Assignment 3 - Part 1 - Voice In Schizophrenia

Studygroup 4 - Kristine, Nanna, Julie, Sarah, Martine

07-10-2020

Assignment 3 - Part 1 - Assessing voice in schizophrenia

Individuals with schizophrenia (SCZ) tend to present voice atypicalities. Their tone is described as "inappropriate" voice, sometimes monotone, sometimes croaky. This is important for two reasons. First, voice could constitute a direct window into cognitive, emotional and social components of the disorder, thus providing a cheap and relatively non-invasive way to support the diagnostic and assessment process (via automated analyses). Second, voice atypicalities play an important role in the social impairment experienced by individuals with SCZ, and are thought to generate negative social judgments (of unengaged, slow, unpleasant interlocutors), which can cascade in more negative and less frequent social interactions.

Several studies show *significant* differences in acoustic features by diagnosis (see meta-analysis in the readings), but we want more. We want to know whether we can diagnose a participant only from knowing the features of their voice.

The corpus you are asked to analyse is a relatively large set of voice recordings from people with schizophrenia (just after first diagnosis) and matched controls (on gender, age, education). Each participant watched several videos of triangles moving across the screen and had to describe them (so you have several recordings per person). We have already extracted the pitch once every 10 milliseconds as well as several duration related features (e.g. number of pauses, etc).

N.B. For the fun of it, I threw in data from 3 different languages: 1) Danish (study 1-4); 2) Mandarin Chinese (Study 5-6); 3) Japanese (study 7). Feel free to only use the Danish data, if you think that Mandarin and Japanese add too much complexity to your analysis.

In this assignment (A3), you will have to discuss a few important questions (given the data you have). More details below.

Part 1 - Can we find a difference in acoustic features in schizophrenia? 1) Describe your sample number of studies, number of participants, age, gender, clinical and cognitive features of the two groups. Furthemore, critically assess whether the groups (schizophrenia and controls) are balanced. N.B. you need to take studies into account.

2) Describe the acoustic profile of a schizophrenic voice: which features are different? E.g. People with schizophrenia tend to have high-pitched voice, and present bigger swings in their prosody than controls. N.B. look also at effect sizes. How do these findings relate to the meta-analytic findings?

3) Discuss the analysis necessary to replicate the meta-analytic findings Look at the results reported in the paper (see meta-analysis in the readings) and see whether they are similar to those you get. 3.1) Check whether significance and direction of the effects are similar 3.2) Standardize your outcome, run the model and check whether the beta's is roughly matched (matched with hedge's g) which fixed and random effects should be included, given your dataset? E.g. what about language and study, age and gender? Discuss also how studies and languages should play a role in your analyses. E.g. should you analyze each study individually? Or each language individually? Or all together? Each of these choices makes some assumptions about how similar you expect the studies/languages to be. *Note* that there is no formal definition of replication (in statistical terms).

Your report should look like a methods paragraph followed by a result paragraph in a typical article (think the Communication and Cognition paper)

Part 2 - Can we diagnose schizophrenia from voice only? 1) Discuss whether you should you run the analysis on all studies and both languages at the same time You might want to support your results either by your own findings or by that of others 2) Choose your best acoustic feature from part 1. How well can you diagnose schizophrenia just using it? 3) Identify the best combination of acoustic features to diagnose schizophrenia using logistic regression. 4) Discuss the "classification" process: which methods are you using? Which confounds should you be aware of? What are the strength and limitation of the analysis?

Bonus question: Logistic regression is only one of many classification algorithms. Try using others and compare performance. Some examples: Discriminant Function, Random Forest, Support Vector Machine, Penalized regression, etc. The packages caret and glmnet provide them. Tidymodels is a set of tidyverse style packages, which take some time to learn, but provides a great workflow for machine learning.

Learning objectives

- Critically design, fit and report multilevel regression models in complex settings
- Critically appraise issues of replication

Overview of part 1

In the course of this part 1 of Assignment 3 you have to: - combine the different information from multiple files into one meaningful dataset you can use for your analysis. This involves: extracting descriptors of acoustic features from each pitch file (e.g. mean/median, standard deviation / interquartile range), and combine them with duration and demographic/clinical files - describe and discuss your sample - analyze the meaningful dataset to assess whether there are indeed differences in the schizophrenic voice and compare that to the meta-analysis

There are three pieces of data:

1- Demographic data (https://www.dropbox.com/s/e2jy5fyac18zld7/DemographicData.csv?dl=0). It contains

- Study: a study identifier (the recordings were collected during 6 different studies with 6 different clinical practitioners in 2 different languages)
- Language: Danish, Chinese and Japanese
- Participant: a subject ID
- Diagnosis: whether the participant has schizophrenia or is a control
- Gender
- Education
- Age
- SANS: total score of negative symptoms (including lack of motivation, affect, etc). Ref: Andreasen, N. C. (1989). The Scale for the Assessment of Negative Symptoms (SANS): conceptual and theoretical foundations. The British Journal of Psychiatry, 155(S7), 49-52.
- SAPS: total score of positive symptoms (including psychoses, such as delusions and hallucinations): http://www.bli.uzh.ch/BLI/PDF/saps.pdf
- VerbalIQ: https://en.wikipedia.org/wiki/Wechsler_Adult_Intelligence_Scale
- NonVerbalIQ: https://en.wikipedia.org/wiki/Wechsler_Adult_Intelligence_Scale
- TotalIQ: https://en.wikipedia.org/wiki/Wechsler_Adult_Intelligence_Scale

2. Articulation.txt (https://www.dropbox.com/s/vuyol7b575xdkjm/Articulation.txt?dl=0). It contains, per each file, measures of duration:

- soundname: the name of the recording file
- nsyll: number of syllables automatically inferred from the audio
- npause: number of pauses automatically inferred from the audio (absence of human voice longer than 200 milliseconds)
- dur (s): duration of the full recording
- phonationtime (s): duration of the recording where speech is present
- speechrate (nsyll/dur): average number of syllables per second
- articulation rate (nsyll / phonationtime): average number of syllables per spoken second
- ASD (speakingtime/nsyll): average syllable duration

3. One file per recording with the fundamental frequency of speech extracted every 10 milliseconds (excluding pauses): https://www.dropbox.com/sh/bfnzaf8xgxrv37u/AAD2k6SX4rJBHo7zzRML7cS9a?dl=0

- time: the time at which fundamental frequency was sampled
- f0: a measure of fundamental frequency, in Herz

NB. the filenames indicate: - Study: the study, 1-6 (1-4 in Danish, 5-6 in Mandarin Chinese) - D: the diagnosis, 0 is control, 1 is schizophrenia - S: the subject ID (NB. some controls and schizophrenia are matched, so there is a 101 schizophrenic and a 101 control). Also note that study 5-6 have weird numbers and no matched participants, so feel free to add e.g. 1000 to the participant ID in those studies. - T: the trial, that is, the recording ID for that participant, 1-10 (note that study 5-6 have more)

```
#Loading packages
library(pacman)
pacman::p_load(tidyverse, fs, readr, EnvStats, DescTools, pastecs)
```

Getting to the pitch data

You have oh so many pitch files. What you want is a neater dataset, with one row per recording, including a bunch of meaningful descriptors of pitch. For instance, we should include "standard" descriptors: mean, standard deviation, range. Additionally, we should also include less standard, but more robust ones: e.g. median, iqr, mean absoluted deviation, coefficient of variation. The latter ones are more robust to outliers and non-normal distributions.

