

Emil Trenckner Jessen (EJ)  
(au604547@uni.au.dk)  
Cognitive Science, Aarhus University  
Jens Chr. Skous Vej 2,  
8000 Aarhus, Denmark

Voice atypicalities in Schizophrenia; replication of machine learning approaches

Contents

[Abstract 2](#_Toc55908539)

[1. Introduction 3](#_Toc55908540)

[1.1 Research into voice atypicalities in schizophrenia 3](#_Toc55908541)

[1.2. Practical appliance of ML 4](#_Toc55908542)

[1.3 What this paper aims to do 4](#_Toc55908543)

[1.4 Notes for all sections below, here 5](#_Toc55908544)

[2. Materials and Methods 7](#_Toc55908545)

[2.1 Literature search (finding a study to replicate) 7](#_Toc55908546)

[2.1 Data 10](#_Toc55908547)

[2.1.1 Data sources 10](#_Toc55908548)

[2.1.2 Participants 10](#_Toc55908549)

[2.2.2 Procedure/task 11](#_Toc55908550)

[2.3 Preprocessing 12](#_Toc55908551)

[2.4.1 Cleaning of audio files 12](#_Toc55908552)

[2.4.2 Feature extraction from audio files 12](#_Toc55908553)

[2.4.3 Partitioning 13](#_Toc55908554)

[2.4.4 Normalization 13](#_Toc55908555)

[2.4 Feature selection using LASSO 14](#_Toc55908556)

[2.4.1 Motivation for feature selection using LASSO 14](#_Toc55908557)

[2.4.2 LASSO – L2 regularization 14](#_Toc55908558)

[2.4.3 Cross-validation on the training data 15](#_Toc55908559)

[3.6 Model tuning 16](#_Toc55908560)

[3.7 Predicting the holdout and evaluation metrics 16](#_Toc55908561)

[3.7 Software used 17](#_Toc55908562)

[4. Results 19](#_Toc55908563)

[4.1 Model predictions 19](#_Toc55908564)

[5. Discussion: 19](#_Toc55908565)

[5.1 ML results: 19](#_Toc55908566)

[5.2 Differences in original vs. this 19](#_Toc55908567)

[5.2.1 Feature selection methods 19](#_Toc55908568)

[5.3 Feature extraction method 20](#_Toc55908569)

[5.3 Task 20](#_Toc55908570)

[5.4 Sex bias 21](#_Toc55908571)

[5.4 In general 21](#_Toc55908572)

[5.4.1: Hard to replicate, given the sparse information given 21](#_Toc55908573)

[5.4.2 What new knowledge has been acquired? 21](#_Toc55908574)

[5.6 Where should the research go from here? 22](#_Toc55908575)

[6. Conclusion 22](#_Toc55908576)

[7. References 22](#_Toc55908577)

To do list:  
  
1. Methods er bade I nutid og datid. Ret det til det rigtige??

# Abstract

**Actual paper:**  
  
**Notes for section:**

**(Fabio exercise):**

Can machine learning (ML) applied to voice data be used as a tool to help diagnose people with schizophrenia? Numerous studies have shown high ML accuracy when classifying schizophrenia, but the ways in which they do so differ widely, as concluded in the latest metastudy within the field. Little work has been done to replicate these previous ML methods on new data, and there is currently no consensus on which methods should be used.  
This study replicated two promising ML studies on new data, using an improved validation technique and an inclusion of sensitivity and specificity rates. Accuracy rates found through replication were dissimilar to the original studies, with study X\* and study Y\* having overall accuracy rates for classification at 60% and 67%. In other words a drop of 6 and 3 percentage points for the two studies, respectively. Through discussion, this study has found that the difference in scores in the replication points toward low ecological validity and robustness. The rest of the litterature was also discussed, and I found that the widely heterogeneous results within the field indicate similar trends.  
As a consequence, this study has attempted to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research. Finally this paper advocates for a more open and cumulative scientific community.

