Modelling of Speech Aspects in Parkinson's Disease by Multitask Deep Learning

Modellieren von Sprachaspekten bei Parkinson mittels Multitask Deep Learning

Master's Thesis in Medical Engineering

submitted
by

Martin Korbinian Heinrich Strauß

born December 7, 1990 in Bad Aibling

Written at

Lehrstuhl für Mustererkennung (Informatik 5)
Department Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg.

Advisor: Prof. Dr.-Ing. Elmar Nöth
M. Sc. Juan Camilo Vásquez-Correa
PD Dr.-Ing. Tino Haderlein

Started: March 12, 2019

Finished: May 10, 2019

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Studien- und Diplomarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 9. Mai 2019

Abstract

Parkinson's disease is a progressive neurodegenerative disorder with a variety of motor and non-motor symptoms. Although also other factors are influenced by the disease, the current evaluation process relies mostly on motor aspects and is often subjective. While speech deficits can be found in a majority of patients, its analysis is still underrepresented in the clinical assessment. To increase objectivity and enable long-term monitoring of the patient's status, several computational methods have been been proposed in the literature. Along with the success of deep learning, multitask techniques received more and more attention in recent years. Hence, this Master's thesis proposes the use of a multitask neural network-based approach in order to assess multiple aspects of Parkinsonian speech. The data set included various recordings in numerous sessions obtained from 94 Parkinson patients and 87 healthy controls. A defined set of statistical features was extracted for each utterance to be used as input to the model. The multitask setting was defined with three tasks regarding the distinction between diseased and healthy, as well as, two common Parkinson rating scales, namely the Movement Disorder Society – Unified Parkinson's Disease Rating Scale and the modified Frenchay Dysarthria Assessment. These tasks were supposed to be optimized together compared to individual networks. In order to get a deeper understanding with regard to the influence of each task and the specific recording settings, several experiments with different focuses were conducted. Additionally, the multitask setting was expanded with four additional tasks to exploit the variability of this method. The experimental results demonstrate the classification capabilities with accuracy values of 81.73%, 52.45% and 43.56% for the respective three tasks based on a per session evaluation. These results improve the outcome of individually trained networks for values between 3 and 16 percent points. Further comparison against an Adaboost baseline does not show a clear improvement, however, the proposed model delivers competitive results, especially with focus on other neural network approaches. Thus, this work gives new insights to the application of multitask deep learning to Parkinsonian speech and builds the basis for further research in the field.

Übersicht

Die Parkinson-Krankheit ist eine fortschreitende neurodegenerative Erkrankung mit einer Vielzahl von motorischen und nicht motorischen Symptomen. Obwohl auch andere Faktoren von dieser Krankheit beinflusst werden, basiert die aktuelle Evaluation überwiegend auf motorischen Aspekten und ist oft subjektiv. Während Sprachdefizite bei einer Mehrzahl der Patienten auftreten, ist deren Analyse in der aktuellen klinischen Begutachtung noch immer unterrepräsentiert. Um mehr Objektivität und eine Langzeitüberwachung zu ermöglichen, wurden bereits einige computerbasierte Methoden in der Fachliteratur empfohlen. Zusammen mit dem Erfolg von Deep Learning haben auch neuronale Netze mit einem Multitask-Ansatz mehr und mehr Aufmerksamkeit erhalten. Dementsprechend schlägt diese Masterarbeit die Anwendung von neuronalen Netzen mittels Multitask vor, um mehrere Sprachaspekte bei Parkinson zu begutachten. Der Datensatz beinhaltet unterschiedliche Aufnahmen in mehreren Sitzungen von 94 Parkinsonpatienten und von 87 Teilnehmern einer gesunden Kontrollgruppe. Aus den Äußerungen wurde ein vordefinierter Satz an statistischen Merkmalen gewonnen, um die Eingangsdaten des Modells zu erstellen. Der Multitask-Rahmen wurde anhand von drei Aufgaben definiert. Diese beinhalteten sowohl die Unterscheidung zwischen krank und gesund, als auch zwei geläufige Parkinson Bewertungsskalen, namentlich die Movement Disorder Society – Unified Parkinon's Disease Rating Scale und die modifizierte Version des Frenchay Dysarthria Asessment. Diese Aufgaben sollten gemeinsam optimiert werden und im Anschluss mit individuellen Netzwerken verglichen werden. Um ein tieferes Verständnis des Einflusses einer jeden Aufgabe und der verschiedenen Aufnahmebedingungen zu erhalten, wurden mehrere Experimente mit unterschiedlichem Fokus durchgeführt. Zudem wurde der Multitaskansatz um vier Aufgaben erweitert, um die Variabilität dieser Methode auszutesten. Die experimentellen Ergebnisse demonstrieren diese Klassifizierungsfähigkeiten mit Erkennungswerten von 81.73%, 52.45% und 43.56% für die entsprechenden Aufgaben basierend auf einer Evaluierung pro Sitzung. Diese Ergebnisse verbessern den Ausgang der einzeln trainierten Netzwerke um Werte zwischen 3 und 16 Prozentpunkten. Ein weiterer Vergleich mit einer Adaboost Baseline-Methode liefert keine klare Verbesserung, jedoch zeigt das vorgeschlagene Modell konkurrenzfähige Ergebnisse, vor allem mit dem Fokus auf andere Ansätze mit neuronalen Netzen. In diesem Sinn gibt diese Arbeit neue Einblicke in die Anwendung von Deep Learning mittels Multitask für Sprache von Parkinson und bildet die Basis für weiterer Forschung auf diesem Gebiet.

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Accuracy
ANN	Artificial neural network
AUC	Area Under the receiver operating Curve
BP	Backpropagation
CNN	Convolutional neural network
CV	Cross validation
DDK	Diadochokinetic
DL	Deep Learning
DNN	Deep neural network
FDA	Frenchay Dyarthria Assessment
GMM-UBM	Gaussian Mixture Model - Universal background model
HC	Healthy controls
i-vectors	Identity vectors
LLD	Low level descriptors
LR	Learning rate
m-FDA	modified Frenchay Dyarthria Assessment
MDS-UPDRS	Movement Disorder Society - Unified Parkinson's Disease Rating Scale
MDS-UPDRS-III	Third section of the Movement Disorder Society - Unified Parkinson's Disease Rating Scale
MLP	Multi-layer Perceptron
MTL	Multitask learning
openSMILE	open Speech and Music Interpretation by Large-space Extraction

PD	Parkinson's disease
RBD	Rapid eye movement disorder
ReLU	Rectified Linear Unit
SVR	Support Vector Regression
TW	Task weight
UAR	Unweighted average recall

Chapter 1

Introduction

1.1 Motivation

Parkinson's disease (PD) is a chronic, progressive neurological disorder characterised by a continuous loss of dopaminergic neurons in the substantia nigra of the basal ganglia section in the midbrain [Hor98]. Reportedly, after Alzheimer's disease it is the most common neurodegenerative disorder affecting approximately 1% of the world population over 60 years [Tys17]. In Europe the prevalence for this age group is estimated to be around 1280 to 1500 per 100,000 inhabitants [vC05]. Typical symptoms include motor and non-motor related signs leading to a rapid decrease in the quality of life [Sch06]. Thereby, motor impairments are typically described by shuffling gait patterns, resting tremor or postural instability, while non-motor aspects vary, among others, from issues with the sensory system, to problems with sleep or emotions [Jan08]. Besides the mentioned aspects speech is another factor effected by PD. In fact, Ramig et al. [Ram08] reported that up to 89% of diagnosed PD patients show speech deficits. The observed deficits are reduced loudness, monopitch, monoloudness, breathy hoarse voice quality and imprecise articulation. These symptoms are usually summarized under the term *hypokinetic dysarthria* [Log78]. As a growth of aging population in industrial nations is expected in the following years, the incidence of people with PD will most likely increase as well [Ree14]. However, since there is no cure available for PD up to date, the moment of initial diagnosis and onset of early treatment is substantial for the maintenance of a patient's quality of life.

The *Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS-UPDRS)* [Goe08] is the most common assessment tool to evaluate and rate the status of PD patients. However, the MDS-UPDRS requires the individual to be personally inspected by a physician in clinical interviews and observations. On the one hand, this potentially leads to a subjective

evaluation, since not all symptoms are judged uniformly between all clinicians. On the other hand, to reach the clinic on a regular basis for strongly affected patients or for people living in rural and remote areas may be difficult. In addition, drug prescriptions, the recommended dosage as well as treatment strategies thereby depend heavily on the current state of the patient as it is presented to the physician. Further, only one item of the MDS-UPDRS is related to speech. In fact, Ramig et al. [Ram08] stated out that only 2–3% of PD patients receive speech treatment of any kind. These aspects illustrate the under-representation of speech in the current PD evaluation process. As a consequence, further evaluation methods have been proposed to include speech into the PD assessment. One frequent example here is the *Frenchay Dysarthria Assessment (FDA)* [End08]. This procedure includes several aspects related to dysarthria, such as reflexes, respiration, lips, tongue and palate movement, intelligibility, among others. However, it still requires the patient and doctor to be present in the same room.

Correspondingly, it is an active field of research to include speech based computational methods into the PD assessment to increase objectivity and to make it more widely accessible. The advantage is, for instance, that speech exercises can easily be included into a smartphone application and hence, a large amount of data can be generated on a regular basis. With the use of machine learning algorithms, these data can be processed and provide additional support to clinical experts. In recent years artificial neural networks (ANN) and deep learning (DL) methods have been very successful and achieved state-of-the-art results in many fields, including speech processing [vdO16], computer vision [Sze16] and natural language processing [Dev18]. These techniques usually require a large amount of labelled data to be able to learn from examples. However, to hold a sufficient amount of labelled samples is often very expensive and hard to receive in healthcare applications, due to the sensitive nature of the data or rare disease types. One approach to overcome these issues could be to include multitask learning (MTL) into the learning strategy. In MTL multiple aspects are supposed to be solved at the same time, in contrast to building a model for each task individually. In theory, this leads to a better performance due to shared representations, from which each single task may benefit [Car97]. This enables the model to learn from a smaller total amount of examples.

In this sense, the main aim of this Master's thesis is to create and evaluate a multitask deep learning framework applied to Parkinson's disease speech data. Therefore, a defined set of features obtained from different recorded exercises performed by Colombian native speakers with PD and healthy controls (HC) is processed and fed to the network. The outcome will be compared to the respective single task models and to a baseline Adaboost approach using different evaluation metrics. Applying this method to Parkinsonian speech could potentially provide additional information to

physicians and clinical experts for better, more objective assessment of Parkinson's disease. This chapter continues with a description of speech deficits that are being found in PD. Following, a review of the current state of the art and related work will follow. In Chapter 2 the data which were used in this thesis will be introduced. A description of the data set, recording conditions, exercises, as well as, how the samples were labelled will be explained in detail. Chapter 3 outlines the methods and theoretical foundations with the feature computations and an introduction to deep learning with a focus on multitask optimization. Additionally, different evaluation methods are being discussed and how they are used to give feedback about the performance of the chosen strategy. In Chapter 4 the results of all experiments are being presented and discussed in detail in the following Chapter 5. Finally, the outcome of this work is summed up in a conclusion (Chapter 6) and an outlook towards further research possibilities is given in Chapter 7.

1.2 Speech deficits in Parkinson's disease

The speech disorders in Parkinson's disease can be roughly classified into three dimensions: *prearticulatory*, *prelinguistic* and *linguistic* [GL17]. They are manifested in several different aspects, such as *articulation* (prearticulatory), *phonation* (prelinguistic), *prosody* (prearticulatory) and *intelligibility*. Ackermann and Ziegler [Ack91], for instance, described the articulation deficits in PD to be linked to a reduced amplitude and speed in movements of the lip, jaw and tongue. Although the speed rate seems to be preserved, the production of stop consonant suffers from this articulatory imprecision. Further, it was reported that PD patients have a significantly smaller vowel space area [Tja13] and a lower vowel articulation index [Rus13] compared to HC, leading to lower articulatory capacities. In phonation, PD patients tend to have problems with closing their vocal folds properly, which leads to unstable vibration patterns [Han84]. This abnormality might be connected to rigid muscles in the larynx. Further, an increase in average fundamental frequency, jitter [Ban13] or shimmer values [Ken03] has been reported. In contrast, Rusz et al. [Rus11] reported that prosody impairments in PD can be found in various changes in speech rate and pause characteristics. Further, they described monotonicity and monoloudness to be present in the disease. Reduced intelligibility was observed and described, for instance, in Logemann et al. [Log78]. This is also connected to shallower formant slopes [Wei98] which are associated with a reduced articulation velocity [GL17].

1.3 Related work

Parkinson's disease has been studied in various degrees since many years. Although many works conducted in the field still focused on the motor impairments developed by the patients as, for instance, gait problems [SS19], speech deficits in PD were of increasing importance in recent years. Many research groups tried to describe Parkinsonian speech patterns using objective computational methods.

