Usama Khalid
i19-1236

Assignment 4
Natural Language Processing

2020-11-22

# 1 Speech Detection

The task of speech detection is to identify whether a given segment of audio has a person speaking or is just silence and noise. For speech detection we use noise thresholding to classify a given segment of audio to speech or non speech. Noise is a good parameter for this as thresholding noise will automatically also remove silence. To classify an audio as speech two steps are performed. Firstly a threshold value is calculated from the given audio signal and using that threshold if the amount of noise in an audio is greater than a tolerance amount then it is classified as not speech. This is explained in detailed using the equations below.

To calculate the noise threshold the mean of the audio signal is calculated first using equation 1. $\overline{S}$ is the mean amplitude while $S_t$ is the audio signal at time $t$ and $n$ is the length of the audio signal. The noise threshold $T$ is calculated using equation 2 where $N_h$ is the noise threshold. The classification whether an audio segment is speech is calculated using equations 3 & 4. Equation 3 calculates the amount of noise in the audio segment by aggregating if a given signal value at time $t$ is greater than the noise threshold $T$. The Equation 4 outputs $0$ if the amount of noise $N$ is less than the noise tolerance amount $n * N_r$ of the audio where $N_r$ is the noise tolerance percentage and $n$ is the length of the audio.

$$\overline{S} = \frac{1}{n}\sum_{t=1}^{n} |S_t| \tag{1}$$

$$T = \overline{S} * N_h \tag{2}$$

$$N = \sum_{t=1}^{n} \begin{cases} 0 & |S_t| \leq T \\ 1 & |S_t| > T \end{cases} \tag{3}$$

$$Y = \begin{cases} 0 & N \leq n * N_r \\ 1 & N > n * N_r \end{cases} \tag{4}$$

# 2 Speech Segmentation

Given an audio signal the speech segmentation divides the signal into equal chunks of 1 second each. The sampling rate taken when reading an audio is 16000 hz. Therefore each chunk consists of 16000 values of an audio signal. If the chunk remaining at the end is less than 1 second it is padded with zeros until there are exactly 16000 values.

# 3 Embedding Extraction

To extract embeddings for an audio file, firstly the audio is segmented into chunks of one second each as described in section 2. Each segment is then classified as speech or not speech using the steps explained in section 1. For each segment that is classified as speech we extract a list of features for that segment which are chroma short-time-Fourier-transform, root-mean-square, spectral-centroid, spectral-bandwidth, spectral-rolloff, zero-crossing-rate and MFCC. These features are concatenated in a single feature vector.

# 4 Clustering

The fixed sized embeddings extracted for each audio segment are clustered using Kmeans. The hyper-parameter K is selected using the elbow method by calculating the amount of distortion for each K. The total distortion is calculated for each audio file and value of K. The values of K are varied from 1 to 10 and the value of K for which the greatest distortion is seen is picked. An example graph for the distortions for multiple values of K is shown in Figure 1.

The total distortion for a value of $K$ is calculated by aggregating all the euclidean distances between embeddings and cluster centers and then dividing by the number of embeddings. This is given by equation 5 where $D_K$ is the total distortion for a given $K$, $E_t$ is the embedding for segment $t$ and $C_k$ is cluster $k$ for a given number of clusters specified by $K$.

$$D_k = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{r}(E_t - C_k)^2 \tag{5}$$

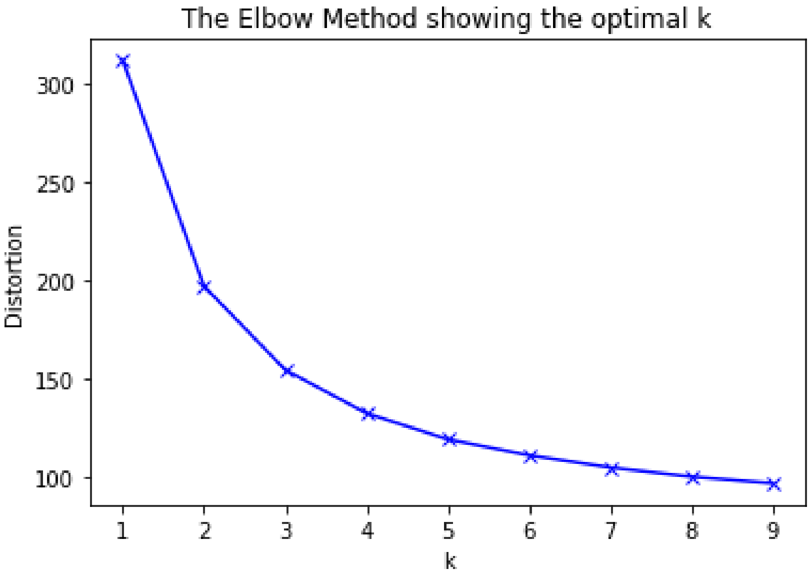

Figure 1: A graph showing the amounts of distortions for different values of k.

# 5 Diarisation Graphs

For better visual analysis the signal of an audio is plotted using matplolib with time in seconds on the x-axis and amplitude of the signal on the y-axis. Each speaker is assigned a colour and the region(s) where the speaker has spoken is highlighted with their colour on the graph. An example graph is shown in Figure 2.

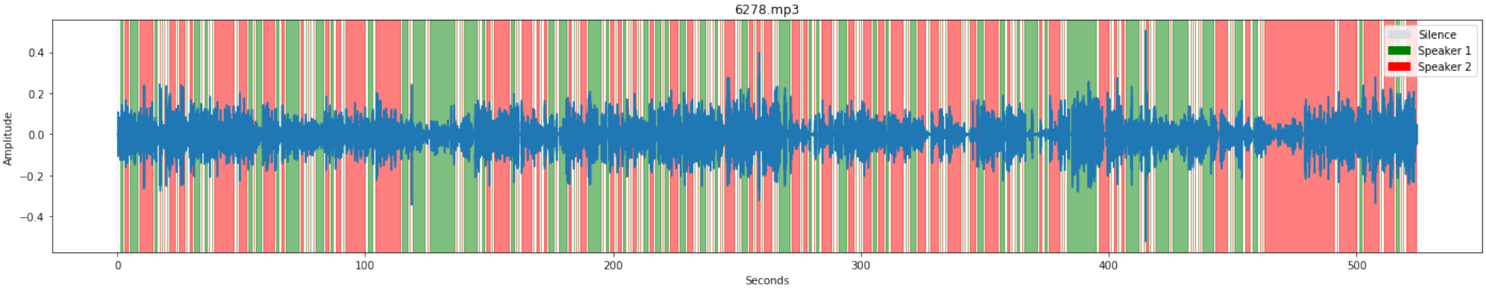

Figure 2: The graph of an audio signal. The white region depicts silence or where no one has spoken while the regions highlighted in green correspond to speaker 1 and the regions in red correspond to speaker 2.