How to write abstract:  
<https://blackboard.au.dk/bbcswebdav/pid-2793891-dt-content-rid-9152972_1/courses/BB-Cou-Hold-36086/L1%20-%20Getting%20started.pdf>  
p. 14 - p. 18

**Keywords:** Schizophrenia, Machine Learning, Voice, SVM

# 1. Introduction

## 1.1 Research into voice atypicalities in schizophrenia

**Actual paper:**  
  
**Notes for section:**

Voice atypicalities in SZ’s have always been known (Bleuler, 1911; Kraepelin, 1919).  
Schizophrenia has certain distinctive features vocally. Qualitatively the atypicalities have been described using numerous different terms (Alogia, blunt affect, "poverty of speech", "latency of speech", increased pauses, distinctive tone, intensity of voice etc.).  
  
  
Effect sizes of acoustic features is partially determined by task (and task difficulty):  
\*\*sas  
  
  
The atypicalities have been studied using 3 methods;  
Qualitative perceptual ratings, quantitative acoustic analysis and ML investigations.  
Qualitative perceptual ratings have found relatively robust differences in voice between SZ and TD. Relying on raters to assess perceptual differences has some limitations. A feature such as “latency of speech” is interpretable and must be rated on the basis of human intuition – this requires comprehensive training for the rater. Moreover, the complex interplay between multiple acoustic features is hardly very accessible, even given proper and rigorous training.  
  
Quantitative acoustic analyses have identified acoustic features on the basis of automated processes, leaving the assessment of the acoustic features more reliable. Using automation, the features of a set of voice data will identical over multiple feature detections, given the same feature detection hard- and software.  
Here, fewer robust differences were found with varying effect sizes and direction, depending on the features investigated (Cohen et al., 2014; https://www.biorxiv.org/content/10.1101/583815v4.full.pdf).  
  
Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

## 1.2. Practical appliance of ML

**Actual paper:**  
  
**Notes for section:**

ML can perhaps help with showing:  
a) Severity of schizophrenic symptoms  
b) Diagnosis, schizophrenia  
  
Practical applications:   
1. Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

2. Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.  
"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)Va [(Bush et al., 1998)](https://www.zotero.org/google-docs/?KFKj12).  
  
but no validation between datasets exist. Moreover, language differences?

## 1.3 What this paper aims to do

**Actual paper:**  
  
**Notes for section:**

Need for replication. Need for clear pipeline within the field

On prediction of severity of clinical features from acoustic measures:  
(Püschel et al., 1998)

## 1.4 Notes for all sections below, here

**Actual paper:**  
  
**Notes for section:**

A nificantly higher brain activation ex

**History of the project:**

1. Schizophrenia has certain distinctive features vocally. (Alogia, blunt affect, "poverty of speech", "latency of speech" etc.). This has been known since forever (Bleuler, 1911; Kraepelin, 1919).

Voice atypicalities have been studied using 3 methods. Qualitative perceptual ratings, quantitative acoustic analysis and ML investigations.

Qualitative perceptual ratings have found robust differences between SZ and TD.

Quantitative acoustic analyses have found fewer robust differences, with varying effect sizes and sometimes direction.

ML investigations have found promising results, but no validation between datasets and/or languages exist.

Overall:

We don't know which features proves to have differences between SZ and TD

The litt. is a mess - results in different directions.

Different ways of conducting studies

Systematic review, for both ASD and Schizo. Finds all acoustic measures (not perceptual). They have found everything that is in the litterature.

Might have been additional new papers.

Replicate in a new corpus with 2 languages - even though metanalysis tells us it is robust. (These are the ones in Google Docs).

Other studies have used ML - but support vector machine, different data, some overfit.

1. There's already a metastudy on Schizophrenia; which found atypicalities on different voice/speaking parameters - with varying effect sizes.

Large heterogeneity between studies.

More demanding tasks meant larger effect sizes.

**Applicability of Bachelors project:**

Meta-science, open science.

Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.

