

Assignment 3 - Part 1 - Voice In Schizophrenia

The Eyes of Kasparov

October 06, 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. Furthermore, 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>
 - VerballIQ: https://en.wikipedia.org/wiki/Wechsler_Adult_Intelligence_Scale
 - NonVerballIQ: 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)

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 absolute 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.

```
library(tidyverse)

## -- Attaching packages -----
----- tidyverse 1.3.0 --

## v ggplot2 3.3.2      v purrr    0.3.4
## v tibble  3.0.3      v dplyr    1.0.0
## v tidyr   1.1.2      v stringr  1.4.0
## v readr   1.3.1      v forcats  0.5.0

## -- Conflicts -----
----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(fs)

get_study_files <- function(path = "./pitch", studies = "123") {
  pattern <- paste0("Study[", studies, "]")
  dir_ls(path = path, regexp = pattern)
}

read_pitch <- function(filename) {
  # Load data
  read_tsv(filename) %>%
    mutate(file_name = basename(filename))
  # parse filename to extract study, diagnosis, subject and trial

  # extract pitch descriptors (mean, sd, iqr, etc)

  # combine all this data in one dataset
}
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

## Parsed with column specification:
## cols(
##   time = col_double(),
##   f0 = col_double()
## )

summarised_pitch_data <- pitch_data %>%
  group_by(file_name) %>%
  summarise(freq_mean = mean(f0),
            freq_sd = sd(f0),
            freq_iqr = IQR(f0)) %>%
  mutate(study_nr = parse_number(str_extract(file_name, regex("Study\\d"))),
         diagnosis = str_extract(file_name, regex("D\\d")),
         Participant = parse_number(str_extract(file_name, regex("S\\d+"))),
         trial_nr = parse_number(str_extract(file_name, regex("T\\d+")))) %>%
  mutate(soundname = str_extract(file_name, "Study\\dD\\dS\\d+T\\d+"))

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

demo_data <-
read_delim("https://www.dropbox.com/s/e2jy5fyac18zld7/DemographicData.csv?dl=1",
delim=";") %>%
  filter(Study < 5) %>%
  mutate(real_ID = if_else(Diagnosis == "Control", paste0(Participant, "C"),
paste0(Participant, "S")))

## Parsed with column specification:
## cols(
##   Study = col_double(),
##   Language = col_character(),
##   Diagnosis = col_character(),
##   Participant = col_double(),
##   Gender = col_character(),
##   Age = col_double(),
##   Education = col_double(),
##   SANS = col_double(),
##   SAPS = col_double(),
##   VerbalIQ = col_double(),
##   NonVerbalIQ = col_double(),
##   TotalIQ = col_double()
## )

art_data <-
read_delim("https://www.dropbox.com/s/vuyol7b575xdkjm/Articulation.txt?dl=1",
delim=",")

## Parsed with column specification:
## cols(
##   soundname = col_character(),

```

```
## ` nsyll` = col_character(),
## ` npause` = col_character(),
## ` dur (s)` = col_character(),
## ` phonationtime (s)` = col_character(),
## ` speechrate (nsyll/dur)` = col_character(),
## ` articulation rate (nsyll / phonationtime)` = col_character(),
## ` ASD (speakingtime/nsyll)` = col_character()
## )

View(demo_data)

master_data <- summarised_pitch_data %>%
  left_join(art_data) %>%
  # create ID
  mutate(real_ID = if_else(diagnosis == "D0", paste0(Participant, "C"),
paste0(Participant, "S"))) %>%
  left_join(demo_data, by="real_ID")

## Joining, by = "soundname"

write_csv(master_data, "finalish_data.csv")
```