Tip: Load one file (as a sample) and: - write code to extract the descriptors - write code to extract the relevant information from the file names (Participant, Diagnosis, Trial, Study) Only then (when everything works) turn the code into a function and use map_df() to apply it to all the files. See placeholder code here for help.

```
# Function to extract study, diagnosis, subject and trial from the file name + adding columns
read_pitch <- function(filename) {
    # getting filenames and subsetting the relevant parts
    files = path_file(path = filename)

    for (file in filename){
    Study = substr(files, 6,6)
    Diagnosis = substr(files, 8,8)
    Subject = substr(files, 10, 12)
    Trial = substr(files, 14,15)
    }
    # creating dataframes, loading data and and merging the df's
    df = data_frame(Study, Diagnosis, Subject, Trial)
    df1 = read.delim(filename)
```

```
    data = merge(df, df1)

    # extract pitch descriptors (mean, sd, iqr, etc)
    data$pitch_mean = mean(data$f0)
    data$pitch_sd = sd(data$f0)
    data$pitch_min = min(data$f0)
    data$pitch_max = max(data$f0)
    data$pitch_median = median(data$f0)
    data$pitch_IQR = IQR(data$f0)
    data$pitch_meanAD = MeanAD(data$f0)
    data$pitch_cv = cv(data$f0)

    #extracting time descriptors
    data$time_mean = mean(data$time)
    data$time_sd = sd(data$time)
    data$time_min = min(data$time)
    data$time_max = max(data$time)
    data$time_iqr = IQR(data$time)
    data$time_median  = median(data$time)
    data$time_meanAD = mad(data$time)
    data$time_cv = cv(data$time)
    data = slice(data,(1))

    data = data %>% mutate(
        Trial = str_replace_all(data$Trial, '[[:punct:]]', ''),
        Subject = as.factor(Subject),
        Study = as.numeric(Study),
        Diagnosis = as.factor(Diagnosis),
        Diagnosis = recode(Diagnosis,
                           '0' = 'Control',
                           '1' = 'Schizophrenia')
        )

    # combine all this data in one dataset
    return(data)
}

# test it on just one file while writing the function
#test_data = read_pitch("Pitch/Study1D0S101T1_f0.txt")
    #it works

# when you've created a function that works, you can
#pitch_data <- list.files(path = 'Pitch/', pattern = '.txt', all.files = T, full.names = T) %>%
    #purrr::map_df(read_pitch)

# save the new dataset as a csv file
#write_csv(pitch_data, 'pitch_data.csv')
    #We have saved it
```

Now you need to merge demographic/clinical, duration and pitch data

```
# Let's start with the demographic and clinical data
demo <- read.csv('DemographicData.csv', sep = ';', header = T)

#Removing columns that have no data
demo <-demo[-c(387:391),]

#Filter study 5, 6, 7
demo <- demo %>%
  filter(Study <= 4) %>%
  rename(Subject = Participant)


# Then duration data
art <- read.delim('Articulation.txt', sep = ',', header = T)

# Cleaning duration/articulation data
art <- art %>% mutate(
  Study = str_extract(soundname, '\\d'),
  Diagnosis = str_sub(soundname, 8, 8),
  Subject = str_extract(soundname, '\\d{3}'),
  Trial = str_extract(soundname, '.\\d$'),
  Trial = str_replace(Trial, pattern = 'T', ''),
    Subject = as.factor(Subject),
    Study = as.numeric(Study),
    Diagnosis = as.factor(Diagnosis),
    Diagnosis = recode(Diagnosis,
                           '0' = 'Control',
                           '1' = 'Schizophrenia')
  )

art$soundname <- NULL

#Removing observations that are not from the danish study
art <- art %>%
  filter(Study <= 4, Subject != 342)

# Finally the pitch data
pitch <- read.csv('pitch_data.csv')

#Removing observations that are not from the danish study
pitch <- pitch %>%
  filter(Study <= 4, Subject != 342)

# Now we merge them
    # But first we make sure that everything is in the same class

#For demo
demo <- demo %>% mutate(
    Study = as.factor(Study),
    Diagnosis = as.factor(Diagnosis),
    Subject = as.factor(Subject))
```

```
#For art
art <- art %>% mutate(
    Study = as.factor(Study),
    Diagnosis = as.factor(Diagnosis),
    Subject = as.factor(Subject),
    Trial = as.factor(Trial))

#For pitch
pitch <- pitch %>% mutate(
    Study = as.factor(Study),
    Diagnosis = as.factor(Diagnosis),
    Subject = as.factor(Subject),
    Trial = as.factor(Trial))

#Merging
  #Making new column that has a unique ID
demo$ID <- paste0(demo$Subject, demo$Diagnosis)
art$ID <- paste0(art$Subject,art$Diagnosis)
pitch$ID <- paste0(pitch$Subject, pitch$Diagnosis)

  # create a surrogate key by adding trial to iD in a new column?
# don't know if this is necessary or not
art$ID2 <- paste0(art$Subject,art$Diagnosis, art$Trial)
pitch$ID2 <- paste0(pitch$Subject, pitch$Diagnosis, pitch$Trial)

#assesing that ID2 is unique
x <- art %>% count(ID2) %>% filter(ID2 > 1)
y <- pitch %>% count(ID2) %>% filter(ID2 > 1)

#try both left join and full join to see what happens.
pitch_art <- pitch %>%
  left_join(art)

pitch_art$Subject <- as.factor(pitch_art$Subject)

df <- pitch_art %>% left_join(demo)

# Now we save them

#write_csv(df, "SchizophreniaData.csv")
```

Now we need to describe our sample

```
#Loading data from files

df <- read_csv("SchizophreniaData.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Diagnosis = col_character(),
```

```
##   ID = col_character(),
##   ID2 = col_character(),
##   Language = col_character(),
##   Gender = col_character()
## )

## See spec(...) for full column specifications.
```

```
#correcting dataset classes
df <- df %>% mutate(
  Trial = as.factor(df$Trial),
  Study = as.factor(df$Study),
  ID = as.factor(df$ID),
  Diagnosis = as.factor(df$Diagnosis),
  ASD..speakingtime.nsyll. = as.numeric(df$ASD..speakingtime.nsyll.),
  Gender = as.factor(df$Gender),
)
```

