Automatically diagnosing mental disorders from voice

A deep learning approach

Ludvig Renbo Olsen

Study no.: 201507847

Supervised by Riccardo Fusaroli

Bachelor Thesis, Cognitive Science, Aarhus University

2018-06-01

Contents

Chapter 1

Abstract

Mental disorders, such as autism, schizophrenia and depression, are very heterogeneous collections of symptoms that are difficult to diagnose and often go undetected. Various voice features have been proposed in the literature as potential markers, with promising discriminative accuracies. **Results**: This thesis presents a convolutional multiclass classifier that fairly accurately discriminate between autism spectrum disorder (*0.85/0.85/0.93/0.78/0.93*), schizophrenia (*0.94/0.83/0.90/0.78/0.98*), major- and persistent depressive disorder (*0.98/0.85/0.90/0.81/0.99*), and neurotypical controls (*0.81/0.65/0.53/0.82/0.81*) (metrics: *accuracy/f1/precision/recall/specificity*) (average accuracy: *0.9*; overall accuracy: *0.79*) from spectrograms of 6 seconds of speech. **Data**: 523 Danish-speaking, adult participants (mean age: 29.06y ; sd: 10.19y) from 9 datasets went through up to 10 trials of the Frith–Happé animations task (Abell et al., 2000). The voice activity was collected, thereby removing pauses, and sliced to 6-second windows with a 1-second stride, resulting in 48.645 audio slices. **Challenges**: In speech classification, a common confounding factor is room noise and reverberation (Bone et al., 2013). This was dealt with through extensive preprocessing (removal of noise and reverb) and augmentation (application of new noise and reverb) of

the audio files. **Literature**: A summary of the diagnostic criteria is included, along with a review of current studies on voice as marker for the selected mental disorders. **Take-away**: Based on 6 seconds of speech, it is possible to decently discriminate the 4 classes with a single model, using automatic feature extraction on spectrograms.

Chapter 2

Introduction

Mental diagnoses such as autism spectrum disorder (ASD), schizophrenia, major depressive disorder (MDD) and persistent depressive disorder (PDD) are very heterogeneous collections of symptoms which often overlap and are easily confused (Adam, 2013; Cuthbert and Insel, 2013; Fried and Nesse, 2015; Goldstein et al., 2002; Rapoport et al., 2009). Misdiagnosis, or lack of diagnosis, can have negative effects on patients (Fitzgerald, 2012; Van Schalkwyk et al., 2015), and thus the search for unique and reliable markers for these disorders has intensified (Singh and Rose, 2009). Since the first descriptions of autism, atypical voice patterns have been reported (Asperger, 1944; Kanner et al., 1943). Speech acoustics have since shown promise as predictors of the mentioned diagnoses (Cummins et al., 2015; Fusaroli et al., 2017; Martínez-Sánchez et al., 2015; Parola et al., 2018), potentially allowing for fast and cheap discovery of people with neuropsychiatric conditions, who would benefit from support. Such predictors could also be useful in clinical assessment for decreasing risk of misdiagnosis. So far, research has mainly been focused on manually extracted features, based on hypotheses about cognitive language processing or differences in biological speech production. As an additional approach to finding vocal markers for mental disorders, this thesis instead uses automatic feature extraction, namely a deep convolutional neu-

ral network (CNN) based on spectrograms of voice recordings from each of the aforementioned diagnoses, along with recordings from typically-developed (TD) patients. It is likely, that such a network can identify useful features that we would otherwise not think to extract, and that a symbiosis of automatic detection of features and theory-driven research will be the most fruitful approach going forward.

The thesis first summarizes the DSM-5 (American Psychiatric Association, 2013) definitions of the diagnoses, along with the current research into speech acoustics as diagnostic markers. It then criticizes the discrete nature of diagnoses, given their vastly continuous and heterogeneous nature. It then presents the case for machine learning based feature extraction methods, before giving a detailed description of the trained CNN model and the carefully designed data preprocessing and augmentation steps. The results are discussed along with certain ethical aspects of such a system and ideas for future work.

2.1 Defining disorders - DSM-5

This section summarizes the criteria for diagnosing Autism Spectrum Disorder (ASD), Schizophrenia and Depression (MDD and PDD), as defined by the Diagnostic and Statistical Manual of Mental Disorders, fifth edition, (DSM-5) (American Psychiatric Association, 2013), along with current knowledge about voice markers for each diagnosis. The purpose of DSM-5 is to ensure consistency in diagnostics, and is used by clinicians and researchers across countries. Since the publication of DSM-5, many have argued that the diagnoses are too categorical for research purposes as they often do not map onto the neurobiological and genetic underpinnings (Adam, 2013; London, 2014). These critiques and suggested ways forward will also be described in this section. Given the results in this thesis, it will later be argued that the diagnoses are not completely arbitrary.

2.1.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with symptoms typically starting to show after 12-24 months of age. Patients diagnosed with ASD have deficits in social communication and interaction. These include deficits in social-emotional reciprocity, nonverbal communicative behaviours (with atypical speech intonation as a possible manifestation), and in developing, maintaining, and understanding relationships. Another criterion is the presence of restricted, repetitive patterns of behaviour, interests, or activities, such as hand flipping or stereotyped use of words. Other specifiers, recorded in diagnostics, are intellectual impairment, language impairment, associated genetic or medical condition or environmental factor, associations to other neurodevelopmental, mental or behavioural disorders, and catatonia. These specifiers can help individualize the support given to patients. About 1% of the population is diagnosed with ASD[1]. Of those, about 70% may have one comorbid mental disorder, while 40% may have two or more comorbid mental disorders. (American Psychiatric Association, 2013)

2.1.1.1 Voice as marker for ASD

Fusaroli et al. (2017) did a review and meta-analysis of empirical studies seeking to map voice features to autism. The studies attempting to correlate ASD with single voice features (univariate studies), such as pitch mean and pitch range, were inconclusive about such relationship, with little replication and contradictory findings. The statistical estimates from the 2 conducted meta-analyses obtained accuracies of only 61-64% when discriminating between ASD and TD. Speech duration, i.e. length in seconds of either syllables, words or full utterances, was also inconclusive, with 7 out of 15 studies finding longer duration in ASD. This could be task-specific though, as a larger proportion of the studies with a spontaneous

[1]Baxter et al. (2015) estimates that about one in 132 persons, or 0.76%, are diagnosed with ASD worldwide.

production task found longer duration than those with a social interaction task. The multivariate studies, using machine learning to predict if participants were ASD or TD from multiple voice features, were more successful in discriminating between classes, reporting accuracies between 63% and 93%. It should be noted that the studies used different methods and that results were not followed up with attempts of replication.

2.1.2 Schizophrenia

Schizophrenia is a schizophrenia spectrum disorder (SSD) with a typical onset between late teens and into the late-20s. Patients can suffer from delusions, hallucinations, disorganized speech, grossly disorganized or catatonic motor behaviour, and/or negative symptoms, such as asociality and diminished emotional expression, e.g. in speech prosody. Other associated features are dysphoric mood (e.g. anger, anxiety or depression), inappropriate affect, a disturbed sleep pattern, cognitive deficits (e.g. memory, language, theory of mind, attention), significant social dysfunction, with more. Post onset, there is a decline in level of functioning, or failure to achieve expected level of functioning, and continuous signs of the disturbance for at least 6 months. Schizophrenia is considered a heterogeneous clinical syndrome, meaning that patients with schizophrenia vary vastly on most diagnostic features. About 0.3% - 0.7 % of the population has schizophrenia in their lifetime (lifetime prevalence). 20% of schizophrenics attempt suicide, with 5-6% of schizophrenics dying as a consequence. (American Psychiatric Association, 2013)

90% have one or more additional diagnoses, with childhood-onset schizophrenia being preceded by, and co-occuring with, ASD in 30 – 50% of cases. (London, 2014)

2.1.2.1 Voice as marker for schizophrenia

In their review and meta-analysis of studies on voice as marker in schizophrenia, Parola et al. (2018) found clear effects of reduced speech production, reduced pitch variability and increased pause duration. There were no effects of pitch mean, mean intensity, mean utterance duration, or number of pauses. No multivariate studies were reported on. Martínez-Sánchez et al. (2015) could discriminate schizophrenic patients from TDs with an accuracy of 93.8%, while Kliper et al. (2015) obtained an accuracy of 76.19% and Rapcan et al. (2010) obtained an accuracy of 79.4%.

