Early Biomarkers of Parkinson's Disease Based on Natural Connected Speech

Anja Probst[1]

[1] University of Geneva

Abstract

Parkinson's Disease is a degenerative disorder of the nervous system that globally affects more than 6 million people (doi:10.1016/S0140-6736(16)31678-6). While the most well-recognized symptoms of the disease are motor-related, such as shaking and instability, a further group of symptoms, which is only partially motor-related and occurs in a majority of patients, are speech-altering symptoms (https://pubmed.ncbi.nlm.nih.gov/30223711/). While the disease is well-recognizable at a later stage, it is exceptionally hard to diagnose and differentiate in its early stages and appropriate treatment is often delayed. In 2017, Hlavnička et al. have published a study suggesting that automated analysis of connected speech can reveal early biomarkers in subjects with REM sleep behaviour disorder, who are at high risk of developing Parkinson's disease. In this project I analyse the data set published by the authors that contains experimental evaluation of healthy controls (HC, $n = 50$), subjects with REM sleep behaviour disorder (RBD, $n = 50$), and subjects with Parkinson's Disease (PD, $n = 30$). While the constraints of this project limit the scope of analysis, I will show that interesting insights into the data can be gained nontheless.

Keywords: keywords
Word count: X

The authors made the following contributions. Anja Probst: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Anja Probst, 24 rue du Général-Dufour, 1211 Genève 4. E-mail: anja.probst@etu.unige.ch

Introduction

Context of the Project

Patients with the neurodegenerative disease Parkinson's have numerous symptoms ranging from cognitive impairments to motor symptoms. Those symptoms may appear relatively late in the disease when the neurodegeneration has already widely spread in different areas of the brain (mainly Basal Ganglia). Main symptoms of PD are motor dysfunctions including abnormalities in the production and sound of speech of such patients (up to 90%). These abnormalities in speech and voice are called hypokinetic dysarthria which is characterized by a decreased quality of the speech, where the voice, sound formation as well as the articulation is impaired. As I mentioned before, often motor impairments are detected relatively late in the disease. To improve diagnostics and to detect the disease in a much earlier stage, the detection of biomarkers related to neurodegeneration could lead to a better prognosis and therapy of PD.

Therefore, the investigation of prodromal speech changes could be an appropriate and suitable approach. To investigate this approach, an automated speech monitoring system was developed, that uses a segmentation method for the precise estimation of voiced and unvoiced segments of speech, respirations, and pauses. Further proposed was a set of acoustic speech features based on the segmentation algorithm applicable to connected speech, allowing the description of complex vocal disturbances due to neurodegeneration including respiratory deficits, dysphonia, imprecise articulation, and dysrhythmia.

In this data analysis project, the main focus was to explore, if there are any speech patterns that support the usage of an automated speech monitoring system to detect prodromal parkinsonian neurodegeneration based on natural connected speech.

130 subjects were tested. 30 subjects with early, untreated Parkinson's disease (PD) where the disease is already manifested. 50 subjects with REM sleep behaviour disorder (RBD), which is a disease where its relatively likely to develop PD in a later phase. As a control group, 50 healthy subjects (HD) were included.

