# Feature Selection

## Main process

1. We divide the variables in 8 main groups that were given in the csv data file:
   * Baseline Features
   * Intensity Parameters
   * Formant Frequencies
   * Bandwidth Parameters
   * Vocal Fold
   * MFCC
   * Wavelet Features
   * TQWT
2. For some groups we are able to get the correlation immediately.
3. However, some groups are very large and so they will be splitted again based on patterns on the variable’s names, like “mean” or “std”. Then, we will check the variable’s correlation in these subgroups.
4. After having all the simple correlations, we identify groups of variables that are more correlated and, calculate the correlation between the groups and a new variable that is given by the mean of the variables in the group.
5. After having all the groups correlations, we will create several datasets according to the ratio of correlation between all the group’s variables. For instance, if we define the minimum ratio for removal as 0.95, then in a group:
   * All variables that don’t have a ratio equal or higher than 0.95 with any variable will be kept.
   * If there is a set of variables that are correlated with each other, we either replace this set by its mean or pick one variable from the set that fits best, depending on the decision that has a better average correlation.

We are now trying to find correlations between gender and all the groups.