2.1.3 Depression

Patients with major depressive disorder (MDD) have a range of symptoms, such as depressed mood (e.g. sadness, hopelessness, emptiness), diminished interest or pleasure in daily activities, change in weight and appetite, sleep problems (insomnia or hypersomnia), fatigue, psychomotor agitation (e.g. unable to sit still) or retardation (e.g. speaking slower or in lower volume), feelings of worthlessness or inappropriate guilt, inability to concentrate or make decisions, and recurrent thoughts of death and suicide. There are attempts to distinguish symptoms from normal responses to loss.

The patient must experience depressed mood or loss of interest or pleasure in most daily activities, along with four (or more) of the other symptoms, for a 2-week period, before it counts as a major depressive episode. In the United States, twelve-month prevalence is approximately 7%, with large differences between gender (higher in females) and age (higher in young adults). The more recent the onset, at time of intervention, the more likely they are to experience near-term recovery. (American Psychiatric Association, 2013) After 2 years (for adults) with either depressed mood or diminished interest or pleasure in daily activities, along with two or more of the other symptoms, and with no symptom-free periods

longer than 2 months, the patient is diagnosed with Persistent Depressive Disorder (Dysthymia), previously known as chronic depressive disorder and dysthymia. In the United States, twelve-month prevalence of *chronic* depressive disorder is approximately 1.5%. (American Psychiatric Association, 2013)

2.1.3.1 Voice as marker for depression

In a review on voice as marker for depression and suicidality, Cummins et al. (2015) looked at 4 categories of speech features. *Prosodic* features, which are long-time, phoneme-level variations in perceived intonation, stress and rythm, such as pitch and speech rate. *Source* features, which describe the speech producer (e.g. air flow, vocal fold movements, laryngeal control), such as jitter and shimmer. *Formant* features, which describe the acoustic resonances of the vocal tract, such as the location of the various formants. *Spectral* features, which describe the distribution of frequencies in the audio recordings over time, such as Mel Frequency Cepstral Coefficients (MFCCs). Of prosodic features, speech rate related features were the most promising, with reports of slower speech rates in depressed participants, differences in speech pause duration in free speech, and correlations between speech pause duration and Hamilton Rating Scale for Depression (HAMD) score. Of source features, jitter, shimmer and harmonic-to-noise ratio were found to correlate with depression, though a range of confounding factors were reported, regarding the feature extraction process. The authors also note that only a small number of papers have studied source features in depression. Other mentioned features were spirantization (Flint et al., 1993), teager energy operator autocorrelation (Low et al., 2011; Teager and Teager, 1990), normalised amplitude quotient (Scherer et al., 2014), and quasi-open-quotient (Scherer et al., 2014). Of formant features, the frequencies, bandwidths and dynamics of formants were highlighted, with Helfer et al. (2013) obtaining an AUC[2] of 0.73 when discriminating between constructed low vs. high depression classes from formant

[2]Cummins et al. (2015) incorrectly reports accuracy instead of AUC.

frequencies and formant dynamics (velocity and acceleration). Cummins et al. (2015) notes that anti-depressant medication might affect formants by drying out the vocal tract and mouth. Of spectral features, shifts of spectral energy in the 0-500 Hz and 500-1000Hz bands has been observed. (Cummins et al., 2015)
Predicting the presence of depression, Moore II et al. (2008) obtained an accuracy of 91% for male speakers and 96% for female speakers, though with a small dataset. Low et al. (2011) obtained accuracies up to 87% for males and 79% for females, using teager energy operator based features. Kliper et al. (2015) obtained an accuracy of 87.5 % comparing depressed patients with healthy patients and 71.43% comparing schizophrenic patients with depressed patients.

2.1.4 Review of voice as marker

Good results have been obtained when using voice features to discriminate either ASD, schizophrenia or depression from TD. The most successful approaches included multiple features in the model (multivariate), with examples of discriminative accuracies above 80% for all diagnoses, with some reaching more than 90%. In univariate studies, differences in speech rate, speech pause duration and pitch variability were commonly reported. Low-level source features, such as jitter and shimmer, seemed underinvestigated, considering their potential to describe more ingrained aspects of vocalization, possibly less affected by choice of task, mental effort to conform or similar. The effect of speech task should be investigated further as well.
The large variability of methods and investigated features makes comparing studies difficult. Sharing of code[3], replication of studies, and the inclusion of an (updatable) common set of promising base features (besides the more exotic exploratory experiments), would likely bring more clarity to the topic.
None of the studies used multiclass classification to discriminate between all three diagnoses at once. Such approach, which is taken in this thesis, should decrease

[3]We intend to make the code for this thesis available in connection with a future paper.

the risk of simply detecting abnormalities and find more disorder-specific features. To avoid task-specific differences, all patients were presented with the same task (see: Data).

2.2 Misdiagnosis

In a meta-analysis of 10 studies, Marín et al. (2018) investigated the comorbidity of schizophrenia spectrum disorders (SSD) in average-IQ adults diagnosed with ASD and found 6.4% prevalence of SSD in ASD adults. They argued that ASD patients might be more vulnerable to develop psychotic disorders. Matsuo et al. (2015) found autistic-like traits in almost half of adult participants diagnosed with MDD (not including remitted), bipolar disorder (remitted and unremitted) and schizophrenia (remitted and unremitted). For bipolar disorder and schizophrenia this was independent of symptom severity. For MDD patients, prevalence and degree of autistic-like traits were associated with the depressive symptom severity with remitted patients not differing from TDs. Conversely, Kim et al. (2000) found a greater rate of depression and anxiety problems in children with high-functioning autism than in a sample of the general population, while surveys found that US psychiatrists discover symptoms of depression in approximately one third of schizophrenic patients (Siris et al., 2001, Addington et al. (2002)).

Given this heterogeneity in patients with, often comorbid, mental disorders, it is likely that some patients are misdiagnosed. Based on 5 cases, Van Schalkwyk et al. (2015) argued that the subjective framing of experienced symptoms by clinicians and family members can be the main difference between an ASD diagnosis and an SSD diagnosis. This can lead to the use of wrong, and potentially harmful, medications or affect available treatments, insurance and support (Van Schalkwyk et al., 2015).

2.3 Beyond discrete diagnoses

Just like a headache can have multiple, biologically very different causes (Landtblom et al., 2002; Levine, 1991), it can be argued that behavioural and self-reported symptoms are not necessarily very good descriptors of mental disorders. Instead of grouping patients together in discrete diagnoses, it has been proposed to create more individual mental profiles (London, 2014). One approach, the Research Domain Criteria (RDoC) framework from the National Institute of Mental Health (NIMH), is to to describe patients by a set of functional dimensions, such as cognition, genetics and social abilities (Cuthbert and Insel, 2013). This would provide more granular insight to clusters and combinations of neurobiological and genetic conditions, observable behaviours, and self-reported symptoms, across and within diagnoses, and allow for more individualized treatments.

When evaluating a new intervention, it may be found to work for patients with similar scores on some dimensions, while not affecting the other patients. Against discrete diagnoses, this intervention could be discarded due to insignificant results, though a subset of the patients would have benefitted from a more granular evaluation approach. With the computing power available today and in the near future, machine learning approaches to understand these disorders would likely benefit greatly from an increased number of descriptive dimensions, potentially allowing us to predict the effects of an intervention and list the common side-effects for patients with similar profiles.