For instance, Dimauro et al. [Dim17] focused in their work on an objective assessment of intelligibility in PD. By using Google®'s speech-to-text system they analyzed falsely recognized words from recordings of PD patients to obtain several measures being compared to the original MDS-UPDRS value for speech. A similar approach was used by Orozco-Arroyave et al. [OA16] in 2016 when they used Google's ® automatic speech recognition API to describe intelligibility in PD using speech data from three different languages. They concluded, that it is possible to estimate the neurological state of a PD patient with a Spearman correlation of 0.72 compared to the neurological expert's evaluation. Smith and Quatieri [Smi17] analyzed articulatory speech markers and phonemic timing in PD speech. They found a moderate correlation towards the motor and cognitive symptoms and weaker correlation with depressive symptoms. The authors concluded that it is possible to discriminate the impact of non-motor symptoms based on PD speech. In 2013 Rusz et al. [Rus13] proposed an acoustic analysis based on extracted vowels from sustained phonations, sentences, passage reading and monologue. Their results indicate that monologue is the most suitable task to differentiate PD from HC. Just recently, Montaña et al. [Mon18] proposed a novel algorithm to detect voice onset time segments of diadochokinetic syllables. Their method reached 94.4% in distinguishing between PD and HC using a leave-one-out cross validation. Tsanas et al. [Tsa14] used phonations of sustained vowels for an objective assessment of PD. 156 phonations of the /a/ vowel were characterised with 309 dysphonia measures to select relevant features. These features were used to classify the voice assessment as acceptable or unacceptable, based on a clinical expert rating. With an overall accuracy of 90% the authors concluded, that their method is suitable for the task. However, no speaker independence between the train and test set was guaranteed in the study, which makes the results very optimistic. Hlavnička et al. [Hla17] showed that PD speech deficits can also be found in people with rapid eye movement disorder (RBD). They used speech data from 50 RBD patients, 30 untreated PD patients and 50 HC to provide an automated speech analysis for respiration, phonation, articulation and timing based on acoustic features. Their results indicate that such acoustic evaluation could be beneficial to identify subjects with high risk of developing neurodegenerative diseases like PD. Galaz et al. [Gal16] focused on the prosodic analysis of PD speech to

identify hypokinetic dysarthria. They applied a sequential floating feature selection algorithm and a random forest classifier to data from 98 PD patients and 51 HC. As a result the authors found a reduced fundamental frequency variability in PD and proposed develpoment of prosodic features. A different approach was followed by Oung et al. [Oun18], when they used features obtained from speech and motion data using wavelet transforms. Instead of solving a binary problem they defined a multiclass problem by differentiating the PD patients into mild, moderate and severe disease states compared to HC. The experimental results showed good classification accuracy with 95% when both signal types were combined. In the work of Arias-Vergara et al. [AV18] individual speaker models were created based on a Gaussian Mixture Model – Universal background model (GMM-UBM) and an i-vectors approach. The authors conducted longitudinal and at-home tests for spontaneous speech and readtext recordings obtained from different recording conditions to evaluate the dysarthria level of the patients. The authors concluded, that their work indicates the possibility of tele-monitoring of the neurological state of a PD patient using their method. Similarly, GMM-UBM and i-vectors were used by Moro-Velázquez et al. [MV18] to perform speaker recognition with an accuracy of up to 87% and an area under the receiver operating curve (AUC) of 0.93. In addition, the authors proposed d to include kinetic information in future work. Considering that smartphones are now available around the world, new possibilities in so-called e-medicine or mobile health emerge [Ste15]. In fact, remote monitoring of patients with certain diseases may be one of the major medical markets in the upcoming future [Mal19]. In recent years, there were a few approaches showing their application to the investigation of Parkinson speech. Zhan et al. [Zha16] introduced a high-frequency platform to monitor symptoms of PD patients. They called their application *HopkinsPD* and collected data for a duration of six month from 226 participants. Their dataset includes voice, gait, dexterity and reaction time recordings. As a classifier they used the random forest algorithm to distinct between measurements taken before and after medication input. With an accuracy of 71.0%, the authors concluded, that their technique enables remote monitoring of PD patients. Schwab and Karlen [Sch19] proposed *PhoneMD*, a smartphone application for long-term data collection of PD patients. They included walking, voice, tapping and memory tests into the application and gathered data from 1853 patients. With a result of 0.85 for the AUC the authors concluded, that smartphone data obtained over a long period can be used potentially as support for PD diagnosis. Another smartphone-based approach was presented in 2018 by Zhang et al. [Zha18]. Their application called *DeepVoice* included a DNN based on convolutional neural networks (CNN) from spectrogram images. The accuracy obtained from 10-second long /a/ phonations from the mPower dataset [Bot16] was at best 90.45%. As a consequence, the authors proposed to use the application for PD identification. However, those

three examples lack of deep speech analysis, since they only use sustained vowel /a/ utterances. Further, Wan et al. [Wan18] presented an approach combining speech and movement pattern recorded by a smartphone. They aimed to evaluate the severity of symptoms of PD patients using several classifiers, such as logistic regression, random forest, k-nearest neighbours, M5P and a Deep Multi-Layer Perceptron (DMLP). The study included data from 20 PD subjects and 20 HC. The experiments indicate that DMLP performs best out of all selected classifiers with an accuracy of 80 % doing a leave-one-out cross-validation.
With the success of deep learning, it also became a main research topic in PD speech. The approach of Grósz et al. [Gró15] was the winner of the Interspeech 2015 Computational Paralinguistics Challenge [Sch15]. They applied a deep neural network and a Gaussian processes regression algorithm on the given task and found out that both approaches are showing superior performance to the support vector machine baseline. Further, they achieved an increased performance by using a clustering method to identify the files belonging to one speaker. In the work of Gaballah et al. [Gab18] deep neural networks processed extracted speech features from 11 PD and 10 HC to measure the perceived quality. Seven different speech amplifiers were applied to the speech signals and given to subjective raters together with the original recordings. With their ANN they reached a correlation of 0.81 for a comparison between objective and subjective scores. Further, Tu et al. [Tu17] proposed to add an intermediate interpretable layer into the neural network which can be used as a bottleneck feature extractor. The aim of this was to be able to better understand the decisions made by the network. Their data contained different types of dysarthria including Huntington's disease and PD. The results indicated that learning the bottleneck features and predicting the dysarthria level with a combined approach performs better then training them sequentially. On the contrary, Cernak et al. [Cer17] focused in their work on the phonation composition in PD speech. They proposed a two-stage procedure, where in the first stage statistical features were learned to differentiate modal versus non-modal phonations with a neural network. Secondly, statistics were learned from speech data comprised from 50 PD patients and 50 HC, respectively. The obtained features were used to predict the dysarthria severity of the patients. Another aspect of speech was taken into account by Vásquez-Correa [VC17] et al. in 2017. The authors focused on the start and stop movements of the vibration folds to assess articulation issues in PD. Their method was based on a CNN applied to time-frequency spectrograms of the extracted voiced/unvoiced transition segments. With 89% accuracy the study showed promising results in distinguishing PD from HC. Just recently, Berus et al. [Ber18] used multiple ANNs on features selected from 26 different voice samples for 20 PD patients and 20 healthy controls. They did a feature selection beforehand based on Pearson's and Kendall's correlation coefficient, as well

as, a principal component analysis and self-organizing maps. The decision for “HC” or “PD” is done by combining multiple classifiers and performing a majority vote. After fine-tuning a test accuracy of 81.33% using Kendall’s correlation based features is reached.
In Vásquez-Correa et al. [VC18a] the authors used a multitask learning approach to assess the severity of PD patient’s dysarthria. Therefore, eleven aspects regarding speech were taken into account and were optimized simultaneously by a CNN model. Thus, the voiced and unvoiced segments of speech were taken as an input for their network. The obtained results from up to 4 percent points in increased accuracy indicated that using a multitask learning approach is beneficial for this problem compared to train individual networks for each task. The authors explained their outcome with more generalizable feature maps created by the multitask CNN.

Chapter 2

Data

2.1 Data set

2.1.1 Participant details

Table 2.1: Participant details. Demographic information for age, disease duration, m-FDA and MDS-UPDRS-III rates of the Parkinson's disease patients and healthy controls, divided into male and female participants.

	PD patients		**Healthy speakers**	
	male	**female**	**male**	**female**
Number of speakers	47	47	44	43
Age [years] (mean $\pm$ std)	62.21 $\pm$ 10.27	59.79 $\pm$ 11.28	64.93 $\pm$ 10.5	60.93 $\pm$ 8.32
Range of age [years]	33 – 81	29 – 83	42 – 86	49 – 83
Disease duration [years] (mean $\pm$ std)	7.92 $\pm$ 5.26	10.56 $\pm$ 10.64	-	-
Range of disease duration [years]	0.4 – 20	0 – 43	-	-
MDS-UPDRS-III (mean $\pm$ std)	36.47 $\pm$ 20.16	36.85 $\pm$ 18.38	-	-
Range of MDS-UPDRS-III	6 – 92	9 – 106	-	-
m-FDA (mean $\pm$ std)	26.4 $\pm$ 7.73	26.29 $\pm$ 7.93	9.04 $\pm$ 8.24	6.64 $\pm$ 7.03
Range of m-FDA	12 – 41	10 – 47	0 – 25	0 – 23

The data used for this work consisted of speech recordings from native speakers of Spanish living in Colombia. Details about the data can be found in Table 2.1. In total they included 94 subjects diagnosed with Parkinson's disease and 87 healthy controls. 47 of the PD patients where female and 47 male, while 43 of the HC participants were male and 44 female. The mean age in years was 62.21 ± 10.27 for PD and 64.93 ± 10.5 for HC male participants. Females were in average 59.79 ± 11.28 (PD) and 60.93 ± 8.35 (HC) years old. The time in years since the diagnosis was

10.56 ± 10.64 for the female patients and 7.92 ± 5.56 for males.
Each PD participant was recorded in up to 8 recording sessions performing several speech tasks, though not every subject did all sessions. Consequently, the amount of sessions recorded from PD patients was 276 and 135 for HC. In addition, seven patients were recorded for a time span of four month in so-called *at-home* sessions. This took place once a month with four sessions on that day. Hence, 16 additional recording sessions were obtained for these subjects. Overall, this led to a total of 427 recording sessions included into the data set. Further, all patients were recorded in the ON-state meaning no more then 3 hours after the morning medication was taken.

2.1.2 Speech recordings

Each participant was required to fulfil a set of speech tasks during one session which were designed to investigate articulation and prosody aspects in speech of PD patients [OA14]. Details about the exercises are explained in the following:

Diadochokinetic (DDK) Exercises:

A rapid repetition of 6 words or syllables: */pa-ta-ka/*, */pa-ka-ta/*, */pa-te-ka/*, */pa/*, */ta/*, */ka/*. These phrases are used to evaluate the articulation capabilities of a participant, since one is required to use certain muscular parts and the velum to form these syllables.

Sentences:

One task for the evaluation of prosody contains the repetition of 10 syntactically simple and complex sentences and sentences with emphasis on specific words (see capital letters) e.g. *"Mi casa tiene tres cuatros."* (Translation: My house has three rooms.), *"Juan se ROMPIÓ una PIERNA cuando iba en la MOTO."* (Translation: Juan BROKE his LEG when was driving his MOTORCYCLE.).

Reading text:

Reading a predefined patient and doctor dialog is another evaluation method concentrating on prosody. The text is phonetically balanced and contains all Spanish sounds.

Monologue:

Performing a monologue speaking about daily habits to evaluate spontaneous speech. This is also a prosody evaluation task.

These tasks recorded in 427 sessions led to a total amount of 5145 utterances to be included into the data set. Table 2.2 shows the distribution of all exercises performed by PD patients and HC.

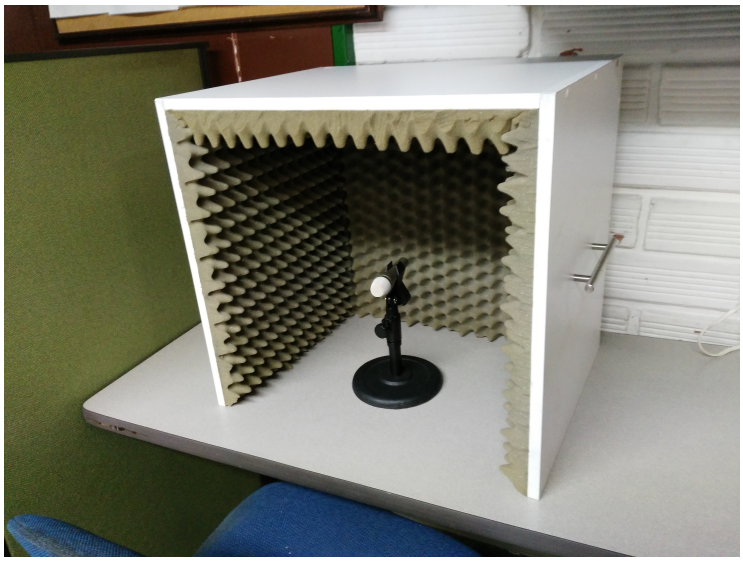

Figure 2.1: Portable soundproof booth.

The average duration in seconds of the monologues was 95.48 ± 51.32 for PD and 54.59 ± 33.44 for healthy controls. A complete set of all spoken sentences and the patient-doctor dialog can be found in the Appendix A.

The recordings were executed in different acoustic conditions, namely in a professional *soundproof booth* environment, a *portable soundproof booth*, a *headset* and a so-called *at-home* environment. For the soundproof booth a dynamic omnidirectional *Shure SM63L* microphone was used. The sampling rate was $f_s = 44.1\,\mathrm{kHz}$ with a 16 bit resolution, but the recordings were resampled to 16 kHz. Further, an *M-Audio Fast Track C400* audio card with 24 bits and a support for up to 96 Kbps sampling rate was used. In Figure 2.1 the portable soundproof booth is displayed. The headset was a *Logitech H390*, with a sampling frequency of 16 kHz and 16 bits depth. In the *at-home* setting, the same headset was used. More details about the speech exercises and the acoustic conditions of the soundproof booth can be found in [OA14].

Table 2.2: Distribution of the performed exercises

Exercise	**PD**	**HC**	
DDK	1411	552	1933
Sentences	1588	870	2458
Read text	292	87	379
Monologue	289	86	375
Total	3580	1565	5145

2.2 Parkinson evaluation scales

2.2.1 MDS-UPDRS-III

The most widely used evaluation method for PD is the *Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS)* [Goe08]. It consists of 4 sections concerning "non-motor experiences of daily living" (Part I), "motor experiences of daily living" (Part II), "motor examination" (Part III) and "motor complications" (Part IV). In fact, the progression of PD can be documented with section 3 of the scale which is usually referred to as MDS-UPDRS-III. Overall the MDS-UPDRS-III has a volume of 33 items based on 18 aspects with specific instructions for each task. It should be performed in no longer than 15 minutes. Each question is rated with a value from 0 to 4 referring to the clinical terms 0 = normal, 1 = slight, 2 = mild, 3 = moderate and 4 = severe.
Figure 2.2 shows the distribution of MDS-UPDRS-III values of the PD patients included into the data set of this work. The average MDS-UPDRS-III value for the male participants was 36.47 ± 20.16 and 36.85 ± 18.38 for the female ones. The overall range of values was between 6 and 92 for males and 9 to 106 for females, respectively. The statistics of the data set are also summarized in Table 2.1.