Manual Variable Selection

Due to the constraints of this project, I reduced the data set from originally 62 variables to the best fitting 7. As I am looking specificially into the aspect of speech, and to evaluate if speech is a good predictor for PD, I chose speech related variables that were assessed empirically and were reported to have the most significant differences between healthy controls and subjects with early stages of Parkinson's Disease. Note that patient group will be extracted from the variable Participant_code. The resulting data set is summarized in Table 1

```
cols.to.keep <- c(
    "Participant_code", "Age", "Gender", "Rate_of_speech_timing",
    "Rate_of_speech_timing.1", "Duration_of_pause_intervals",
```

```
    "Duration_of_pause_intervals.1"
)

# Above columns will be renamed to
rename.cols.to <- c(
    "Participant_code", "Age", "Gender", "Speech.Timing.Rate.Reading",
    "Speech.Timing.Rate.Monologue", "Pause.Interval.Duration.Reading",
    "Pause.Interval.Duration.Monologue"
)

csv.path <- "BiomarkersPD.csv"
df <- read.csv(csv.path, sep = ",", header = TRUE)

# Only keep required columns and rename them
df <- df[cols.to.keep]
colnames(df) <- rename.cols.to

# Replace "-" with NA
df[df == "-"] <- NA

# Get groups from participant codes by replacing numerical values
df$Group <- gsub("[[:digit:]]+", "", df$Participant_code)

# Participant codes no longer required, remove
df <- subset(df, select = -c(Participant_code))

# Convert columns to factors
col.names <- c("Group", "Gender")
df[col.names] <- lapply(df[col.names], as.factor)
```

Data Description

For each sample in this data set ($n = 130$), we have the following information:

- Demographic information:
  - Age (years)
  - Gender (M for male, F for female)
- Speech examination - Speaking task of reading passage: speakers read a standardized, phonetically-balanced text of 80 words twice
  - Duration_Of_Pause_Intervals_Reading: Duration of pause intervals (DPI) describes the quality of speech timing, as pauses can be heavily influenced by the ability to properly initiate speech, it is measured in miliseconds (ms)

  - Rate_Of_Speech_Timing_Reading: Rate of speech time (RST) includes voiced, unvoiced and pause intervals, it is measured in intervals per minute (-/min)

- Speech examination - Speaking task of monologue: participants were instructed to provide monologue about their interests, job, family or current activities for approximately 90 seconds
  - Duration_Of_Pause_Intervals_Monologue: Duration of pause intervals (DPI) describes the quality of speech timing, as pauses can be heavily influenced by the ability to properly initiate speech, it is measured in milliseconds (ms)
  - Rate_Of_Speech_Timing_Monologue: Rate of speech time (RST) includes voiced, unvoiced and pause intervals, it is measured in intervals per minute (-/min)

- Group: based on Participant Code
  - PD: subjects with Parkinson's disease
  - RBD: subjects with REM sleep behaviour disorder
  - HC: healthy controls

Table 1
Summary of the Data Set used in this Analysis

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	130	64.331	10.134	34	58.25	72	83
Gender	130						
... F	27	20.8%					
... M	103	79.2%					
Speech.Timing.Rate.Reading	130	327.277	47.385	140	297.25	358.75	457
Speech.Timing.Rate.Monologue	130	288.338	52.892	112	258	328.75	412
Pause.Interval.Duration.Reading	130	166.646	46.488	96	138.25	185	388
Pause.Interval.Duration.Monologue	130	229.069	79.697	117	177	263.25	611
Group	130						
... HC	50	38.5%					
... PD	30	23.1%					
... RBD	50	38.5%					

```
'data.frame':   130 obs. of  7 variables:
 $ Age                               : int  58 68 68 75 61 58 79 59 73 66 ...
 $ Gender                            : Factor w/ 2 levels "F","M": 1 1 2 2 2 2 2 1 2 2 ...
 $ Speech.Timing.Rate.Reading        : int  354 340 211 140 269 317 269 338 374 281 ...
 $ Speech.Timing.Rate.Monologue      : int  333 285 247 112 230 181 289 370 288 258 ...
 $ Pause.Interval.Duration.Reading   : int  146 173 377 360 211 186 214 145 117 213 ...
 $ Pause.Interval.Duration.Monologue: int  158 295 280 397 206 611 251 118 194 246 ...
 $ Group                             : Factor w/ 3 levels "HC","PD","RBD": 2 2 2 2 2 2 2 2 2
```

Data Pre-Processing

As an initial step, I created boxplots to check the distribution of the numerical data per group in detail (Figure 1). At first glance, parts of the data show skewed distributions, as the mean (shown as a orange point) differs substantially in many cases. This might prompt data transformations such as the *log*-transform. Additionally, within each variable, the distributions between the groups were assessed for significant differences. Here, the data showed significant differences between healthy controls (HC) and Parkinson's (PD) and REM sleep behaviour disorder subjects (RBD), but no significant differences between PD and RBD. Based on this, I decided to split the data anlysis part into two sections: (1) Creating a logistic regression model using `glm` to discriminate between the two groups HC and PD and (2) creating a multinomial regression model which discriminates between all three groups (HC, PD, and RBD).