Now we need to describe our sample

First look at the missing data: we should exclude all recordings for which we do not have complete data. Then count the participants and recordings 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

```
master_data <- read_csv("finalish_data.csv")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   file_name = col_character(),
##   diagnosis = col_character(),
##   soundname = col_character(),
##   real_ID = col_character(),
##   Language = col_character(),
##   Diagnosis = col_character(),
##   Gender = col_character()
## )

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

filtered_master_data <- master_data %>%
  drop_na()
```

```

filtered_master_data %>%
  group_by(Diagnosis) %>%
  summarise(num_male = sum(Gender == "M"),
            num_female = sum(Gender == "F"),
            mean_age = mean(Age),
            sd_age = sd(Age),
            mean_SANS = mean(SANS),
            sd_SANS = sd(SANS),
            mean_SAPS = mean(SAPS),
            sd_SAPS = sd(SAPS))

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

## # A tibble: 2 x 9
##   Diagnosis num_male num_female mean_age sd_age mean_SANS sd_SANS mean_SAPS
##   <chr>      <int>      <int>    <dbl> <dbl>    <dbl>  <dbl>    <dbl>
## 1 Control      317        215    23.0   3.35      0      0      0
## 2 Schizoph~    308        206    23.0   3.39     10.2   4.65    11.9
## # ... with 1 more variable: sd_SAPS <dbl>

plot_df <- master_data %>%
  select(Diagnosis, Age, Gender, study_nr) %>%
  drop_na()

age_plot <- ggplot(plot_df, aes(x = Diagnosis, y = Age)) +
  geom_boxplot() +
  stat_summary(fun="mean") +
  facet_grid(~study_nr) +
  theme_bw() +
  labs(title = "Age distribution", subtitle = "By study and diagnosis")
age_plot

## Warning: Removed 2 rows containing missing values (geom_segment).

## Warning: Removed 2 rows containing missing values (geom_segment).

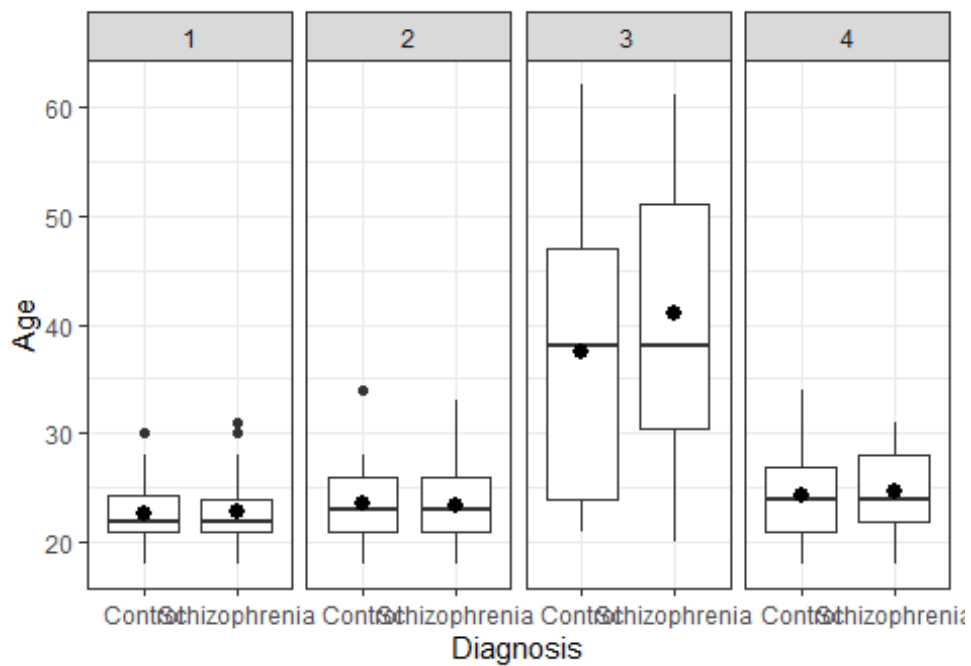
## Warning: Removed 2 rows containing missing values (geom_segment).

## Warning: Removed 2 rows containing missing values (geom_segment).

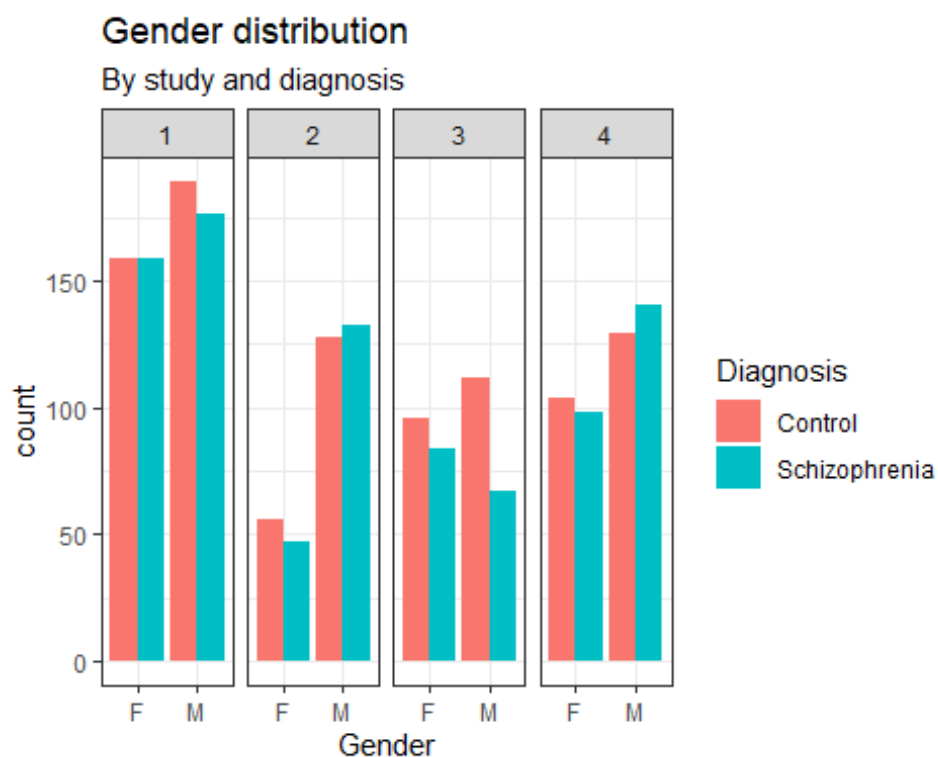
```