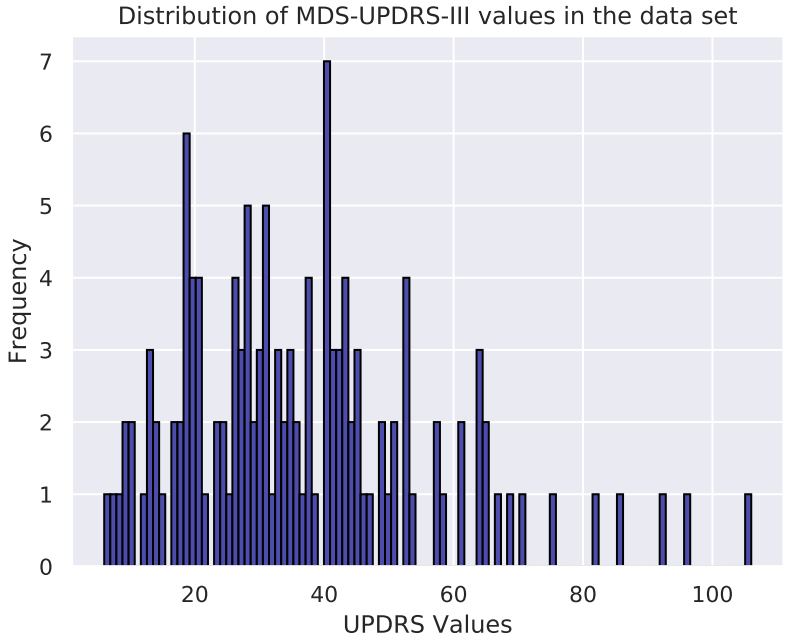

Figure 2.2: Histogram of the distribution of MDS-UPDRS-III values throughout the data

2.2.2 m-FDA

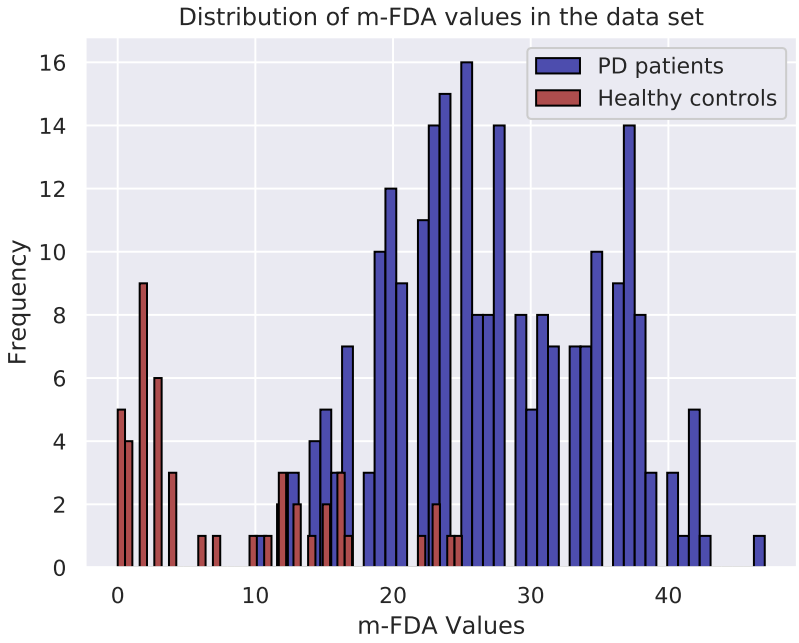

Figure 2.3: Histogram showing the data set distribution according to the modified Frenchay Dysarthria Assessment (m-FDA) for healthy controls (red) and Parkinson's disease patients (blue).

To overcome the fact that speech aspects are under-represented in PD disease state evaluation, different approaches were taken into account to assess merely speech deficits in PD patients. One common example is the *Frenchay Disarthria Assessment (FDA)* introduced and modified by Enderby et al. in 1980 and 2008 [End08]. It was originally created to evaluate the dysarthria level of patients including items of a wide range like reflexes, respiration, lips movement, palate movement, laryngeal capacity, as well as, tongue movement or intelligibility, among others. However, as in MDS-UPDRS-III, a major disadvantage of this process is that the examiner has to be with the patient to perform the assessment, which is difficult for some patients due to their reduced mobility.

As a consequence, a modified version of the FDA (m-FDA) was introduced by Vásquez-Correa et al. [VC18b] in 2018. The goal was to create an objective and reproducible version, which only relies on speech recordings. Therefore, in theory no examiner would have to be present regularly and speech recordings could be obtained remotely, for instance, with a smartphone. The assessment focuses on several aspects, namely *Respiration*, *Lips*, *Palate/Velum*, *Laryngeal*, *Tongue*, *Monotonicity* and *Intelligibility*. In total it has 13 items ranging from 0 (healthy or normal)

to 4 (very impaired). Thus, the final score ranges from 0 to 52. In Figure 2.3 the distribution of m-FDA values in the current data set is illustrated for HC and PD patients.
The average m-FDA value for male and female PD patients was 26.40 ± 7.73 and 26.29 ± 7.93, while it had a value of 9.04 ± 8.24 and 6.64 ± 7.03 for the HC participants. The range of values was 0 to 25 for males and 0 to 23 for females (see Table 2.1).

2.3 Labelling procedure

2.3.1 Task description

For the supervised learning procedure and evaluation of the model the samples had to be labelled and the tasks needed to be defined. In Table 2.3 the target tasks which were included into the multitask learning are displayed. Each one of them consisted of multiple classes. The first task was to differentiate between a PD patient and healthy controls (PD vs. HC). Next, the MDS-UPDRS-III, m-FDA and Age tasks were supposed to assign each utterance to one of four classes. Further, one task was to distinguish between the exercise performed and, in addition, in which acoustic condition it was recorded. The last task was to differentiate between male and female (Gender) leading to a 24 dimensional label vector per utterance. The vector was one-hot encoded for each task.

Table 2.3: The tasks to be learned in the multitask setting. Each task included several classes. In the main setting the first three tasks are considered and the others are included into the features. In another setting all seven were optimized at the same time.

Task number	Task name	Number of classes
1.	PD vs. HC	2
2.	MDS-UPDRS-III	4
3.	m-FDA	4
4.	Acoustic condition	4
5.	Exercise	4
6.	Gender	2
7.	Age	4

To investigate the influence of different amounts of tasks two experimental settings were introduced. The first one included the tasks to differentiate PD vs. HC, MDS-UPDRS-III and m-FDA, while the remaining tasks were added to the features as so-called *knowledge* features. This assumption is reasonable in terms of application, since the information, like gender, age or which

exercise was performed would also be known, for instance, in a smartphone application. In the second setup all tasks were included into the multitask learning approach.

2.3.2 Data imputation

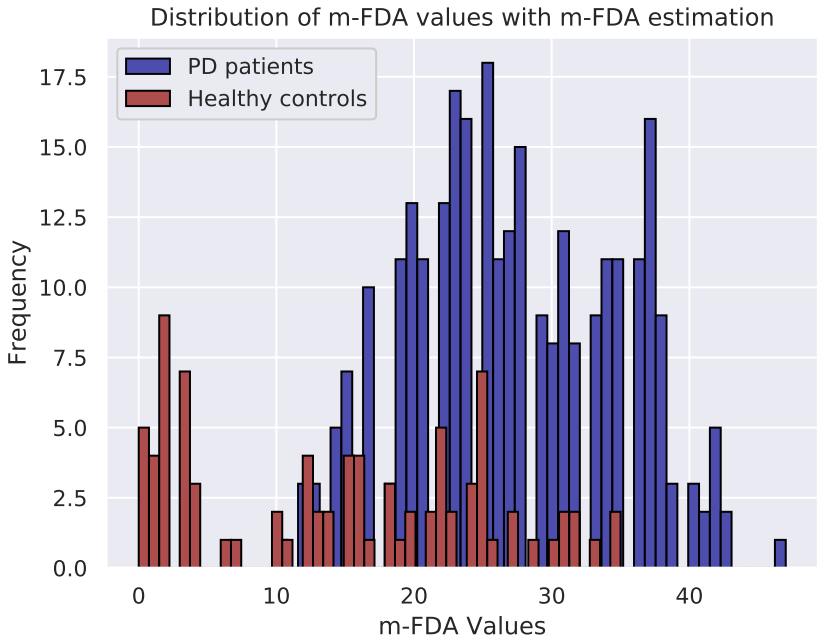

Figure 2.4: The m-FDA values distribution after the missing value estimation using a support vector regression for PD patients (blue) and healthy controls (red).

Several utterances in the data set did not include a MDS-UPDRS-III or m-FDA score. Since the amount of data is limited, a value estimation strategy was followed to be able to include the missing samples into the training process.
Healthy controls were not rated according to the MDS-UPDRS-III during the recording sessions. Hence, the assumption was made, that all HC subjects would reach a lower score than PD patients. As a consequence, their scores were sampled from a range between 0 and 6, which was the minimum occurring value in the PD patients. The remaining PD participants without a MDS-UPDRS-III score were not included into the estimation, since the rating outcome can vary from one recording session to another and no reliable value would have been available. As a consequence, these samples were labelled with –1 in order to detect them during training.
To estimate the missing m-FDA values an approach following the work of Vásquez et al. [VC18b] was followed. It is based on a Support Vector Regression (SVR) with the existing scores as

training items to estimate the missing values. The Spearman's correlation between the estimated and the real m-FDA scores in the training process was up to 0.66, which guaranteed the quality of the estimated labels. In Figure 2.4 the m-FDA value distribution is illustrated, as it was obtained after the estimation process.

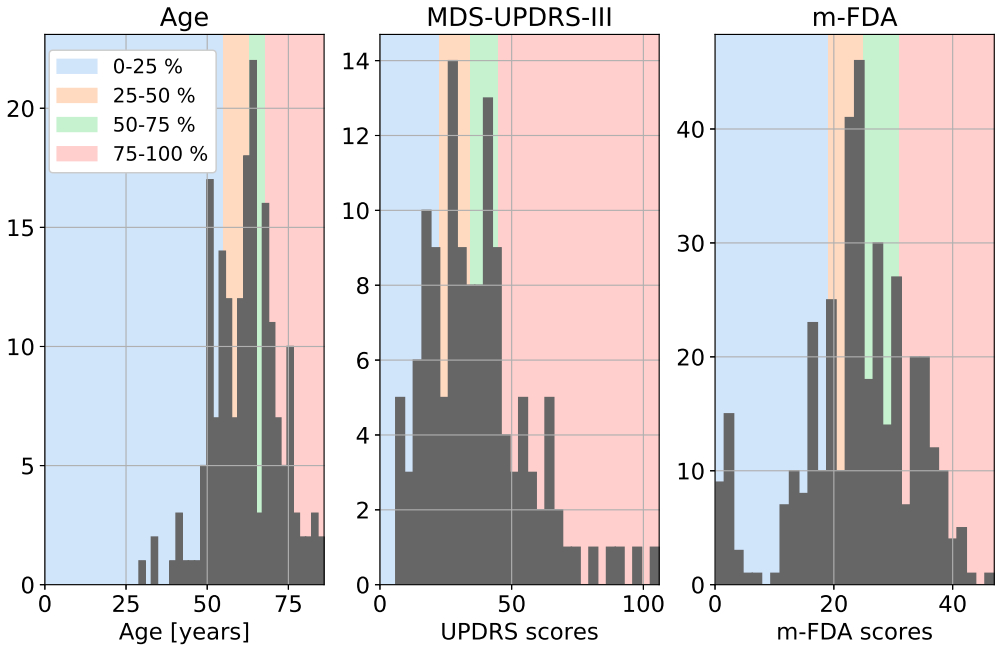

Figure 2.5: Class assignments for Age, MDS-UPDRS-III and m-FDA. These tasks were partitioned into 4 classes, each according to the percentiles at 25%, 50% and 75%. The single classes are illustrated with different background colors.

2.3.3 Class assignments

To create class labels for the MDS-UPDRS-III, m-FDA and Age tasks the different value distributions needed to be partitioned into different classes.

Table 2.4: The value range for each class in Age, MDS-UPDRS-III and m-FDA according to the percentile range.

Task	**0–25**	**25–50**	**50–75**	**75–100**
Age	< 55	55–63	63–68	$68 <$
MDS-UPDRS-III	< 22	22–34	34–45	$45 <$
m-FDA	< 19	19–25	25–31	$31 <$

Table 2.5: The amount of utterances associated with each class according to the percentile ranges.

Task	0–25	25–50	50–75	75–100
Age	1246	1507	1253	1079
MDS-UPDRS-III	2112	542	549	490
m-FDA	1586	1165	1412	982

In Figure 2.5 a histogram of the occurring values for each of the named task is displayed. The background colors illustrate the class parts according to the 25%, 50% and 75% percentile. This led to 4 classes per task. Table 2.4 shows the value ranges for each class and the amount of utterances assigned to each class can be found in Table 2.5. Note that the imbalance in the first class of the MDS-UPDRS-III class is resulting from the fact that all HC were assigned to this range. Also, in this task the number of utterances do not sum up to the total amount of 5145 utterances, since the samples without a value were not included.

Chapter 3

Methods

3.1 The openSMILE feature set

In classical machine learning methods a set of usually hand crafted features is taken and fed to a classifier to make a decision. One big advantage of neural networks and deep learning is that the optimal features to make this decision are supposed to be learned by the network and do not have to be defined beforehand. However, this approach requires a large amount of data to be available for the training process, so it is able to generalize well for unseen data. In medical applications large data sets are hard to get mostly due to the sensitive nature of the data and the difficult labelling process. Although there exists a trend to make these data increasingly available, until now there are no large scale data sets in the range of, for instance, the ImageNet [Den09] or WordNet [Fel98]. In this work the idea was to provide some assistance to the deep learning model. For that reason, a larger amount of features should be precomputed beforehand to be used as an input and let the network learn which of the features are useful to fulfil the task. This would reduce the amount of input data drastically, due to the fact that instead of images or audio signals the network only needs to handle a vector of a fixed size.
A popular collection of features for speech analysis is the *openSMILE* feature set [Eyb13]. It is part of an open source software project, which also provides a toolkit to quickly compute a large amount of audio features. In this work specifically the feature set of the 2010 INTERSPEECH ComPare challenge [Sch10] was used. Although there exist more recent versions of the feature set with a larger amount of included features, initial tests showed the 2010 version to be sufficient. Accordingly, 1428 features are computed using 21 functionals on 34 low level descriptors (LLD) and their corresponding delta coefficients. Additionally, 152 features are obtained using 19 functionals on pitch based LLDs, namely the smoothed fundamental frequency contour, the

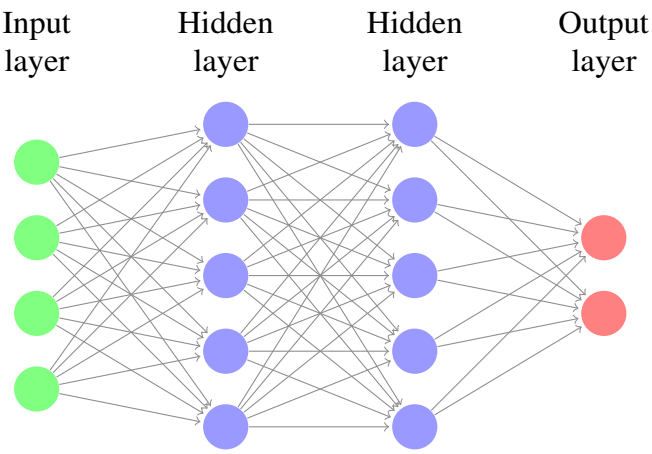

Figure 3.1: Basic neural network with a depth of three layers. It consists of one *input layer*, two *hidden layers* and one *output layer*. All nodes are fully connected from the previous layer to the next one.

local (frame-to-frame) jitter, the differential frame-to-frame jitter and the local (frame-to-frame) shimmer with their four delta coefficients. Further, the last two features are the number of total pitch onsets and the total duration of the input leading to a total amount of 1582 computed features per utterance. The entire list of functionals used to compute the features can be found in the Appendix B.