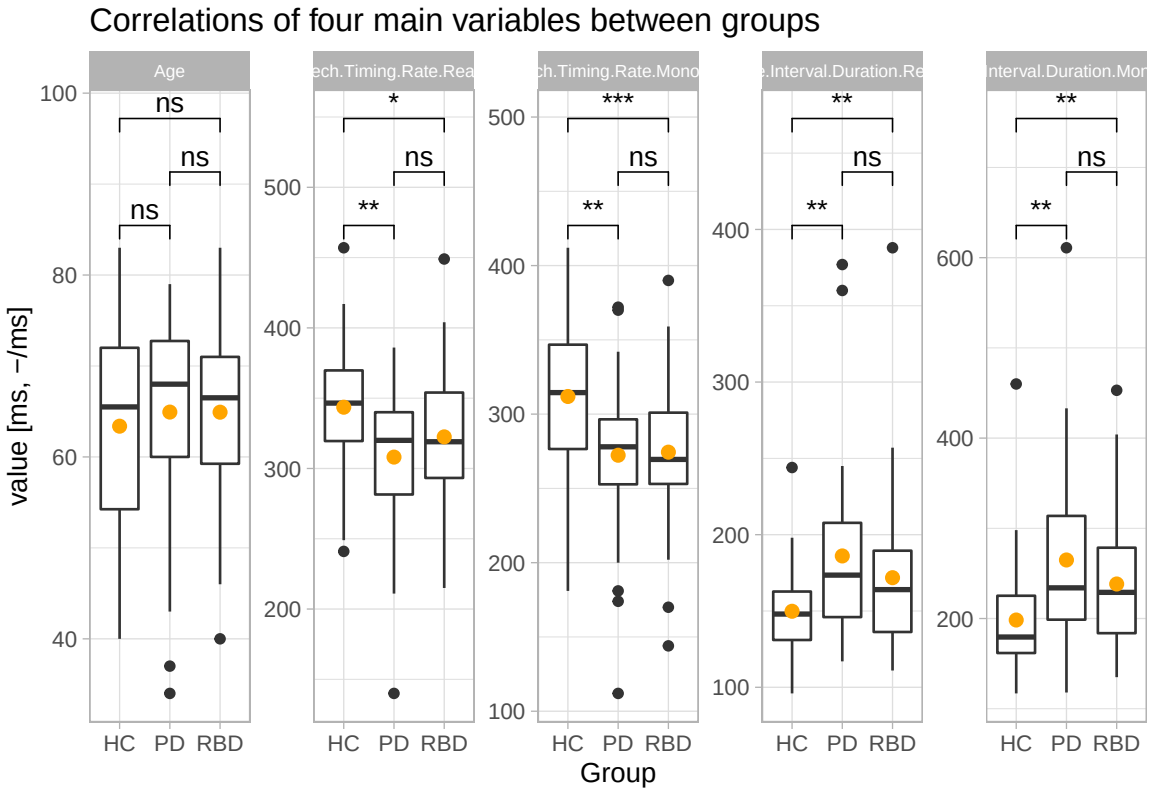

Figure 1. Distributions of data within variables and between groups. Some of the data shows skewed distributions (mean is represented by orange point), especially within the variable Age. While there is significant difference (t-Test) between healthy controls (HC) and subjects with Parkinson's disease (PD) as well as REM sleep behavior disorder (RBD), there are no significant differences between PD and RBD

