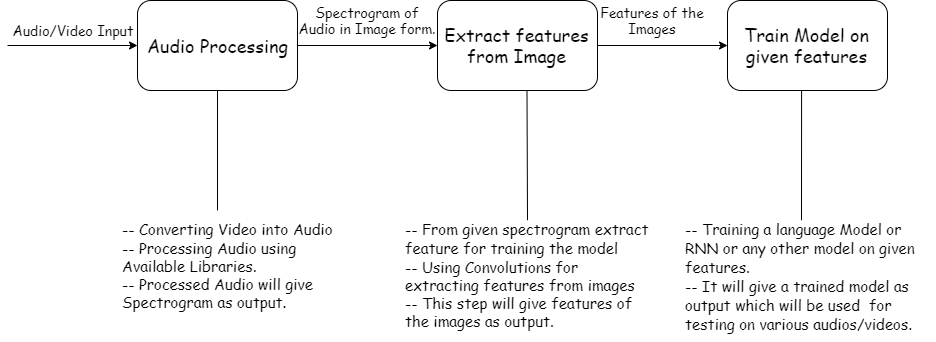
Subtitle Generator Prototype



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.

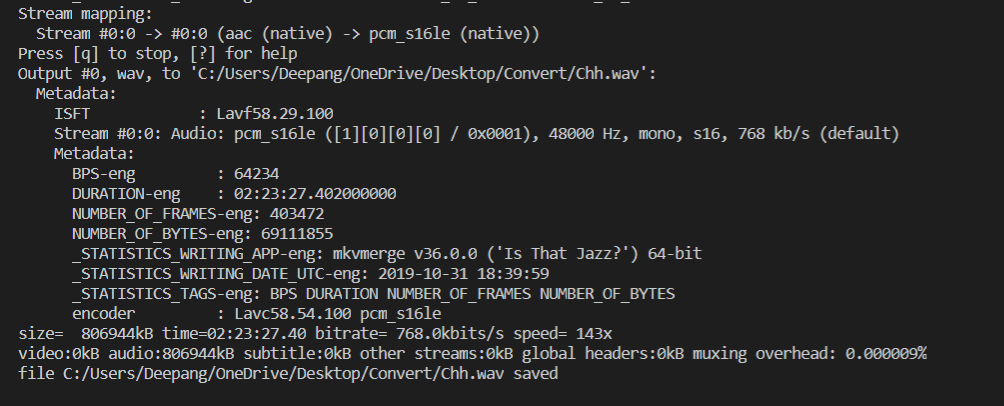
The total size of this dataset is 3.8 GB.

Field Structure for labels:

* Sentence id [tab] Lang [tab] Text

Video to Audio Processing:0

For video to audio conversion we used [FFMPEG](https://www.ffmpeg.org/) library with python sub process command to achieve good quality audio file in less amount of time from videos. The ou0t0000000put 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 –

Mel Scale -

It is based on the Mel Scale, mathematically speaking, is the result of some non-linear transformation of the frequency scale.

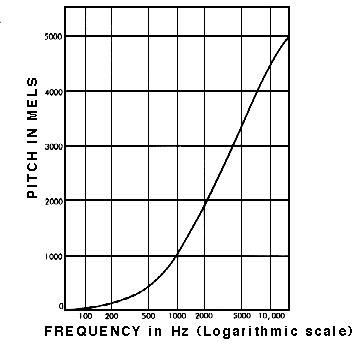
Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another.

The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels.

Above about 500 Hz, increasingly large intervals are judged by listeners to produce equal pitch increments.

Which basically means that it divides any range above 1 kHz into equal bins which are audible to the human ears.

Mel Spectrogram, is, rather surprisingly, a Spectrogram with the Mel Scale as its y axis.



As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the mel scale.

Divided not by distance on the frequency dimension, but distance as it is heard by the human ear.

Cepstrum –

The cepstrum can be seen as information about the rate of change in the different spectrum bands.

Power spectra of windows of speech signals contain information about the most important features of speech signals like the identity of vowels.

Cepstrum pitch determination is particularly effective because the effects of the vocal excitation (pitch) and vocal tract (formants) are additive in the logarithm of the power spectrum and thus clearly separate.

A second useful piece of information in the cepstrum is the harmonic structure of the log-spectrum.

Mel-Frequency Cepstral Coefficients (MFCCs) –

To further improve on the cepstral representation, we can include more information about auditory perception into the model. Specifically, by introducing information about human perception, we focus the model on that part of the information which human listeners would find important.

The log-spectrum already takes into account perceptual sensitivity on the magnitude axis, by expressing magnitudes on the logarithmic-axis. The other dimension is then the frequency axis.

The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

MFCCs are commonly derived as follows:

Take the Fourier transform of (a windowed excerpt of) a signal.

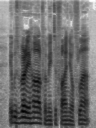
Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.

Take the logs of the powers at each of the mel frequencies.

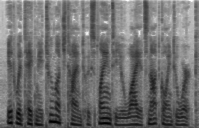
Take the discrete cosine transform of the list of mel log powers, as if it were a signal.

The MFCCs are the amplitudes of the resulting spectrum.

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.