Age distribution

By study and diagnosis



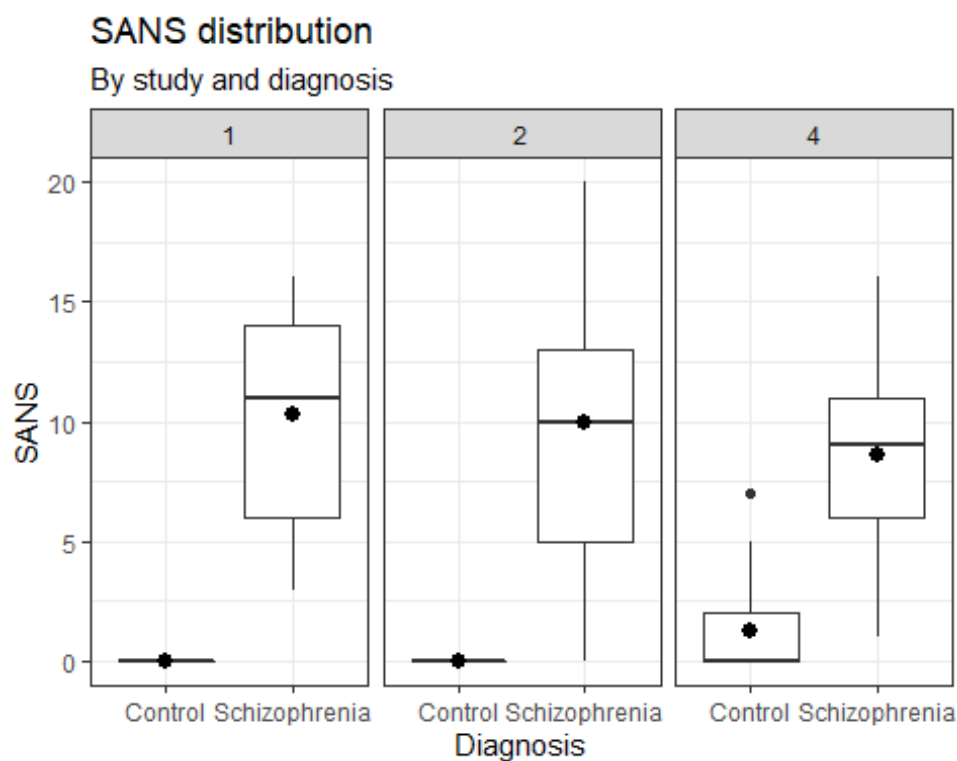
```
ggplot(plot_df, aes(y = Gender, fill = Diagnosis)) +
  geom_bar(position="Dodge") +
  facet_grid(cols = vars(study_nr)) +
  theme_bw() +
  coord_flip() +
  labs(title = "Gender distribution", subtitle = "By study and diagnosis")
```



```
new_plot_df <- master_data %>%
  select(Diagnosis, Age, Gender, study_nr, SANS, SAPS) %>%
  drop_na()

SANS_plot <- ggplot(new_plot_df, aes(x = Diagnosis, y = SANS)) +
  geom_boxplot() +
  stat_summary(fun="mean") +
  facet_grid(~study_nr) +
  theme_bw() +
  labs(title = "SANS distribution", subtitle = "By study and diagnosis")
SANS_plot

## Warning: Removed 2 rows containing missing values (geom_segment).
## Warning: Removed 2 rows containing missing values (geom_segment).
## Warning: Removed 2 rows containing missing values (geom_segment).
```