Based on the empirical variables chosen, I excpected them to correlate. Indeed, Figure 2 shows a relatively strong correlation between these variables. Based on visual inspection of the boxplots (Figure 1), I chose to remove outliers in the following way:

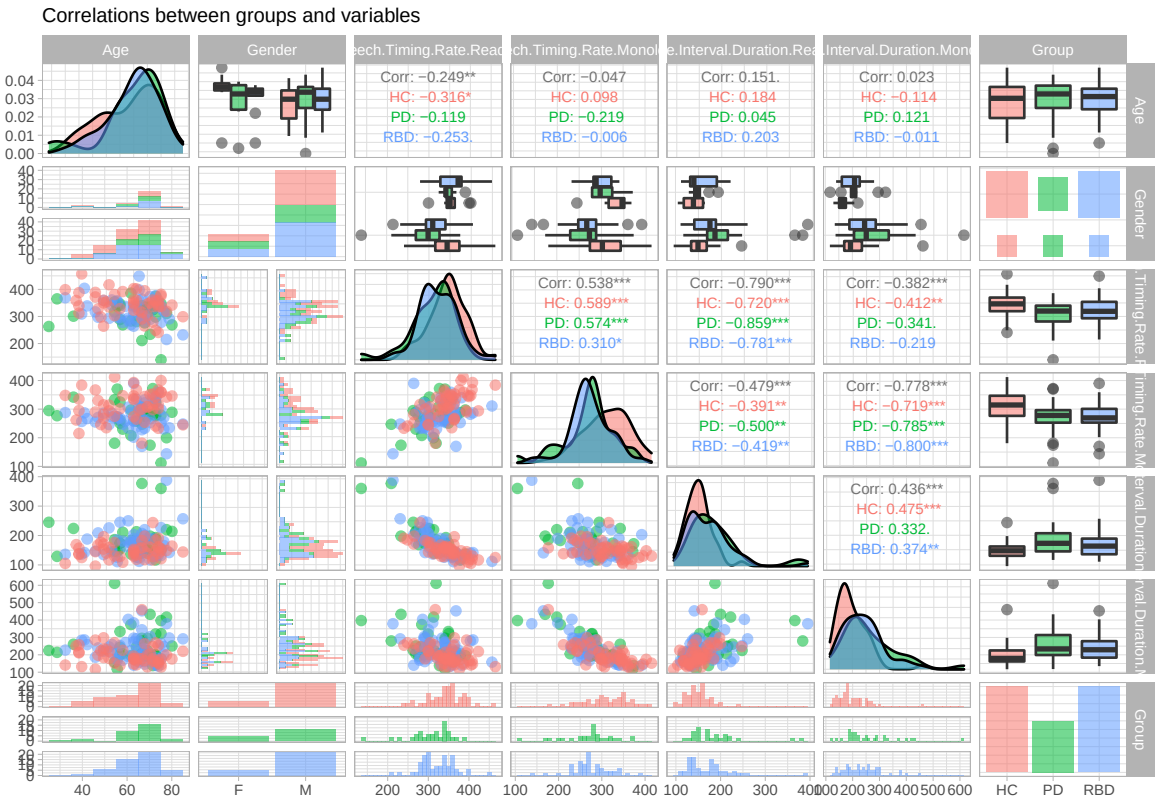

Figure 2. Plot based on ggpairs, colored by the response variable Group. The empirically collected speech data shows strong correlations (both positive and negative). In addition the density plots show the skewed distributions that were already seen in the boxplots.

```
df <- df[df$Pause.Interval.Duration.Monologue < 600, ]
df <- df[(df$Group != "HC" | df$Pause.Interval.Duration.Monologue < 450), ]
```

Given a lack of correlation between age and any of the speech-related variables, I chose to not remove outliers based on the variable age.