"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)

Companies interested in this (Lasse Hansen), Switzerland Internship on this in depression

**Thesis statement idea 1 (Maries):**

This thesis aims to investigate the capabilities of existing machine-learning classifying individuals with ASD from acoustic features. We will review previous literature, extract strong voice-features and machine-learning models, and validate models on new data. We predict that support vector machine will achieve higher accuracy but will have less x and that naive bayes will x. Additionally, we predict that validation methods x,y,z will make results stronger in specific case/weaker generalization. By this, we will attempt to establish a procedure for machine-learning studies that achieve the most robust and ecologically valid measures.

**Thesis statement idea 2:**

This thesis aims to replicate two promising findings of machine learning classification of schizophrenia, using voice data. Since the litterature on the area has very hetereogeneous findings, I expect worse performance given the new data that I will test on. Given the inrobustness and low ecological validity of ML attempts, I will attempt to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research.

# 2. Materials and Methods

**Actual paper:**s

**Notes for section:**

General on how everything was carried out, just in a small condensed way.

Look at Ludvigs

## 2.1 Literature search (finding a study to replicate)

**Actual paper:**s

**Notes for section:**

A literature search for papers, dissertations and unpublished manuscripts was conducted for finding the paper(s) \* to replicate.

First and foremost, the complete list of papers listed in the meta-analysis \*\*\* (Fusaroli et al, 201?).

As their search was last updated as of April 12 2018, the search was continued from that date and forward in time by the use of search using Google Scholar on the Sep 15 2020, using the same search terms as Fusaroli et al.

(schizo\* AND machine learning AND prosody OR inflection OR intensity OR pitch OR fundamental frequency OR speech rate OR voice quality OR acoustic OR intonation OR vocal)

The complete list of articles were then manually screened - first by title, and since by content;

**Their criteria:**The literature search was conducted on Pubmed and Google Scholar, the latter including dissertations and unpublished manuscripts..  
  
prosody OR inflection OR intensity OR pitch OR fundamental frequency OR speech rate OR voice quality OR acoustic OR intonation OR vocal) AND (schizo\*).   
The search was conducted on August 21 2017, and updated on **April 12 2018**.   
  
We complemented the list by performing a backward and forward literature search: we screened the bibliography of the papers found, and the papers citing them as identified by Google Scholar. Articles were screened for eligibility by two authors (AP and AS).   
  
Study selection was conducted according to the following inclusion criteria:   
  
(a) empirical study, (b) quantification of acoustic features in the vocal production of participants with schizophrenia or schizoaffective disorder1 (c) sample including at least two individuals with schizophrenia or schizoaffective disorder, (d) inclusion of a non-clinical comparison group, or an assessment of variation in acoustic features in relation to severity of clinical features.   
  
Clinical comparison groups (e.g. with depression) were excluded because the limited number of studies did not permit meta-analytic estimations.   
  
Fig. 1 shows the flow-diagram of study selection. We report the assessment of the risk of bias in the Supplementary Materials.

Machine generated alternative text:
m 'Ching 
re 
In • 174) 
Fun-text 
In • 160) 
in qualitative 
Studi« in 
uded 
and abstract Stag 
Article not 
null-tut "ticks with 
. No Acoustic In = 
No data (n 16) 
he I thx 
no relation With In = 
study In 
- t" (n = 241 

**My search:**  
  
Google scholar search, conducted on 15-09-2020:

schizo\* AND machine learning AND prosody OR inflection OR intensity OR pitch OR fundamental frequency OR speech rate OR voice quality OR acoustic OR intonation OR vocal

## 2.1 Data

### 2.1.1 Data sources

**Actual paper:**The data used in this paper consists of speech recordings gathered from 4 separate studies, namely \*\*\*, \*\*\*, \*\* and \*\*\*.  
Although the data was acquired in separate studies, the speech data has a number of qualities which makes it suitable for combining into a single study.

Participants from all studies went through the same tasks; namely the Frith Happé animations task (FHA) (\* cite), (for description, see “ 2.2.2 Task “). All participant went through 8 such trials, with the exception of study 1 where the participants were presented with 10 trials.

Moreover, recording equipment and recording setting was unique to each study but remained constant within study. This results in data corpora of diverse speech recordings suitable for testing whether implementation of a certain method\* of machine learning proves to be versatile in its predictions when replicated.