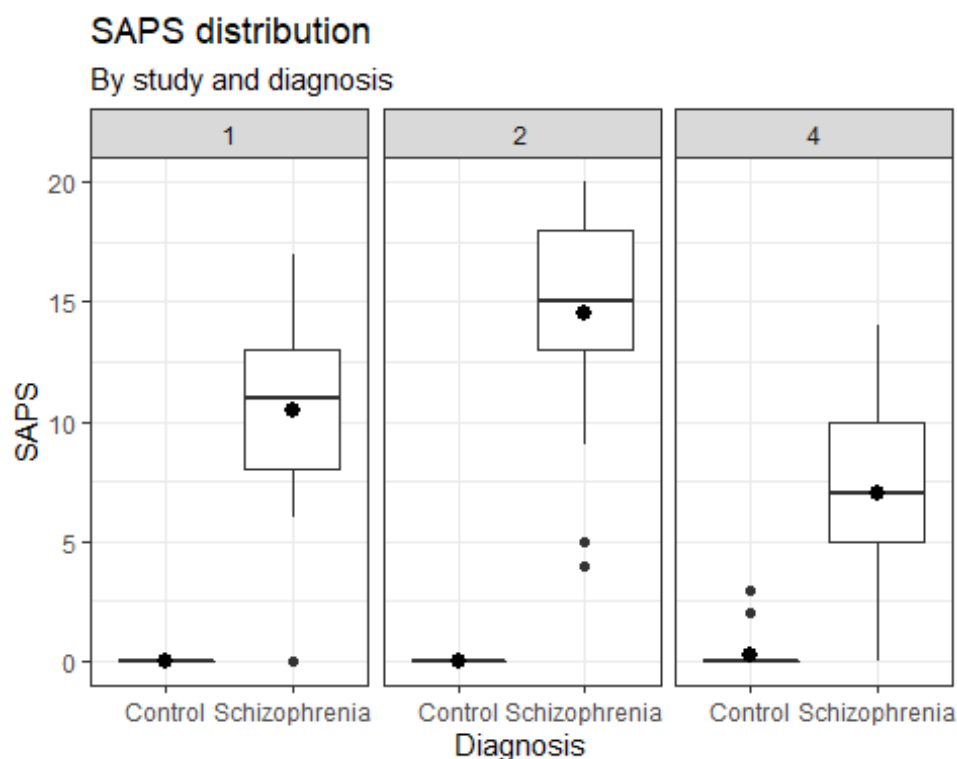
```
SAPS_plot <- ggplot(new_plot_df, aes(x = Diagnosis, y = SAPS)) +
  geom_boxplot() +
  stat_summary(fun="mean") +
  facet_grid(~study_nr) +
  theme_bw() +
  labs(title = "SAPS distribution", subtitle = "By study and diagnosis")
```

SAPS_plot

Warning: Removed 2 rows containing missing values (geom_segment).

Warning: Removed 2 rows containing missing values (geom_segment).

Warning: Removed 2 rows containing missing values (geom_segment).



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.

```
normalize <- function(x) {
  (x - mean(x, na.rm = T)) / sd(x, na.rm = T)
}

model_data <- master_data %>%
  select(articulation_rate = "articulation rate (nsyll / phonationtime)",
         real_ID, Diagnosis, diagnosis,
         Participant = Participant.x,
         pitch_variability = freq_iqr,
         npause,
         spoken_duration = "phonationtime (s)",
         speech_rate = "speechrate (nsyll/dur)",
         duration = "dur (s)", study_nr, trial_nr
  ) %>%
  mutate(pause_duration = (duration - spoken_duration) / npause,
         pause_duration = replace(pause_duration, is.infinite(pause_duration),
NA),
         study_nr = as_factor(study_nr),
         trial_nr = as_factor(trial_nr),
         proportion_spoken = spoken_duration / duration) %>%
  # Normalize all numeric columns
  mutate(across(where(is.numeric), normalize)) %>%
  mutate(trial_nr = as.integer(trial_nr))
```