3.2 Multitask Deep Learning

3.2.1 Deep Learning Fundamentals

Deep learning (DL) is a machine learning technique which saw immense growth, success and public attention in recent years. With examples like the AlexNet [Kri17] or BERT [Dev18] it surpassed the previous state of the art in the respective fields and became one of the main topics in current research and industry. DL usually comes in the form of *Artificial neural networks*, which are roughly inspired by the way how neurons are connected in the human brain. The goal of a neural network is to approximate a function f which is able to map an input $\boldsymbol{x}$ to a desired output $\boldsymbol{y}$ [Goo16]. More formally the mapping is described as follows with $\boldsymbol{\theta}$ being the function parameters:

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{3.1}$$

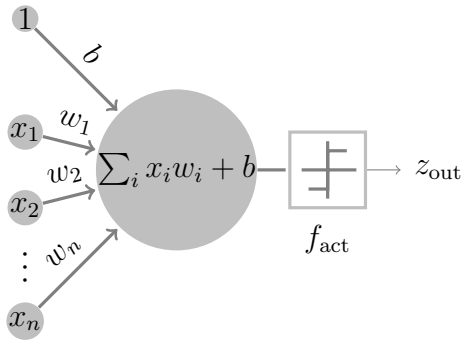

Figure 3.2: Single neuron. A weighted sum of input values receives a specific weight and is passed trough an activation function. b indicates the bias term. z_{out} is the neuron's output after being passed element wise through the activation function (f_{act}).

3.2.2 Feed-Forward Neural Network

The most common form of ANNs are so-called feed-forward neural networks or *Multi-layer Perceptrons*. In MLPs the input is propagated forward through each layer to create an output. The nodes in these layers are fully connected (See Figure 3.1). Figure 3.2 shows an exemplary single node or neuron. A weighted sum of the n input values of x_i is computed by using the weights w_i. b represents the bias term. The weighted sum is forwarded to an elementwise non-linear function, the so-called *activation function (f_{act})*. This activation function creates the actual output z_{out}. Formally, with a set of multiple neurons in one layer this computation is essentially a vector-matrix multiplication as it is described in Equation 3.2.

$$f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{W}^\intercal \boldsymbol{x} + \boldsymbol{b} \tag{3.2}$$

The activation will then be computed as follows:

$$z_{\text{out}} = f_{\text{act}}(f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b})) = f_{\text{act}}(\boldsymbol{W}^\intercal \boldsymbol{x} + \boldsymbol{b}) \tag{3.3}$$

Optimization

The goal of a DL training process is to find an optimal set of parameters $\boldsymbol{\theta}^*$ which minimizes the cost function $J(\boldsymbol{\theta})$. Formally, this can be expressed as follows:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta}) \tag{3.4}$$

As it was pointed out in Goodfellow et al. [Goo16] the cost function can be described as the

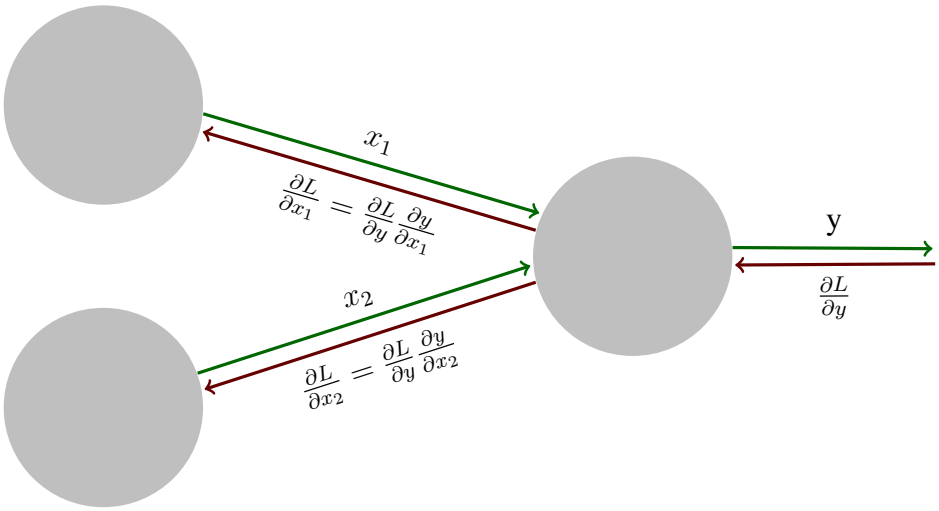

Figure 3.3: Illustration of the backpropagation algorithm idea. The green lines denote the activations being computed in the forward pass. In the backward pass (red lines) the chain rule is used to compute the gradients. Adapted from the Stanford CS231n lecture notes [Li18].

average over the training set computing the loss function L (see Equation 3.5) and ends up essentially in reducing the empirical risk (see Equation 3.6).

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\hat{p}_{\text{data}}(\boldsymbol{x},y)}\, L(f(\boldsymbol{x};\boldsymbol{\theta}), y) \tag{3.5}$$

$$\mathbb{E}_{\boldsymbol{x},y\sim\hat{p}_{\text{data}}(\boldsymbol{x},y)}\left[L(f(\boldsymbol{x};\boldsymbol{\theta}), y)\right] = \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x^{(i)}};\boldsymbol{\theta}), y^{(i)}) \tag{3.6}$$

The loss L is computed between the input $\boldsymbol{x}$ and the target output y. These values are taken from the empirical distribution $\hat{p}_{\text{data}}(\boldsymbol{x}, y)$ of the training set. The variable m describes the amount of training samples.
The optimization step is performed following the negative gradient direction of the loss function (see Equation 3.7). This is usually called a *gradient descent*.

$$\boldsymbol{\theta_{k+1}} = \boldsymbol{\theta_k} - \eta\nabla J(\boldsymbol{\theta_k}) \tag{3.7}$$

The variable η indicates the *learning rate* (LR) which denotes the step size taken at iteration k during the current optimization step [Dud00].

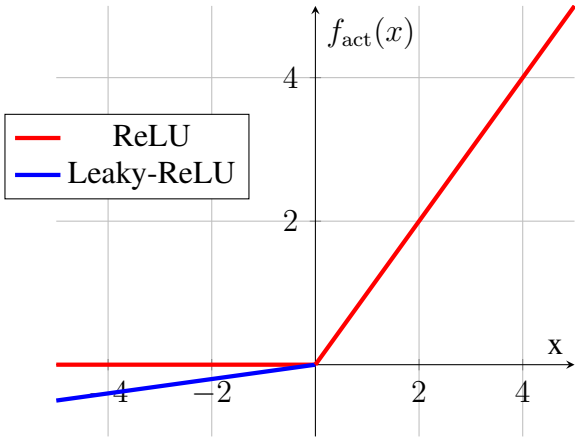

Figure 3.4: ReLU and Leaky-ReLU activation functions.

Backpropagation

An essential point in training is to update the weights dependent on the current loss. The *backpropagation algorithm (BP)* is an efficient way to compute the gradients with respect to the weights based on the chain rule and dynamic programming. The basic procedure is as follows:

1. Forward pass: Propagate the input through the network in a layer-by-layer fashion to compute all activations and get the loss at the output layer.

2. Backward pass: Recursively apply the chain rule to propagate backwards through the network and compute all gradients. In this way each neuron weight is updated according to the amount it contributes to the error.

In Figure 3.3 the basic concept of BP is illustrated. The green lines indicate the activations computed in the forward pass until the final error is computed in loss L. Consequently, the chain rule is applied to compute the gradient with respect to y and is propagated backwards through the network (red lines).

Activation functions

There are many different kinds of activation functions available. While at the beginning the $\mathrm{sigmoid}$ or $\tanh$ functions were very popular, the introduction of the *Rectified Linear Unit (ReLU)* function [Nai10] has brought significant improvement to DL. The ReLU function is piecewise linear, which makes it easy and fast to optimize (see Equation 3.8). In fact, Krizhevsky et al. [Kri17] reported a speed-up by a factor of six compared to a standard $\tanh$ in their application. In addition, it helps to prevent the vanishing gradient problem, which means that no further updates

can be made due to a neglectable gradient. However, if the input distribution tends to be more negative, the ReLU sets all activations to 0 which prevents further learning. This problem can be overcome using a modified version of the ReLU, the so-called *Leaky-ReLU* [Maa13]. Here a slight slope with factor α is applied (see Equation 3.9 and 3.10 for its derivation) in the negative value range which prevents the gradient from becoming 0. In Figure 3.4 illustrations of the ReLU and the Leaky-ReLU can be found.

$$f_{\text{act}}(x) = \max(0, x) \tag{3.8}$$

$$f_{act}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{3.9}$$

$$f'_{act}(x) = \begin{cases} 1 & \text{if } x > 0 \\ \alpha & \text{otherwise} \end{cases} \tag{3.10}$$

In the output layer very often a decision needs to be made according to the most likely output. Hence, it is beneficial to map the output values to a probability distribution to find the most probable output class. To achieve this a common activation function for the last layer is the so-called *softmax* function. The mathematically expression can be found in Equation 3.11.

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.11}$$

The softmax function converts the values from each element x_i to a range of 0 to 1, so they sum up to 1. Hence, this function maps the output to a probability distribution.

3.2.3 Multitask Learning

Although the majority of work in deep learning is aimed to optimize a single outcome, multitask learning has gained more and more attention in recent years [Yan17a]. Essentially, a network does multitask learning as soon as more than one loss function is optimized at the same time. The overall idea in MTL is that multiple tasks share representations among them, creating more general feature maps for each task. Potentially, this leads to a better performance than learning every task individually.

Basic architectures in multitask learning can roughly be separated into two different approaches, namely hard and soft parameter sharing. In Figure 3.5 these basic architectures are illustrated. Hard parameter sharing exists of common layers and parameters for all tasks combined, as well as individual layers for each task on top to produce the outcome [Car97]. In soft parameter sharing each task has its own layers and parameters, but the distance of the parameters are regularized to share common knowledge, for instance, with an ℓ_2 distance [Duo15].

There exist several intuitions why multitask learning works in deep learning. Most of them were described in the work of R. Caruana [Car97] already in 1997. For instance, it adds a so-called *inductive bias* to the model [Bax11]. The inductive bias is an effect that causes a learner to prefer a specific hypothesis over another one. In MTL the additional tasks would serve as such a bias and essentially lead to preferring certain samples which helps to generalize better. Further, MTL leads to *implicit data augmentation*. It implicitly increases the training samples, because the examples to train each task may contain information to help other tasks. In addition, it may occur that some features are easier to learn for one task than the other. However, this feature could also be important for the second task. Hence, the tasks can *eavesdrop* each other to benefit from their knowledge [AM90]. In this sense, MTL has multiple theoretical advantages which can reduce the risk of overfitting the training data and lead to a better performance on unseen data. Usually the tasks to be shared are required to be somewhat related, however uncorrelated tasks may also help each other when they act as a form of noise [Car97].
In this thesis a hard parameter sharing will be used (see Figure 3.5 I.)), since the objective is to constrain the common layers to learn more general feature representations which can be used in between all tasks. On top more classification layers will be added for each task separately.

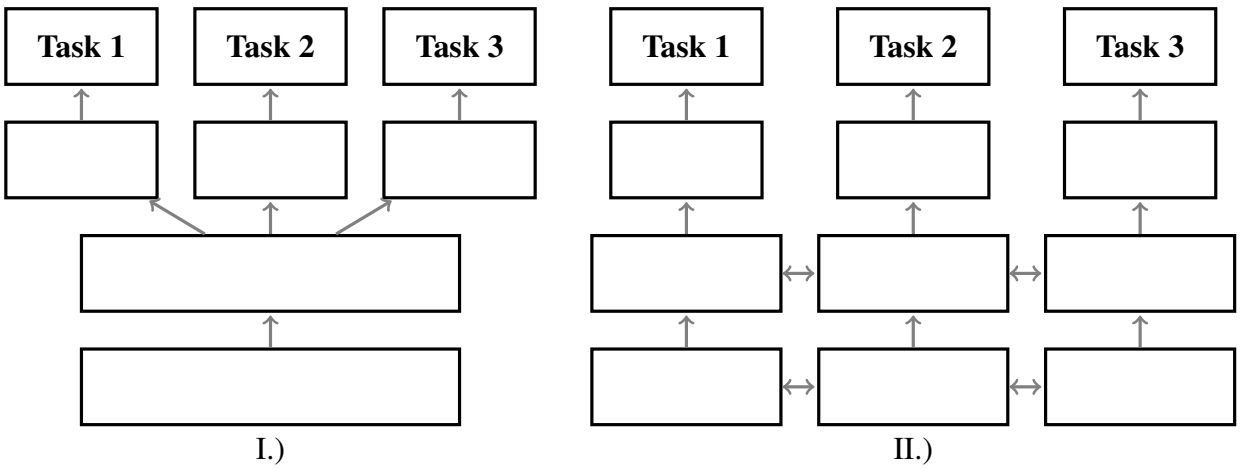

Figure 3.5: I.) Hard parameter sharing multitask learning. Several layers are shared in between all tasks ending in individual layers for each task. II.) Soft parameter sharing multitask learning. Each task has its own network, but several lower layers are constrained connected. The graph is adapted from Ruder [Rud17].

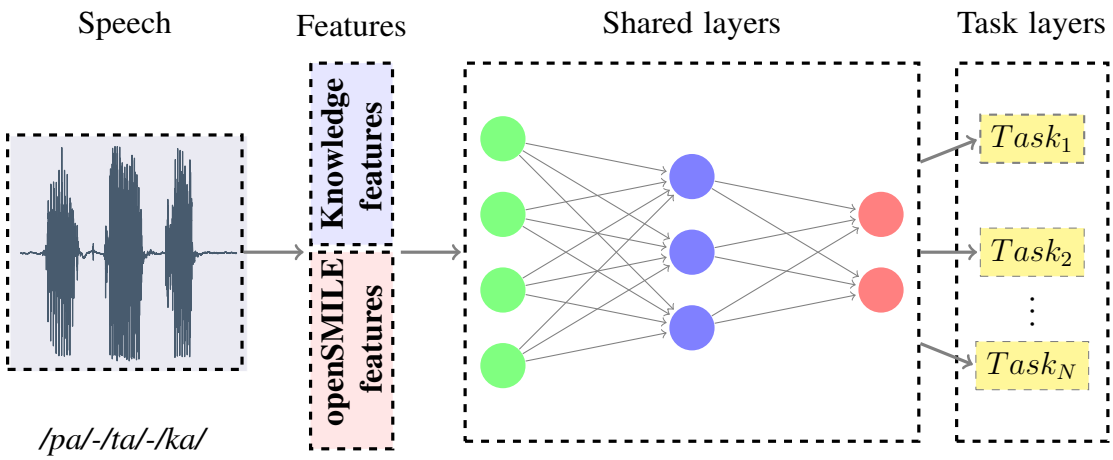

Figure 3.6: The basic model architecture. The speech signal is forwarded to the openSMILE feature extractor to pre-compute a large set of features. In addition, the so-called knowledge features are added, which contain information about the age, gender, the exercise being performed and the acoustic condition. A fully connected neural network with shared layers receives these features as input and individual task layers are added on top for each task.