**Notes for section:**

### 2.1.2 Participants

**Actual paper:**222 Danish participants were included in the 4 studies combined; Study 1 had 70 participants, study 2 had 46 participants, while study 3 and study 4 had 48 and 58 participants respectively.  
Out of the 222 participants \*\* were clinically diagnosed with either schizophrenia or schizoaffective disorder and will for the remainder of this paper be referred to as the patient group, while the remaining \*\* participants were typically developed individuals with no history of psychological disorders who were matched to the patient group on \*, \* and \*. The typically developed group will be referred to as controls. For further information on the participants, see table 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study | N | Diagnosis | N(females) | N(males) | Mean(age) | SD(age) | Range(age) |
| 1 | 70 | SZ | 16 | 18 | 22.8 | 3.13 | 18-31 |
| TD | 17 | 19 | 22.7 | 3.19 | 18-30 |
| 2 | 46 | SZ | 6 | 17 | 23.3 | 3.94 | 18-33 |
| TD | 7 | 16 | 23.7 | 3.61 | 18-34 |
| 3 | 48 | SZ | 11 | 8 | 40.8 | 12.4 | 20-61 |
| TD | 13 | 16 | 37.5 | 13.1 | 21-62 |
| 4 | 58 | SZ | 12 | 18 | 24.8 | 3.66 | 18-31 |
| TD | 13 | 15 | 24.4 | 4.65 | 18-34 |
|  |  |  |  |  |  |  |  |
| **Total** | **106** | **SZ** | **45** | **61** | **26.7** | **9.02** | **18-61** |
| **116** | **TD** | **50** | **66** | **26.7** | **9.22** | **18-62** |

Table 1:

*Demographic data on the gender and diagnosis within each of the different studies. N means number and SD standard deviation.*

**Notes for section:**

A

### 2.2.2 Procedure/task

**Actual paper:**The participants went through the Frith Happé animations task. This task consisted of watching a 2D top-view video of animated triangles. There were two distinct triangles; one large red and one small blue, both of which moved around on the screen and most videos furthermore contained an enclosure in the center of the video. There were three conditions with multiple videos for each condition:

1. Random movement sequences. There was no obvious interaction between the triangles and movement appears random. 2. Goal-directed (G-D) movement sequences. An interaction between the triangles in which actions are directed toward each other in order to achieve specific goals. 3. Mental interaction (ToM). An interaction between the triangles involving the manipulation of the emotions and thoughts of one triangle by the other.

After watching an animation from one of these conditions, the participants were interviewed and asked to describe what happened in the animation.

**Notes for section:**

Storytelling of triangle moving meaningfully or sometimes not. Storytelling monologues, no social component in this data. (Mid-level)

Study 1: 10 trials

Study 2: 8 trials

Study 3: 8 trials

Study 4: 8 trials

For each trial of each of the experiments, the participants were presented with a 2D top-view video of animated triangles. There were two distinct triangles; one large red and one small blue, both of which moved around on the screen and most videos furthermore contained an enclosure in the center of the video.

3 types of videos, over 8-10 trials.  
The types of videos were:

* Random movement sequences
  + (There is no obvious interaction between the triangles and movement appears random)
  + 2 levels
* Goal-directed movement sequences
  + (An interaction between the triangles in which actions are directed toward each other in order to achieve specific goals.)
  + 4 levels
* Mental interaction
  + (An interaction between the triangles involving the manipulation of the emotions and thoughts of one triangle by the other.)
  + 4 levels

“*A series of computer-presented animations was designed with the help of a graphic artist using the MacroMedia Director 4.0 programme for Power Macintosh. The animations showed one large red and one small blue triangle moving around the screen, which on most trials contained an enclosure. There were three conditions: Random, G-D, and ToM with four animations each in the G-D and ToM conditions, and two in the Random condition. These were taken from a larger set of sixteen animations. Animations that failed to convey the intended meaning to adults at initial piloting were discarded. Two of the unused animations were used as practice stimuli.*”