Modelling: pitch variability

```
pitch_variability <- lmerTest::lmer(pitch_variability ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr), data=model_data)
```

```
summary(pitch_variability)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: pitch_variability ~ diagnosis + (1 | study_nr) + (1 | real_ID) +
## (1 | trial_nr)
## Data: model_data
##
## REML criterion at convergence: 4578.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.0462 -0.1916 -0.0702  0.0292 13.0259
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## real_ID  (Intercept)  0.508332  0.71297
```

```
## trial_nr (Intercept) 0.002006 0.04478
## study_nr (Intercept) 0.009713 0.09855
## Residual            0.496623 0.70471
## Number of obs: 1900, groups: real_ID, 222; trial_nr, 10; study_nr, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   0.13760    0.08679   7.34788   1.585   0.1549
## diagnosisD1  -0.26209    0.10142  216.49165  -2.584   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr)
## diagnosisD1 -0.551
```

modeling proportion of spoken time (PST)

```
PST_model <- lmerTest::lmer(proportion_spoken ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr), control=lme4::lmerControl(optimizer="bobyqa"),
data=model_data)
```

```
summary(PST_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: proportion_spoken ~ diagnosis + (1 | study_nr) + (1 | real_ID) +
##          (1 | trial_nr)
## Data: model_data
## Control: lme4::lmerControl(optimizer = "bobyqa")
##
## REML criterion at convergence: 4417
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.0137 -0.5437  0.0291  0.5581  4.1301
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## real_ID  (Intercept)  0.485243  0.6966
## trial_nr (Intercept)  0.002116  0.0460
## study_nr (Intercept)  0.101646  0.3188
## Residual                    0.452606  0.6728
## Number of obs: 1900, groups: real_ID, 222; trial_nr, 10; study_nr, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   0.08169    0.17410   3.43891   0.469   0.667
## diagnosisD1  -0.16121    0.09900  216.96116  -1.628   0.105
```

```
##
## Correlation of Fixed Effects:
##           (Intr)
## diagnosisD1 -0.268
```

Speech rate

```
speech_rate_model <- lmerTest::lmer(speech_rate ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr), control=lme4::lmerControl(optimizer="bobyqa"),
data=model_data)
```

```
summary(speech_rate_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: speech_rate ~ diagnosis + (1 | study_nr) + (1 | real_ID) + (1 |
##   trial_nr)
##   Data: model_data
## Control: lme4::lmerControl(optimizer = "bobyqa")
##
## REML criterion at convergence: 4593.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.4566 -0.5746 -0.0185  0.5590  4.0812
##
## Random effects:
##   Groups   Name                Variance Std.Dev.
##   real_ID  (Intercept)  0.41523   0.64438
##   trial_nr (Intercept)  0.00917   0.09576
##   study_nr (Intercept)  0.08136   0.28524
##   Residual                    0.50890   0.71337
## Number of obs: 1900, groups:  real_ID, 222; trial_nr, 10; study_nr, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   0.1577     0.1595    3.7009   0.989  0.38299
## diagnosisD1  -0.2967     0.0930   217.4500  -3.190  0.00163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## diagnosisD1 -0.275
```

modelling pause duration

```
pause_duration_model <- lmerTest::lmer(pause_duration ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr), control=lme4::lmerControl(optimizer="bobyqa"),
data=model_data)
```

```
summary(pause_duration_model)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: pause_duration ~ diagnosis + (1 | study_nr) + (1 | real_ID) +
## (1 | trial_nr)
## Data: model_data
## Control: lme4::lmerControl(optimizer = "bobyqa")
##
## REML criterion at convergence: 4619.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8877 -0.3891 -0.1262  0.1969 12.4058
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## real_ID  (Intercept)  0.372728  0.61051
## trial_nr  (Intercept)  0.011362  0.10659
## study_nr  (Intercept)  0.009017  0.09496
## Residual                    0.643396  0.80212
## Number of obs: 1761, groups: real_ID, 222; trial_nr, 10; study_nr, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  -0.12757    0.08586    7.15957  -1.486  0.17999
## diagnosisD1   0.28184    0.09109   204.26426   3.094  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr)
## diagnosisD1 -0.496
```