First look at the missing data: we should exclude all recordings for which we do not have complete data. Then count the participants and recordinsgs by diagnosis, report their gender, age and symptom severity (SANS, SAPS and Social) Finally, do the same by diagnosis and study, to assess systematic differences in studies. I like to use group_by() %>% summarize() for quick summaries

```
#We found that study 3 and 4 have no data for several of the clinical observations,
  #as we don't use any of those columns in the models, we decided to keep the observations

#Overview of the dataset
df %>%
  split(df$Diagnosis) %>%
  map(summary)
```

```
## $Control
##  Study           Diagnosis      Subject          Trial          time
##  1:348   Control      :989   Min.   :101.0   1      :115   Min.   :0.0100
##  2:184   Schizophrenia:  0   1st Qu.:126.0   2      :115   1st Qu.:0.1660
##  3:224                       Median :219.0   3      :115   Median :0.3550
##  4:233                       Mean   :253.4   4      :115   Mean   :0.5075
##                              3rd Qu.:343.0   5      :115   3rd Qu.:0.7180
##                              Max.   :448.0   6      :115   Max.   :4.4960
##                                              (Other):299
##        f0            pitch_mean       pitch_sd         pitch_min
##  Min.   :  46.62   Min.   : 53.1   Min.   :  2.861   Min.   : 42.43
##  1st Qu.: 119.44   1st Qu.:110.5   1st Qu.: 14.975   1st Qu.: 75.71
##  Median : 153.13   Median :142.7   Median : 21.408   Median : 90.14
##  Mean   : 181.71   Mean   :160.2   Mean   : 34.421   Mean   :105.67
##  3rd Qu.: 225.93   3rd Qu.:202.3   3rd Qu.: 31.013   3rd Qu.:136.73
##  Max.   :1100.82   Max.   :780.1   Max.   :375.416   Max.   :484.79
##
##    pitch_max        pitch_median      pitch_IQR        pitch_meanAD
##  Min.   :  56.51   Min.   : 54.25   Min.   :  3.382   Min.   :  2.334
##  1st Qu.: 195.24   1st Qu.:106.30   1st Qu.: 15.018   1st Qu.: 10.697
##  Median : 249.32   Median :126.64   Median : 21.545   Median : 15.201
##  Mean   : 283.70   Mean   :151.61   Mean   : 44.491   Mean   : 27.335
```

```
##  3rd Qu.: 334.06   3rd Qu.:193.27   3rd Qu.: 32.745   3rd Qu.: 23.012
##  Max.   :1100.82   Max.   :890.48   Max.   :808.500   Max.   :354.551
##
##     picth_cv         time_mean          time_sd           time_min
##  Min.   :0.04018   Min.   : 0.3796   Min.   : 0.1214   Min.   :0.0100
##  1st Qu.:0.10823   1st Qu.: 4.3968   1st Qu.: 2.5352   1st Qu.:0.1660
##  Median :0.14129   Median : 7.4185   Median : 4.3551   Median :0.3550
##  Mean   :0.19572   Mean   : 8.8908   Mean   : 5.1306   Mean   :0.5075
##  3rd Qu.:0.19617   3rd Qu.:11.7321   3rd Qu.: 6.8490   3rd Qu.:0.7180
##  Max.   :1.10867   Max.   :47.4656   Max.   :27.7028   Max.   :4.4960
##
##     time_max         time_iqr       time_median      time_meanAD
##  Min.   : 0.680   Min.   : 0.175   Min.   : 0.345   Min.   : 0.1038
##  1st Qu.: 9.025   1st Qu.: 4.190   1st Qu.: 4.229   1st Qu.: 3.0097
##  Median :15.266   Median : 7.605   Median : 7.074   Median : 5.2929
##  Mean   :18.118   Mean   : 8.799   Mean   : 8.723   Mean   : 6.3662
##  3rd Qu.:24.112   3rd Qu.:11.910   3rd Qu.:11.738   3rd Qu.: 8.7325
##  Max.   :97.843   Max.   :48.360   Max.   :48.033   Max.   :35.9234
##
##     time_cv                ID          ID2                nsyll
##  Min.   :0.1153   101Control: 10   Length:989         Min.   :  1.00
##  1st Qu.:0.5288   102Control: 10   Class :character   1st Qu.: 28.00
##  Median :0.5779   103Control: 10   Mode  :character   Median : 49.00
##  Mean   :0.5690   104Control: 10                      Mean   : 60.87
##  3rd Qu.:0.6172   105Control: 10                      3rd Qu.: 81.00
##  Max.   :0.9108   106Control: 10                      Max.   :401.00
##                   (Other)   :929
##      npause           dur..s.       phonationtime..s. speechrate..nsyll.dur.
##  Min.   : 0.000   Min.   : 1.10   Min.   : 0.48     Min.   :0.670
##  1st Qu.: 3.000   1st Qu.: 9.55   1st Qu.: 5.66     1st Qu.:2.640
##  Median : 7.000   Median :15.85   Median : 9.73     Median :3.200
##  Mean   : 7.925   Mean   :18.71   Mean   :12.24     Mean   :3.169
##  3rd Qu.:11.000   3rd Qu.:24.60   3rd Qu.:16.03     3rd Qu.:3.710
##  Max.   :45.000   Max.   :97.96   Max.   :81.08     Max.   :5.950
##
##  articulation.rate..nsyll...phonationtime. ASD..speakingtime.nsyll.
##  Min.   :1.740                             Min.   :0.1120
##  1st Qu.:4.610                             1st Qu.:0.1850
##  Median :5.020                             Median :0.1990
##  Mean   :4.968                             Mean   :0.2062
##  3rd Qu.:5.400                             3rd Qu.:0.2170
##  Max.   :8.930                             Max.   :0.5760
##
##    Language          Gender          Age          Education          SANS
##  Length:989         F   :423   Min.   :18.00   Min.   : 8.00   Min.   :0.0000
##  Class :character   M   :558   1st Qu.:21.00   1st Qu.:13.00   1st Qu.:0.0000
##  Mode  :character   NA's:  8   Median :24.00   Median :15.00   Median :0.0000
##                                Mean   :26.47   Mean   :14.87   Mean   :0.3922
##                                3rd Qu.:27.00   3rd Qu.:17.00   3rd Qu.:0.0000
##                                Max.   :62.00   Max.   :23.00   Max.   :7.0000
##                                NA's   :24      NA's   :8       NA's   :224
##       SAPS            VerbalIQ      NonVerbalIQ       TotalIQ
##  Min.   :0.00000   Min.   : 64.0   Min.   : 60.0   Min.   : 61.0
##  1st Qu.:0.00000   1st Qu.: 94.0   1st Qu.: 93.0   1st Qu.: 93.0
```

```
##  Median :0.00000   Median :103.0   Median :105.0   Median :102.0
##  Mean   :0.07712   Mean   :102.1   Mean   :102.2   Mean   :102.3
##  3rd Qu.:0.00000   3rd Qu.:113.0   3rd Qu.:112.0   3rd Qu.:112.0
##  Max.   :3.00000   Max.   :135.0   Max.   :132.0   Max.   :135.0
##  NA's   :224       NA's   :457     NA's   :457     NA's   :457
##
## $Schizophrenia
##  Study           Diagnosis      Subject          Trial          time
##  1:335   Control      :  0   Min.   :103.0   2      :105   Min.   :0.0090
##  2:179   Schizophrenia:903   1st Qu.:125.0   3      :105   1st Qu.:0.1150
##  3:151                       Median :215.0   4      :105   Median :0.2830
##  4:238                       Mean   :251.4   1      :104   Mean   :0.4644
##                              3rd Qu.:402.0   5      :104   3rd Qu.:0.6295
##                              Max.   :446.0   6      :104   Max.   :6.9860
##                                              (Other):276
##        f0          pitch_mean        pitch_sd         pitch_min
##  Min.   : 51.3   Min.   : 81.81   Min.   :  1.937   Min.   : 40.71
##  1st Qu.:119.9   1st Qu.:111.39   1st Qu.: 12.025   1st Qu.: 79.81
##  Median :156.1   Median :133.22   Median : 17.734   Median : 96.70
##  Mean   :174.2   Mean   :154.23   Mean   : 23.892   Mean   :110.43
##  3rd Qu.:225.0   3rd Qu.:201.12   3rd Qu.: 25.203   3rd Qu.:140.84
##  Max.   :569.2   Max.   :536.65   Max.   :364.479   Max.   :240.70
##
##    pitch_max       pitch_median      pitch_IQR       pitch_meanAD
##  Min.   : 99.61   Min.   : 57.06   Min.   :  1.77   Min.   :  1.661
##  1st Qu.:167.13   1st Qu.:107.61   1st Qu.: 12.71   1st Qu.:  8.731
##  Median :226.18   Median :127.09   Median : 18.37   Median : 12.873
##  Mean   :247.74   Mean   :150.18   Mean   : 28.00   Mean   : 18.361
##  3rd Qu.:303.76   3rd Qu.:196.97   3rd Qu.: 26.68   3rd Qu.: 18.257
##  Max.   :918.16   Max.   :814.03   Max.   :726.90   Max.   :359.186
##
##     picth_cv         time_mean          time_sd            time_min
##  Min.   :0.01687   Min.   : 0.2121   Min.   : 0.09289   Min.   :0.0090
##  1st Qu.:0.08759   1st Qu.: 3.0655   1st Qu.: 1.66977   1st Qu.:0.1150
##  Median :0.11776   Median : 6.0750   Median : 3.46587   Median :0.2830
##  Mean   :0.15189   Mean   : 8.5508   Mean   : 4.90902   Mean   :0.4644
##  3rd Qu.:0.15887   3rd Qu.:11.2606   3rd Qu.: 6.60762   3rd Qu.:0.6295
##  Max.   :0.97617   Max.   :80.4376   Max.   :45.22373   Max.   :6.9860
##
##     time_max          time_iqr       time_median      time_meanAD
##  Min.   :  0.330   Min.   : 0.090   Min.   : 0.230   Min.   : 0.07413
##  1st Qu.:  5.797   1st Qu.: 2.730   1st Qu.: 2.986   1st Qu.: 1.86066
##  Median : 12.170   Median : 5.925   Median : 5.915   Median : 4.24024
##  Mean   : 17.179   Mean   : 8.355   Mean   : 8.463   Mean   : 6.05984
##  3rd Qu.: 22.927   3rd Qu.:11.523   3rd Qu.:11.054   3rd Qu.: 8.22843
##  Max.   :164.580   Max.   :81.110   Max.   :81.560   Max.   :80.01592
##
##     time_cv                       ID          ID2                nsyll
##  Min.   :0.03528   103Schizophrenia: 10   Length:903         Min.   :  1.00
##  1st Qu.:0.50880   104Schizophrenia: 10   Class :character   1st Qu.: 17.00
##  Median :0.57110   105Schizophrenia: 10   Mode  :character   Median : 36.00
##  Mean   :0.55750   106Schizophrenia: 10                      Mean   : 50.32
##  3rd Qu.:0.61490   107Schizophrenia: 10                      3rd Qu.: 70.00
##  Max.   :1.22472   108Schizophrenia: 10                      Max.   :464.00
```

```
##                    (Other)         :843
##      npause          dur..s.       phonationtime..s. speechrate..nsyll.dur.
##  Min.   : 0.000   Min.   :  0.70   Min.   : 0.44     Min.   :0.110
##  1st Qu.: 2.000   1st Qu.:  6.21   1st Qu.: 3.54     1st Qu.:2.410
##  Median : 5.000   Median : 12.71   Median : 7.27     Median :2.920
##  Mean   : 7.592   Mean   : 17.64   Mean   :10.35     Mean   :2.929
##  3rd Qu.:10.000   3rd Qu.: 23.78   3rd Qu.:14.35     3rd Qu.:3.495
##  Max.   :97.000   Max.   :164.83   Max.   :85.57     Max.   :6.520
##
##  articulation.rate..nsyll...phonationtime. ASD..speakingtime.nsyll.
##  Min.   :1.200                             Min.   :0.1250
##  1st Qu.:4.380                             1st Qu.:0.1900
##  Median :4.830                             Median :0.2070
##  Mean   :4.793                             Mean   :0.2157
##  3rd Qu.:5.270                             3rd Qu.:0.2285
##  Max.   :7.980                             Max.   :0.8320
##
##    Language          Gender       Age          Education          SANS
##  Length:903         F:388   Min.   :18.00   Min.   : 8.00   Min.   : 0.000
##  Class :character   M:515   1st Qu.:21.00   1st Qu.:10.00   1st Qu.: 6.000
##  Mode  :character           Median :24.00   Median :12.50   Median :10.000
##                             Mean   :26.49   Mean   :12.89   Mean   : 9.669
##                             3rd Qu.:28.00   3rd Qu.:15.00   3rd Qu.:13.000
##                             Max.   :61.00   Max.   :19.00   Max.   :20.000
##                                                             NA's   :151
##       SAPS          VerbalIQ       NonVerbalIQ        TotalIQ
##  Min.   : 0.00   Min.   : 48.00   Min.   : 45.00   Min.   : 45.00
##  1st Qu.: 7.00   1st Qu.: 74.00   1st Qu.: 75.00   1st Qu.: 78.00
##  Median :11.00   Median : 87.00   Median : 93.00   Median : 88.00
##  Mean   :10.33   Mean   : 89.18   Mean   : 88.64   Mean   : 87.56
##  3rd Qu.:14.00   3rd Qu.:103.00   3rd Qu.:100.00   3rd Qu.:101.00
##  Max.   :20.00   Max.   :129.00   Max.   :119.00   Max.   :124.00
##  NA's   :151     NA's   :389      NA's   :389      NA's   :389
```

```
#Descriptive statistics
  #Creating dataframe grouping by diagnosis
report <- df

  #Creating dataframe with only schizophrenics
Schizo <- subset(report, Diagnosis == "Schizophrenia")

  #Sd and mean age schizophrenics
summarize(Schizo, sd(Age), mean(Age))
```

```
## # A tibble: 1 x 2
##   `sd(Age)` `mean(Age)`
##       <dbl>       <dbl>
## 1      8.82        26.5
```

```
    #Only females
schizo_fe <- subset(Schizo, Gender == "F")
    #Only males
schizo_ma <- subset(Schizo, Gender == "M")
```

```r
  #Control dataframe
control <- subset(report, Diagnosis == "Control")
  #removing na's
control_Age <- na.omit(control)
summarize(control_Age, sd(Age), mean(Age))
```

```
## # A tibble: 1 x 2
##   `sd(Age)` `mean(Age)`
##       <dbl>       <dbl>
## 1      3.35        23.0
```

```r
  #Only females
control_fe <- subset(control, Gender == "F")
  #Only males
control_ma <- subset(control, Gender == "M")

#Clinical features
  #Removing na's and thereby exluding study 3 and 4, since none of them has the clinical observations
df_naomit <- na.omit(df) %>% group_by(Diagnosis)
  #SANS
summarize(df_naomit, sd(SANS), mean(SANS))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 3
##   Diagnosis     `sd(SANS)` `mean(SANS)`
##   <fct>              <dbl>        <dbl>
## 1 Control             0             0
## 2 Schizophrenia       4.65         10.2
```

```r
  #SAPS
summarize(df_naomit, sd(SAPS), mean(SAPS))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 3
##   Diagnosis     `sd(SAPS)` `mean(SAPS)`
##   <fct>              <dbl>        <dbl>
## 1 Control             0             0
## 2 Schizophrenia       4.55         11.9
```

```r
  #VerbalIQ
summarize(df_naomit, sd(VerbalIQ), mean(VerbalIQ))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 3
##   Diagnosis     `sd(VerbalIQ)` `mean(VerbalIQ)`
##   <fct>                 <dbl>            <dbl>
## 1 Control                16.1             102.
## 2 Schizophrenia          18.7              89.2
```

```
  #Non-verbalIQ
summarize(df_naomit, sd(NonVerbalIQ), mean(NonVerbalIQ))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Diagnosis    `sd(NonVerbalIQ)` `mean(NonVerbalIQ)`
##   <fct>                    <dbl>               <dbl>
## 1 Control                   13.0                102.
## 2 Schizophrenia             18.3                 88.6
```

```
#Making a t-test to see if the two groups are significantly independent
t.test(df$VerbalIQ ~ df$Diagnosis)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$VerbalIQ by df$Diagnosis
## t = 11.99, df = 1009.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.82798 15.06579
## sample estimates:
##       mean in group Control mean in group Schizophrenia
##                   102.12782                   89.18093
```

```
t.test(df$NonVerbalIQ ~ df$Diagnosis)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$NonVerbalIQ by df$Diagnosis
## t = 13.704, df = 923.51, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.58843 15.46250
## sample estimates:
##       mean in group Control mean in group Schizophrenia
##                   102.16165                   88.63619
```