2.4 Automatic feature extraction

Using machine learning-based automatic feature extraction (AFE), such as with convolutional neural networks[4] (CNN; Li et al., 2010, 2014), we often obtain a higher predictive power[5] than with theory-based, manually extracted features

[4]For an introduction to Convolution Neural Networks see Wu (2017)

[5]I.e. lower error when predicting previously unseen data.

(Martinez et al., 2013). Being able to automatically identify important features and combinations of features, instead of testing a few at a time, bear great promise of more efficient research. For such a system to contribute markedly to theory though, we must first solve the lack of transparency in decision-making, surrounding current systems. Explainable Artificial Intelligence (XAI) is a field devoted to solving this challenge (Gunning, 2017; Ribeiro et al., 2016).

Where manual features can be chosen based on theory, an inherent risk of AFE systems is the extraction of noise features instead of features related to the actual signal (e.g. speech). In classification of speech, the reverb of the recording room, along with various room noises, can often be big confounding factors (Bone et al., 2013). E.g., if all schizophrenic participants were recorded in one room, separate from where the other participants were recorded, we could likely obtain a very high discriminative accuracy just by identifying these room characteristics. As the objective of such a classifier is to generalize to out-of-sample data[6], we need to minimize the risk of overfitting to noise features, e.g. through extensively considered preprocessing and augmentation[7] of the audio recordings.

[6]E.g. new particants that we wish to screen for mental disorders.

[7]Audio **preprocessing** is done *once* to each file; Audio **augmentation** is done differently in *every* training round.

Chapter 3

Methods

This section describes the data and methods used. The data was first preprocessed, including noise removal, collection of voice activity and temporal slicing. Within the TensorFlow model, these audio slices were then augmentated differently for each epoch[1] by applying new noise and convolutional reverb, before being converted to spectrograms and inputted to the convolutional feature extraction pipeline. A fully connected neural network layer then used the extracted features to classify the input as one of 4 classes (autism, schizophrenia, depression, or control). These predictions were evaluated by a selected set of metrics, also described in this section.

[1] An epoch refers to the model seeing the entire training set once.

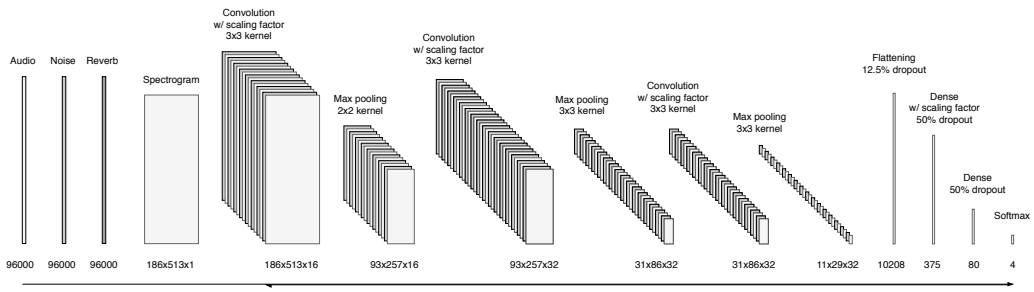


Figure 3.1: Visualization of the TensorFlow model architecture, covering data augmentation, convolutional feature extraction and final classification.

3.1 Data

Speech recordings were gathered from 9 datasets, totalling 523 participants with 302 participants from (Bliksted et al., 2014, 2017, 2018; Bang, 2009; Ladegaard et al., 2014) and 221 from unpublished studies performed by Line Gebauer, Emma Fowler, Heine Lund Pedersen, and Vibeke Bliksted. All schizophrenic participants were 1st episode patients. Depressed participants were a mix of 1st episode and chronicly depressed patients, as defined by DSM-IV-TR (American Psychiatric Association, 2000).

Participants went through up to 10 trials of the Frith–Happé animations (FHA) task (Abell et al., 2000) with varying response lengths[2]. The animations were developed to assess the ability to attribute mental states to others, also referred to as theory of mind (ToM). They contain two triangles, one large red and one small blue triangle, moving around the screen. There are three conditions: **Random**: where triangles do not interact and move about purposelessly; **Goal-directed (G-D)**: where triangles appear to respond to each other's movements, either following, chasing, fighting or dancing; and **ToM**: where triangles appear to respond to each other's the mental state. Depending on the condition, the upcoming animation is introduced with different roles. In the Random condition, the triangles are introduced as "just triangles". In the G-D condition, the triangles are given animal roles, and in the ToM condition, the triangles are introduced as persons. After watching an animation, the participant is asked to describe what happened in the animation (Abell et al., 2000).

3.1.1 Partitioning

In order to validate the model, the dataset was partitioned into train/test sets. To avoid leakage between these datasets, all audio files from a participant were placed

[2]Due to practicalities, statistics on number of trials and lengths of responses are not included.

Table 3.1: Participants per study.

Study	N	F	F/M	M	Mean.Age	SD.Age	Min.Age	Max.Age
Dan Bang (2009)	24	6	0	18	28.29	7.29	20.0	47.0
Line Gebauer (unpublished)	40	17	1	22	26.73	6.06	19.0	45.0
Emma Fowler (unpublished)	11	4	0	7	28.64	10.23	17.0	46.0
Heine Lund Pedersen (unpublished)	24	15	0	9	25.21	5.88	19.0	38.0
Nicolai Ladegaard (2014)	160	121	1	38	33.86	11.81	18.1	62.6
Vibeke Bliksted (2014)	72	34	0	38	22.67	3.14	18.0	31.0
Vibeke Bliksted (2018)	46	13	0	33	23.50	3.74	18.0	34.0
Vibeke Bliksted (unpublished)	86	45	0	41	34.07	12.58	19.0	67.0
Heine Lund Pedersen (unpublished)	60	26	0	34	24.75	4.26	18.0	35.0
TOTAL	**523**	**281**	**2**	**240**	**29.06**	**10.19**	**17.0**	**67.0**

Number of participants (N), genders (F, F/M, M), and ages for each study.

Table 3.2: Participants per diagnosis.

Diagnosis	N	F	F/M	M	Mean.Age	SD.Age	Min.Age	Max.Age
Autism	54	22	1	31	27.69	7.25	17.0	46.0
Control	263	145	0	118	28.70	10.15	18.0	62.1
Depression	72	55	1	16	35.02	11.64	18.5	62.6
Schizophrenia	134	59	0	75	27.17	9.37	18.0	67.0
TOTAL	**523**	**281**	**2**	**240**	**29.06**	**10.19**	**17.0**	**67.0**

Number of participants (N), genders (F, F/M, M), and ages for each diagnosis.

in only one of the datasets. The test set was randomly appointed recordings from 5 participants from each class using the groupdata2 R package (Olsen, 2018). As audio files had different lengths, and as there were a variable number of trials per participant, the test set ended up with a different number of 6 second audio slices per class.

Table 3.3: Number of audio slices and participants per class for train and test sets.

Diagnosis	Partition	Slices	IDs
Autism	Train	11392	49
Control	Train	21775	258
Depression	Train	3183	67
Schizophrenia	Train	9390	129
TOTAL	**Train**	**45740**	**503**
Autism	Test	1509	5
Control	Test	617	5
Depression	Test	251	5
Schizophrenia	Test	528	5
TOTAL	**Test**	**2905**	**20**

3.2 Preprocessing

The audio files were converted to 16 bit Wav files with a 16k samplerate, meaning that information above 8kHz was removed. After being denoised (see *Noise Removal*), these were sliced into 6-second files using a sliding window with a stride of 1 second, in order to generate a lot more data samples and position the various voice features in different spots on the time-axis of the spectrogram. Audio files shorter than 6 seconds were discarded.

In order to include as much voice information as possible in the 6-second audio slices, the voice activity was collected, removing the pauses between speech. Voice activity detection was done with Google's WebRTC Voice Activity Detector, using the python interface webrtcvad (Wiseman, 2016) with an aggressiveness setting of 1, a frame duration of 20ms, and 350ms of padding, to avoid artificially sounding transitions. Patterns in speech pauses have been associated with different disorders (Cummins et al., 2015; Parola et al., 2018), but given the short duration of the audio slices, these was unlikely to be useful predictors.