Data Analysis

Logistic Regression

As stated previously, I have seen that there are no significant differences between the groups PD and RBD. Based on this observation, I will limit my initial investigation to creating a logistic regression model predicting between the groups HC and PD. Indeed, the paper from which the data was extracted explicitly discusses the hard problem of differentiating PD from RBD, which might very well be impossible with generalised linear models. I will revisit this problem in the section Multinomial Regression.

As a first step, a subset is created that does not contain any observations from the group RBD.

```
df.binom <- data.frame(df[df$Group != "RBD", ])
```

Based on this subset, I first create simple logistic regression models with one response variable for each of the selected variables (Figure 3). For simplicity they were created using the `ggplot2` function `stat_smooth`. As can be seen by visual inspection of the data points (red), none of the predictors is sufficient to predict the response variable (Group) on its own, given the respective overlap between the two groups.

Given that a single predictor is clearly not sufficient, a series of multiple logistic regression models have to be built and evaluated. As I would have to test 64 models (all possible combinations plus intercept only) to be certain to have found the best one, I instead chose to use the automated model selection function `dredge` from the R package `MuMIn`. Starting from the global binomial model `Group ~ .` as an input, `dredge` enumerates all possible models and evaluates them based on their AIC.

```
m.full <- glm(
    data = df.binom, Group ~ .,
    family = binomial,
    na.action = "na.fail"
)

d <- dredge(m.full, rank = "AIC")

m.best.no.interactions <- get.models(d, 1)[[1]]
summary(m.best.no.interactions)
```

Importantly, the above automated model selection did not consider interactions between the predictors. Given the strong collinearity of the model (based the relatively strong correlation between the speech-related variables) it would be interesting to see whether solely the interaction between two variables would provide a better model. Indeed, the model below shows that solely focusing on the interactions provides a better result.

