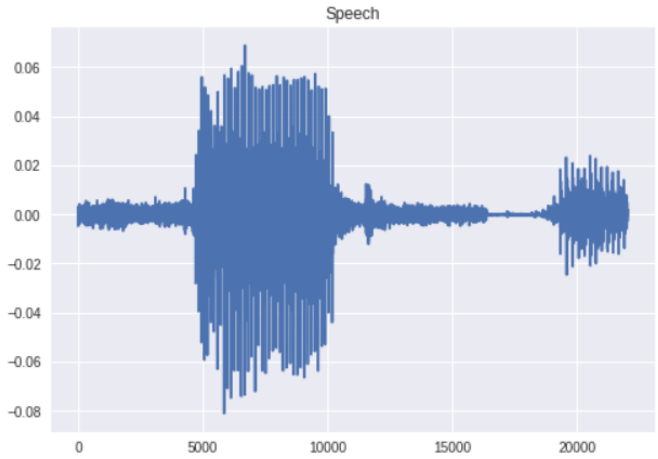
Voice Activation Detection in noisy environment

By,

SANDESH GANGADHAR
D17129109
MSc in Computing (Data Analytics)

Concept

- Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected.



Voice activity detection. (2019). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Voice_activity_detection&oldid=882571384

Problem

- Implementation in real world scenario with the presence of additive noise.
- Detection of segments of audio file where the,

  Signal = Voice + Noise
- Simulation of real world scenario,

  Signal = Voice + Exercise_Bike_Noise

Voice Activity Detection in Noise Using Deep Learning - MATLAB & Simulink - MathWorks India. (n.d.). Retrieved April 9, 2019, from https://in.mathworks.com/help/audio/examples/voice-activity-detection-in-noise-using-deep-learning.html#d117e11310

Dataset

- Speech commands dataset by Google AI blog posted by Pete Warden, Software Engineer, Google Brain Team.

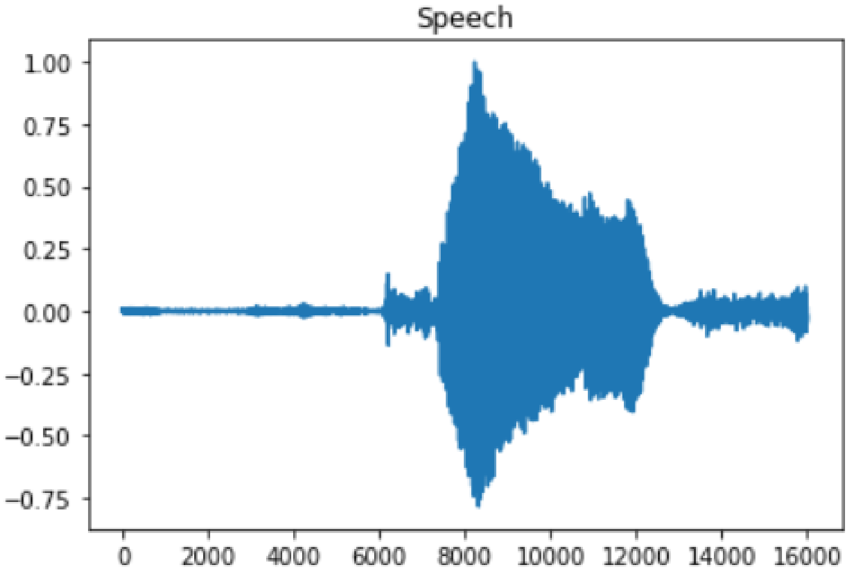
https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

The dataset has **65,000 one-second long** utterances of **30 short words**, by **thousands of different people**, contributed by members of the public through the AIY website.