Audio I/O was handled with the Librosa python package (McFee et al., 2017). Data augmentation was done as part of the TensorFlow (Abadi et al., 2015) model. Spectrograms for visualizing the preprocessing and augmentation steps in this chapter were plotted with Librosa (McFee et al., 2017) and matplotlib (Droettboom et al., 2017).

3.2.1 Noise removal

Room noise[3] and hum was greatly reduced by stacking multiple instances of the Voice De-noise and De-hum tools in the iZotope RX 6 audio editor (iZotope, 2017). A small equalizer tilt was applied at 1085Hz with the Fabfilter Pro-Q2 equalizer (FabFilter, 2014) to bring back some brightness to the signal. The signal was

[3]Including room reverberation.

normalized to peak at -1dB before and after these preprocessing steps. A noise-removal preset was sculpted from a small subset of the data[4] and used to batch process all audio files at once[5].

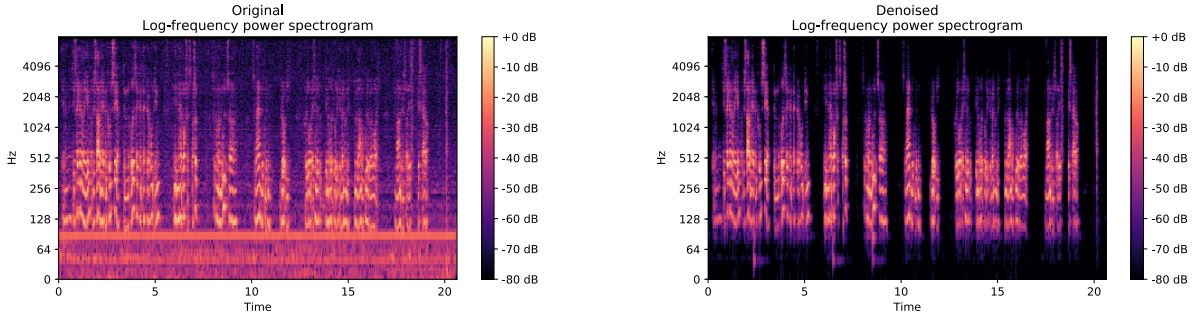


Figure 3.2: Spectrograms of the original audio file (left) and the denoised audio file (right). The signal to noise ratio is much larger in the denoised version.

3.3 Data augmentation

As part of the input pipeline for the convolutional neural network, random noise and convolutional reverb was applied to the audio slices before they were converted to spectrograms. In this section, these augmentation steps will be explained and visualized with spectrograms. The augmentation was different for each audio slice at each epoch, to allow for a more robust and generalized model.

3.3.1 Blending two signals

When blending the speech signal with the noise signal or reverb signal, the following function was used:

[4]With representives from each study and diagnosis.

[5]Note that some signal processing operations automatically adapted to the individual audio files

$$blend(x, y, amount) = x \odot (1 - amount) + y \odot amount \tag{3.1}$$

Where $\odot$ means element-wise multiplication, and x and y are signals to blend together.

3.3.2 Global modulator

To ensure that the data was also seen with tiny amounts of overall augmentation in training, *amount* parameters were multiplied with a global *augmentation amount modulator* when applying noise and reverb.

$$globalModulator \sim\ Uniform(min = 0.0, max = 1.0) \tag{3.2}$$

Where $\sim$ is read as "samples from".

3.3.3 Applying noise

When choosing the noise distribution to blend with the signal, the objective was to change the numerical values of the signal while still keeping the overall structure. To accomplish this, a Gaussian distribution based on the median and interquartile range (IQR) of the signal was chosen.

$$\begin{aligned} noise &\sim\ Gaussian(\mu, \sigma) \\ \mu &\leftarrow median(originalSignal) \\ \sigma &\leftarrow IQR(originalSignal) \end{aligned} \tag{3.3}$$

Where μ is the mean of the Gaussian distribution and σ is the standard deviation.

Alternatives explored were uniform noise and Gaussian noise based on the mean and standard deviation of the signal.

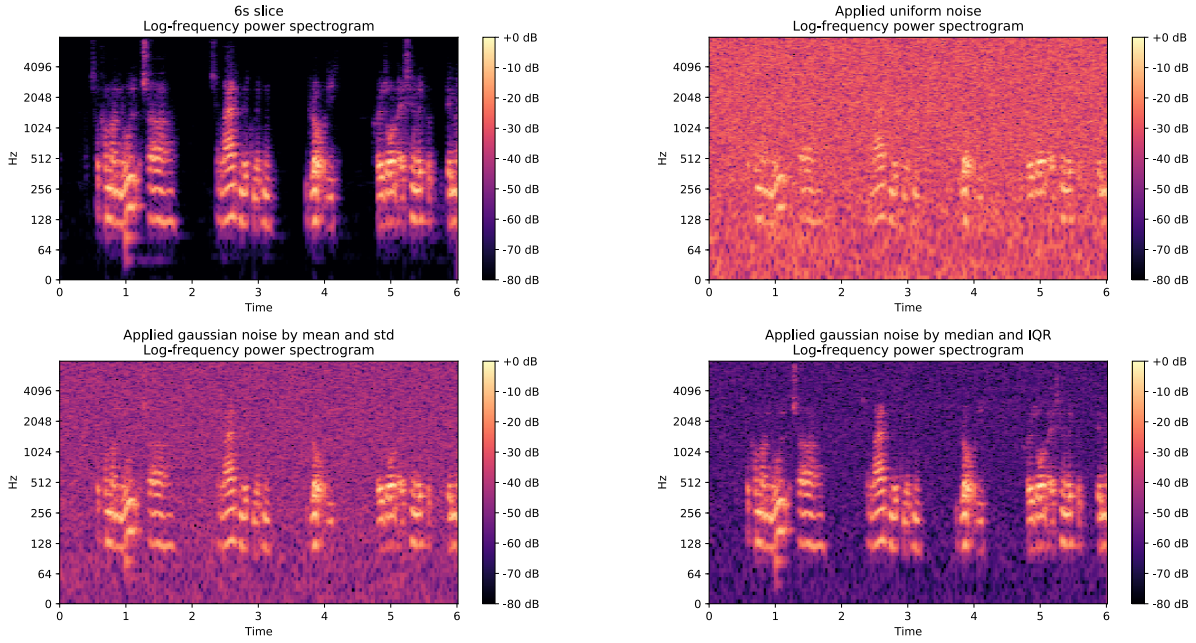


Figure 3.3: Noise distribution spectrograms with ratio 70% signal / 30% noise. First row: Pure signal (left) and uniform noise (right). Second row: Gaussian noise based on mean and standard deviation (left) and Gaussian noise based on median and IQR (right).

The noise was then blended with the original signal.

$$\begin{aligned} & blend(signal, noise, amount) \\ amount \sim\ & Uniform(min = 0.0, max = 0.35) \cdot globalModulator \end{aligned} \tag{3.4}$$

3.3.4 Applying convolutional reverb

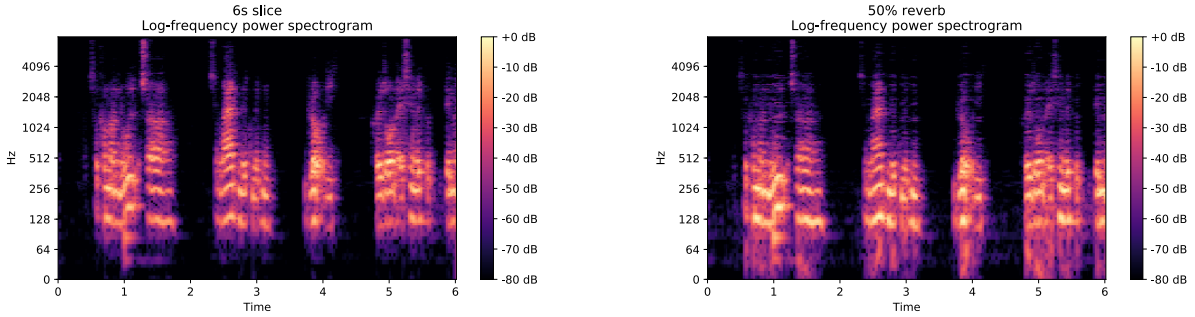
As mentioned, a likely confound for classification of speech recordings is the reverberation of the recording room (Bone et al., 2013). To make the model more

robust to different reverbs and amounts of reverb, convolutional reverb was applied to the signal, randomly selecting one of 20 impulse responses[6].