3.3 Model architecture

3.3.1 Basic architecture

The basic model architecture used in this thesis is illustrated in Figure 3.6. It was based entirely on a fully connected neural network with several shared layers and individual task layers. The amount of shared layers was part of the hyperparameter optimization (see Section 3.5.1). Each task received two additional layers with the amount of nodes being kept per layer. The intuition here was that by using the shared layers the network should learn the features which generalize best for all tasks combined, but additional individual layers would be necessary to choose the correct features for classification. Thus, these layers were named “classification layers”. For the shared hidden layers the nodes were reduced at each step to the next lower power of two, which was supposed to ensure that the feature dimensions decrease to only keep important ones. The weights are initialized doing Xavier initialization [Glo10]. The activation function after each layer was a Leaky-ReLU function as described in section 3.2.2. The batch size used for training was 8 and the maximum number of training epochs was set to 100. Setting the learning rate properly is crucial to the training process to reach a good convergence. A too high learning rate could miss the optimal point, while a too small learning rate does not decrease fast enough or leads to inconsistent loss computations and being stuck at a local minimum. A common way to overcome this problem is to adapt the learning after a certain amount of learning steps [Ben12]. Hence, a learning rate

decay by a factor of 0.1 after every 20th epoch was implemented. As an optimizer the RMSProp [Hin12] was chosen. The RMSProp was designed to increase performance in non-convex settings by using an exponentially decaying average [Goo16]. The choice was based on initial tests, where it showed a better performance compared to other optimizers like a pure Momentum [Sut13] or Adam [Kin15].

3.3.2 Regularization methods

The most common issue in deep learning is the risk of overfitting the training data. Hence, several regularization techniques have been proposed throughout recent years. The methods, which were used in this work are outlined shortly in the following.

Early stopping

Early stopping has been shown to be an effective way to prevent overfitting from too many training epochs [Bis06]. Therefore, one must keep track of the validation loss on the development set and stop the training as soon as the loss does not decrease for a certain amount of training steps called *patience*. For this thesis a early stopping mechanism with a patience of 10 epochs showed to be beneficial.

Dropout

Dropout is another common regularization technique. It was introduced by Srivastava et al. [Sri14] in 2014 and was since then widely used to reduce the risk of overfitting. It proposes the removal of neurons based on a probability $1 - p$ during training, which is supposed to prevent the neurons to adapt too much to neighbouring units. However, the choice of p has a large influence on the training, which is why it was included into the hyperparameter optimization.

Batch Normalization

One disadvantage in using a ReLU-based activation function is that it is not zero-centered. This leads to a so-called *internal co-variance shift* [Iof15], which slows down the training and increases the risk of overfitting, because the network constantly has to adapt. Hence, the layers should be normalized to have zero mean and a unit standard deviation. In Equation 3.12 the computation of the batch norm is displayed. x_i is the activation at layer i. $\mu_{B,i}$ and $\sigma_{B,i}$ are the mean and standard deviation of the current minibatch. In Equation 3.13 the input activation are scaled and shifted by

the parameters δ_i and β_i.

$$\widetilde{x_i} = \frac{x_i - \mu_{B,i}}{\sqrt{\sigma_{B,i}^2 + \epsilon}} \tag{3.12}$$

$$\hat{y} = \delta_i \widetilde{x_i} + \beta_i \tag{3.13}$$

Class imbalance weighting

If the data set is imbalanced, meaning that single classes are over-represented compared to others, there is a risk that the network does not learn how to differentiate these classes, but tends to decide for the most represented class. One approach to address this issue is to put different weights to each class based on their occurrences in the training set.

3.3.3 Loss function

The loss function for a multitask problem is a linear combination of the individual losses for each task (see Equation 3.14). The parameter γ is called a *weight factor* and can be a hyperparameter to be learned in the learning process.

$$L(\theta) = \gamma L_1(\theta) + (1 - \gamma) L_2(\theta) \tag{3.14}$$

If there are more than two tasks the loss is calculated by a weighted sum of each individual losses as it is illustrated by Equation 3.15. The weight factor is required to sum up to 1 (see Equation 3.16).

$$L(\theta) = \sum_i \gamma_i L_i(\theta) \tag{3.15}$$

$$\sum_i \gamma_i = 1 \tag{3.16}$$

3.4 Experimental setup

Table 3.1: Experiment setting for the task weight factors in the loss functions which were performed. The goal was to find out based on which task focus the best results are being reached.

	Loss function weight factor γ		
Experiment	**PD vs. HC**	**MDS-UPDRS-III**	**m-FDA**
1	0.8	0.1	0.1
2	0.1	0.8	0.1
3	0.1	0.1	0.8
4	learned	learned	learned

Several experiments have been conducted to evaluate the MTL approach.

3.4.1 Adaboost Baseline

To get a first performance estimation an *Adaboost algorithm* [Sch99] was applied to the data. Adaboost is a machine learning algorithm which trains sequentially several weak classifiers, meaning that their performance should be slightly better than random guessing, and combines them into a weighted sum to receive a final prediction. The classifier used in this work was a decision tree. The number of estimators was set to a value of 50 with a learning rate of 1. An individual Adaboost model was trained for the tasks of PD vs. HC, MDS-UPDRS-III and m-FDA leading to a baseline result for each of them.

3.4.2 Single task neural network

As further experiments a single task neural network was trained for each task to evaluate the performance of a DL approach. As in the MTL setup, the amount of hidden layers was included into the parameter optimization, although two "classification layers" were set to be fixed to keep it comparable.

3.4.3 Multitask neural network with three tasks

To evaluate the multitask framework several MTL settings were created. The setups are displayed in Table 3.1. On each of the target tasks was once put the main focus with a loss function weight factor of 0.8. The other two tasks received weight of 0.1. This was done to see whether it is favourable to concentrate on a single task but also learn the others with a smaller focus.

One additional experiment was done when all task weights were learned in the hyperparameter optimization.

3.4.4 Multitask neural network with seven tasks

The last experiment was conducted in including all seven tasks into the learning. Therefore, the knowledge features were removed and the tasks were included into the network. The task weights of the loss function were learned for all tasks.

3.5 Evaluation

3.5.1 Hyperparameter optimization

A 10-fold cross validation (CV) approach was used for the parameter optimization and performance evaluation of the model. Therefore, the data were split into a training (80%), development (10%) and test set (10%) with balancing for PD and HC. The train set was used to train the model and the development set to tune the hyperparamaters. The best set of parameters was used on the test set. This procedure was repeated 10 times to make sure each patient was used for testing once. Since the performance of a deep learning system is heavily affected by the choice of parameters, the optimal set of these should be found. A common process to do that is performing a so-called grid search, where a set of possible parameters is chosen beforehand and tried in all combinations on the development set. At the end the parameter combination giving the best results is selected for final performance tests. However, it was reported that pure grid search might miss the optimal values due to the preceding selection [Ber12]. A possible solution could be to select the parameters randomly using a Bayesian optimization approach [Sno12]. Here, the evaluation of the validation loss is defined as a posterior distribution of functions in a Gaussian process which is sought to be maximized. Hence, in this example the objective is to maximize the negative validation loss to get the best performing parameter set. The Gaussian process then tries to find an optimum from the chosen set of boundaries for each parameter. To guide the process three initial starting point have been chosen here which seemed to be promising in initial tests of the network. A total of 20 maximization steps has been performed to find the optimal set of parameters. The set of hyperparameters and the corresponding ranges of possible selection points is displayed in Table 3.2.

Table 3.2: Hyperparameter search space for the Gaussian process optimization. The weight factor for the loss function was optimized for each task resulting in seven different factors.

Parameter	**Values**	
	min	max
Learning rate	0.001	0.1
Dropout probability	0.1	0.9
Number of hidden layers	2	5
Loss function weight factor γ_i	0.1	0.9

3.5.2 Performance measures

Accuracy

The most common used metric to evaluate classification outputs is the accuracy. Accuracy describes the correct classified samples divided by the total amount of available samples as it is described in Equation 3.17.

$$\mathrm{ACC} = \frac{1}{N} \sum_{i=1}^{C} \mathrm{TP}_i \cdot 100\% \tag{3.17}$$

TP_i indicates the correct classified samples in a set of C classes. The sum is finally divided by the number of N test samples. The accuracy is computed for each task separately to monitor the progress per tasks during training and evaluation.

Unweighted average recall

Another performance measure used in this thesis was the unweighted average recall (UAR) [Sch14]. This metric is used if the occurrences of classes are imbalanced and a simple classification accuracy would lead to a bias towards the more frequent classes. Equation 3.18 shows how the UAR is calculated. In contrast to the normal accuracy the number of correctly assigned samples is divided by the amount of samples occurring in this class (T_i). Last, it is divided by the number of classes.

$$\mathrm{UAR} = \frac{1}{C} \sum_{i=1}^{C} \frac{\mathrm{TP_i}}{\mathrm{T_i}} \cdot 100\% \tag{3.18}$$

Session based evaluation

In a real application it is very likely that a participant would perform several speech exercises and after all tasks have been finished an assumption is made if, for instance, he is classified as PD or HC. In this sense, an additional metric was introduced to combine all decisions of one session to a final one. This ensemble-based approach was named *session evaluation*. In the classification between PD and HC a majority vote was performed to decide for a final class. On the contrary, in the tasks involving the MDS-UPDRS-III and m-FDA scales the median of the decisions in one session was used. As a final score the average over all decisions is calculated.

3.5.3 Architecture comparison

To be able to see if the multitask approach is beneficial, each task is also learned in a single task matter. To keep is comparable the general structure for the single task network is the same. However, for each task the whole parameter optimization process is repeated, not to include a bias towards the multitask approach. On top of that the DL approached will be compared to the Adaboost baseline.

Chapter 4

Results

4.1 Adaboost baseline results

Table 4.1: The test results for using the Adaboost algorithm.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	80.70	77.21	85.71
MDS-UPDRS-III	53.29	32.14	63.77
m-FDA	36.56	33.78	40.28

The Adaboost baseline test results are illustrated in Table 4.1. *ACC* represents the accuracy, *UAR* the unweighted average recall and *Session* the session based evaluation as described in Section 3.5.2. In PD vs. HC a total accuracy of 80.70% and 77.21% UAR was reached. The session based evaluation showed an outcome of 85.71%. For the MDS-UPDRS-III and m-FDA tasks the session evaluation was 63.77% and 40.28%, respectively. The m-FDA classification reached 36.56% accuracy and 33.78% UAR. For MDS-UPDRS-III the UAR achieved a value of 32.14% and 53.29% for the accuracy.

4.2 Hyperparameter search single task learning

For the single task framework each network was optimized individually. Table 4.2, 4.3 and 4.4 show the results obtained during the hyperparameter optimization. Hence, the models created with these parameters have been used for testing of the corresponding folds. Note that in all tables *LR* represents the learning rate, *Layer* the amount of hidden layers and *Dropout* the dropout

probability as explained in section 3.3.2.

4.2.1 PD vs. HC single task

Table 4.2: The results per fold of the hyperparameter optimization for the PD vs. HC task in a single task framework.

Fold	LR	Layer	Dropout
1	0.1	3	0.50
2	0.01	3	0.20
3	0.01	3	0.20
4	0.0148	2	0.34
5	0.1	5	0.10
6	0.01	3	0.20
7	0.0346	3	0.25
8	0.042	4	0.53
9	0.0346	3	0.25
10	0.0148	2	0.34

The outcome for the PD vs. HC single task network is illustrated in Table 4.2. The parameter optimization for fold 1 resulted in a learning rate of 0.1, 3 hidden layers and a dropout probability of 0.5. Equal results were obtained for fold 2 and 3 with three layers, 0.2 dropout probability and a learning rate of 0.01. In fold 4 the optimal learning rate was 0.0148 with 2 hidden layers and a dropout probability of 0.34. Learning rate and dropout had a value of 0.1 in fold 5, while the best model used 5 hidden layers. Fold 6 showed the same results than fold 2 and 3. In fold 7 once more 3 hidden layers were shown to be the best number, but the learning rate changed to 0.0346 and the dropout probability to 0.25. A dropout probability of 0.53, a learning rate of 0.042 and 4 layers showed to be the best for fold number 8, while in fold 9 this changed to 0.25, 0.0346 and 3. The best model of the last fold was created with a learning rate of 0.0148, 2 hidden layers and a dropout of 0.34.

4.2.2 MDS-UPDRS-III single task

The outcome of the parameter optimization for the single task network of MDS-UPDRS-III is displayed in Table 4.3. As can be seen in the table, several folds showed the same results. Fold 1, 3 and 9 resulted in a learning rate of 0.01, a dropout probability of 0.2 and 3 hidden layers. In fold 2 and 10 the results were equivalent with two hidden layers and a learning rate of 0.0001, as well

Table 4.3: The results per fold of the hyperparameter optimization for the MDS-UPDRS-III task in a single task framework.

Fold	LR	Layer	Dropout
1	0.01	3	0.20
2	0.0001	2	0.10
3	0.01	3	0.20
4	0.0148	2	0.34
5	0.0148	2	0.34
6	0.0005	4	0.30
7	0.0001	2	0.90
8	0.0005	4	0.30
9	0.01	3	0.20
10	0.0001	2	0.10

as, a dropout of 0.1. Further, the models of fold numbers 4 and 5 are at optimum with 2 hidden layers and a dropout probability of 0.34. The learning rate for these folds was found to be the best with 0.0148. Fold 6 obtained a minimal validation loss for a learning rate of 0.0005, 4 hidden layers and 0.3 dropout probability. The same accounts for fold number 8. Apart from this, fold 7 gave the same model parameters for learning rate and amount of layers as fold 2 and 10. However, the dropout probability here was found to be 0.9.

4.2.3 m-FDA single task

Table 4.4: The results per fold of the hyperparameter optimization for the m-FDA task in a single task framework.

Fold	LR	Layer	Dropout
1	0.0148	2	0.34
2	0.0001	3	0.36
3	0.1	4	0.10
4	0.0148	2	0.34
5	0.1	5	0.10
6	0.042	4	0.53
7	0.042	4	0.53
8	0.01	3	0.50
9	0.0346	3	0.25
10	0.01	3	0.20

Table 4.4 shows the hyperparameter settings that were found for the m-FDA task. As it is shown in the table, models using a learning rate of 0.0148, with 2 hidden layers and a dropout of 0.34 are the test settings for fold 1 and 4. On the contrary, fold 2 has as optimum learning rate a value of 0.0001, three layers and a dropout probability of 0.36. Folds 3 and 5 had an optimal parameter outcome with 0.1 learning rate and dropout probability, but 4 and 5 hidden layers, respectively. Equal results were obtained for fold 6 and 7 with 4 hidden layers, a dropout of 0.53 and 0.042 initial learning rate. Fold 8 and 10 only differed in the dropout probability with 0.5 for fold 8 and 0.2 for fold 10. The learning rate and amount of hidden layers was 0.01 and 3 for both. Last, in fold 9 the minimum validation loss was reached with a model using 0.0346 learning rate, 3 layers and 0.25 dropout.