https://aiyprojects.withgoogle.com/open_speech_recording

Data pre-processing

One second audio files

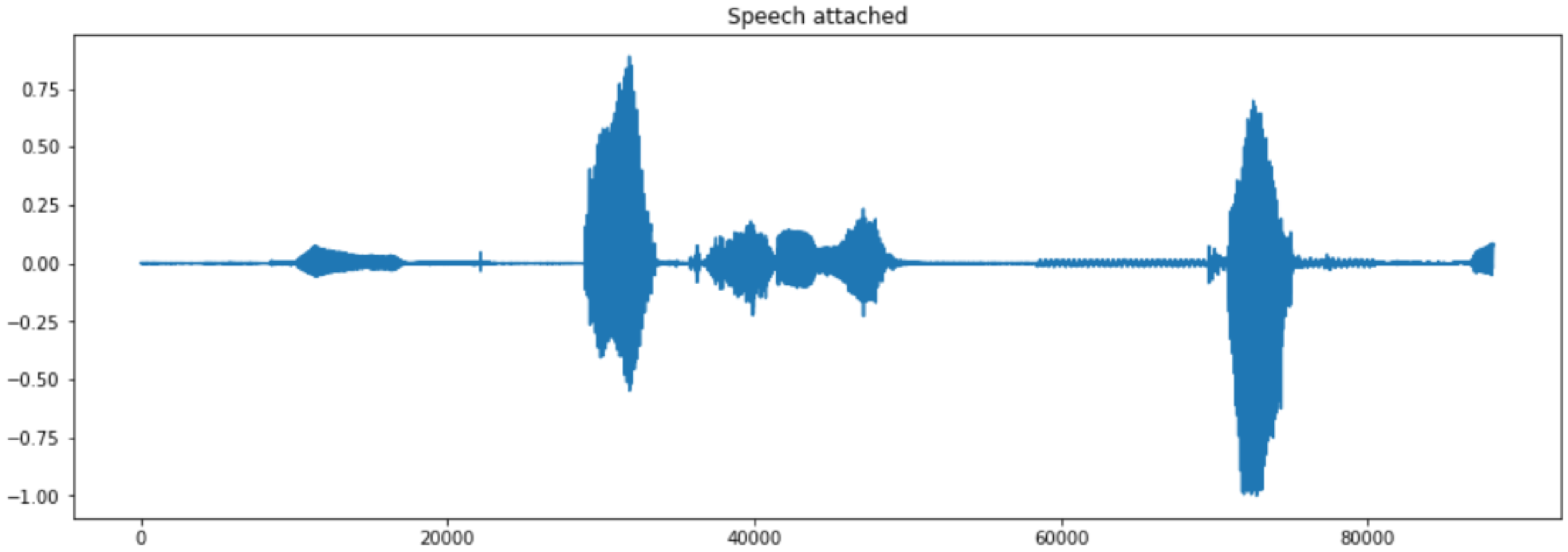


Tree

fs = 22050
Frame_size = 40ms = 0.04 x 22050 = 882
Frame_overlap = Frame_size / 2 = 441
Number_of_Files = 65,000