Now we can analyze the data

If you were to examine the meta analysis you would find that the differences (measured as Hedges' g, very close to Cohen's d, that is, in standard deviations) to be the following - pitch variability (lower, Hedges' g: -0.55, 95% CIs: -1.06, 0.09) - proportion of spoken time (lower, Hedges' g: -1.26, 95% CIs: -2.26, 0.25) - speech rate (slower, Hedges' g: -0.75, 95% CIs: -1.51, 0.04) - pause duration (longer, Hedges' g: 1.89, 95% CIs: 0.72, 3.21). (Duration - Spoken Duration) / PauseN

We need therefore to set up 4 models to see how well our results compare to the meta-analytic findings (Feel free of course to test more features) Describe the acoustic profile of a schizophrenic voice *Note* in this section

you need to describe the acoustic profile of a schizophrenic voice and compare it with the meta-analytic findings (see 2 and 3 in overview of part 1).

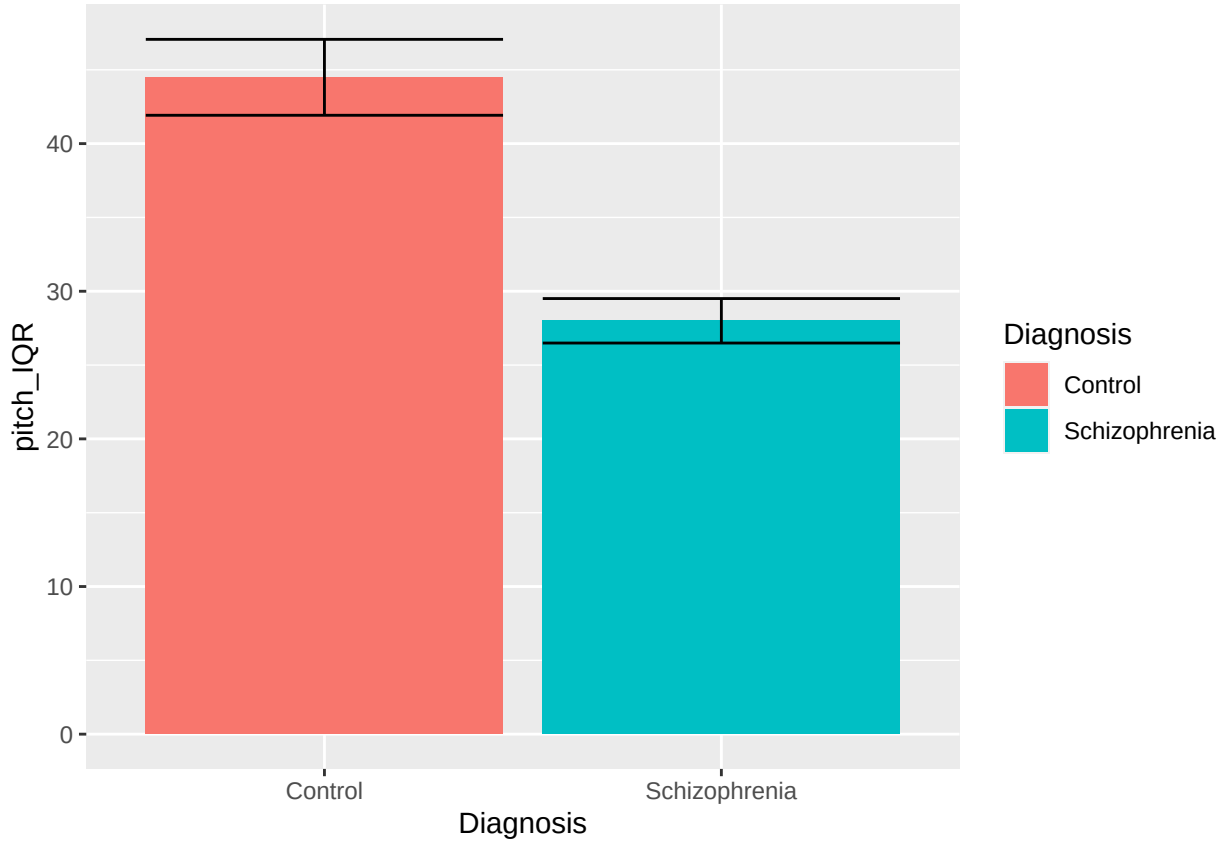
N.B. the meta-analytic findings are on scaled measures. If you want to compare your results with them, you need to scale your measures as well: subtract the mean, and divide by the standard deviation. N.N.B. We want to think carefully about fixed and random effects in our model. In particular: how should study be included? Does it make sense to have all studies put together? Does it make sense to analyze both languages together? Relatedly: does it make sense to scale all data from all studies together? N.N.N.B. If you want to estimate the studies separately, you can try this syntax: Feature ~ 0 + Study + Study:Diagnosis + [your randomEffects]. Now you'll have an intercept per each study (the estimates for the controls) and an effect of diagnosis per each study