Calculating effect sizes (Hedge's g)

```
#install.packages("esc")
library(esc)
library(broom.mixed)

## Registered S3 method overwritten by 'broom.mixed':
##   method      from
## tidy.gamlss broom

library(magrittr)

##
## Attaching package: 'magrittr'
```

```

## The following object is masked from 'package:purrr':
##
##   set_names

## The following object is masked from 'package:tidyr':
##
##   extract

calc_hedge_g <- function(mmodel, group="diagnosis") {
  # Get fixed effects
  feffects <- mmodel %>%
    tidy %>%
    filter(effect == "fixed" & term != "(Intercept)") %>%
    select(term, estimate, std.error)

  #
  group_counts <- mmodel %>%
    augment %>%
    group_by(eval(parse(text=group))) %>%
    summarise(group_count = n()) %>%
    spread(key=1, value=2) %>%
    mutate(total = rowSums(select(., everything())))

  final_data <- feffects %>%
    cbind(group_counts) %>%
    mutate(sd_val = std.error * sqrt(total))

  final_data %$%
    esc_beta(estimate, sd_val, D0, D1, es.type = "g")
}

calc_hedge_g(pitch_variability)

## `summarise()` ungrouping output (override with `.groups` argument)
##
## Effect Size Calculation for Meta Analysis
##
##   Conversion: standardized regression coefficient to effect size Hedges' g
##   Effect Size: -0.5433
##   Standard Error: 0.0468
##   Variance: 0.0022
##   Lower CI: -0.6350
##   Upper CI: -0.4516
##   Weight: 457.0018

calc_hedge_g(PST_model)

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

```

```
##
## Effect Size Calculation for Meta Analysis
##
##      Conversion: standardized regression coefficient to effect size Hedges' g
##      Effect Size:  -0.3268
##      Standard Error:  0.0462
##      Variance:  0.0021
##      Lower CI:  -0.4174
##      Upper CI:  -0.2361
##      Weight: 467.6059
```

```
calc_hedge_g(speech_rate_model)
```

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

```
##
## Effect Size Calculation for Meta Analysis
##
##      Conversion: standardized regression coefficient to effect size Hedges' g
##      Effect Size:  -0.6215
##      Standard Error:  0.0470
##      Variance:  0.0022
##      Lower CI:  -0.7137
##      Upper CI:  -0.5293
##      Weight: 452.0488
```

```
calc_hedge_g(pause_duration_model)
```

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

```
##
## Effect Size Calculation for Meta Analysis
##
##      Conversion: standardized regression coefficient to effect size Hedges' g
##      Effect Size:  0.5888
##      Standard Error:  0.0488
##      Variance:  0.0024
##      Lower CI:  0.4931
##      Upper CI:  0.6846
##      Weight: 419.2786
```

Plotting the difference

```
plot_g_data <- tibble(model_name = c("pitch_variability", "percent duration
spoken", "speech_rate_model", "pause_duration"),
                      our_effect = c(-0.5433, -0.3268, -0.6215, 0.5888),
                      our_lower = c(-0.6350, -0.4174, -0.7137, 0.4931),
                      our_upper = c(-0.4516, -0.2361, -0.5293, 0.6846),
                      meta_effect = c(-0.55, -1.26, -0.75, 1.89),
                      meta_lower = c(-1.06, -2.26, -1.51, 0.72),
```