Data pre-processing

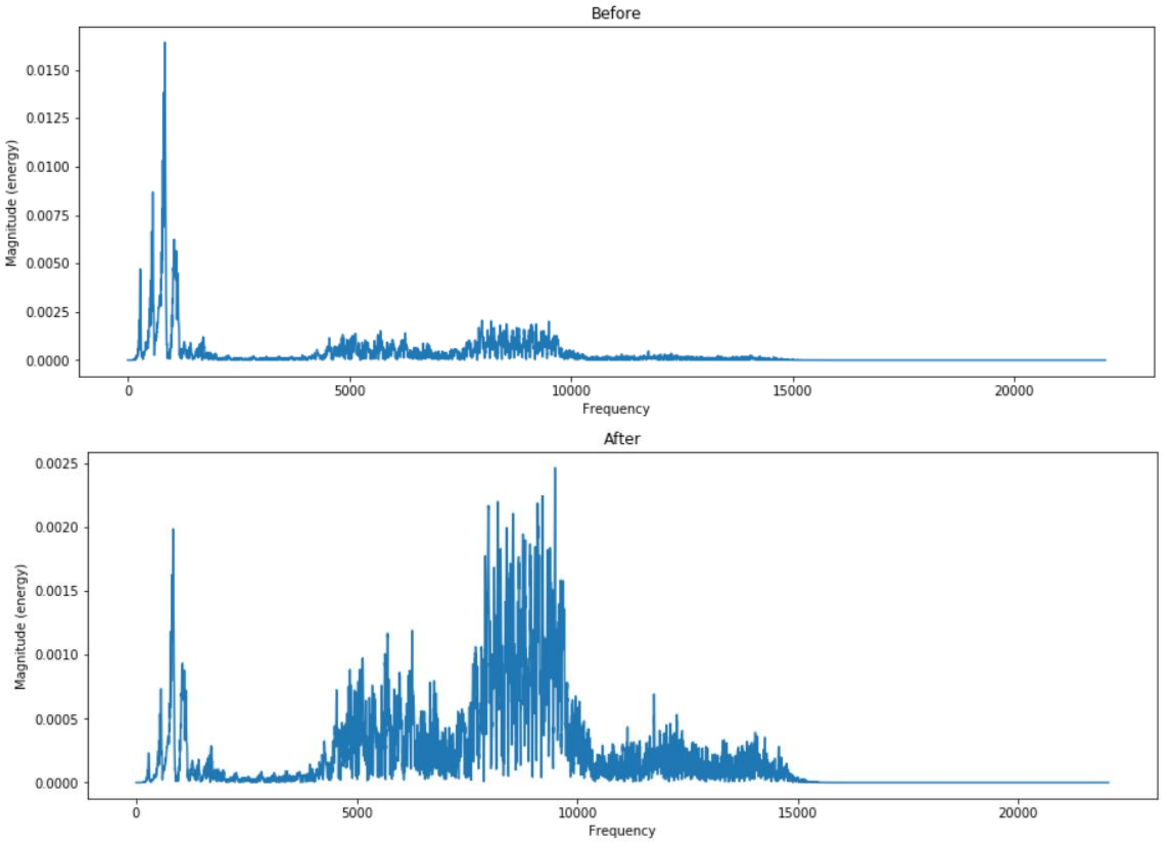
One long audio file



Duration of the attached files depends on the sample size chosen.

Data pre-processing

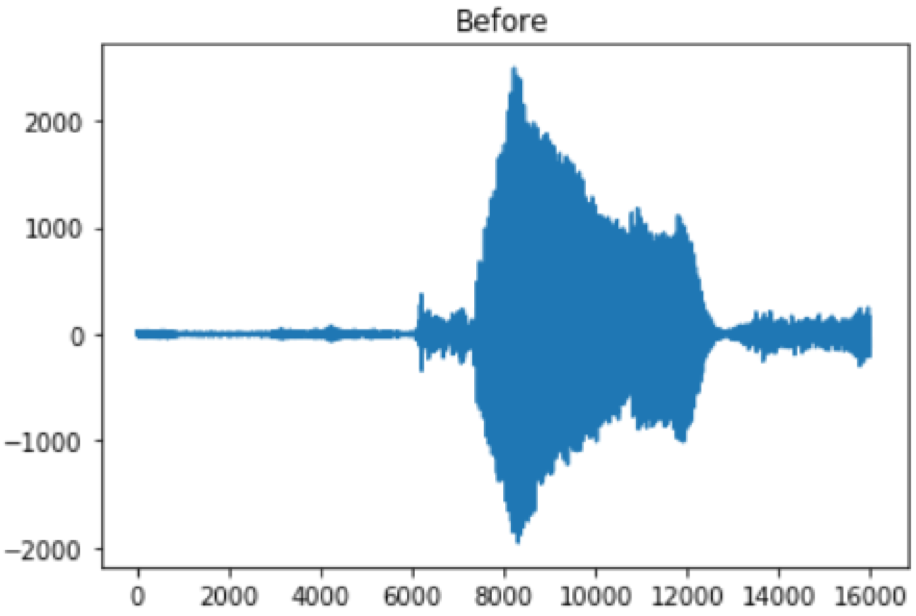
Apply pre-emphasis

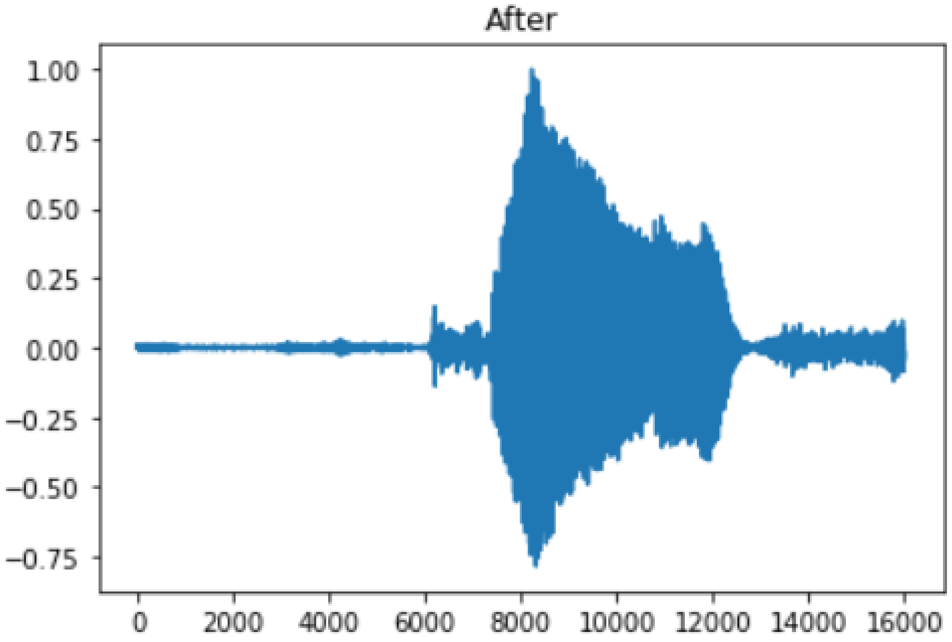


Vergin, R & O'Shaughnessy, D. (1995). Pre-emphasis and speech recognition. 2. 1062 - 1065 vol.2. 10.1109/CCECE.1995.526613.