<https://www-sciencedirect-com.ez.statsbiblioteket.dk:12048/science/article/pii/S0885201400000149?casa_token=R0-Bnt_WS90AAAAA:_AYLWBR2cCCbJrZA6hIfFLnnwtv2QkyT-EK0LRufvfT0gfFEP_Ami4AAF1i8U1tziaIzQB9cLGaV>

## 2.3 Preprocessing

### 2.4.1 Cleaning of audio files

**Actual paper:**  
  
**Notes for section:**

<https://www.dropbox.com/s/nl1eoeeqr6ptp44/Ludvig%20Renbo%20Olsen_364085_278504_Hand-in_bachelor-thesis-ludvig-renbo-olsen.pdf?dl=0&fbclid=IwAR0s41VaiexJvHd4kCIZgpDZLNAQBcv9fOVNZyjhu4EHRCDQeqDWiAF7K3U>  
  
Eller i bachelormappen under filnavnet:  
"Ludvig Renbo Olsen"

Also ”cut” into a single answer for each. Not interviewer/interlocutor noise (dissimilarly to Chakraborty)

### 2.4.2 Feature extraction from audio files

**Actual paper:**The toolkit openSMILE 2.3.0 was used for extracting the features needed for the SVM classification algorithm. From within the openSMILE software package, the base-set configuration file of emotion recognition features called ‘emobase’ was chosen for feature extraction.

The feature set specified by emobase contains 988 features used for emotion recognition:

Intensity, Loudness, 12 MFCC’s, F0 Pitch, Probability of voicing, F0 envelope, 8 LSFs (Line

Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are then computed from all these previously mentioned low-level descriptors (LLD). Both the LLDs and their delta coefficients are smoothed by a moving average window that filters with a window size of 3 seconds. Furthermore, the following functionals are applied to the LLDs and the delta coefficients:

Max./Min. values and their respective relative position within input, range, arithmetic mean, 2 linear

regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis,

quartile 1-3, and 3 inter-quartile ranges.

This results in the feature set consisting of 988 features. In other words; 26 LLDs, a delta regression coefficient for each LLD and 19 functionals for each of the LLDs and for each of the delta regression coefficients (26 \* 2 \* 19 = 988).

The process of feature extraction was executed on each of the speech recordings, yielding a single feature vector for each trial of each participant.

**Notes for section:**

Måske til introduction?:

Open-Source Media Interpretation by Large feature-space Extraction (openSMILE) is a toolkit for audio feature extraction and classification of speech and music signals and is especially widely applied within the field of automatic recognition of emotion within speech.

### 2.4.3 Partitioning

**Actual paper:**To be able to evaluate the performance of the model the dataset was partitioned into a training set and a test set consisting of 80% and 20% of the total data, respectively. The partitioning was carried out using the package groupdata2 and was done semi-randomly. The partitioning kept each participant ID only within one of the two resulting training and test sets. This prevented leakage of information from the training set to the test set, which otherwise would have led to overfitting and finally unprecise values for the evaluation.  
Moreover, to avoid a skewed distribution of gender or diagnosis between sets (e.g. ending up with only males/controls in the test set as a result of a random partitioning), genders and controls/patients were evenly distributed in the partitioning.

**Notes for section:**

S

### 2.4.4 Normalization

**Actual paper:**All feature parameters were normalized using the min-max feature scaling formula in order to achieve a dataset with a common scale without losing information or distorting differences in the range of values.



To avoid overfitting as a result of carrying data from the test set to the training set, the normalization was carried out separately for the training and the testing set. The scaling used the min. and the max. value for each feature, only from the training set, both for the training and for the testing set. This had the advantage of having both the training and the test features on the same scale, while not letting information from the test set flow to the training set and is common practice when doing SVM machine learning classification.

**Notes for section:**

## 2.4 Feature selection using LASSO

### 2.4.1 Motivation for feature selection using LASSO

**Actual paper:**As the 988 acoustic features from the ‘emobase’ package were originally designed to distinguish emotions from speech, some of the features were bound to be redundant for the purpose of distinguishing between patients and controls. As a measure to counterfeit this, a rigorous feature selection method was applied to rid the model of superfluous features. This was done in order to simplify the model and thereby reduces both complexity, computational power needed to run the model and in order to improve both predictive power and interpretability of the classifier.