- Bonus points: cross-validate the models and report the betas and standard errors from all rounds to get an idea of how robust the estimates are.

```
#Making new columns for predictors
  #spoken time / duration
        #making a column for proportion of spoken time
df$prop_spoken_time <- df$phonationtime..s. / df$dur..s.

  #making new column with pause duration -> (Duration - Spoken Duration) / PauseN
df$pause_duration <- (df$dur..s. - df$phonationtime..s.) / df$npause

  #removing inf number when pause duration is 0
df$pause_duration <-  ifelse(df$npause == 0, 0, df$pause_duration)

#Making plots to eyeball the data

#Pitch variability

ggplot(df, aes(x = Diagnosis, y = pitch_IQR, fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

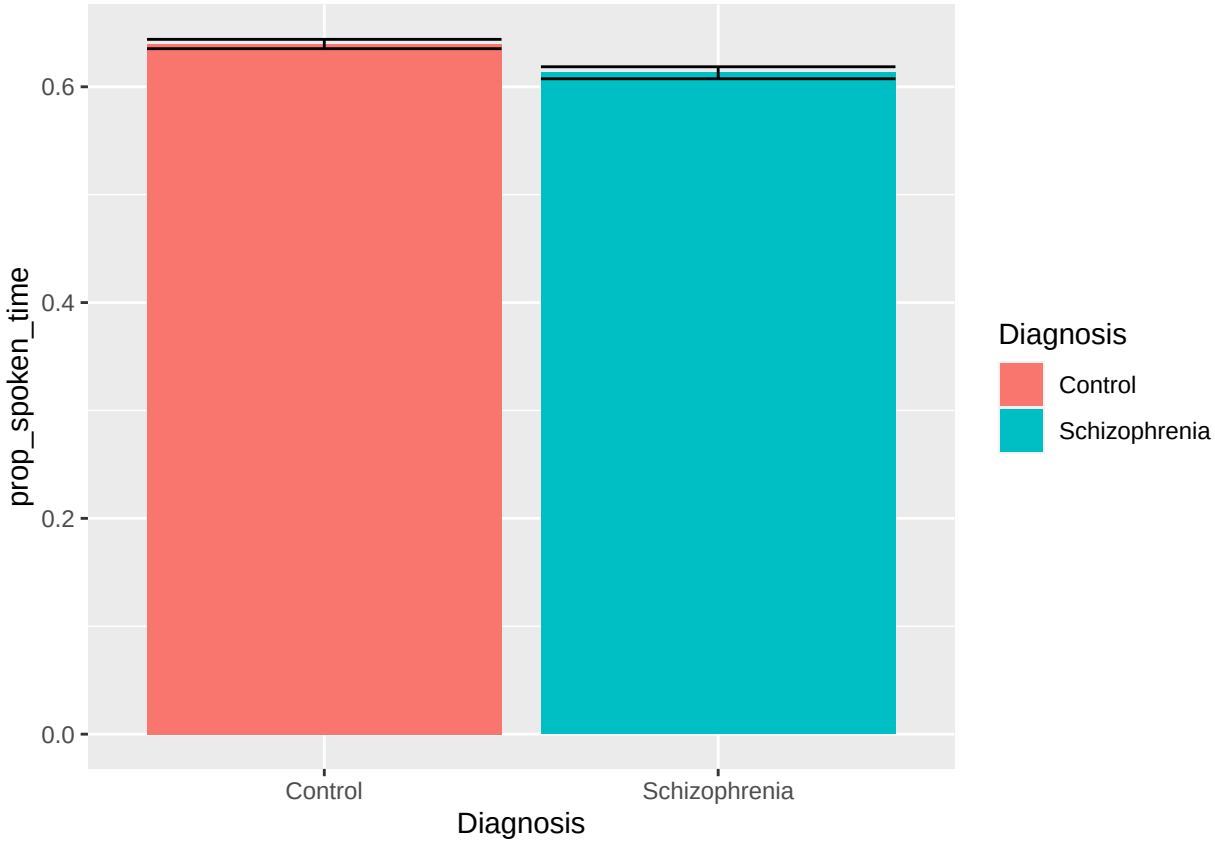
```
## No summary function supplied, defaulting to `mean_se()`
```




```
#proportion of spoken time

ggplot(df, aes(x = Diagnosis, y = prop_spoken_time, fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

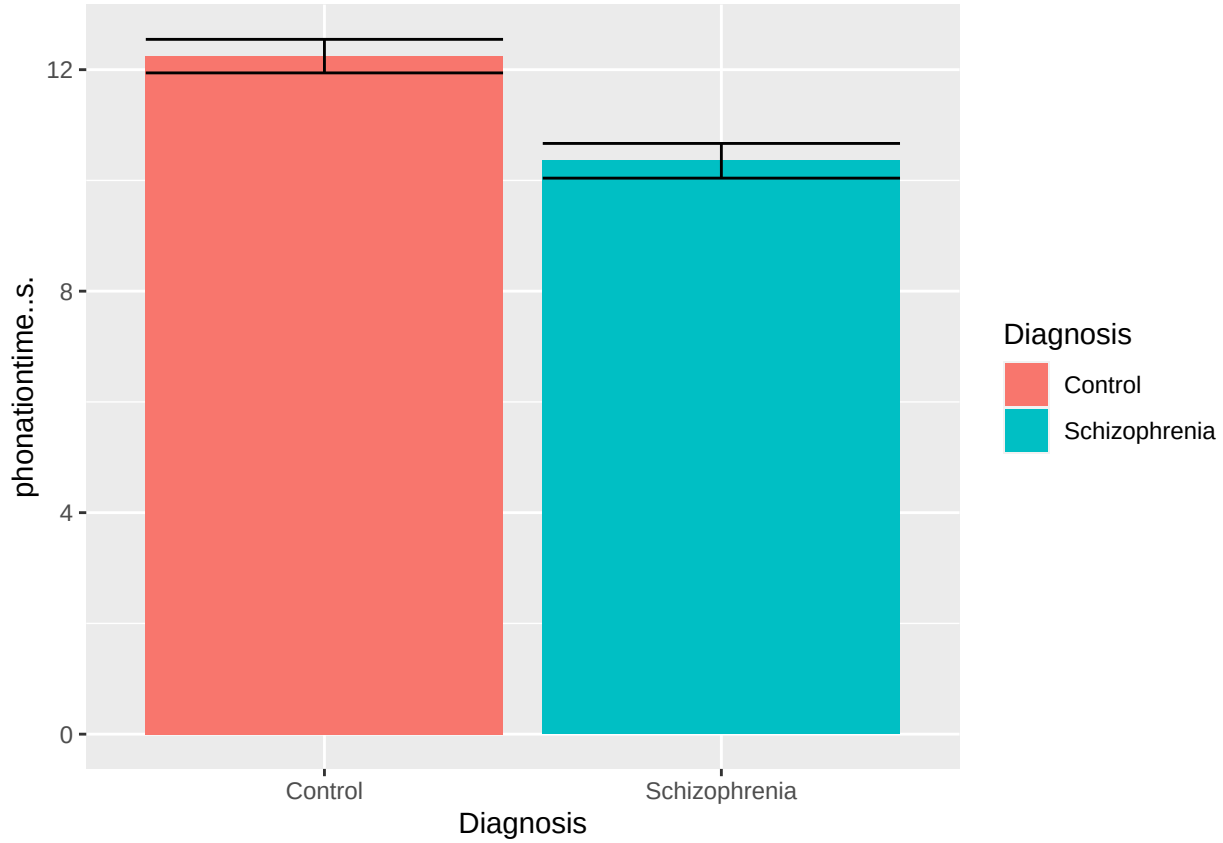
```
## No summary function supplied, defaulting to `mean_se()`
```




```
#Duration of Utterance

ggplot(df, aes(x = Diagnosis, y = phonationtime..s., fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

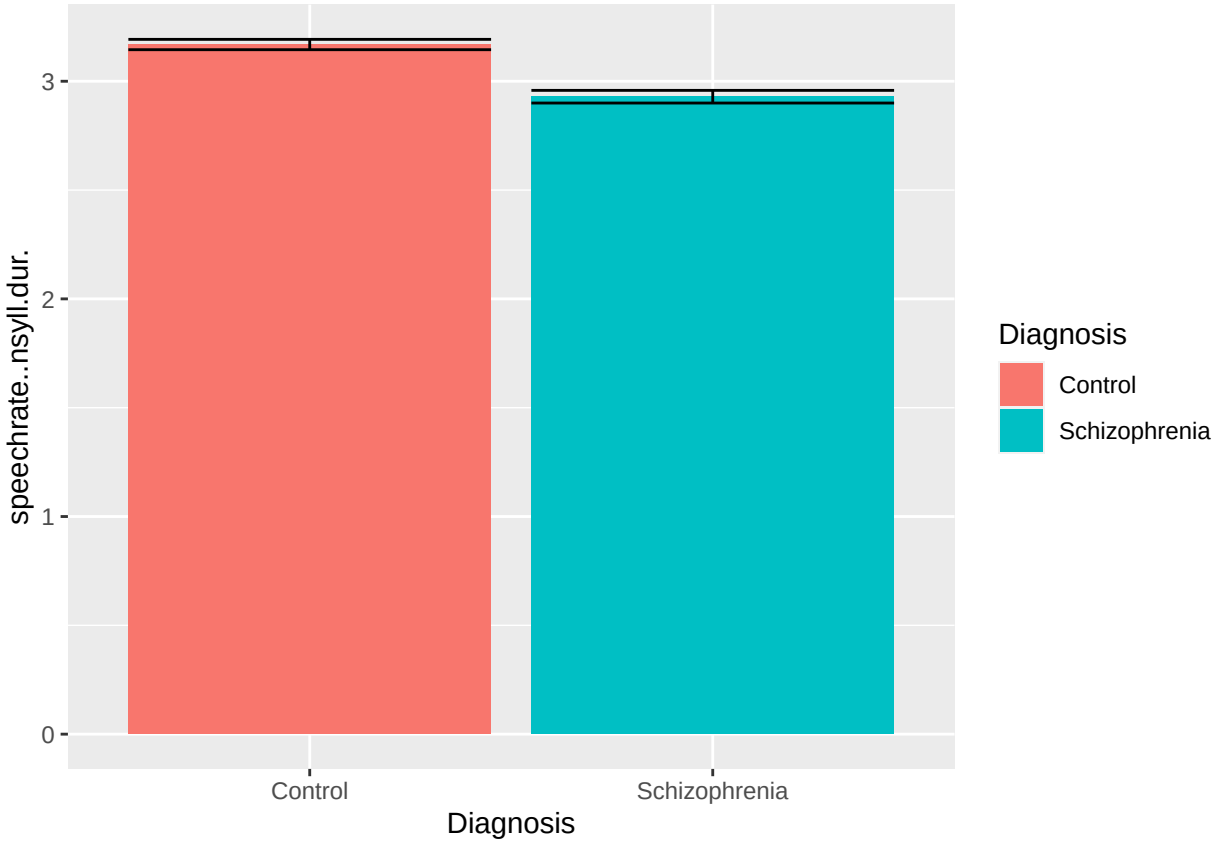
```
## No summary function supplied, defaulting to `mean_se()`
```




