

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](#_Toc55224972)

[1. Introduction 2](#_Toc55224973)

[1.1 Research into voice atypicalities in schizophrenia 2](#_Toc55224974)

[1.2. Practical appliance of ML 3](#_Toc55224975)

[1.3 What this paper aims to do 4](#_Toc55224976)

[1.4 ? 4](#_Toc55224977)

[2. Materials and Methods 6](#_Toc55224978)

[2.1 Litterature search (finding a study to replicate) 6](#_Toc55224979)

[2.1 Data collection 6](#_Toc55224980)

[2.2 Participants 6](#_Toc55224981)

[2.3 Procedure/task 6](#_Toc55224982)

[2.1 Design and stimuli 6](#_Toc55224983)

[2.5 Analysis 6](#_Toc55224984)

[3.1 Cleaning of audio files 6](#_Toc55224985)

[3.2 Feature extraction from audio files 7](#_Toc55224986)

[3.3 Partitioning 7](#_Toc55224987)

[3.4 Normalization 7](#_Toc55224988)

[3.5 Feature selection LASSO 7](#_Toc55224989)

[3.6 Model tuning 8](#_Toc55224990)

[3.7 Predicting the holdout 8](#_Toc55224991)

[3.7 Software used 8](#_Toc55224992)

[4. Results 9](#_Toc55224993)

[4.1 Model predictions 9](#_Toc55224994)

[5. Discussion: 9](#_Toc55224995)

[5.1 ML results: 9](#_Toc55224996)

[5.2 Differences in original vs. this 9](#_Toc55224997)

[5.2.1 Feature selection methods 9](#_Toc55224998)

[5.3 Feature extraction method 10](#_Toc55224999)

[5.3 Task 10](#_Toc55225000)

[5.4 Sex bias 10](#_Toc55225001)

[5.4 In general 10](#_Toc55225002)

[5.4.1: Hard to replicate, given the sparse information given 10](#_Toc55225003)

[5.4.2 What new knowledge has been acquired? 10](#_Toc55225004)

[5.6 Where should the research go from here? 11](#_Toc55225005)

[6. References 11](#_Toc55225006)

# Abstract

**(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

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

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

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 ?

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

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

The data used in this study was collected through four different studies;

From another study

Voice data, from matched participants of SZ and TD

Task the same – describing the video with different shapes behaving as if though they had intentions

## 2.1 Data collection

The data used in this study was collected through four different studies;

From another study

Voice data, from matched participants of SZ and TD

Task the same – describing the video with different shapes behaving as if though they had intentions

### 2.2 Participants

T Age, Gender, etc. (demographic data)

### 2.3 Procedure/task

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

### 2.1 Design and stimuli

Participants went through

D

## 2.5 Analysis

### 3.1 Cleaning of audio files

<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"

### 3.2 Feature extraction from audio files

openSMILE,

emobase

**Getting the basics down**  
 ./bin/Win32/SMILExtract\_Release.exe -C config/emobase\_kopi.conf -I ../danish\_denoised/td/Study1D0S101T1.wav -O ../danish\_denoised/td\_extracted\_features/Study1D0S101T1\_output.arff

**Kode der virker, TD:**

for filename in ../danish\_denoised/td/Study\*.wav; do name\_path=$filename; echo $name\_path; no\_wav=$"${name\_path%.\*}"; echo $no\_wav; base\_name=$(basename $no\_wav); output\_name=$"${base\_name}\_td\_output.arff"; echo $output\_name; ./bin/Win32/SMILExtract\_Release.exe -C config/emobase\_kopi.conf -I $filename -O ../danish\_denoised/td\_extracted\_features/$output\_name; done

**Kode der virker, SZ:**

for filename in ../danish\_denoised/sz/Study\*.wav; do name\_path=$filename; echo $name\_path; no\_wav=$"${name\_path%.\*}"; echo $no\_wav; base\_name=$(basename $no\_wav); output\_name=$"${base\_name}\_sz\_output.arff"; echo $output\_name; ./bin/Win32/SMILExtract\_Release.exe -C config/emobase\_kopi.conf -I $filename -O ../danish\_denoised/sz\_extracted\_features/$output\_name; done

### 3.3 Partitioning

Split the data into 2 different sets; train and test 20/80

(Groupdata2, partitioning)

### 3.4 Normalization

Normalizing both train and test, remembering to use max and min from train, to normalize also test. To avoid overfitting.

### 3.5 Feature selection LASSO

Removing all variables with no variance

Use cv.glmnet to select optimal number of features for the Machine Learning

Penalizing on the basis of misclassification error

LASSO, not ridge

Using lambda.1se instead of lambda.min (to user simpler almost equally good model)

Subset train and test, so they only contain the best features for the ML

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

.s

### 3.7 Predicting the holdout

SVM, linear kernel

### 3.7 Software used

Python

Visual Studio Coder

R

Rstudio

OpenSMILE, emobase

# 4. Results

## 4.1 Model predictions

s

# 5. Discussion:

## 5.1 ML results:

## 5.2 Differences in original vs. this

### 5.2.1 Feature selection methods

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

**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

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

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

### 5.4.2 What new knowledge has been acquired?

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?

More replications, to better understand

# 6. Conclusion

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)