The signal and impulse response (IR) were converted to 64 bit complex tensors. The signal was sliced to windows of 128 samples, moving 32 samples per window to create overlap. The IR was cropped to 128 samples, equivalent to one window in the signal. A Fast Fourier Transformation was applied to both tensors, before multiplying the IR with all the signal windows elementwise. Inverse Fast Fourier Transform was applied to the convoluted windows and the windows were merged, using Hann windowing[7]. This signal was converted back to a 32 bit float tensor and blended with the original signal. The following is a simplified notation, *not taking the windowing into account*:

$$\begin{aligned} blend(signal, convoluted, amount) \\ convoluted \leftarrow \mathcal{F}^{-1}(\mathcal{F}(signal) \odot \mathcal{F}(IR)) \\ amount \sim \ Uniform(min = 0.0, max = 0.38) \cdot globalModulator \end{aligned} \tag{3.5}$$

Where $\mathcal{F}$ is the Fast Fourier Transform and $\mathcal{F}^{-1}$ is the Inverse Fast Fourier Transform.



[6]AK-SROOMS by Kristoffer Ekstrand at https://www.adventurekid.se/akrt/free-reverb-impulse-responses/

[7]As described in the TensorFlow guide: https://www.tensorflow.org/api_guides/python/contrib.signal#Framing__variable_length_sequences; The merged signal has the dimensionality of the original signal.

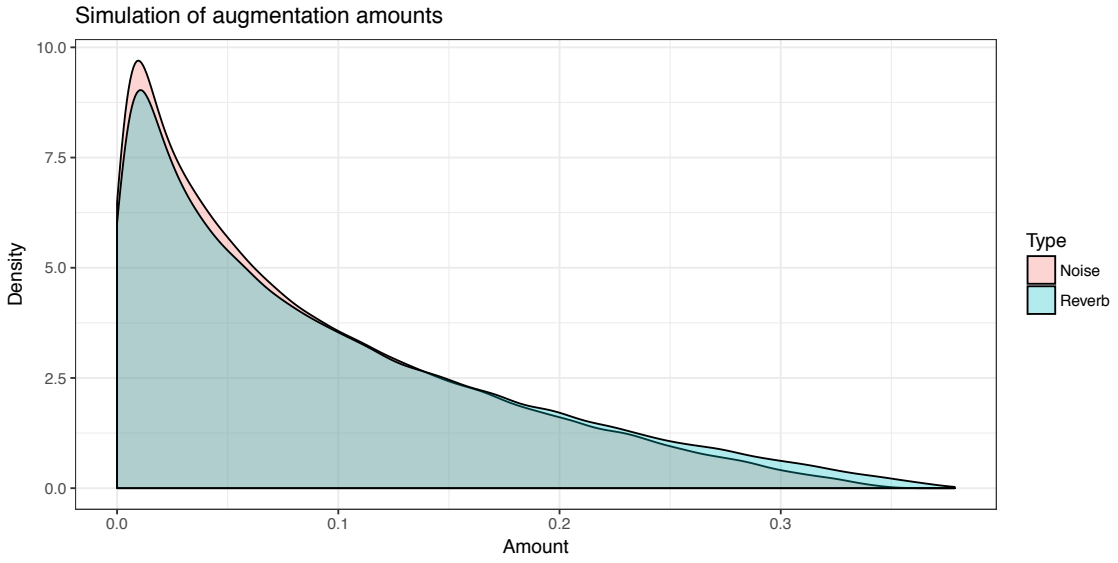


Figure 3.5: Simulation of the amounts of data augmentation applied. Uses 100,000 samples.

Figure 3.4: Illustration of convolutional reverb with the clean slice (left) and the slice with 50% convolutional reverb (right). Notice the added activity in low frequencies and the small dip at ~1024Hz. Note: Usually the noise is added before the reverb, but the reverb effect is easier to illustrate on the clean slice.

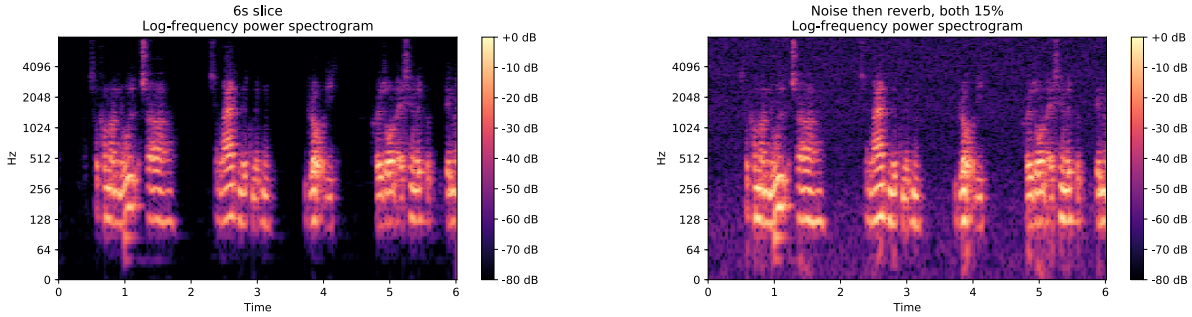


Figure 3.6: The clean slice (left) and the slice with 15% noise and 15% reverb (right).

3.3.5 Spectrogram

The augmented audio signal was converted to a dB-scaled log-frequency spectrogram of shape 186x513x1, before being rescaled to the $[0, 1]$ range using fixed min/max values of $[-80.0, 0.0]$.

The code for converting the audio to a spectrogram was ported from the python package Librosa (McFee et al., 2017) into TensorFlow (Abadi et al., 2015) code.

3.4 Model architecture

An iterative approach was taken to find a good model architecture, optimized for the following criteria: I) it should not overfit to the training data, II) it should be a good overall model measured by macro f1 and III) there should be high precision in the non-TD classes so errors fall into the TD base class.

The final model (see figure 3.1) consisted of 3 convolutional layers with max pooling and local response normalization in between, totalling 80 convolutional feature maps. These were followed by 2 fully connected layers with 50% dropout and finally a softmax classifier. Non-linearities were introduced using the rectified linear unit (ReLU) activation function and multiple layers used a scaling factor to allow for faster training. Training time was between 15 and 24 hours with a batch size of 100 audio slices, from which the best epoch test set evaluation was saved and reported. Local response normalization was used as it appeared to help with memory usage, allowing for a larger batch size.

Weighted multiclass cross-entropy was used as loss function (Hinton, 1990; Hinton and Salakhutdinov, 2006; Mirowski et al., 2010), where classes were initially weighted inversely proportional to class size and then tuned as hyperparameters. Since the softmax classifier sums the predicted class probabilities to 1 and we only have one true class per sample, the formula for weighted multiclass cross-entropy

Table 3.4: Loss weight per class.

Diagnosis	Weight
Schizophrenia	1.35
Autism	1.06
Control	0.55
Depression	1.22

is the following:

$$H_{y'}(y) = -\sum_i y'_i \cdot log(y_i) \cdot w_i \tag{3.6}$$

Where y_i is the predicted probability for class i, y'_i is the true probability for class i (either 0.0 or 1.0), and w_i is the weight for class i[8].

