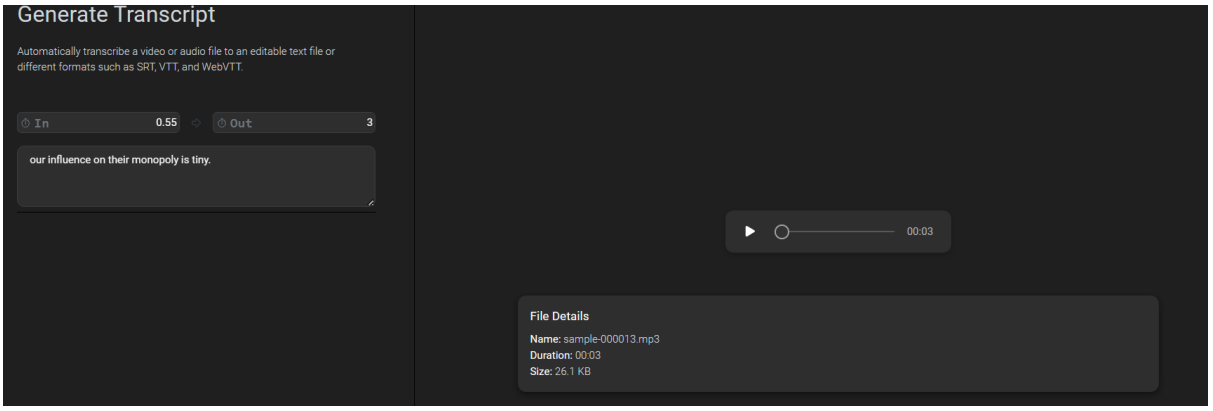
```
import csv
import os

csv_file = 'english_final_output.csv'  # Your CSV file containing file paths
source_folder = 'source_folder/'  # Folder where the files are currently located
destination_folder = 'cv-other-dev/'  # Folder where you want to move the files

# Create the destination folder if it doesn't exist
if not os.path.exists(destination_folder):
    os.makedirs(destination_folder)

# Open the CSV file and retrieve file paths
with open(csv_file, 'r', newline='') as file:
```

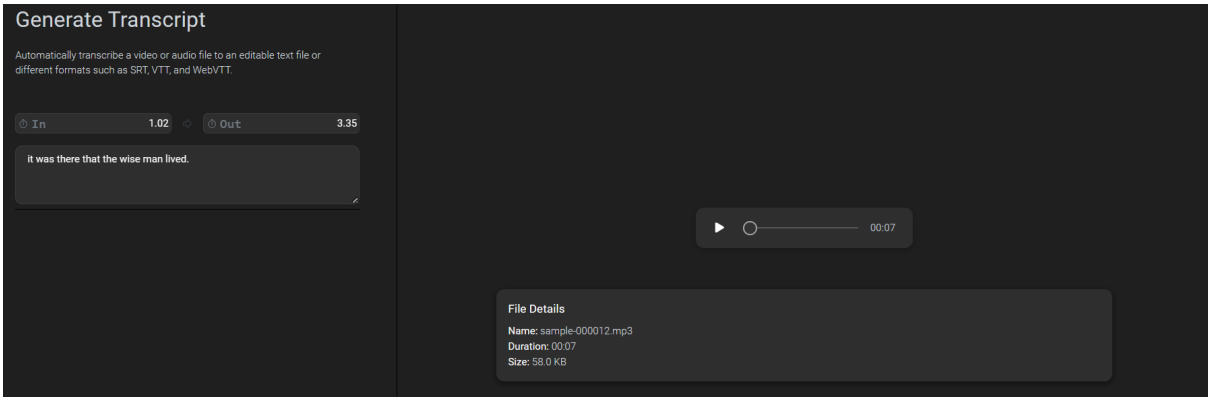
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\gvvr2\OneDrive\Desktop\DE\Data collection and pre processing>
```



And for the dataset the verification process for audio to be transcripted i used generate transcript.





We can see from above screenshots that both audio sample and transcript are matching.