In the development of a speech emotion analyzer, preprocessing is the crucial phase to tune the audio data to create feature vectors used as an input to the model. The audio data obtained from the raw dataset has non-uniform dimensions varying in audio length. This raw data is sampled to have uniform standardized dimensions by adding zero paddings to shorter audio clips. Furthermore, it is checked for the signal-to-noise ratio to make sure there is less ambient distorted noise in the data.

Mel Spectrograms are generated using tuned data clips and are converted to mel scales so that the perpetual difference becomes smaller to deal with human frequencies. Furthermore, the log-mel spectrogram, given as input to the model, will be normalized by the mean and variance of the training set. Moreover, we may compress the mel spectrogram into a short-time spectrum to extract only the essential coefficients, called Mel Frequency Cepstral Coefficients (MFCC), which only corresponds to human frequency ranges.