4.3 Hyperparameter search for multitask learning

In the parameter optimization for the MTL framework, PD vs. HC, MDS-UPDRS-III and m-FDA are jointly optimized. For three optimization rounds a main focus layed on one of the tasks using a higher task weight in the loss function. Additionally, in one setting the task weights were learned and included into the optimization. The results of the parameter optimization for multitask can be seen in the Tables 4.5, 4.6, 4.7 and 4.8. As in Section 4.2 *LR*, *Layer* and *Dropout* stand for learning rate, number of hidden layers and the dropout probability, respectively. In the setting, where the task weights for the loss function should be learned and optimized, *TW* is an abbreviation for *task weight*, meaning the weight applied to each task in the loss function as described in Section 3.3.3.

4.3.1 MTL with focus on PD vs. HC

The best parameters obtained for each fold when putting a focus on the PD vs. HC task are illustrated in Table 4.5. Fold 1, 4 and 5 resulted in an optimal model for the same set of parameters. The learning rate here was 0.01, dropout 0.2 and the amount of hidden layers 3. The same accounts for fold 2, 3 and 7 with a 0.0005 learning rate, 4 hidden layers and 0.3 dropout probability. Fold 6 and 8 appeared to also use the same parameters for the best model choice. Dropout was set to 0.25 with 3 hidden layers and a learning rate value of 0.0346. On the contrary, a parameter set of 0.042, 4 and 0.53 for learning rate, hidden layers and dropout was found to minimize the validation loss in fold 9. Finally, the parameter setting of the testing model for fold 10 is 0.0148 for learning rate, 2 for hidden layers and 0.34 for dropout, respectively.

Table 4.5: The results per fold of the hyperparameter optimization for putting the focus on PD vs. HC task in a multitask framework.

Fold	LR	Layer	Dropout
1	0.01	3	0.20
2	0.0005	4	0.30
3	0.0005	4	0.30
4	0.01	3	0.20
5	0.01	3	0.20
6	0.0346	3	0.25
7	0.0005	4	0.30
8	0.0346	3	0.25
9	0.042	4	0.53
10	0.0148	2	0.34

4.3.2 MTL with focus on MDS-UPDRS-III

Table 4.6: The results per fold of the hyperparameter optimization for putting the focus on MDS-UPDRS-III task in a multitask framework.

Fold	LR	Layer	Dropout
1	0.042	4	0.53
2	0.01	3	0.20
3	0.0148	2	0.34
4	0.0148	2	0.34
5	0.0346	3	0.50
6	0.0148	2	0.34
7	0.01	3	0.20
8	0.0148	2	0.34
9	0.01	3	0.20
10	0.0001	3	0.18

As it is displayed in Table 4.6 the parameter optimization obtained the same learning rate, amount of hidden layers and dropout probability in several folds, namely fold 3, 4, 6 and 8. The learning rate was 0.0148, dropout 0.34 with 3 hidden layers. Similarly, in fold 2, 7 and 9 a learning rate of 0.01 with 3 layers and 0.2 dropout appeared to be the best model setting. In contrast, fold 1 reached a learning rate of 0.042, 4 hidden layers and a value of 0.53 for dropout. In fold 5 the parameter optimization reached a minimum for 3 hidden layers, 0.5 dropout probability and 0.0346 learning rate. Last, a learning rate of 0.0001 with three hidden layers and 0.18 dropout

showed to be the best parameters for the model in fold 10.

4.3.3 MTL with focus on m-FDA

Table 4.7 shows the parameters found for testing in focusing on the m-FDA task. Fold 1 resulted in a learning rate of 0.0001, 4 hidden layers and a dropout of 0.1. In fold 2, 6 and 8 equal parameters were found. The learning rate was 0.01, with a dropout 0.2 and 3 hidden layers. A minimum validation loss was reached using a 0.1 learning rate, 4 hidden layers and 0.1 for dropout in fold 3. On the contrary, a value of 0.0005 for the learning rate with 4 layers and 0.3 dropout are the parameter set in the testing model for fold 4. In the test for fold 5 a parameter set with 0.0148 learning rate, 2 hidden layers and 0.34 dropout will be used. Fold number 7 has the same learning rate than fold 3, but uses 3 hidden layers and 0.5 probability for dropout. In contrast, a learning rate of 0.0346, 0.25 dropout and three hidden layers were found to be optimal in the model for fold 9. Although fold 10 has the same amount of layers, a different learning rate and dropout probability was found with values of 0.0171 and 0.18.

Table 4.7: The results per fold of the hyperparameter optimization for putting the focus on m-FDA task in a multitask framework.

Fold	LR	Layer	Dropout
1	0.0001	4	0.10
2	0.01	3	0.20
3	0.1	4	0.10
4	0.0005	4	0.30
5	0.0148	2	0.34
6	0.01	3	0.20
7	0.1	3	0.50
8	0.01	3	0.20
9	0.0346	3	0.25
10	0.0171	3	0.18

4.3.4 MTL with learned task weights

Table 4.8 shows the results of optimizing the parameters for all 10 folds when the individual task weights are included into the learning process. It is displayed that fold 2 to 10 achieved equal task weights with 0.09 for PD vs. HC and m-FDA, but 0.82 for MDS-UPDRS-III. In fold 1 these values changed to 0.17 for PD vs. HC and 0.74 for MDS-UPDRS-III. The weight for m-FDA was

Table 4.8: The results per fold of the hyperparameter optimization when the task weights are learned.

Fold	LR	Layer	Dropout	TW PD vs. HC	TW UPDRS	TW m-FDA
1	0.1	4	0.1	0.17	0.74	0.08
2	0.0001	3	0.1	0.09	0.82	0.09
3	0.0001	4	0.1	0.09	0.82	0.09
4	0.0001	4	0.1	0.09	0.82	0.09
5	0.0001	5	0.1	0.09	0.82	0.09
6	0.0001	5	0.9	0.09	0.82	0.09
7	0.0001	4	0.1	0.09	0.82	0.09
8	0.0001	4	0.1	0.09	0.82	0.09
9	0.0001	3	0.1	0.09	0.82	0.09
10	0.1	5	0.1	0.09	0.82	0.09

0.08. The dropout probability was 0.1 for all folds, instead of fold 6 with a value of 0.9. The learning rate was 0.0001 for fold 2 to 9, while it was 0.1 for fold number 1 and 10. The amount of hidden layers varied from 3 for fold 2 and 9 to 4 layers for in fold 1, 3, 4, 7 and 8. Fold 5, 6 and 10 resulted in 5 hidden layers.

4.3.5 MTL with seven tasks

The result table regarding the parameter optimization for the 7 task problem can be found in the Appendix C. The learning rate parameter resulted in the same value for all folds instead of number 2. For this fold it was 0.1 and for the others 0.0001. The dropout probability was learned to be 0 in all combinations. In the amount of hidden layers it can be seen that fold 7 to 9 ended up with 4 and fold 2 with 3. All the other folds took 4 hidden layers as optimal value. In fold 1 and 5 a weight of 0.39 was found for MDS-UPDRS-III and the exercise task. The others received a weight of 0.04. Similarly, the weights in fold 6 had the same values, but the higher weight of 0.39 was put on PD vs. HC instead of the exercise. On the contrary, in fold 6 a weight of 0.6 was found to be optimal for the MDS-UPDRS-III task, but 0.06 for all other tasks. In the remaining folds always three tasks received a 0.29 weight with 0.03 for the others in different combinations.

4.4 Results obtained with the neural networks

The choice of best parameters led to model selections for each fold. In the following, the results over all folds combined are illustrated as final performance test.

4.4.1 Single task test results

Table 4.9: Test results from the single task networks.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	71.82	71.12	78.22
MDS-UPDRS-III	29.46	29.13	36.23
m-FDA	36.83	34.50	40.52

The test results for all single task frameworks can be seen in Table 4.9. To discriminate between PD and HC the accuracy was 71.82% and the UAR 71.12%. Using the evaluation per session a value of 78.22% was achieved. Testing the network trained solely on the MDS-UPDRS-III task a total accuracy of 29.46% with 29.13% UAR was reached. The session evaluation resulted in 36.23%. The m-FDA classification resulted in a 36.83% accuracy, 34.50% UAR and 40.52% session evaluation.

4.4.2 Multitask test results

Table 4.10: Test results for using a multitask framework and putting a high weight on PD vs. HC.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	73.16	72.49	81.73
MDS-UPDRS-III	46.79	34.45	52.45
m-FDA	37.30	35.94	43.56

As it can be seen in Table 4.10, putting a high loss weight on the PD vs. HC task reached a test accuracy for this task of 73.16% and an UAR of 72.49%. A value of 34.45% was achieved for UAR in MDS-UPDRS-III and 35.94% in m-FDA. 46.79% was the test accuracy for MDS-UPDRS-III and 37.30% in m-FDA. For the sessions evaluation 43.56% were achieved for m-FDA, 52.45% for MDS-UPDRS-III and 81.73% for PD vs. HC.
The results for MTL in focusing on MDS-UPDRS-III can be seen in Table 4.11. To differentiate between PD and HC reached an accuracy of 64.70 %, UAR of 64.35 % and a session evaluation of 70.49 %. A 33.90 % accuracy and 31.81 % UAR were achieved for the MDS-UPDRS-III, while the session evaluation resulted in 65.47 %. For m-FDA the UAR outcome was 31.20% and the accuracy 32.91 %. Session based evaluation for this task resulted into 35.83 %.
The session based evaluation values varied for focusing on m-FDA from 70.73% to 38.87% and 38.64% for PD vs. HC, MDS-UPDRS-III and m-FDA. PD vs. HC had an accuracy of 66.39% and

Table 4.11: Test results for using a multitask framework and putting a high weight on MDS-UPDRS-III.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	64.70	64.35	70.49
MDS-UPDRS-III	33.90	31.81	35.47
m-FDA	32.91	31.20	35.83

Table 4.12: Test results for using a multitask framework and putting a high weight on m-FDA.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	66.39	66.14	70.73
MDS-UPDRS-III	35.28	27.88	38.87
m-FDA	37.47	34.41	38.64

66.14% UAR. For MDS-UPDRS-III and m-FDA the accuracy values were 35.28%, as well as, 27.88%, while the UAR showed results of 27.88% and 34.41%, respectively. These numbers are also displayed in Table 4.12.

Table 4.13: Test results for using a multitask framework and learning the task weights.

Task	ACC (%)	UAR (%)	Session (%)
PD vs. HC	60.49	64.25	63.47
MDS-UPDRS-III	37.21	28.80	40.38
m-FDA	31.70	30.19	33.26

The MTL results when all task weights are learned can be seen in Table 4.13. For m-FDA the accuracy reached a value of 31.70%, 30.19% for UAR and 33.26% for session evaluation. On the contrary, the MDS-UPDRS-III outcome was 28.80% in UAR, 37.21% in accuracy and 40.38% on the session evaluation metric. To distinguish between PD patients and HC gave an accuracy of 60.49%, with 64.25% UAR and 63.47% session metric result.
The last test being performed was to include all seven tasks into the MTL set up. As illustrated in Table 4.14 the accuracy for PD vs. HC, MDS-UPDRS-III and m-FDA was 68.05%, 38.26% and 32.15%. The UAR for these tasks varied from 66.02%, 26.46% and 31.55%, while the session metric resulted in 74.71%, 41.51% and 36.53%. In classifying the acoustic condition, the accuracy of the model was 55.78% with an UAR of 47.95%. Differentiating the exercises resulted in 85.89% accuracy and 78.31% UAR. Another task was to recognise if a male or female person did the recording. The outcome was 79.18% for accuracy and 79.14% for the UAR. The age

Table 4.14: Test results for using a multitask framework with seven tasks.

Task	**ACC (%)**	**UAR (%)**	**Session (%)**
PD vs. HC	68.05	66.02	74.71
MDS-UPDRS-III	38.26	26.46	41.51
m-FDA	32.15	31.55	36.53
Acoustic condition	55.78	47.95	-
Exercise	85.89	78.31	-
Gender	79.18	79.14	-
Age	28.36	28.92	-

classification ended up with 28.36% a accuracy and a 28.92% UAR.

Chapter 5

Discussion

The baseline results obtained by the Adaboost algorithm show that it is possible to differentiate PD and HC with an accuracy of 80.7 % and an UAR of 77.21 %. However, it is important to notice that these results are also related to the imbalance of the data. As it can be seen in the confusion matrix in Table 5.1, 32% of HC are classified as PD. This indicates that the imbalance has a strong influence on the outcome of the algorithm. This gets even clearer if the confusion matrix of the MDS-UPDRS-III task is considered (see Table 5.2). As it is displayed most samples are classified to be in class 1. This leads to the assumption, that the algorithm learns to put a higher weight on the first class to reach solid results, but the other classes are not classified reliably. The UAR value of 32.14% compared to 53.29% in accuracy demonstrates that the algorithm has problems with the different representations. Nonetheless, with a value of 85.71% in session based evaluation for PD vs. HC the Adaboost baseline conducts promising results if multiple utterances are considered. The confusion matrices of the m-FDA baseline and all other experiments can be found in the Appendix E.

Table 5.1: Confusion matrices of the PD vs HC task for the Adaboost baseline.

		Prediction	
		PD	HC
Reference	PD	3083	479
	HC	496	1069

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.86	0.14
	HC	0.32	0.68

(b) Normalized confusion matrix

The parameter optimization of training a neural network on each task individually shows, that the network for the PD vs. HC and m-FDA task chooses more complex models. While for the MDS-UPDRS-III task two hidden layers are selected in five of ten folds, the other two tasks

Table 5.2: Confusion matrices of the MDS-UPDRS-III task for the Adaboost baseline. The values from 1 to 4 represent the respective classes.

		Prediction			
		1	2	3	4
Reference	1	1719	165	112	116
	2	311	76	76	79
	3	234	109	98	108
	4	264	41	110	75

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.81	0.08	0.05	0.05
	2	0.57	0.14	0.14	0.15
	3	0.43	0.20	0.18	0.20
	4	0.54	0.08	0.22	0.15

(b) Normalized confusion matrix

choose mainly a more complex architecture with three or more layers. The testing results show a slight improvement in the m-FDA task in all metrics compared to the baseline. However, the performance in MDS-UPDRS-III and PD vs. HC decreases in all fields. Interestingly, when looking at the classifications, the single task network for MDS-UPDRS-III puts most samples into the third class, although the imbalance exists towards the first one. An explanation for this could be, that the network overfitted the training data or the optimization rested in a local minimum and missed the optimal solution.