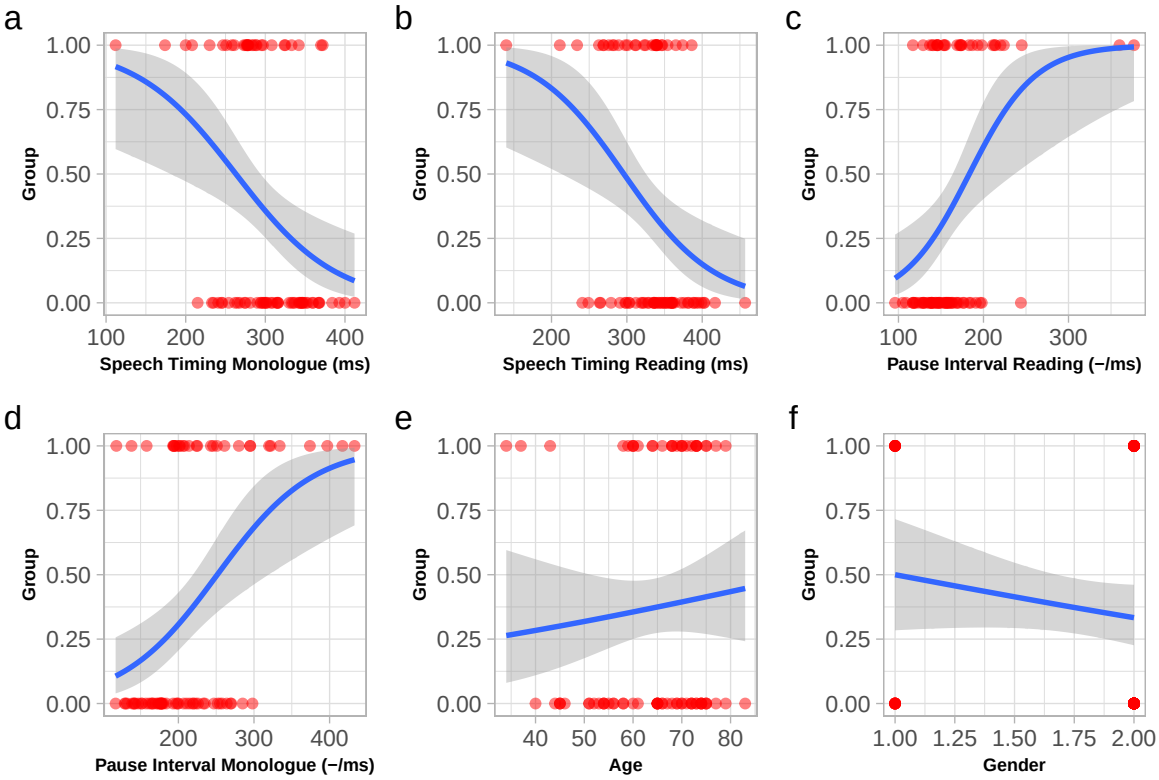

Figure 3. Simple logistic regression models with one predictor each. For variables a to d, we can see a clear sigmoid curve, while variables e, and of course f, which is a factor, do not show such a curve.

Table 2
A full regression table of the best model (selected using dredge) without interactions.

Predictor	*b*	95% CI	*z*	*p*
Intercept	-5.53	[-8.93, -2.73]	-3.50	< .001
GenderM	-1.79	[-3.22, -0.49]	-2.61	.009
Pause Interval Duration Monologue	0.01	[0.00, 0.03]	2.33	.020
Pause Interval Duration Reading	0.02	[0.00, 0.04]	1.98	.048

Table 3
A full regression table of a manually created model with based on an interaction between the variables Pa Interval Duration Reading and Pause Interval Duration Monologue.

Predictor	*b*	95% CI	*z*	*p*
Intercept	-2.29	[-3.94, -0.82]	-2.90	.004
GenderM	-1.72	[-3.10, -0.44]	-2.57	.010
Pause Interval Duration Reading × Pause Interval Duration Monologue	0.00	[0.00, 0.00]	3.67	< .0

PCA

As there has been significant correlation between the predictors in the ggpairs plot as well as some extreme changes in coefficients when adding additional variables, there exists the possbility of collinearity negatively affecting the models. Indeed, we observe variance inflation factors of more than 2.5 between all experimental predictors. This warrants and attempt at solving the potential collinearity issue.

```
                              Age                            Gender
                         1.210068                          1.417044
       Speech.Timing.Rate.Reading      Speech.Timing.Rate.Monologue
                         3.099071                          2.417658
  Pause.Interval.Duration.Reading Pause.Interval.Duration.Monologue
                         2.722974                          2.340769
```

```
Importance of components:
                           PC1    PC2     PC3     PC4
Standard deviation     1.7106 0.8041 0.54100 0.36693
Proportion of Variance 0.7315 0.1616 0.07317 0.03366
Cumulative Proportion  0.7315 0.8932 0.96634 1.00000
```

Figure 4 shows the loadings of the PCA.

Figure 5 shows, that the PCA has resolved the correlations between the variables.

A comparison between the interaction model with a model based on PC1 and the variable gender, does not show a significant difference.

```
Analysis of Deviance Table

Model 1: Group ~ PC1 + Gender
Model 2: Group ~ Pause.Interval.Duration.Reading:Pause.Interval.Duration.Monologue +
    Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        75     77.765
2        75     77.694  0    0.071
```

Multinomial Regression

To predict over all three groups (HC, PD, RBD), we have to use a more complex multinomial model. In order to evaluate the multinomial model, I created a train and test set. The training set contains 70% of the observations, while the test set contains the remaining 30%. Using the formula `Group ~ . - Age - Gender` results in the best performance.