```
#speech rate

ggplot(df, aes(x = Diagnosis, y = speechrate..nsyll.dur., fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

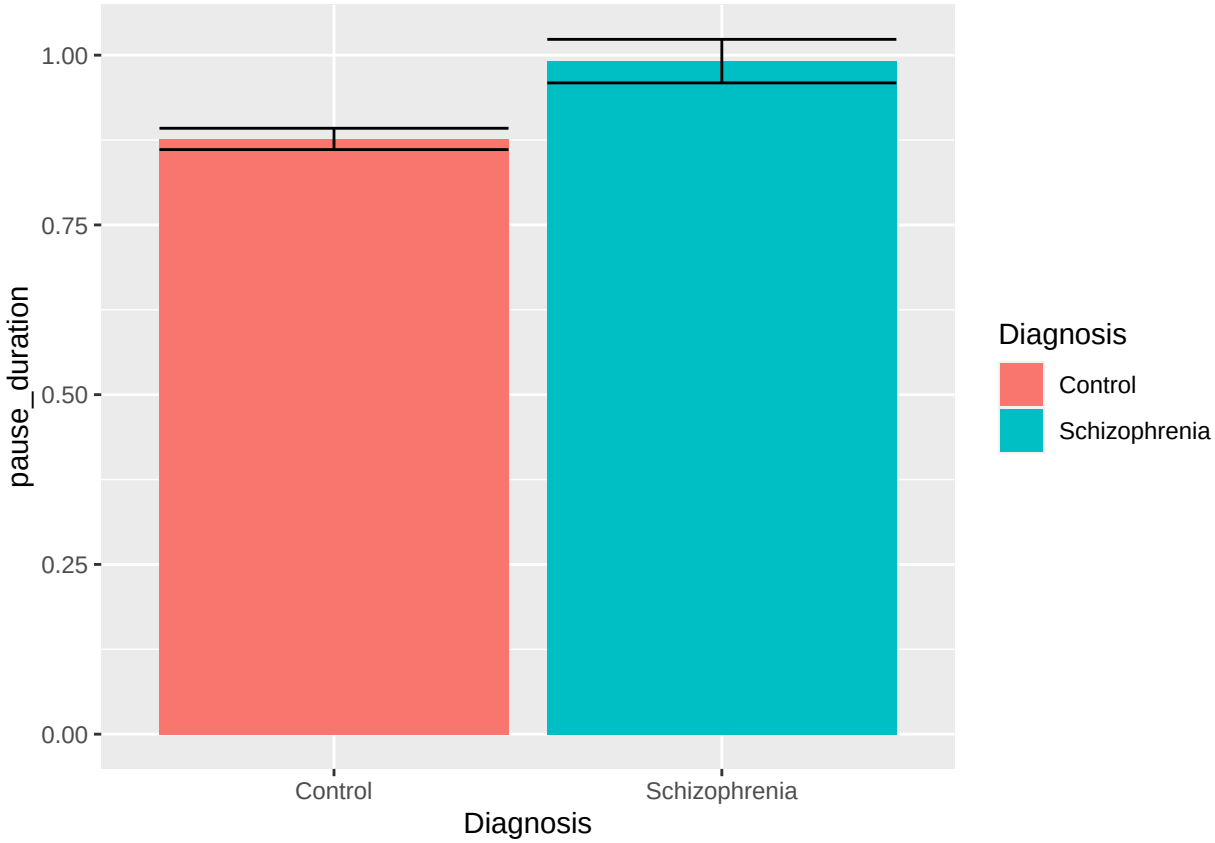
```
## No summary function supplied, defaulting to `mean_se()`
```




```
#duration of pauses
ggplot(df, aes(x = Diagnosis, y = pause_duration, fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

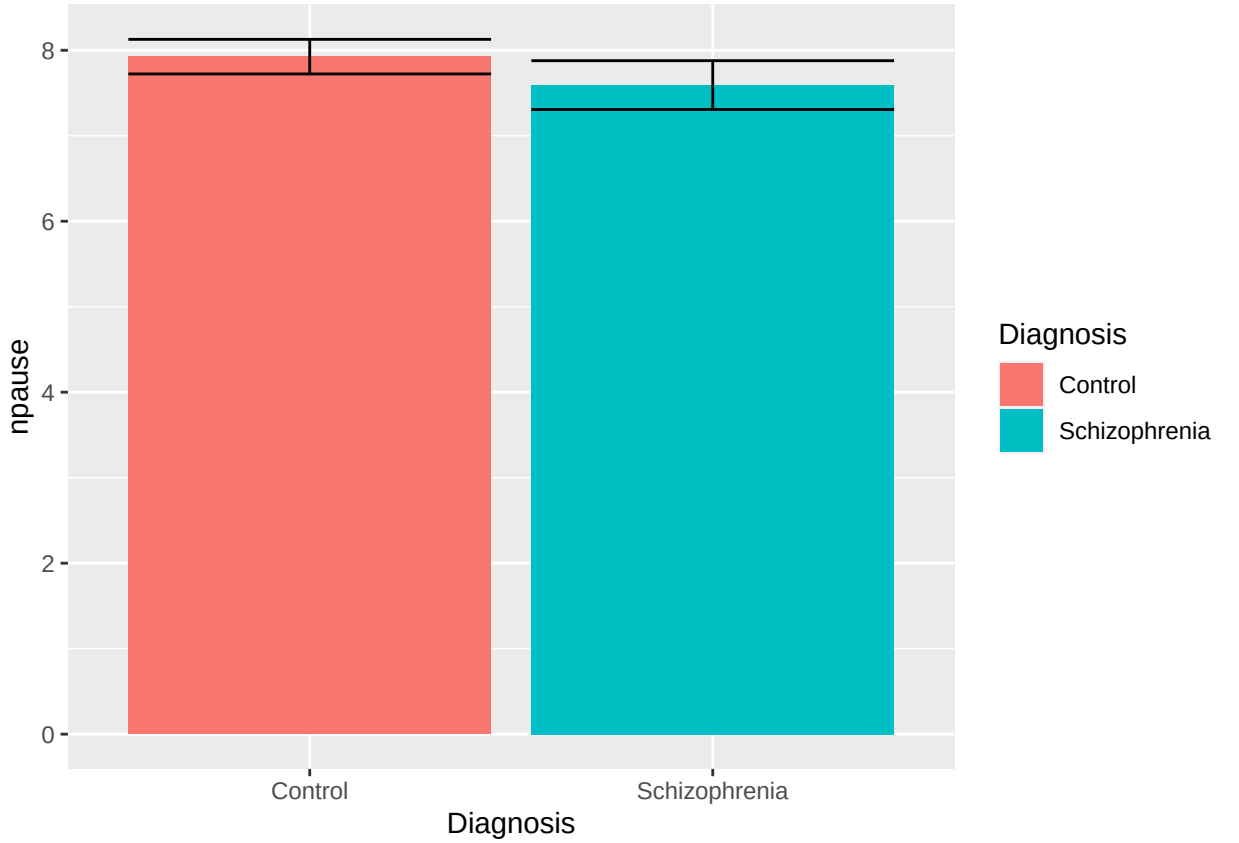
```
## No summary function supplied, defaulting to `mean_se()`
```




