Report: This must include a description of how you arrived at your final model, the way you handled silence, the performance you observed, and a plot of your model. Provide other information such as the hyperparameters and preprocessing method you employed (if any). Provide other useful information as well.

Before training, I choose three videos from Hindi, English and mandarin respectively, as test data. I created three labels, [1,0,0], [0,1,0], [0,0,1] for three languages. To remove the silence in different languages, I use librosa.effects.trim() after load video. I extracted 64 features from each 10ms in a video. Because the shortest video is only 8min 30 sed, I keep all videos as 8min 30sed. Then I reshape the video data as a three-dimensional array. The first dimension means sequence number, there are 98 sequence in each video, the sequence length is 600. Every one video was processed, first concatenate it with its label, then concatenate the feature sequence and the labels together to the whole data set. So shuffle the dataset won't mess up the feature sequence and its corresponding labels. Finanlly, there are 23030 sequences.

During training, I tried to use different number of dense layers(1,2,3) and different number of units(4,16,24,32) in RRN network. Finally, I use 24 units and GRU network. For dense layers, I use softmax activation for the last layer. I also tried use regularization and dropout, but there is no significant improvement, so, the final model doesn't include dropout. Cause I use one hot label with length 3, the loss function should be Category cross entropy.

Finally, for sequence model, I got about 60% in training accuracy. And the test accuracy is 61.23% For stream model, the test accuracy is 55.69%.

Change the 'path' to the local video location. Then run test_streaming_model.py, the language variable is the predict language.