Data pre-processing

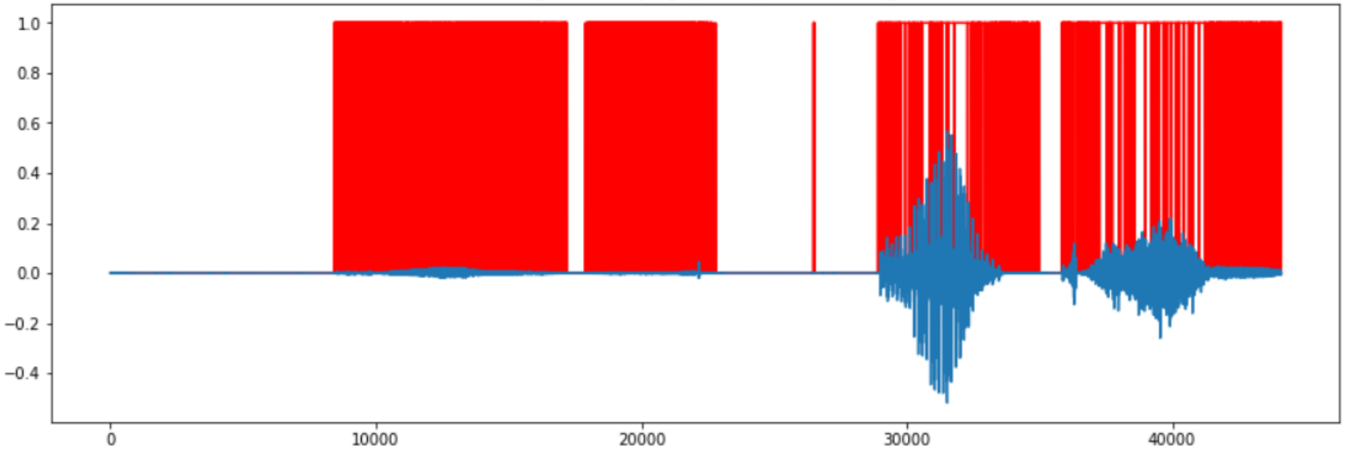
Normalization

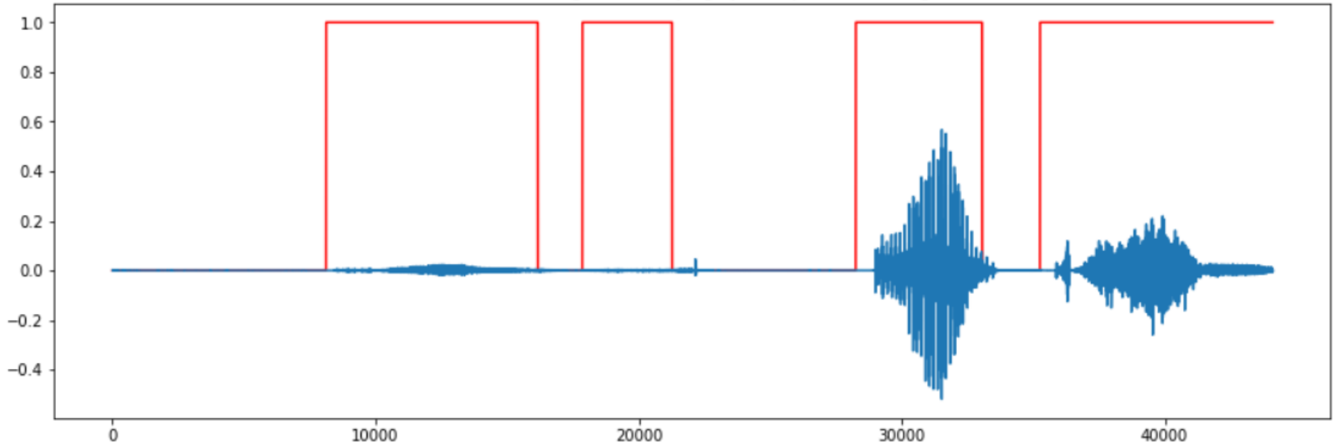

Before


After

Data pre-processing

Detect voice in speech


Mask signal identifying speech parts of the signal

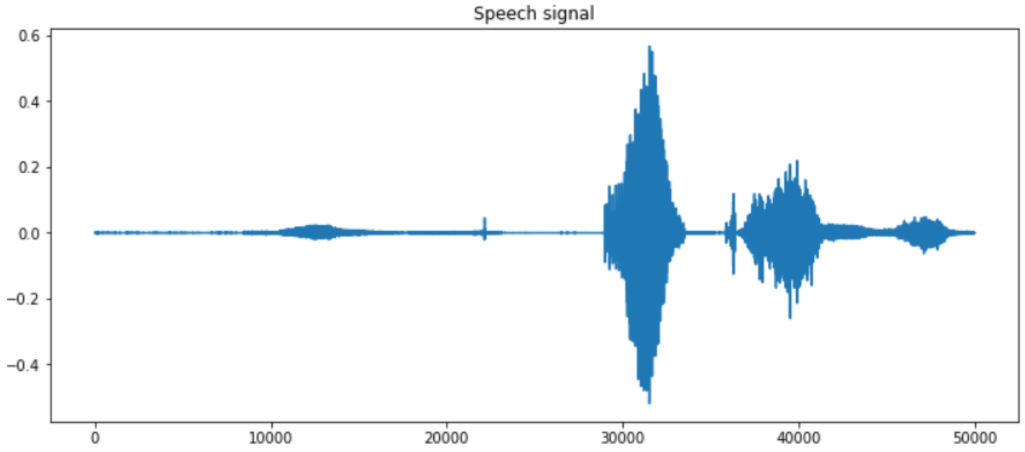

Smooth mask signal identifying speech parts of the signal

Data pre-processing

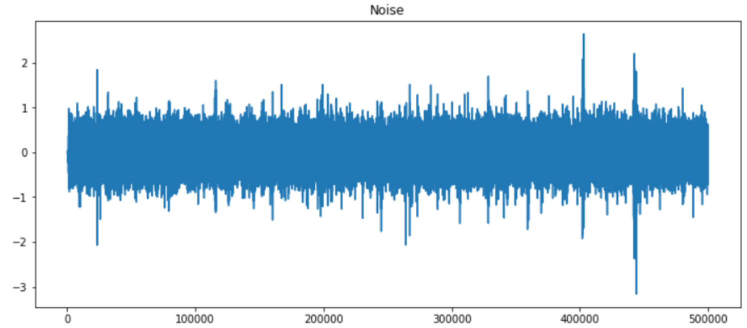
Process to create the mask using thresholding technique:

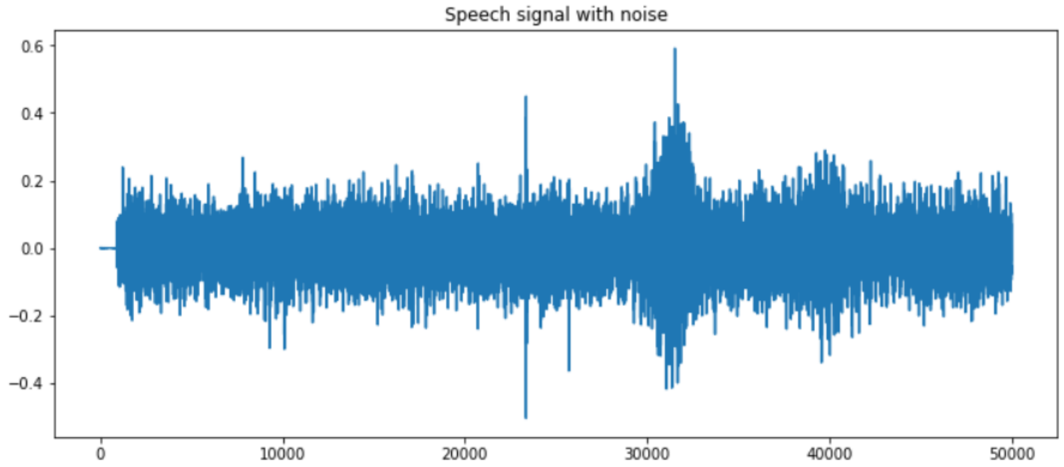
- Frame the signal.
- Calculate median energy of the frame.
- Calculate the centroid of the frequency for the frame.
- Create a dummy mask signal with length == frame length.
- Make mask signal = 1 if signal energy at the point within frame is greater than median and the centroid frequency of the frame < 5000 Hz.
- Apply smoothing for the signal using 1500 point moving average technique. (Was acceptable)
- Multiplication of Mask signal and speech signal yield signal with noise removed during silence in the signal.