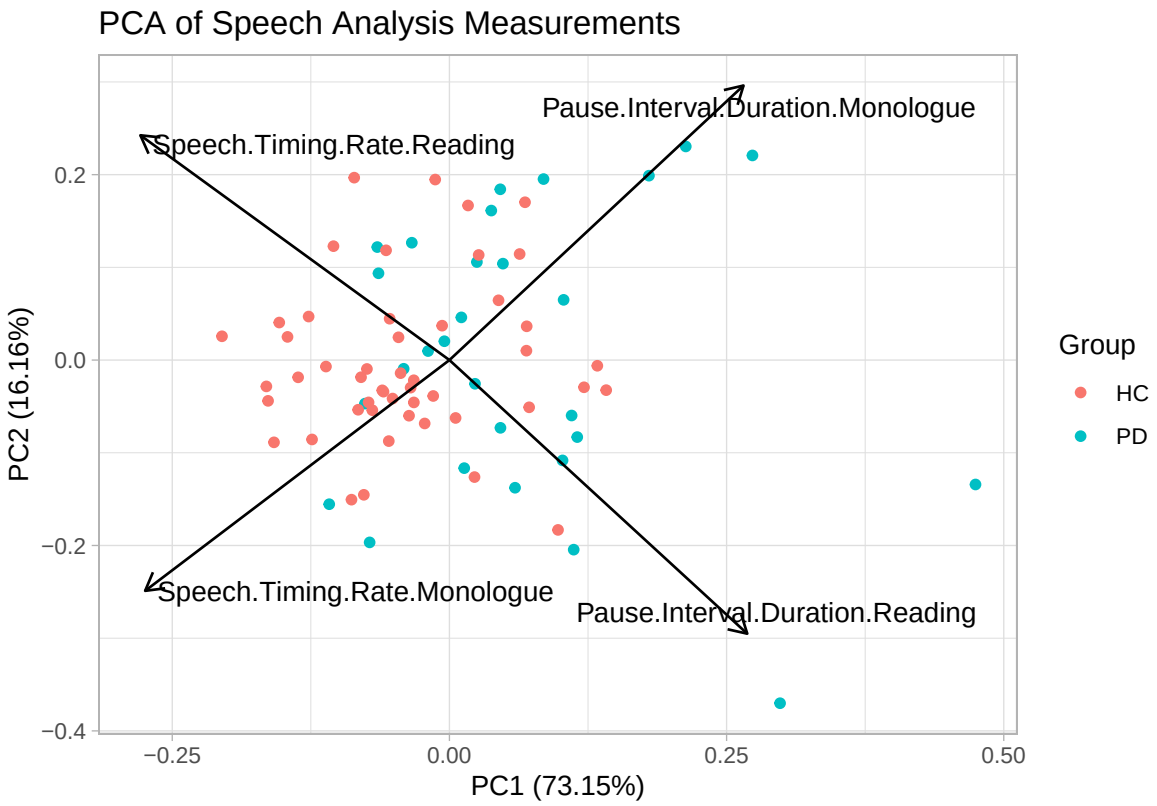

Figure 4. PCA autoplot of PCs 1 and 2. It once again shows that there exists a strong correlation between the variables.

```
# Set reference level explicitly
df$Group <- relevel(df$Group, ref = "HC")

# Train / test split
set.seed(123)
sample <- sample.int(
    n = nrow(df), size = floor(0.7 * nrow(df)),
    replace = FALSE
)

df.train <- df[sample, ]
df.test <- df[-sample, ]

model.all <- multinom(
    Group ~ . - Age - Gender,
    data = df.train
)
```

```
# weights:  18 (10 variable)
```