The sum of cross-entropy scores in the batch was minimized using an Adam optimizer with an initial learning rate of 0.001.

3.4.1 Scaling factor

During experiments, it was found that using a scaling factor on dense layers sped up convergence on toy problems while not affecting the final evaluation. As it is a very simple extension of the traditional dense layer, it likely already exists in the literature under a different name. The version used in the this model was a learned scalar s, truncated to $[0.2, 2]$ and initialized to 1.0. It was multiplied with the layer before the activation function, like the following:

$$A_i = \sigma((w_i \cdot A_{i-1} + b_i) \odot s_i)$$

Where A_i are the new activations, A_{i-1} are the activations of the previous layer, w_i are the weights of the current layer, b_i is the bias term for the current layer, and s_i is the scaling factor for the current layer.

[8]The exact TensorFlow implementation may differ slightly.

It is inspired from the experience, that proper scaling of data, e.g. normalization or rescaling to $[0, 1]$, makes it easier for machine learning algorithms to converge. By letting it learn its own scaling factor for each layer, it should be able to converge faster. An interesting future experiment would be to compare it to various layer normalization methods, such as batch normalization and local response normalization.

3.5 Evaluation metrics

The trained model's predictions of the audio slices in the test set are evaluated both overall and per diagnosis.
First, one-vs-all evaluations are performed for each diagnosis, using the following metrics:

$$\begin{aligned} accuracy &= \frac{tp + tn}{tp + tn + fp + fn + \epsilon} \\ precision &= \frac{tp}{tp + fp + \epsilon} \\ recall &= \frac{tp}{tp + fn + \epsilon} \\ specificity &= \frac{tn}{tn + fp + \epsilon} \\ f1 &= \frac{2 * precision * recall}{precision + recall + \epsilon} \end{aligned} \tag{3.7}$$

Where the constant ϵ is added to the denominator when dividing, to avoid zero-division.

$$\epsilon = 1.0\text{e-}6$$

To evaluate the multiclass classifier, the macro-averaged metrics are computed along with the overall accuracy.

$$
\begin{aligned}
averageAccuracy &= \sum_{i=1}^{n} \frac{accuracy_i}{n} \\
overallAccuracy &= \sum_{i=1}^{n} \frac{tp_i}{totalPredictions} \\
macroPrecision &= \sum_{i=1}^{n} \frac{precision_i}{n} \\
macroRecall &= \sum_{i=1}^{n} \frac{recall_i}{n} \\
macroSpecificity &= \sum_{i=1}^{n} \frac{specificity_i}{n} \\
macrof1 &= \sum_{i=1}^{n} \frac{f1_i}{n}
\end{aligned}
\tag{3.8}
$$

Where n is the number of classes.

Chapter 4

Results

This section presents the best results obtained during training, measured on the test set. The overall results of the classifier are presented in table 4.1. Table 4.2 shows the performance for each diagnosis, while table 4.3 is the confusion matrix of predictions, showing true labels[1] vs. predicted labels.
Given the imbalance of number of audio slices per diagnosis in the test set, the diagnosis-level results along with the average accuracy are the most telling of the model's performance. These results are in the high end of previously reported results in the literature, even though this model distinguish between all 4 classes at once.

Table 4.1: Overall classifier evaluation.

Average.Accuracy	Overall.Accuracy	M.f1	M.Precision	M.Recall	M.Specificity
0.9	0.79	0.79	0.81	0.8	0.93

The objective was to have high precision for the three disorders, with wrong predictions being placed in the TD (Control) base class. This was achieved with precisions ≥ 0.9 for ASD, schizophrenia and depression and with by far the greatest number of false positives placed in the TD class. In a clinical context, this

[1]True labels refer to the targets we want our model to optimize towards.

Table 4.2: Evaluation for each diagnosis.

Diagnosis	Accuracy	f1	Precision	Recall	Specificity	Correct	Incorrect	TP	FP	TN	FN
Schizophrenia	0.94	0.83	0.90	0.78	0.98	2740	165	411	48	2329	117
Autism	0.85	0.85	0.93	0.78	0.93	2476	429	1174	94	1302	335
Control	0.81	0.65	0.53	0.82	0.81	2354	551	509	443	1845	108
Depression	0.98	0.85	0.90	0.81	0.99	2834	71	203	23	2631	48

Control: typically-developed; Depressed: Mix of Major- and Persistent Depressive Disorder.

would mean that fewer patients would be diagnosed with a wrong disorder, with the trade-off that some (442 out of 951 slices predicted as TD) would not be detected and potentially not get the needed support.

Table 4.3: Confusion Matrix.

T.P	Schizophrenia	Autism	Control	Depression
Schizophrenia	411	15	102	0
Autism	28	1174	305	2
Control	14	73	509	21
Depression	6	6	36	203

Control: typically-developed; Depressed: Mix of Major- and Persistent Depressive Disorder. True labels vertically and predicted labels horizontally.

Six seconds of speech might not always hold enough information to correctly detect a complex mental disorder, even though the results of this thesis indicate that it often does. To increase the likelihood that a patient is diagnosed correctly, a future clinical application would aggregate the predictions of all the patient's audio slices and present clinicians with a probability for each diagnosis, along with measures of uncertainty.

Chapter 5

Conclusion

For this thesis, we successfully built a multiclass classifier with automatic feature extraction, that can discriminate participants with ASD, schizophrenia, depression and healthy controls, based on six seconds of speech. This included important steps of data preprocessing and augmentation, to generalize the model, and is a foundation on which a larger-scale research and clinical application could be built. The thesis includes a literature review on vocal markers for each of these mental disorders and a discussion of the premise of categorical diagnostics and its effects on patients and the possibility to conduct useful science.

Chapter 6

Discussion

With the level of accuracy that the model discriminated between diagnoses, it seems that the categorical diagnoses in DSM-5 (American Psychiatric Association, 2013) are not completely arbitrary. The participants appear to share features, or combinations of features, within diagnoses, although it remains to be investigated exactly which features the model uses. For the automatic feature extraction model to be truly useful for informing theory, better explanation of its decision-making should be built in to a future system. The rest of this section discusses: a) the possibilities to improve and expand the model, and b) aspects of the ethics of such a system. As responsible scientists, these should be considered together.

6.1 Future development

The current system is limited in few ways. 1. The participants are all Danish-speaking adults with little variation of dialect. It remains to be tested, if the model generalizes to other dialects, languages, and cultures. 2. Due to the heterogenous nature of the diagnoses, a large-scale data collection would be required with a strict protocol for choice of severity tests and measurement of functional dimensions,

before the system could be applied clinically. 3. Due to practicalities, the gender balance in the test set was not reported. Gender representation within some diagnoses is skewed (Van Wijngaarden-Cremers et al., 2014; Whiteley et al., 2010), and it is important to know how the model reflects this. 4. Currently, we do not evaluate the model on a participant level. Wrong predictions could stem from a lack of information in some of the six second audio slices, or it is possible that wrong predictions are clustered in a few of the participants. This kind of information would be useful for designing a better system.

Some of the questions we would like future research on voice markers to answer are: **a**) Are there any cross-linguistic and cross-cultural markers? **b**) Can we detect comorbidities, specific symptoms or neurobiological and genetic underpinnings? **c**) Do voice markers only appear in certain contexts, e.g. social interaction? **d**) Can we track severity of disorders through time, e.g. to measure the effect of intervention? **e**) Can we predict the effect of specific treatments or recommend helpful treatments?

To address **a**, we could use a cross-validation approach, leaving out one language or cultural region at a time as test set and training a model on the others. Detecting comorbidities and symptoms (**b**) could be represented as a multilabel classification task. **c** could be investigated by evaluating the model's performance on different tasks. **d** and **e** could possibly be solved with machine learning methods but would depend on the quality of severity metrics and amount of long-term participant data.