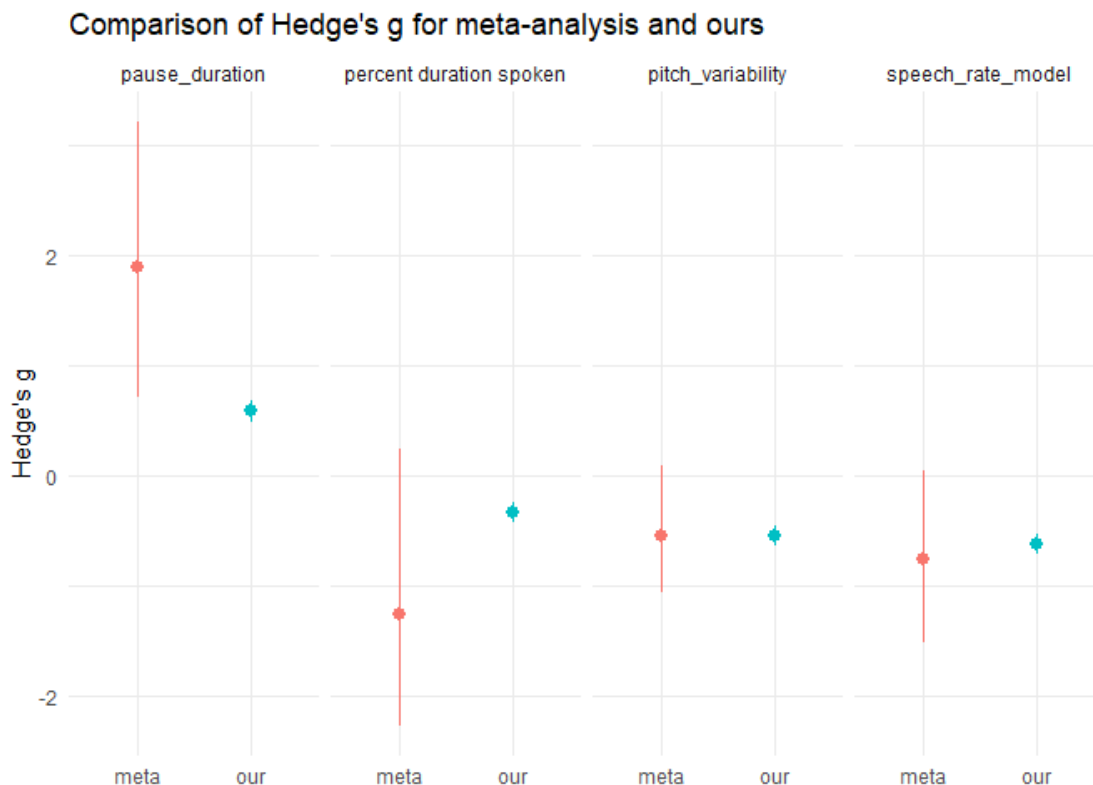
```

meta_upper = c(0.09, 0.25, 0.04, 3.21))

new_plot_data <- plot_g_data %>%
  pivot_longer(cols = 2:7) %>%
  separate(name, c("who", "type")) %>%
  pivot_wider(names_from=type, values_from=value)

ggplot(new_plot_data, aes(x=who, y=effect, colour=who)) +
  geom_point() +
  geom_pointrange(aes(ymin=lower, ymax=upper)) +
  facet_grid(cols=vars(model_name)) +
  theme_minimal() +
  labs(title = "Comparison of Hedge's g for meta-analysis and ours", x=NULL, y =
"Hedge's g") +
  theme(legend.position = "none")

```



Cross validation

```

library(groupdata2) # Shout out to Ludvig Olsen
library(cvms) # Even more shout out to Ludvig Olsen!

fold_model_data <- model_data %>%
  mutate(real_ID = as_factor(real_ID)) %>%

```



```

fold(
  data = ., k = 10,
  cat_col = 'diagnosis',
  id_col = 'real_ID',
  num_fold_cols = 3,
  handle_existing_fold_cols = "keep"
)

mixed_model_formulas <- c("speech_rate ~ diagnosis + (1|study_nr) + (1|real_ID) +
(1|trial_nr)",
                          "pitch_variability ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr)",
                          "pause_duration ~ diagnosis + (1|study_nr) + (1|real_ID)
+ (1|trial_nr)",
                          "proportion_spoken ~ diagnosis + (1|study_nr) +
(1|real_ID) + (1|trial_nr)")

CV5 <- cross_validate(
  data = fold_model_data,
  formulas = mixed_model_formulas,
  control = lme4::lmerControl(optimizer="bobyqa"),
  fold_cols = paste0(".folds_", 1:3),
  family = 'gaussian',
)

## ---
## cross_validate(): boundary (singular) fit: see ?isSingular
## Note: Boundary (Singular) Fit Message
## For:
## Formula: pause_duration ~ diagnosis + (1|study_nr) + (1|real_ID) + (1|trial_nr)
## Fold column: .folds_2
## Fold: 7
## Hyperparameters: REML : FALSE, control : list(list(optimizer = "bobyqa",
restart_edge = TRUE, boundary.tol = 1e-05, calc.derivs = TRUE, use.last.params =
FALSE, checkControl = list(check.nobs.vs.rankZ = "ignore", check.nobs.vs.nlev =
"stop", check.nlev.gtreq.5 = "ignore", check.nlev.gtr.1 = "stop",
check.nobs.vs.nRE = "stop", check.rankX = "message+drop.cols", check.scaleX =
"warning", check.formula.LHS = "stop"), checkConv = list(check.conv.grad =
list(action = "warning", tol = 0.002, relTol = NULL), check.conv.singular =
list(action = "message",
##       tol = 1e-04), check.conv.hess = list(action = "warning", tol = 1e-06)),
optCtrl = list()), model_verbose : FALSE, family : gaussian, is_special_fn : TRUE

## ---
## cross_validate(): boundary (singular) fit: see ?isSingular
## Note: Boundary (Singular) Fit Message