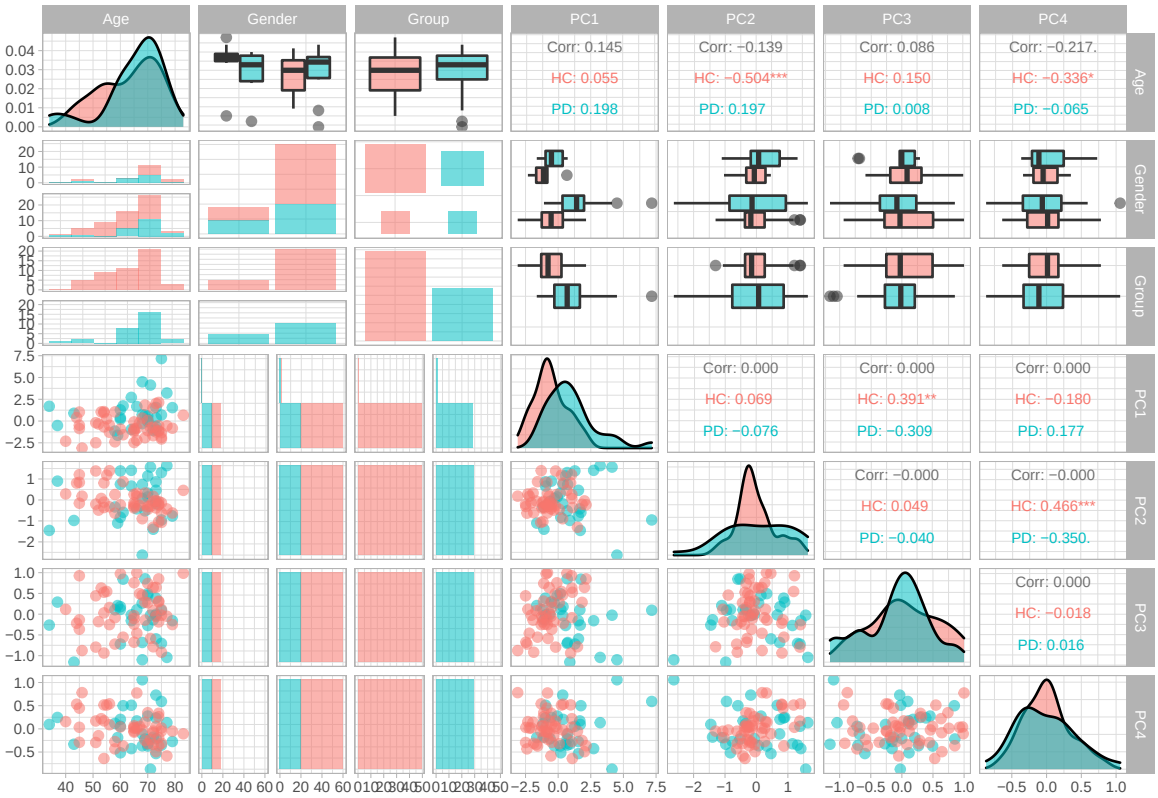

Figure 5. ggpairs plot where the speech-related variables have been replaced by the principal components of a PCA.

```
initial  value 97.776494
iter  10 value 78.995920
iter  20 value 77.591907
final  value 77.591844
converged
```

```
df.train$Group.Predicted <- predict(
    model.all,
    newdata = df.train, "class"
)

tab <- table(df.train$Group, df.train$Group.Predicted)
tab
```

```
      HC PD RBD
  HC  30  0   7
  PD   5  2  10
  RBD 11  1  23
```

```r
# nicer way ot output a confusion matrix
# calculate ratios instead of raw number to asses if True positive and false negative are h
accuracy <- round((sum(diag(tab)) / sum(tab)) * 100, 2)
```

The performance of this simple model is 61.80%, this is especially interesting when inspecting the confusion matrix above. None of the healthy subjects were diagnosed with Parkinson's disease, which is discrimination important in a clinical setting. However, only 2 out of 17 subjects with PD were diagnosed correctly. In addition, only 1 out of 35 subjects with RBD was misdiagnosed with PD.

Conclusion

Given the challenging nature of the data set, as well as the collinearity within it, the multinomial model yielded surprisingly good performance by showing high specifity when diagnosing REM sleep behavior disorder subjects and healthy controls. However, the logistic model, tasked with discriminating between healthy controls and subjects suffering from Parkinson's Disease, did not express satisfactory performance. Chosing difference combinations of predictors only had a small influence on the outcome. While I attribute this to the highly correlated predictors, a PCA-based model could not improve the situation and did not result in significantly better performance.

Manual Model Plot

References