```
#number of pauses

ggplot(df, aes(x = Diagnosis, y = npause, fill = Diagnosis)) +
 geom_bar(stat = "summary", fun.y = mean)+
  geom_errorbar(stat = "summary", fun.data = mean_se)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
#Scaling everything

df <- df %>% mutate_if(is.numeric, scale)

#Models to run

  #pitch variability

pitch_variability <- lmerTest::lmer(pitch_IQR ~ 0 + Diagnosis + (1|ID), df, REML = FALSE)


  #proportion of spoken time

proportion_spoken <- lmerTest::lmer(prop_spoken_time ~ 0 + Diagnosis + (1|ID), df, REML = FALSE)

  #speech rate

speech_rate <- lmerTest::lmer(speechrate..nsyll.dur. ~ 0 + Diagnosis + (1|ID), df, REML = FALSE)

  #pause duration

pause_duration <- lmerTest::lmer(pause_duration ~ 0 + Diagnosis + (1|ID), df, REML = FALSE)


#Summaries
summary(pitch_variability)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: pitch_IQR ~ 0 + Diagnosis + (1 | ID)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   4562.8   4584.9  -2277.4   4554.8     1888
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0664 -0.1822 -0.0751  0.0191 13.0601
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.5106   0.7146
##  Residual             0.4986   0.7062
## Number of obs: 1892, groups:  ID, 221
##
## Fixed effects:
##                          Estimate Std. Error        df t value Pr(>|t|)
## DiagnosisControl          0.13502    0.07010 219.14527   1.926   0.0554 .
## DiagnosisSchizophrenia  -0.12975    0.07365 218.82424  -1.762   0.0795 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             DgnssC
## DgnssSchzph 0.000
```

```
summary(proportion_spoken)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: prop_spoken_time ~ 0 + Diagnosis + (1 | ID)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   4418.0   4440.2  -2205.0   4410.0     1888
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0197 -0.5552  0.0309  0.5443  4.2187
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.5453   0.7385
##  Residual             0.4542   0.6740
## Number of obs: 1892, groups:  ID, 221
##
## Fixed effects:
##                          Estimate Std. Error        df t value Pr(>|t|)
## DiagnosisControl          0.09402    0.07189 220.74153   1.308    0.192
## DiagnosisSchizophrenia  -0.10495    0.07554 220.45833  -1.389    0.166
##
```

```
## Correlation of Fixed Effects:
##             DgnssC
## DgnssSchzph 0.000
```

```
summary(speech_rate)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: speechrate..nsyll.dur. ~ 0 + Diagnosis + (1 | ID)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   4605.0   4627.2  -2298.5   4597.0     1888
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5167 -0.5840 -0.0123  0.5538  4.2390
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.4624   0.6800
##  Residual             0.5172   0.7191
## Number of obs: 1892, groups:  ID, 221
##
## Fixed effects:
##                          Estimate Std. Error        df t value Pr(>|t|)
## DiagnosisControl          0.15682    0.06721 221.17529   2.333   0.0205 *
## DiagnosisSchizophrenia  -0.16313    0.07062 220.81356  -2.310   0.0218 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             DgnssC
## DgnssSchzph 0.000
```

```
summary(pause_duration)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: pause_duration ~ 0 + Diagnosis + (1 | ID)
##    Data: df
##
##      AIC      BIC   logLik deviance df.resid
##   5211.7   5233.9  -2601.9   5203.7     1888
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3622 -0.3534 -0.1225  0.2088 12.0438
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.1822   0.4268
##  Residual             0.8085   0.8991
```

```
## Number of obs: 1892, groups:  ID, 221
##
## Fixed effects:
##                          Estimate Std. Error        df t value Pr(>|t|)
## DiagnosisControl        -0.07543    0.04896 225.24052  -1.541    0.125
## DiagnosisSchizophrenia   0.08089    0.05140 224.10574   1.574    0.117
##
## Correlation of Fixed Effects:
##             DgnssC
## DgnssSchzph 0.000
```

N.B. Remember to save the acoustic features of voice in a separate file, so to be able to load them next time

Reminder of the report to write

Part 1 - Can we find a difference in acoustic features in schizophrenia?

1) Describe your sample number of studies, number of participants, age, gender, clinical and cognitive features of the two groups. Furthemore, critically assess whether the groups (schizophrenia and controls) are balanced. N.B. you need to take studies into account.

2) Describe the acoustic profile of a schizophrenic voice: which features are different? E.g. People with schizophrenia tend to have high-pitched voice, and present bigger swings in their prosody than controls. N.B. look also at effect sizes. How do these findings relate to the meta-analytic findings?

3) Discuss the analysis necessary to replicate the meta-analytic findings Look at the results reported in the paper (see meta-analysis in the readings) and see whether they are similar to those you get. 3.1) Check whether significance and direction of the effects are similar 3.2) Standardize your outcome, run the model and check whether the beta's is roughly matched (matched with hedge's g) which fixed and random effects should be included, given your dataset? E.g. what about language and study, age and gender? Discuss also how studies and languages should play a role in your analyses. E.g. should you analyze each study individually? Or each language individually? Or all together? Each of these choices makes some assumptions about how similar you expect the studies/languages to be.

- Your report should look like a methods paragraph followed by a result paragraph in a typical article (think the Communication and Cognition paper)