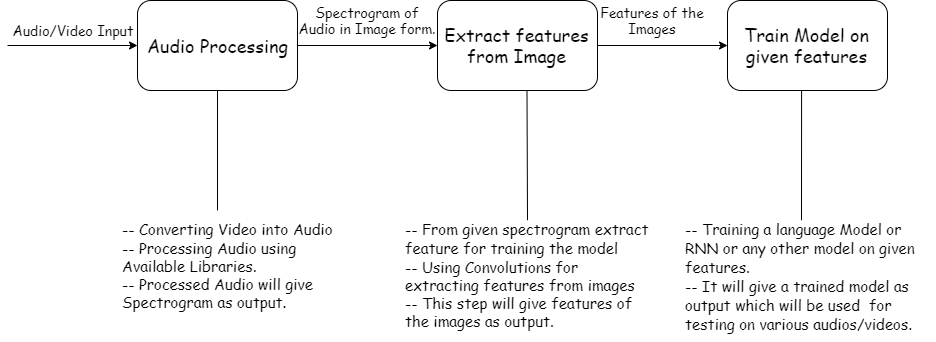
Subtitle Generator Prototype



Abstract:

This software takes the video or audio as input and generates subtitles or transcript respectively for the same. First the video/ audio file is processed using audio processing tools. The output of this function is spectrogram in form of Image. Second the processed audio is used to extract features using convolutions on the given input image.

This step will give features of the respective images. Now these extracted images will be used to train a model such as language model or RNN or any other combination of model for end-to-end speech to text recognition. This step will give trained model as an output which will be used for further testing on various other audios and videos.

After testing the model the given model will be used to give the transcript / subtitles for the given audio which will be synchronized with the given audio and converted into .txt for transcript and .srt file for subtitle.

Dataset:

We used [Tatoeba](https://tatoeba.org/eng/) dataset for prototype model building.

Tatoeba is a large database of sentences, translations, and spoken audio for use in language learning.

This dataset contains 1,265,664 sentences in English with labels of average length 3 seconds recording.

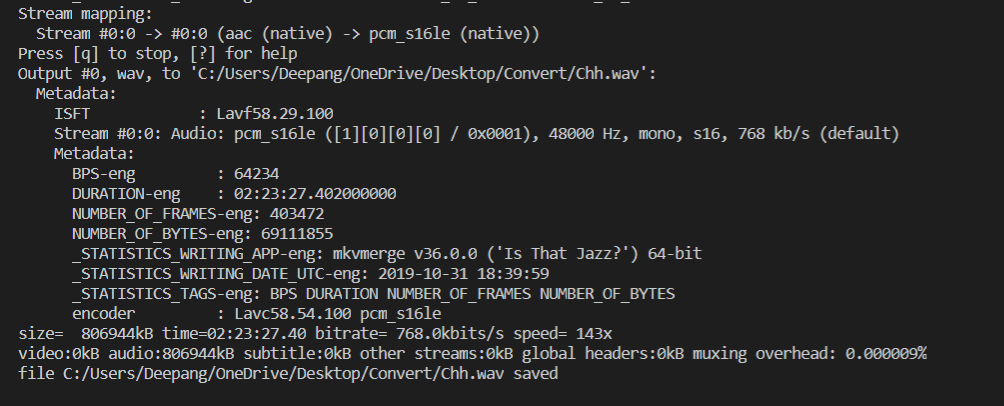
The total size of this dataset is 3.8 GB.

Field Structure for labels:

* Sentence id [tab] Lang [tab] Text

Video to Audio Processing:

For video to audio conversion we used [FFMEG](https://www.ffmpeg.org/) library with python sub process command to achieve good quality audio file in less amount of time from videos. The output format is .wav file.



Now for the given video file of size 1.08 GB and length of 150 minutes, it takes FFMPEG about 1 min to generate its wav file which is of size 800 MB.

This audio will then be used for audio processing.

Audio Processing:

We performed audio processing on the Tatoeba dataset using the [librosa](https://librosa.github.io/librosa/) python library which gives extensive features for working with audio and images.

We plotted different spectrogram to analyze which spectrogram can be used for our images.

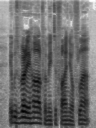
1. Log-Spectrogram –

It is a spectrogram which is having log scale of frequency as its y-axis [log scale of amplitudes.] and time as its x-axis.

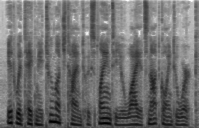
1. Mel-Spectrogram –

It is based on the Mel Scale, mathematically speaking, is the result of some non-linear transformation of the frequency scale. Mel Spectrogram, is, rather surprisingly, a Spectrogram with the Mel Scale as its y axis.

Some saved Images of audios from Tatoeba dataset:



“It was very hard for me to find your flat.”



“It would take me too much time to explain to you why it's not going to work.”

We used Google Colab as our working environment for audio processing which supports high end calculations using the support of GPU.