Data pre-processing

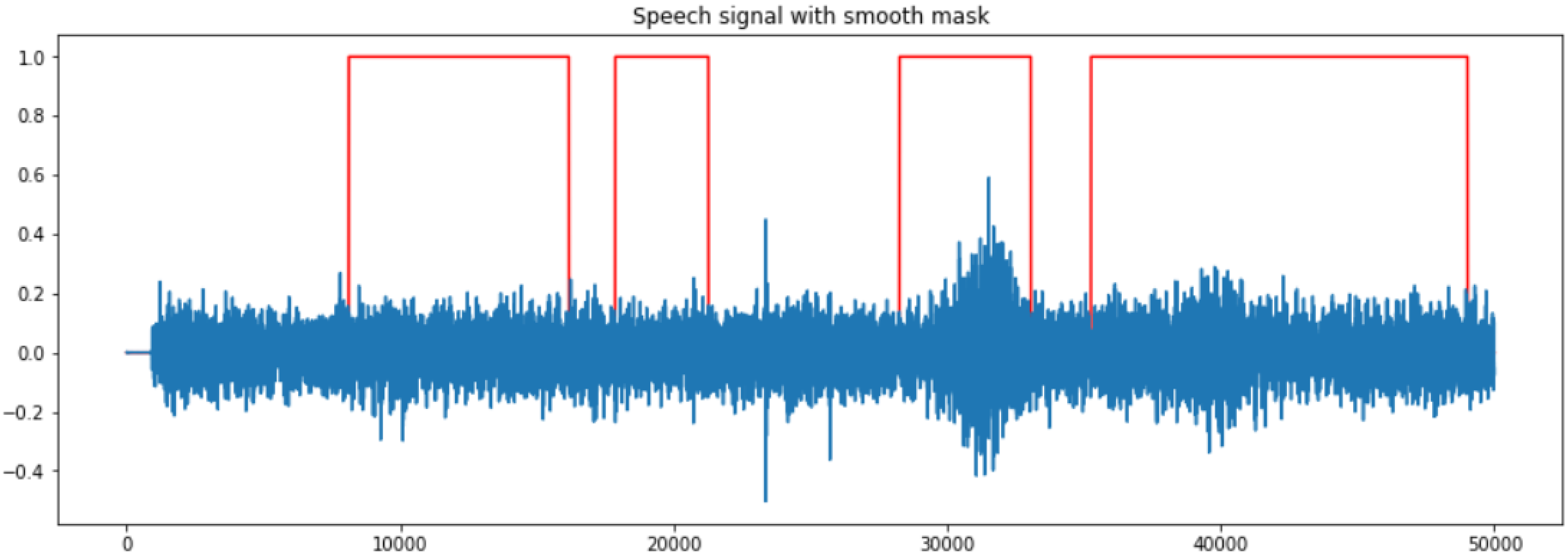


Add noise

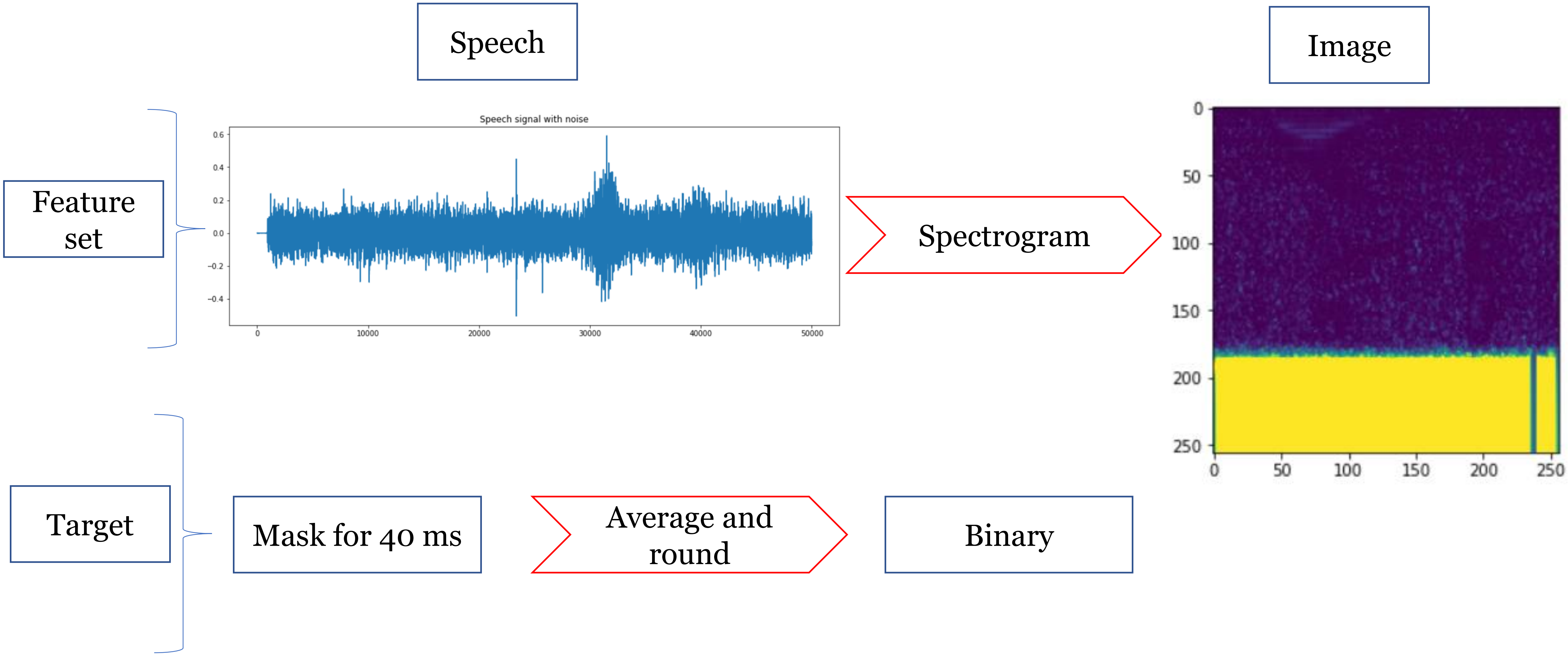




Data pre-processing



Data pre-processing



Speech

Image

Feature set

Spectrogram

Target

Mask for 40 ms