Feature selection was done through the use of L2 regularization. This is also called the Least Absolute Shrinkage and Selection Operator (LASSO) analysis regression, and to carry out this process, the ‘glmnet’ R Package was utilized for the purpose of this paper. ( \* cite (Friedman, Hastie, & Tibshirani, 2010).

\*\*\*\* Why not the others? \*\*\*

### 2.4.2 LASSO – L2 regularization

This method optimizes beta estimates for all parameters not only based on misclassification error but also adds a L2 regularization term. The latter adds a penalty to each beta estimate on the basis of a lambda value multiplied with the beta estimate.

The loss function used for finding parameter estimates using LASSO:  


Since this method requires a lambda value, the optimal lambda value also had to be found.

The lambda value producing the minimum value in the loss function (lambda.min) was first discovered. This was done by testing a range of lambda values using Leave-One-Out CV (LOO-CV). Subsequently the lambda value resulting in the fewest number of parameters within 1 SE from the lambda.min was chosen (lambda.se). The application of LASSO thus generates a list of parameter estimates and all those that have not been shrunken to zero are selected as features for predicting Schizophrenia from typically developing.

### 2.4.3 Cross-validation on the training data

Instead of just using LASSO method on the whole of the training set, cross-validation was applied in order to ensure the optimal number of features. To specify, the training data from the main partitioning was partitioned into 5 folds, with each of them appearing once as a test set in the subsequent cross-validation. The previously mentioned LASSO regularization process was thus carried out 5 times for each step in the cross-validation, resulting in 5 different feature sets.

The feature sets:  
\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

To evaluate which features were optimal, the feature sets were evaluated. As each feature set was based upon 4/5th, and the last 1/5th had not been presented to the model, the last 1/5th was used as a test set using an SVM classifier with a linear kernel. The software Visual Studio Code was used, using the package SciKitLearn.  
  
The test results from the different feature sets:

\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \*

### 2.4.4 Feature selection

f

**Notes for section:**

1. Partition training into 5 folds. (groupdata2 folds)  
  
2. For each training fold, take the other 4 folds combined and do glmnet on these. (using loocv for instance). Extract the lists of features.  
  
3. For each of these 5 lists of features, do SVM linear kernel. For each of these feature sets, the training sets will be the folds that the glmnet was bsaed on, and the test will be the one that wasn't.  
  
4. Get statistics on how well each set of features perform (test-scores)  
  
5. Send the 5 feature lists to Riccardo. Even better to add performance. Even better to also add coefficients of the different features.  
  
6. Combine the feature sets into 1 final feature set, using the help from Riccardo

### 3.6 Model tuning

**Actual paper:**  
  
**Notes for section:**

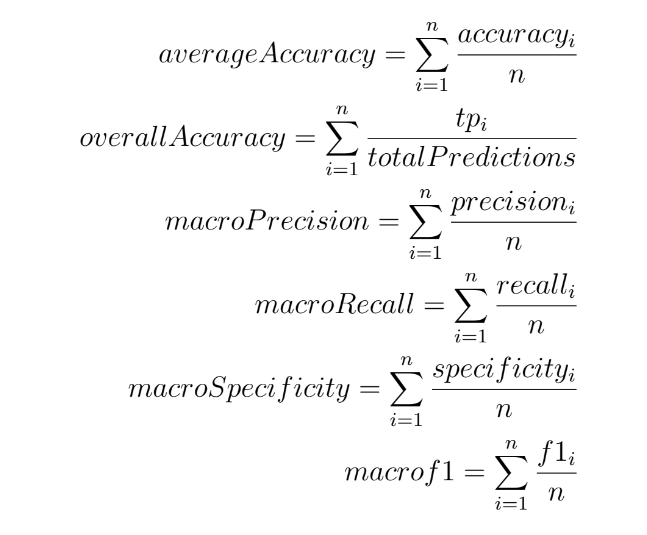
.s

### 3.7 Predicting the holdout and evaluation metrics

**Actual paper:**  
  
**Notes for section:**

SVM, linear kernel

Hav noget lignende det her med:



### 3.7 Software used

**Actual paper:**  
  
**Notes for section:**

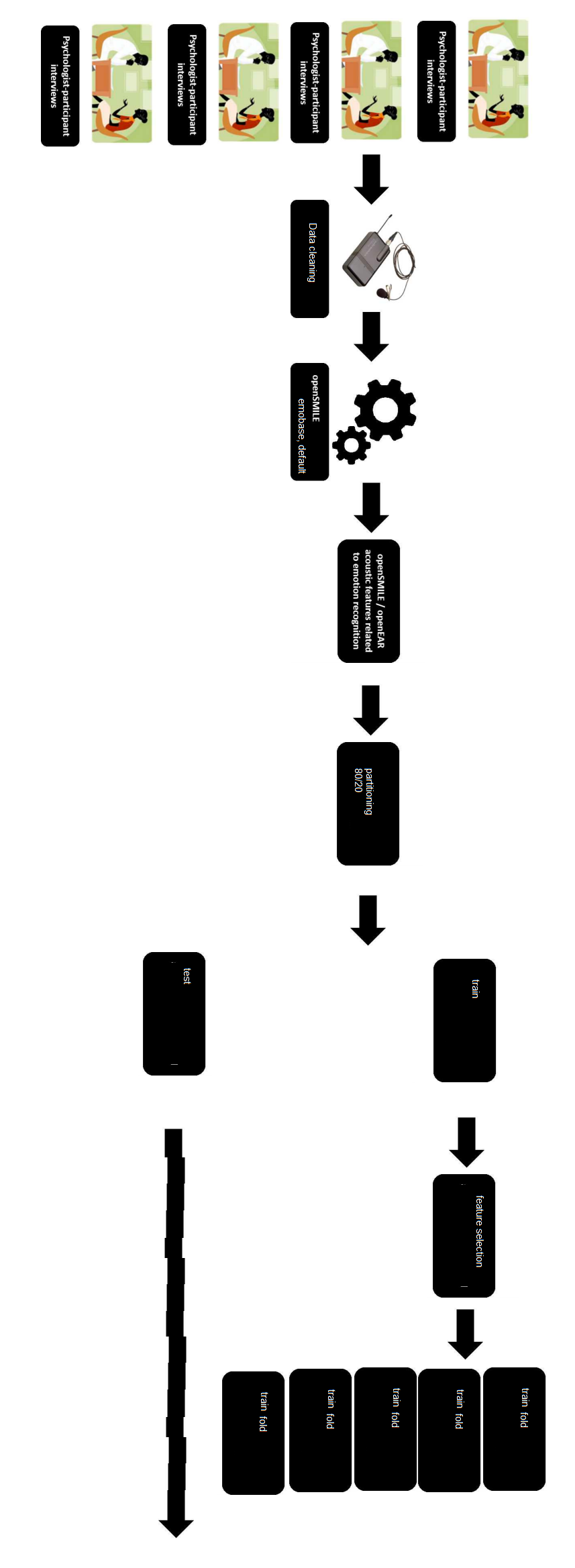
Python

Visual Studio Coder

R

Rstudio

OpenSMILE, emobase



# 4. Results

## 4.1 Model predictions

**Actual paper:**  
  
**Notes for section:**

s

# 5. Discussion:

## 5.1 ML results:

**Actual paper:**  
  
**Notes for section:**

s

## 5.2 Differences in original vs. this

### 5.2.1 Feature selection methods

**Actual paper:**  
  
**Notes for section:**

LASSO, not PCA  
  
Hard to replicate, given the sparse information on how PCA was used to feature select.

“*the features of the training set were ranked using one of the following techniques: F-score (ANOVA), χ 2 , Mutual Information, Pearson correlation, Principal Components, linear SVM, Decision Trees, and Random Forests. Subsequently, the optimal number of features were selected according to the previous ranking methods*”  
PCA used to rank? Most common method is that PCA is used for defining new features, namely PC1 + PC2 + ... +PCn, until some desired threshold of accumulated variance is met.