In the MTL settings different task weights were investigated. If the dropout probabilities of these experiments are compared to the single task models, it is illustrated that the achieved numbers are generally lower for the MTL approach. This could be related to a better generalization capability, since less regularization in using dropout is necessary. Further, the amount of hidden layers is 3 for most of the folds which is coherent with the single task frameworks. When performing the tests, several new insights were obtained. Putting the focus on the m-FDA task, the outcome of the MDS-UPDRS-III task did slightly improve, but showed decreased performance in all other fields compared to the single task network. In comparison to the baseline, the Adaboost reached higher values in all metrics throughout all tasks. Similar results are obtained when focusing on the MDS-UPDRS-III task. This implies, that simply focusing on these two tasks does not bring any benefit in the performance. A reason for this behaviour could be that these tasks are actually the most difficult ones to learn, since they are not closely related to speech. Interestingly, the lowest performance was achieved when the weights of the loss function were learned for all three tasks. For this setting the achieved results stayed also behind the single task network and the Adaboost baseline. At first sight, this is counter-intuitive, since the optimal setting of weights would be learned and not strictly be predefined. One possible reason for this could be, that when learning the task specific loss function weights, it was observed that most of the folds learn to concentrate on the MDS-UPDRS-III task. This could be explained again by the imbalance of the classes in

this task. Classifying for class 1 is comparably cheap in this task so the loss decreases more. In addition, including all weights to be learned increases the complexity of the optimization process, so balancing in between all parameters may not end up in an optimal setting.
The best performance in using a neural network was achieved when focusing on the PD vs. HC task. This setting reached an UAR of 72.49% with a session evaluation of 81.73% for differentiating between diseased and healthy. Thus, these numbers stay slightly below the Adaboost algorithm. In contrast, the MTL network shows a better outcome for the UAR values in the other two tasks compared to the baseline. In the MDS-UPDRS-III and m-FDA tasks the neural network achieved 34.45% and 35.94%, while the Adaboost resulted in 32.14% and 33.78%. Hence, these results indicate that the MTL approach may be more stable towards the data imbalance. However, further investigations are necessary to verify this assumption.

Table 5.3: Confusion matrices of the MDS-UPDRS-III task single task neural network.

		Prediction			
		1	2	3	4
Reference	1	611	73	1384	44
	2	68	15	430	29
	3	64	15	431	39
	4	46	11	402	31

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.29	0.03	0.66	0.02
	2	0.13	0.03	0.79	0.05
	3	0.12	0.03	0.79	0.07
	4	0.09	0.02	0.82	0.06

(b) Normalized confusion matrix

Table 5.4: Confusion matrices of the MDS-UPDRS-III task for focusing on PD vs. HC.

		Prediction			
		1	2	3	4
Reference	1	1351	129	300	332
	2	208	57	90	187
	3	216	65	90	178
	4	211	28	21	230

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.64	0.06	0.14	0.16
	2	0.38	0.11	0.17	0.35
	3	0.39	0.12	0.16	0.32
	4	0.43	0.06	0.04	0.47

(b) Normalized confusion matrix

Furthermore, it is important to notice that this MTL setting shows a better performance than all single task versions. Most notably that is the case for the MDS-UPDRS-III task. This can be demonstrated with the confusion matrices in Table 5.3 and 5.4. While in the single task network the majority of samples was classified to class 3, the MTL results are much more balanced. For instance, only 6% of the samples belonging to class 4 are classified correctly on the single task framework, but 47% for MTL. Additionally, the values for UAR and the session based

evaluation verify this assumption with 34.45% and 52.45%, compared to 29.13% and 36.23% for the individual networks. This illustrates, that the MTL approach helps to generalize better and is beneficial when choosing a neural network approach.

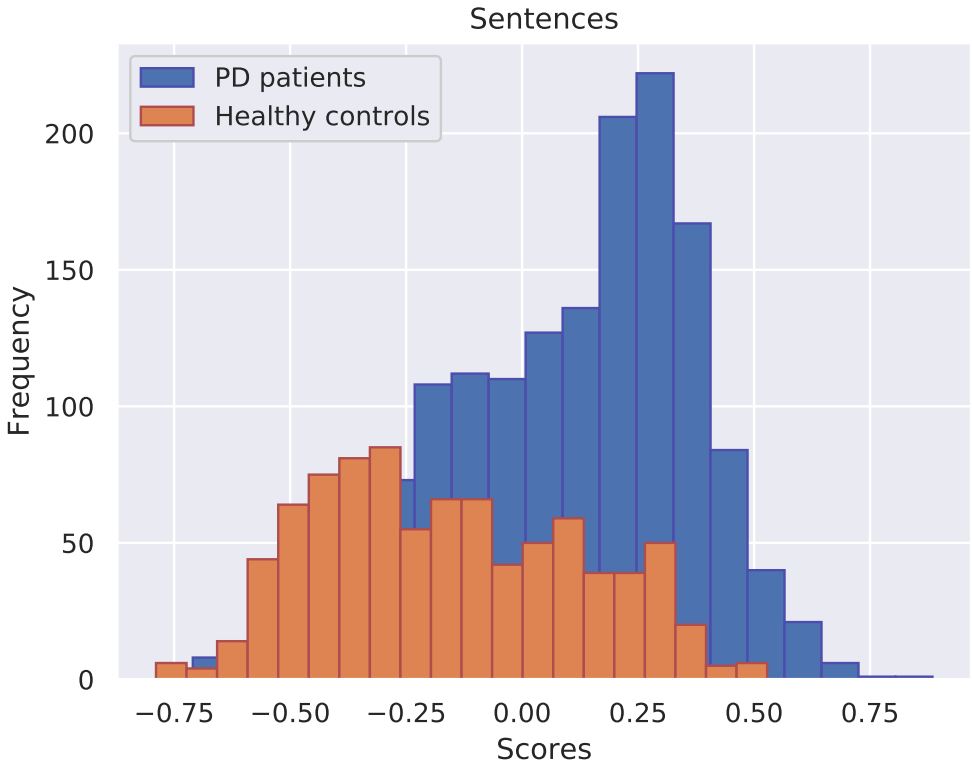

Figure 5.1: Confidence score distribution according to the PD vs. HC. classifications in sentences. The scores are obtained in looking at the softmax output of the PD vs. HC task compared to the true label. −1 means the classifier is totally confident towards HC, 1 represents maximum confidence towards PD. The histograms indicate the distribution of classifications of utterances with the sentence exercise. The figure shows that sentences are nicely separated between PD and HC according to the true values.

In the seven-tasks scenario it can be seen, that the results for the three main tasks are showing lower performance than the previous tests. This leads to the conclusion, that including more tasks does not improve the MTL setting. However, this test gives several new insights. Learning the weights for the loss function shows, that the model learned to mostly concentrate on the MDS-UPDRS-III task. This is consistent with the learned weights in the three task scenario. In addition, a higher weight was learned to be applied to the exercise task. The reason for that could be, that this task is an easy task to learn and, thus, using a higher weight for this task helped to decrease the validation loss to the minimum value. The test results with 78.31% UAR show, that

the model was able to differentiate well between the exercises. The acoustic conditions seem to be more difficult to distinguish. However, this is somewhat as expected, since the headset for the at-home and headset condition was the same type. Looking at the confusion matrix verifies this assumption with these two classes showing the largest confusion. The task to classify the age achieved a comparably weak classification. The reasons for that could rely on the fact that the classes for age did contain a small range of years, so the variance between the individual classes was too small for a reasonable differentiation.

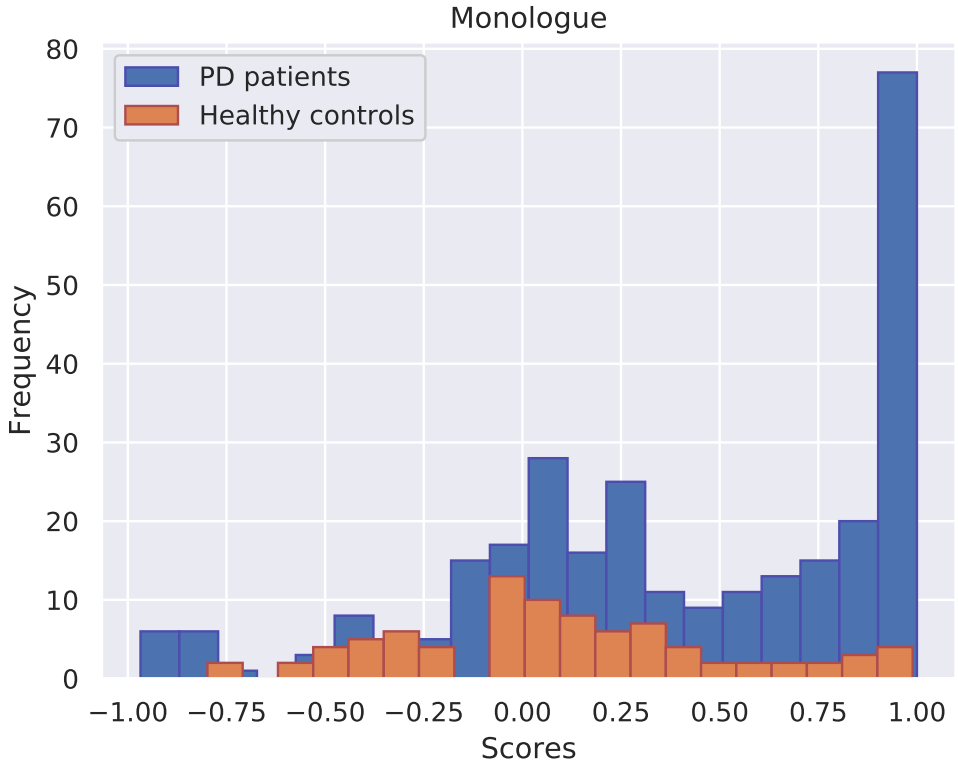

Figure 5.2: Confidence score distribution according to the PD vs. HC. classifications in the monologues. In contrast to the sentences, here multiple HC utterances are classified falsely as PD.

An interesting observation was made when the specific exercise are taken into account. For each exercise the results were categorized by classifying for PD or HC. In the following, the softmax output of the PD vs. HC prediction was taken to calculate confidence value according to the true label. Thus, –1 means total confidence for HC and 1 total confidence towards PD. In Figure 5.1 the outcome of this procedure for the sentence task is illustrated. As can be seen in this figure, the sentence task was mostly classified correctly. The true HC samples tend towards being classified as HC, while the real PD samples are located in the direction of PD. This indicated that the

sentence task could be used as an auxiliary task for a MTL approach. In contrast, in Figure 5.2 it is demonstrated, that the monologue task does not show the same effect. Most of the HC samples have been classified as PD according to the softmax, so no additional information was given by this exercise. However, interestingly a large amount of PD samples show a maximum confidence score in the monologue exercise, which needs to be further investigated. Similar observations can be made with the DDK and the reading text exercise. The figures of these two exercises according to this procedure can be found in the Appendix D.

Looking at the results it can be seen, that the MDS-UPDRS-III and the m-FDA tasks achieve relatively low accuracy values. There are several assumptions, why this could be the case. On the one hand, the rating scale values may not be closely enough related to speech and are hard to learn in a MTL setting, since the tasks are usually required to be of a similar kind [Car97]. On the other hand, as explained earlier, the rating scales are fairly subjective, so the true distribution of values may not be acquired during the data generation process. In addition, a further issue was observed with the ground truth labelling of the data. Several PD patients did not receive a MDS-UPDRS-III value and, hence, were excluded from the training. In the training process, whenever a –1 label occurred, the batch was not included into learning this task. Although a relatively small batch size of 8 was chosen to make sure this phenomena does not occur often, it is possible that the task was not properly learned in several training steps. Additionally, since all HC fell under the threshold for class one, a certain imbalance was created towards this class. Further, many samples in the data did not have a m-FDA value. To overcome this, a data imputation approach was followed. As can be seen in Figure 2.3 and Figure 2.4, the HC subject were more spread out after this process and resulted in a certain shift towards the PD distribution. This could have made it harder for the models to recognize these classes. Also, separating the samples into four different classes may result in a loss of important information coming from the exact scoring values. A solution to this could be to perform a regression rather than a classification. However, a different strategy of calculating the loss function would have to be taken into account for this approach, since the loss for classification and regression do not scale in the same way. Still, it is important to notice that the outcome of these tasks are consistent to the results obtained by previous work [VC18a].

Another insight of this thesis is, that the single task neural network approach achieved a lower performance than the baseline and the MTL methods. One explanation could be that too many parameters were predefined. A DL approach has a large amount of parameters. Thus, to keep the computation time and complexity feasible, several parameters, like the batch size, were set beforehand. Additionally, to keep the approaches comparable the classification layers were fixed to two layers throughout this work. However, this may lead to a non-optimal network,

especially for the single task framework. In fact, adding two additional classification layers as a presumption may already lead to overfitting the data in the individual networks, since the additional regularization due to the tasks is not included. This would have to be considered in future work.
A further outcome of this work is, that using the ensemble approach to evaluate the tasks on a per session basis leads to improved performance values throughout all tests and is able to reach competitive results compared to the related work. Hence, this work proposes the use of this evaluation metric in future research to this topic.

Chapter 6

Conclusion

With an aging population the prevalence of PD in industrial nations is expected to increase. Although a majority of PD patients also develop speech related deficits, it is still a minor factor in PD evaluation. Moreover, common PD assessment tools lack objectivity, long-term monitoring capability and availability for strongly impaired people. Hence, computational methods are proposed to close this gap, for instance, by using e-medicine. In recent years DL has proven to surpassed previous state-of-the-art results in various fields. Specifically, MTL frameworks in DL have promising theoretical properties to overcome the issues of little amounts of training samples, which is usually the case in medical .

In this sense, this Master's thesis investigated the application of a MTL framework based on neural networks to PD speech recordings. Therefore, a set of openSMILE features was precomputed and were used as the network's input. Three tasks were introduced to be optimized jointly. The approach was tested in various experimental settings with focusing on the tasks individually or learning weighted influence of them. The outcome was evaluated and compared with respect to single task settings and an Adaboost baseline.

Based on the experimental results, it can be concluded that the MTL approach is superior compared to the alternative neural network approaches. Hereby, focusing on the task to discriminate between PD and HC was shown to be the most promising setting with an UAR of 72.49% for PD vs. HC, 34.45% for MDS-UPDRS-III and 35.94% for m-FDA. An additionally introduced session based evaluation metric improved the results with values of 81.73% for PD vs. HC, 52.45% for MDS-UPDRS-III and 43.56% for m-FDA. The results were compared to those obtained by individual networks trained for each task, which achieved values of 78.22%, 36.23% and 40.52% for PD vs. HC, MDS-UPDRS-III and m-FDA, respectively. Although the MTL framework does not show a clear advantage over the Adaboost baseline, it was demonstrated that it is preferable to

training a neural network on each single task. Further insights were obtained when including more tasks into the framework to enlarge the regularization effect. The additional tasks do not increase the performance, however, they help to understand the effects of MTL, for instance, regarding the individual exercises.