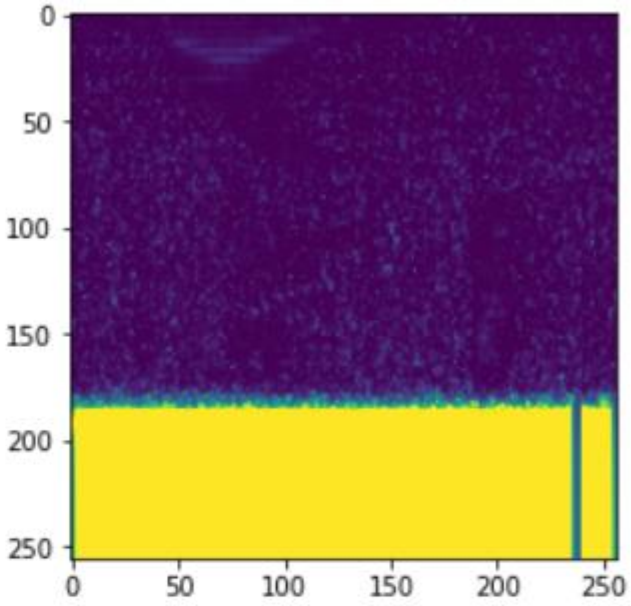
Average and round

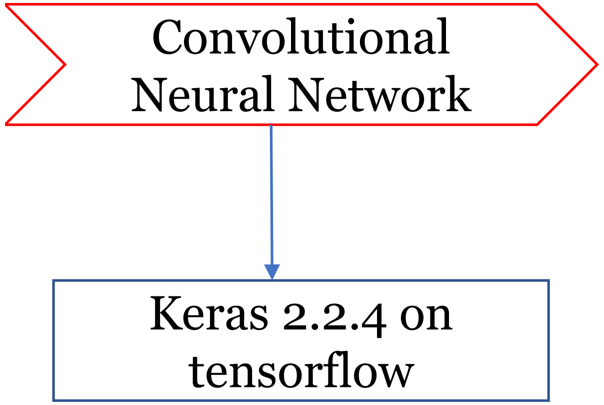
Binary

Model development

Apply CNN for classification of a frame to identify as voice activity or no-voice activity.