There’s also the possibility that it truly was used to rank, e.g. by looking at the features with least shared variance in the different principal components to avoid covarying features, but also here it is not possible to replicate 1-1. The method is still not specified

Regardless, of method used by Chakraborty et al, the method used here is good. And if the method using speech for classification truly is robust, then either would work. If these results truly are reliable and reliable, they shouldn’t be dependent on PCA/LASSO / whatever

Link of idea of PCA for feature selection. (starts at 3:50). It shows that there are different methods (example with gain, here)

[https://www.youtube.com/watch?v=YEDOSOd44bU&list=PLBv09BD7ez\_5\_yapAg86Od6JeeypkS4YM&index=2&frags=wn&ab\_channel=VictorLavrenko](https://www.google.com/url?q=https://www.youtube.com/watch?v%3DYEDOSOd44bU%26list%3DPLBv09BD7ez_5_yapAg86Od6JeeypkS4YM%26index%3D2%26frags%3Dwn%26ab_channel%3DVictorLavrenko&sa=D&ust=1601362784622000&usg=AFQjCNERIVXMRd-YOHtyHEZ-n_Q_viBn-Q)

Link for example of PCA for feature selection (creating new features):

[https://www.quora.com/How-do-you-use-PCA-for-feature-selection](https://www.google.com/url?q=https://www.quora.com/How-do-you-use-PCA-for-feature-selection&sa=D&ust=1601362784622000&usg=AFQjCNEJOcJmuf_2J_jkzAGToPaZN4tLrg)

### 5.3 Feature extraction method

**Actual paper:**  
  
**Notes for section:**

**Feature extraction method:**  
Original paper:   
"*Both the LLDs and their delta coefficients are smoothed by a moving average filter with window size 3*" - it only specifies the frameSize, not the frameStep!  
  
Replication paper:   
Not necessarily done exactly the same way. The above was interpreted as a frameStep of 3 as well.

Machine generated alternative text:
B. Multiple summaries over fixed size (sliding) windows (5 seconds long, shifted forward at 
intervals of 2 seconds): 
frameMode = fixed 
frameSize = S 
frameStep = 2 
frameCenterSpecia1 
= left 

**OBS: EDIT YES IT WAS! It was just the default setting!**

### 5.3 Task

**Actual paper:**  
  
**Notes for section:**

Task the data is from:  
Storytelling of triangle moving meaningfully or sometimes not. Storytelling monologues, no social component in this data. (Mid-level)  
  
Replicated paper is highly difficult - interview. This likely has huge effects.

### 5.4 Sex bias

**Actual paper:**  
  
**Notes for section:**

No information on bias in terms of sex, but balanced datasets in the original.  
Here, perhaps a bit worse training, but balanced test-set

## 5.4 In general

### 5.4.1: Hard to replicate, given the sparse information given

**Actual paper:**  
  
**Notes for section:**

### 5.4.2 What new knowledge has been acquired?

**Actual paper:**  
  
**Notes for section:**

ML methods not robust to language differences and task differences.  
Also not as balanced  
Which of them and how much, is hard to tell.

### 5.6 Where should the research go from here?

**Actual paper:**  
  
**Notes for section:**

More replications, to better understand

# 6. Conclusion

**Actual paper:**  
  
**Notes for section:**

sd

# 7. References

Anderson, J. R. (2015). *Cognitive psychology and it’s implications*. New York, NY. Worth Publishers

[Banich, M. T., Milham, M. P., Jacobson, B. L., Webb, A., Wszalek, T., Cohen, N. J., & Kramer, A. F. (2001). Chapter 29 Attentional selection and the processing of task-irrelevant information: insights from fMRI examinations of the Stroop task. I *Vision: From Neurons to Cognition*: *Bd. 134*. *Progress in Brain Research* (s. 459–470). https://doi.org/10.1016/S0079-6123(01)34030-X](https://www.zotero.org/google-docs/?VsVZwf)