A big focus of the model-development was generalizability. The model should discriminate diagnoses based on relevant features and filter out the noise that is specific to the single recording or study. The current system applies noise and convolutional reverb to the audio signal and uses dropout to avoid overfitting to the training data. Our long-term vision for such a system is to create a good general "listener", which can detect whether it is a human speaking or not (e.g. a

seagull screaming), what gender and age-group the participant belongs to, what language is spoken and what kind of speech task (e.g. social interaction or reading aloud) that was performed. Such a model could be pre-trained on the vast amount of available speech corpora, before being optimized to discriminate diagnoses, allowing for larger networks with potentially higher accuracy.

Three additional data augmentation methods that will be implemented in a future version to increase generalizability are *pitch shifting*, application of *microphone impulse responses* through convolution and *addition of environment noises*.
These will make the model more robust to, respectively, changes on the frequency-axis, choice of microphone when recording new patients, and noisy office sounds likely to appear in clinics or in the training data.

6.1.0.1 Other disorders

Making the model a more general "listener" could involve the inclusion of other disorders, e.g. with similar symptomatology, such as bipolar disorder that often involve autistic traits (Matsuo et al., 2015), as discussed in Misdiagnosis, or disorders that would benefit from early detection, such as Parkinson's disease (Tinelli et al., 2016), where voice features have already shown promise for discriminating patients from controls (Tsanas, 2012).

6.1.0.2 Inclusion of manually extracted features

Some features, like statistics of pauses, need more than 6 seconds of audio to be computable or useful. Others would need more convolutional layers to extract or direct access to the audio data. A solution could be to manually extract these as part of the preprocessing or within the tensorflow model. To establish if a new feature brings additional knowledge to the model, we can test how well the model with automatic feature extraction is able to predict it. If a feature cannot be

predicted with some accuracy, it might be worth including in the model.

6.2 Ethical considerations

If the diagnostic tools described in this thesis are used responsibly by competent professionals who respect patient confidentiality, they could come with great benefits. Quick and cheap access to more accurate diagnostics could increase the likelihood that troubled individuals get the right treatment and support. But, such technologies also come with potential societal risks, if they become available outside psychiatry, such as the reduction of individuals to their diagnoses and fundamental shifts in what is considered private.

Extra-clinical possession of tools for mapping speech recordings[1] to diagnoses (f.i. commonly available as a smartphone app or exclusive to powerful organizations, such as governments or providers of social networking services), allows easy collection and usage of any speaking person's mental profile; information the subject might wish, and believe he has the right, to keep private.

The ability to pre-assess people's candidacy for certain roles, based on their mental profile, might lead to social stigma towards individuals with certain diagnoses, e.g. if we collectively discard schizophrenic job applicants before meeting them, due to an inaccurate understanding of schizophrenia[2] or because it is simply more convenient to choose a neurotypical applicant. Furthermore, the habit of categorical pre-assessment might lead to social reductions, e.g. judging any atypical behaviour as a symptom of a disorder, or ascribing agency only when the actions cannot be explained as purely symptomatic[3].

[1] Large databases are already available to anyone online at f.i. Youtube. Adding to these databases is easy for anyone with a microphone.

[2] Common misunderstandings of schizophrenia are the conflation with dissociative identity disorder and the association with harmful behaviour towards others (Lindberg, 2018).

[3] Further, a diagnosed subject might unconsciously identify with, and conform to, the socially projected typical characterization of his diagnosis.

Finally, it is also worth considering the effect of a move towards a more granular description of mental disorders, e.g. through functional dimensions, on the public conversation and understanding of mental disorders. Dissolving the current diagnostical categories might create a void in the common vocabulary that is unlikely to be filled with detailed descriptors of each functional dimension. Finding the right simplifications could affect the risks presented.

Appendix A

Theory of mind

For all three diagnoses, investigated in the thesis, impairments in theory of mind (ToM) are common. This might have affected the patients' descriptions, given the ToM condition in the Frith–Happé animations task.

A.1 ToM in autism spectrum disorder

In 3 meta-analyses, Yirmiya et al. (1998) compared theory of mind (ToM) abilities in autistic participants, participants with mental retardation (MR) and typically developed participants. They found significantly impaired ToM abilities in the autistic participants and the MR participants, with the autistic participants being significantly worse than MR participants (Yirmiya et al., 1998). A more recent meta-analysis of 932 participants with autism showed an emotion recognition difficulty, with fear being harder to recognize than happiness (Uljarevic and Hamilton, 2013).

A.2 ToM in schizophrenia

Multiple reviews (Brüne, 2005; Harrington et al., 2005) and meta-analyses (Bora et al., 2009; Sprong et al., 2007) has reported considerable ToM impairment in schizophrenic patients, along with, though to a lesser degree, patients in remission.

A.3 ToM in depression

In a meta-analysis of 18 studies, Bora and Berk (2016) found significant ToM impairment in participants with major depressive disorder, compared to healthy controls, and a significant relation between severity of depressive symptoms and ToM impairment. Another meta-analysis of 22 studies by Dalili et al. (2015) found impaired recognition of happiness, surprise, fear, anger, and disgust, but not sadness. Inoue et al. (2006) followed 50 patients for one year during remission and found that patients with ToM impairments in second order false belief during remission relapsed more frequently than those without the impairment.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abell, F., Happé, F., and Frith, U. (2000). Do triangles play tricks? attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, 15(1):1 – 16.

Adam, D. (2013). On the spectrum. *Nature*, 496(7446):416.

Addington, D., Azorin, J., Falloon, I., Gerlach, J., Hirsch, S., and Siris, S. (2002). Clinical issues related to depression in schizophrenia: an international survey of psychiatrists. *Acta Psychiatrica Scandinavica*, 105(3):189–195.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. Author, Washington, DC, 4th ed., text rev. edition.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Author, Washington, DC, 5th ed. edition.

Asperger, H. (1944). Die „autistischen psychopathen" im kindesalter. *Archiv für psychiatrie und nervenkrankheiten*, 117(1):76–136.

Bang, D. (2009). Pronominal use in asd. [Unpublished bachelor thesis; Title not confirmed.].

Baxter, A. J., Brugha, T., Erskine, H., Scheurer, R., Vos, T., and Scott, J. (2015). The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, 45(3):601–613.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., and Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia research*, 153(1):9–17.

Bliksted, V., Frith, C., Videbech, P., Fagerlund, B., Emborg, C., Simonsen, A., Roepstorff, A., and Campbell-Meiklejohn, D. (2018). Hyper-and hypomentalizing in patients with first-episode schizophrenia: fmri and behavioral studies. *Schizophrenia bulletin*.

Bliksted, V., Videbech, P., Fagerlund, B., and Frith, C. (2017). The effect of positive symptoms on social cognition in first-episode schizophrenia is modified by the presence of negative symptoms. *Neuropsychology*, 31(2):209.

Bone, D., Chaspari, T., Audhkhasi, K., Gibson, J., Tsiartas, A., Van Segbroeck, M., Li, M., Lee, S., and Narayanan, S. (2013). Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In *INTERSPEECH*, pages 182–186.

Bora, E. and Berk, M. (2016). Theory of mind in major depressive disorder: a meta-analysis. *Journal of affective disorders*, 191:49–55.

Bora, E., Yucel, M., and Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia research*, 109(1):1–9.

Brüne, M. (2005). "theory of mind"" in schizophrenia: a review of the literature. *Schizophrenia bulletin*, 31(1):21–42.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Cuthbert, B. N. and Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of rdoc. *BMC medicine*, 11(1):126.

Dalili, M., Penton-Voak, I., Harmer, C., and Munafò, M. (2015). Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological medicine*, 45(6):1135–1144.

Droettboom, M., Caswell, T. A., Hunter, J., Firing, E., Nielsen, J. H., Varoquaux, N., Root, B., Elson, P., Dale, D., Lee, J.-J., de Andrade, E. S., Seppänen, J. K., McDougall, D., May, R., Lee, A., Straw, A., Stansby, D., Hobson, P., Yu, T. S., Ma, E., Gohlke, C., Silvester, S., Moad, C., Schulz, J., Vincent, A. F., Würtz, P., Ariza, F., Cimarron, Hisch, T., and Kniazev, N. (2017). matplotlib/matplotlib v2.0.2.