Image

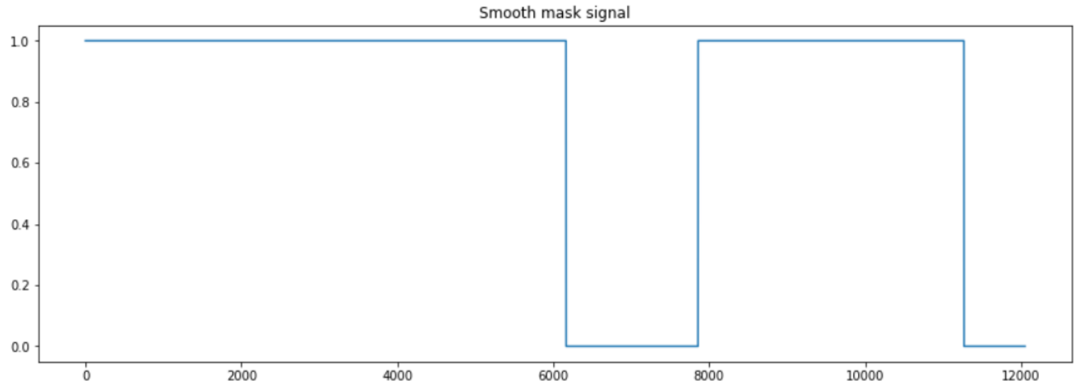




Convolutional Neural Network

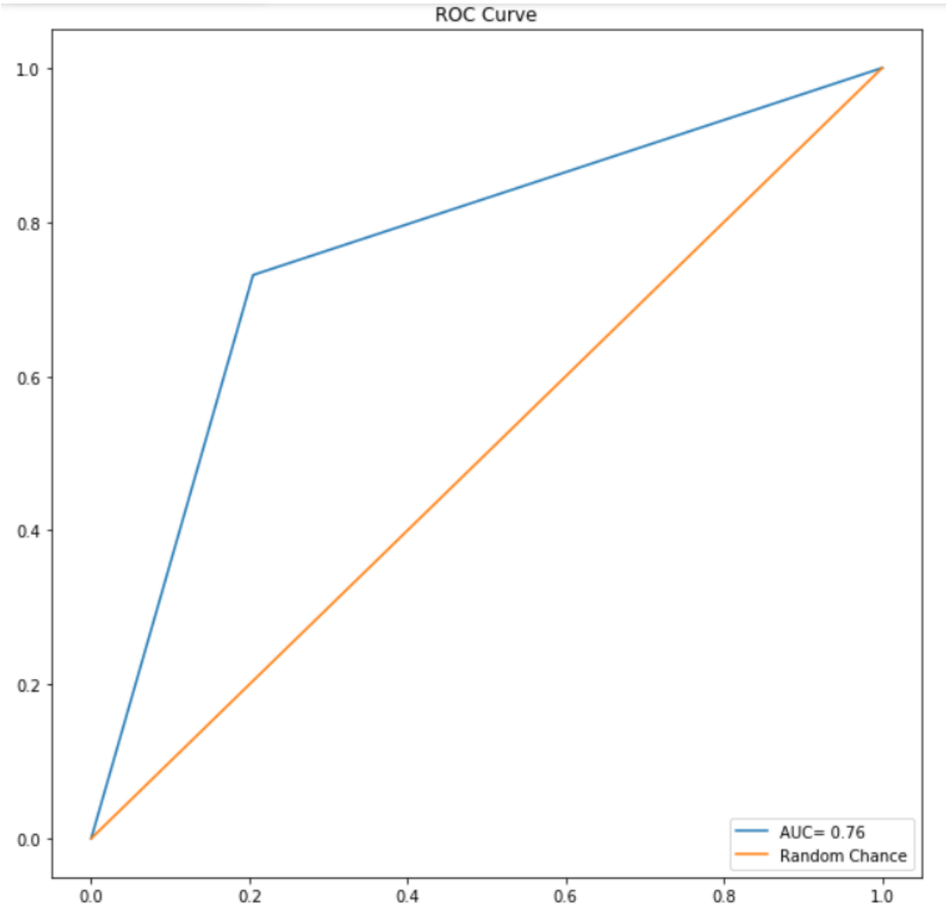
Keras 2.2.4 on tensorflow

Binary / Mask signal

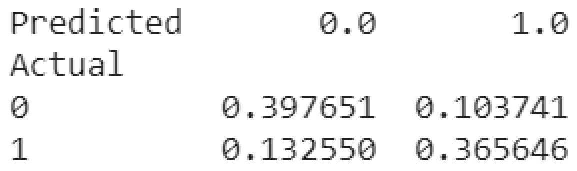


Model Evaluation

Plot of ROC curve



Confusion matrix in proportions



```
Predicted       0.0       1.0
Actual
0          0.397651  0.103741
1          0.132550  0.365646
```

Area Under the Curve achieved

76%

Let’s see it in action.

Q & A

THANK YOU