```

```

## For:
## Formula: pause_duration ~ diagnosis + (1|study_nr) + (1|real_ID) + (1|trial_nr)
## Fold column: .folds_3
## Fold: 2
## Hyperparameters: REML : FALSE, control : list(list(optimizer = "bobyqa",
restart_edge = TRUE, boundary.tol = 1e-05, calc.derivs = TRUE, use.last.params =
FALSE, checkControl = list(check.nobs.vs.rankZ = "ignore", check.nobs.vs.nlev =
"stop", check.nlev.gtreq.5 = "ignore", check.nlev.gtr.1 = "stop",
check.nobs.vs.nRE = "stop", check.rankX = "message+drop.cols", check.scaleX =
"warning", check.formula.LHS = "stop"), checkConv = list(check.conv.grad =
list(action = "warning", tol = 0.002, relTol = NULL), check.conv.singular =
list(action = "message",
##      tol = 1e-04), check.conv.hess = list(action = "warning", tol = 1e-06)),
optCtrl = list())), model_verbose : FALSE, family : gaussian, is_special_fn : TRUE

## ---
## cross_validate(): boundary (singular) fit: see ?isSingular
## Note: Boundary (Singular) Fit Message
## For:
## Formula: pitch_variability ~ diagnosis + (1|study_nr) + (1|real_ID) +
(1|trial_nr)
## Fold column: .folds_1
## Fold: 5
## Hyperparameters: REML : FALSE, control : list(list(optimizer = "bobyqa",
restart_edge = TRUE, boundary.tol = 1e-05, calc.derivs = TRUE, use.last.params =
FALSE, checkControl = list(check.nobs.vs.rankZ = "ignore", check.nobs.vs.nlev =
"stop", check.nlev.gtreq.5 = "ignore", check.nlev.gtr.1 = "stop",
check.nobs.vs.nRE = "stop", check.rankX = "message+drop.cols", check.scaleX =
"warning", check.formula.LHS = "stop"), checkConv = list(check.conv.grad =
list(action = "warning", tol = 0.002, relTol = NULL), check.conv.singular =
list(action = "message",
##      tol = 1e-04), check.conv.hess = list(action = "warning", tol = 1e-06)),
optCtrl = list())), model_verbose : FALSE, family : gaussian, is_special_fn : TRUE

CV5

## # A tibble: 4 x 22
##   Fixed RMSE MAE `NRMSE(IQR)` RRSE RAE RMSLE AIC AICc BIC
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 diag~ 0.964 0.758 0.737 0.982 0.972 NaN 4142. 4142. 4175.
## 2 diag~ 0.942 0.454 4.07 1.01 1.10 0.484 4122. 4122. 4155.
## 3 diag~ 0.928 0.546 1.51 NaN NaN NaN 4157. 4157. 4190.
## 4 diag~ 0.971 0.761 0.810 0.990 0.998 NaN 3983. 3983. 4016.
## # ... with 12 more variables: Predictions <list>, Results <list>,
## # Coefficients <list>, Folds <int>, `Fold Columns` <int>, `Convergence
## # Warnings` <int>, `Singular Fit Messages` <int>, `Other Warnings` <int>,
## # `Warnings and Messages` <list>, Family <chr>, Dependent <chr>, Random <chr>

```

Testing RMSE on training data (BADBADNOTGOOD, but maybe alright)

```
get_train_rmse <- function(model, outcome_var) {  
  pred_dat <- model %>%  
    augment %>%  
    select(outcome_var, .fitted)  
  
  Metrics::rmse(pred_dat[[outcome_var]], pred_dat[[".fitted"]])  
}  
  
get_train_rmse(pause_duration_model, "pause_duration")  
## Note: Using an external vector in selections is ambiguous.  
## i Use `all_of(outcome_var)` instead of `outcome_var` to silence this message.  
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.  
## This message is displayed once per session.  
## [1] 0.7580551  
  
get_train_rmse(PST_model, "proportion_spoken")  
## [1] 0.6355508  
  
get_train_rmse(pitch_variability, "pitch_variability")  
## [1] 0.6660046  
  
get_train_rmse(speech_rate_model, "speech_rate")  
## [1] 0.6745654
```

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. Furthermore, 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)