FabFilter (2014). Fabfilter pro-q 2.

Fitzgerald, M. (2012). Schizophrenia and autism/asperger's syndrome: Overlap and difference. *Clinical Neuropsychiatry*, 9(4).

Flint, A. J., Black, S. E., Campbell-Taylor, I., Gailey, G. F., and Levinton, C. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, 27(3):309 – 319.

Fried, E. I. and Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the star* d study. *Journal of affective disorders*, 172:96–102.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M., and Gaigg, S. B. (2017). Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis. *Autism Research*, 10(3):384–407.

Goldstein, G., Minshew, N. J., Allen, D. N., and Seaton, B. E. (2002). High-functioning autism and schizophrenia: a comparison of an early and late onset neurodevelopmental disorder. *Archives of Clinical Neuropsychology*, 17(5):461–475.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

Harrington, L., Siegert, R., and McClure, J. (2005). Theory of mind in schizophrenia: a critical review. *Cognitive neuropsychiatry*, 10(4):249–286.

Helfer, B. S., Quatieri, T. F., Williamson, J. R., Mehta, D. D., Horwitz, R., and Yu, B. (2013). Classification of depression state based on articulatory precision. In *Interspeech*, pages 2172–2176.

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine Learning, Volume III*, pages 555–610. Elsevier.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Inoue, Y., Yamada, K., and Kanba, S. (2006). Deficit in theory of mind is a risk for relapse of major depression. *Journal of affective disorders*, 95(1):125–127.

iZotope, I. (2017). izotope rx 6 audio editor.

Kanner, L. et al. (1943). Autistic disturbances of affective contact. *Nervous child*, 2(3):217–250.

Kim, J. A., Szatmari, P., Bryson, S. E., Streiner, D. L., and Wilson, F. J. (2000). The prevalence of anxiety and mood problems among children with autism and asperger syndrome. *Autism*, 4(2):117–132.

Kliper, R., Portuguese, S., and Weinshall, D. (2015). Prosodic analysis of speech and the underlying mental state. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 52–62. Springer.

Ladegaard, N., Lysaker, P. H., Larsen, E. R., and Videbech, P. (2014). A comparison of capacities for social cognition and metacognition in first episode and prolonged depression. *Psychiatry Research*, 220(3):883–889.

Landtblom, A.-M., Fridriksson, S., Boivie, J., Hillman, J., Johansson, G., and Johansson, I. (2002). Sudden onset headache: a prospective study of features, incidence and causes. *Cephalalgia*, 22(5):354–360.

Levine, H. L. (1991). Otorhinolaryngologic causes of headache. *The Medical clinics of North America*, 75(3):677–692.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). Medical image classification with convolutional neural network. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 844–848. IEEE.

Li, T. L., Chan, A. B., and Chun, A. (2010). Automatic musical pattern feature extraction using convolutional neural network. In *Proc. Int. Conf. Data Mining and Applications*.

Lindberg, R. (2018). Myths in schizophrenia.

London, E. B. (2014). Categorical diagnosis: a fatal flaw for autism research? *Trends in neurosciences*, 37(12):683–686.

Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B., and Allen, N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586.

Marín, J. L., Rodríguez-Franco, M. A., Chugani, V. M., Maganto, M. M., Villoria, E. D., and Bedia, R. C. (2018). Prevalence of schizophrenia spectrum disorders in average-iq adults with autism spectrum disorders: A meta-analysis. *Journal of autism and developmental disorders*, 48(1):239–250.

Martinez, H. P., Bengio, Y., and Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33.

Martínez-Sánchez, F., Muela-Martínez, J. A., Cortés-Soto, P., García Meilán, J. J., Vera Ferrándiz, J. A., Egea Caparrós, A., and Pujante Valverde, I. M. (2015). Can the acoustic analysis of expressive prosody discriminate schizophrenia? *The Spanish Journal of Psychology*, 18:E86.

Matsuo, J., Kamio, Y., Takahashi, H., Ota, M., Teraishi, T., Hori, H., Nagashima, A., Takei, R., Higuchi, T., Motohashi, N., et al. (2015). Autistic-like traits in adult patients with mood disorders and schizophrenia. *PLoS One*, 10(4):e0122711.

McFee, B., McVicar, M., Nieto, O., Balke, S., Thome, C., Liang, D., Battenberg, E., Moore, J., Bittner, R., Yamamoto, R., Ellis, D., Stoter, F.-R., Repetto, D., Waloschek, S., Carr, C., Kranzler, S., Choi, K., Viktorin, P., Santos, J. F., Holovaty, A., Pimenta, W., and Lee, H. (2017). librosa 0.5.0.

Mirowski, P., Ranzato, M., and LeCun, Y. (2010). Dynamic auto-encoders for semantic indexing. In *Proceedings of the NIPS 2010 Workshop on Deep Learning*, pages 1–9.

Moore II, E., Clements, M. A., Peifer, J. W., and Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107.

Olsen, L. R. (2018). groupdata2: Creating groups from data. R package version 1.0.0.9000.

Parola, A., Simonsen, A., Bliksted, V., and Fusaroli, R. (2018). T138. acoustic patterns in schizophrenia: A systematic review and meta-analysis. *Schizophrenia Bulletin*, 44(suppl_1):S169–S169.

Rapcan, V., D'Arcy, S., Yeap, S., Afzal, N., Thakore, J., and Reilly, R. B. (2010). Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering and Physics*, 32(9):1074–1079.

Rapoport, J., Chavez, A., Greenstein, D., Addington, A., and Gogtay, N. (2009). Autism spectrum disorders and childhood-onset schizophrenia: clinical and biological contributions to a relation revisited. *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(1):10–18.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Morency, L.-P., et al. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Singh, I. and Rose, N. (2009). Biomarkers in psychiatry. *Nature*, 460(7252):202.

Siris, S. G., Addington, D., Azorin, J.-M., Falloon, I. R., Gerlach, J., and Hirsch, S. R. (2001). Depression in schizophrenia: recognition and management in the usa. *Schizophrenia research*, 47(2):185–197.

Sprong, M., Schothorst, P., Vos, E., Hox, J., and Van Engeland, H. (2007). Theory of mind in schizophrenia: meta-analysis. *The British Journal of Psychiatry*, 191(1):5–13.

Teager, H. and Teager, S. (1990). Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer.

Tinelli, M., Kanavos, P., and Grimaccia, F. (2016). The value of early diagnosis and treatment in parkinson's disease. *A literature review of the potetial clinical and socioeconomic impact of targeting unmet needs in Parkinson's disease. London School of Economics*.

Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*. PhD thesis, Oxford University, UK.

Uljarevic, M. and Hamilton, A. (2013). Recognition of emotions in autism: a formal meta-analysis. *Journal of autism and developmental disorders*, 43(7):1517–1526.

Van Schalkwyk, G. I., Peluso, F., Qayyum, Z., McPartland, J. C., and Volkmar, F. R. (2015). Varieties of misdiagnosis in asd: an illustrative case series. *Journal of autism and developmental disorders*, 45(4):911–918.

Van Wijngaarden-Cremers, P. J., van Eeten, E., Groen, W. B., Van Deurzen, P. A., Oosterling, I. J., and Van der Gaag, R. J. (2014). Gender and age differences in the core triad of impairments in autism spectrum disorders: a systematic review and meta-analysis. *Journal of autism and developmental disorders*, 44(3):627–635.

Whiteley, P., Todd, L., Carr, K., and Shattock, P. (2010). Gender ratios in autism, asperger syndrome and autism spectrum disorder. *Autism Insights*, 2:17.

Wiseman, J. (2016). py-webrtcvad. https://github.com/wiseman/py-webrtcvad.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*.

Yirmiya, N., Erel, O., Shaked, M., and Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, 124(3):283.