Even though continuous work on the application of MTL based neural networks to PD speech analysis is needed, this work outlined the benefits of such a framework, especially with regard to pathological speech, where the data is sparse. To the best of the author's knowledge, this is the first time PD speech data was analysed in such a MTL setting with predefined features. The results obtained by this thesis deliver further understanding of the behaviour of the proposed methods. Further, it can build the basis of subsequent studies and potentially help to improve the objectivity in PD assessment.

Chapter 7

Outlook

The results obtained by the MTL approach in this thesis leave room for multiple further investigations. One possible future direction could be to merge the presented approach with frameworks based on spectrograms and CNNs, such as in [VC18a]. It is expected that a compound of these two techniques could lead to an enhanced performance, when their benefits are combined. Additionally, one possible aspect to prevent the imbalance towards a single class could be to add a fifth class in the MDS-UPDRS-III task, which would include the HC samples. Thus, the differentiation would be between a HC class and multiple PD classes representing different disease severity ranges. Another possibility would be, rather than performing a classification for the MDS-UPDRS-III and m-FDA tasks, to convert them into a regression problem. To separate these tasks into four classes potentially leads to a loss of information, since multiple score values are summarized to one class. However, therefore the loss function would have to be adjusted. One possible approach could be to introduce an uncertainty measure as proposed in Kendall et al. [Ken17]. Future work could also include more features. Although larger feature sets of the openSMILE did not show an increased performance in initial test, this approach was not exploited to full extend. Further, several new MTL architectures were presented in the literature in recent years [Yan17b, Has16]. Although more complex models would also need more data in order not to overfit, these techniques show interesting approaches. Additionally, as mentioned earlier, several labels were missing especially for the HC participants, so it would also be important to obtain these labels for future work. Several parameters in the network were predefined to keep the complexity of the network in a feasible range. With the obtained knowledge further investigations to understand the influence of these parameters would give an important additional insight. Also, it would be interesting to concentrate more on the parameter selection process to investigate if optimal network settings are found to be in a similar range. Moreover, since the 10 fold cross validation appeared to create

rather different models for each fold, it would be important to investigate, if one of these settings is a good trade-off for the whole data set.
A long-term goal could be to include the presented method into a mobile application with the goal of progressive monitoring of PD patients, which could be of help for medical professionals. Additionally, it would be interesting to combine the proposed speech based approach with other bio-signals to obtain a more accurate description of the behaviour of the disease.

Bibliography

[Ack91] H. Ackermann and W. Ziegler. "Articulatory deficits in parkinsonian dysarthria: an acoustic analysis". *Journal of neurology, neurosurgery, and psychiatry*, 54(12):1093–1098, 1991.

[AM90] Y. Abu-Mostafa. "Learning from hints in neural networks". *Journal of Complexity*, 6(2):192–198, 1990.

[AV18] T. Arias-Vergara, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. "Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions". *Speech Communication*, 101:11–25, 2018.

[Ban13] Y.-I. Bang, K. Min, Y. H. Sohn, and S.-R. Cho. "Acoustic characteristics of vowel sounds in patients with Parkinson disease". *NeuralRehabilitations*, 32(3):649–654, 2013.

[Bax11] J. Baxter. "A Model of Inductive Bias Learning". *CoRR*, abs/1106.0245, 2011.

[Ben12] Y. Bengio. "Practical recommendations for gradient-based training of deep architectures". *CoRR*, abs/1206.5533, 2012.

[Ber12] J. Bergstra and Y. Bengio. "Random Search for Hyper-parameter Optimization". *Journal of Machine Learning Research*, 13:281–305, 2012.

[Ber18] L. Berus, S. Klancnik, M. Brezocnik, and M. Ficko. "Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks". *Sensors (Basel)*, 19(1):16, 2018.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[Bot16] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister. "The mPower study, Parkinson disease mobile data collected using ResearchKit". *Scientific Data*, 3, 2016.

[Car97] R. Caruana. "Multitask Learning". *Machine Learning*, 28(1):41–75, 1997.

[Cer17] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth. "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features". *Computer Speech & Language*, 46:196–208, 2017.

[Den09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Dev18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *CoRR*, abs/1810.04805, 2018.

[Dim17] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi. "Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System". *IEEE Access*, 5:22199–22208, 2017.

[Dud00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, New York, NY, USA, 2000.

[Duo15] L. Duong, T. Cohn, S. Bird, and P. Cook. "Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser". In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850. Association for Computational Linguistics, 2015.

[End08] P. M. Enderby, R. Palmer, and Pro-Ed (Company). "Frenchay dysarthria assessment", 2008.

[Eyb13] F. Eyben, F. Weninger, F. Gross, and B. Schuller. "Recent Developments in openSMILE, the MunichOpen-Source Multimedia Feature Extractor". In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 835–838. ACM Press, 2013.

[Fel98] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[Gab18] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams. "Assessment of Amplified Parkinsonian Speech Quality Using Deep Learning". In *IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, pages 1–4, 2018.

[Gal16] Z. Galaz, J. Mekyska, Z. Mzourek, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, and D. Berankova. "Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease". *Computer Methods and Programs in Biomedicine*, 127:301–317, 2016.

[GL17] J. I. Godino-Llorente, S. Shattuck-Hufnagel, J. Y. Choi, L. Moro-Velázquez, and J. A. Gómez-García. "Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers". *PLOS ONE*, 12(12):1–35, 2017.

[Glo10] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

[Goe08] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, and N. LaPelle. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results". *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.

[Goo16] I. Goodfellow, Y. Bengio, and A. Courville. *"Deep Learning"*. MIT Press, 2016.

[Gró15] T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth. "Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks". In *Proceedings of Interspeech 2015*, pages 919–923, 2015.

[Han84] D. G. Hanson, B. R. Gerratt, and P. H. Ward. "Cinegraphic observations of laryngeal function in parkinson's disease". *The Laryngoscope*, 94(3):348–353, 1984.

[Has16] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher. "A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks". *CoRR*, abs/1611.01587, 2016.

[Hin12] G. Hinton, N. Srivastava, and K. Swersky. "Neural Networks for Machine Learning. Coursera video lecture 6a". `https://www.youtube.com/watch?v=defQQqkXEfE`, 2012. Accessed 05/02/19.

[Hla17] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz. "Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder". *Scientific Reports*, 7(1):1–13, 2017.

[Hor98] O. Hornykiewicz. "Biochemical aspects of Parkinson's disease". *Neurology*, 51(2 Suppl 2):2–9, 1998.

[Iof15] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *CoRR*, abs/1502.03167, 2015.

[Jan08] J. Jankovic. "Parkinson's disease: clinical features and diagnosis". *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008.

[Ken03] R.D. Kent, H.K. Vorperian, J.F. Kent, and J.R. Duffy. "Voice dysfunction in dysarthria: application of the Multi-Dimensional Voice Program". *Journal of Communication Disorders*, 36(4):281–306, 2003.

[Ken17] A. Kendall, Y. Gal, and R. Cipolla. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics". *CoRR*, abs/1705.07115, 2017.

[Kin15] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". *CoRR*, abs/1412.6980, 2015.

[Kri17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". *Commun. ACM*, 60(6):84–90, 2017.

[Li18] F.-F. Li, J. Johnson, and S. Yeung. "Lecture notes in Convolutional Neural Networks for Visual Recognition", 2018.

[Log78] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky. "Frequency and Co-occurrence of Vocal Tract Dysfunctions in the Speech of a Large Sample of Parkinson Patients". *Journal of Speech and Hearing Disorders*, 43(1):47–57, 1978.

[Maa13] A. L. Maas, A. Hannun, and A. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In *International Conference on Machine Learning (ICML)*, 2013.

[Mal19] L. P. Malasinghe, N. Ramzan, and K. Dahal. "Remote patient monitoring: a comprehensive study". *Journal of Ambient Intelligence and Humanized Computing*, 10(1):57–76, 2019.

[Mon18] D. Montaña, Y. Campos-Roca, and C. J. Pérez. "A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease". *Computer Methods and Programs in Biomedicine*, 154:89–97, 2018.

[MV18] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak. "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease". *Applied Soft Computing*, 62:649–666, 2018.

[Nai10] V. Nair and G. E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress.

[OA14] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth. "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease". In *LREC*, pages 342–347, 2014.

[OA16] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth. "Towards an automatic monitoring of the neurological state of Parkinson's patients from speech". In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6490–6494, 2016.

[Oun18] Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijean. "Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity". *Journal of Medical Systems*, 42(2):1–17, 2018.

[Ram08] L. O. Ramig, C. Fox, and S. Sapir. "Speech treatment for Parkinson's disease". *Expert Review of Neurotherapeutics*, 8(2):297–309, 2008.

[Ree14] A. Reeve, E. Simcox, and D. Turnbull. "Ageing and Parkinson's disease: Why is advancing age the biggest risk factor?". *Ageing Res Rev*, 14(100):19–30, 2014.

[Rud17] S. Ruder. "An Overview of Multi-Task Learning in Deep Neural Networks". *CoRR*, abs/1706.05098:30–43, 2017.

[Rus11] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease". *The Journal of the Acoustical Society of America*, 129(1):350–367, 2011.

[Rus13] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka. "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task". *The Journal of the Acoustical Society of America*, 134(3):2171–2181, 2013.

[Sch99] R. E. Schapire. "A Brief Introduction to Boosting". In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[Sch06] A. Schrag. "Quality of life and depression in Parkinson's disease". *Journal of the Neurological Sciences*, 248(1):151–157, 2006.

[Sch10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Deviller, C. Müller, and S Narayanan. "The INTERSPEECH 2010 Paralinguistic Challenge". In *Proceedings of Interspeech 2010*. ISCA, 2010.

[Sch14] B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2014.

[Sch15] B. W. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. "The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition". In *Proceedings of Interspeech 2015*, pages 478–482, 2015.

[Sch19] P. Schwab and W. Karlen. "PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data". In *33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[Smi17] K. M. Smith, J. R. Williamson, and T. F. Quatieri. "Vocal markers of motor, cognitive, and depressive symptoms in Parkinson's disease". In *2017 Seventh International*

Conference on Affective Computing and Intelligent Interaction (ACII), pages 71–78, 2017.

[Sno12] J. Snoek, H. Larochelle, and R. P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, pages 2951–2959, USA, 2012. Curran Associates Inc.

[Sri14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[SS19] R. San-Segundo, H. Navarro-Hellín, R. Torres-Sánchez, J. Hodgins, and F. De la Torre. "Increasing Robustness in the Detection of Freezing of Gait in Parkinson's Disease". *Electronics*, 8(2), 2019.

[Ste15] S. R. Steinhubl, E. D. Muse, and E. J. Topol. "The emerging field of mobile health". *Sci Transl Med*, 7(283):283rv3, 2015.

[Sut13] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. "On the importance of initialization and momentum in deep learning". In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147. PMLR, 2013.

[Sze16] C. Szegedy, S. Ioffe, and V. Vanhoucke. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". *CoRR*, abs/1602.07261, 2016.

[Tja13] K. Tjaden, J. Lam, and G. Wilding. "Vowel acoustics in Parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions". *Journal of Speech, Language, and Hearing Research*, 56(5):1485–1502, 2013.

[Tsa14] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig. "Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease". *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190, 2014.

[Tu17] M. Tu, V. Berisha, and J. Liss. "Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks". In *Proceedings of Interspeech 2017*, pages 1849–1853, 2017.

[Tys17] O.-B. Tysnes and A. Storstein. "Epidemiology of Parkinson's disease. *Journal of Neural Transmission*, 124(8):901–905, 2017.

[vC05] S. von Campenhausen, B. Bornschein, R. Wick, K. Bötzel, C. Sampaio, W. Poewe, W. Oertel, U. Siebert, K. Berger, and R. Dodel. "Prevalence and incidence of Parkinson's disease in Europe". *European Neuropsychopharmacology*, 15(4):473–490, 2005.

[VC17] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease". In *Proceedings of Interspeech 2017*, pages 314–318, 2017.

[VC18a] J. C. Vásquez Correa, T. Arias, J. R. Orozco-Arroyave, and E. Nöth. "A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease". In *Interspeech 2018*, pages 456–460, ISCA, 2018. ISCA.

[VC18b] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth. "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease". *Journal of Communication Disorders*, 76:21–36, 2018.

[vdO16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". *CoRR*, abs/1609.03499, 2016.

[Wan18] S. Wan, Y. Liang, Y. Zhang, and M. Guizani. "Deep Multi-Layer Perceptron Classifier for Behavior Analysis to Estimate Parkinson's Disease Severity Using Smartphones". *IEEE Access*, 6:36825–36833, 2018.

[Wei98] G. Weismer and J. Wildermuth. "Formant trajectory characteristics in persons with Parkinson, cerebellar, and upper motor neuron disease". *The Journal of the Acoustical Society of America*, 103(5):2892–2892, 1998.

[Yan17a] Q. Yang and Y. Zhang. "An overview of multi-task learning". *National Science Review*, 5(1):30–43, 2017.

[Yan17b] Y. Yang and T. Hospedales. "Deep Multi-task Representation Learning: A Tensor Factorisation Approach". In *International Conference on Learning Representations (ICLR 2017)*, 2017.

[Zha16] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis. "High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection". *CoRR*, abs/1601.00960, 2016.

[Zha18] H. Zhang, A. Wang, D. Li, and W. Xu. "DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification". In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 214–217, 2018.

Appendix A

Speech exercises

A.1 Sentences

The list of sentences included into the exercises includes syntactically simple and complex sentences, as well as sentences with emphasis illustrated in capital letters.

1. Mi casa tiene tres cuartos. (Simple)
   Translation: My house has three rooms.

2. Omar, que vive cerca, trajo miel. (Complex)
   Translation: Omar, who lives near, brought honey.

3. Laura sube al tren que pasa. (Complex)
   Translation: Laura gets on the passing train.

4. Los libros nuevos no caben en la mesa de la oficina. (Simple)
   Translation: The new books do not fit in the table of the office.

5. Rosita Niño, que pinta bien, donó sus cuadros ayer. (Complex)
   Translation: Rosita Niño, who paints well, donated her paintings yesterday.

6. Luisa Rey compra el colchón duro que tanto le gusta. (Complex)
   Translation: Luisa Rey buys the hard mattress that she likes so much.

7. Viste las noticias? Yo vi GANAR la medalla de plata en pesas. Ese muchacho tiene mucha fuerza!

Translation: Did you see the news? I saw to WIN the silver medal in Weightlifting. That boy is very strong!

8. Juan se ROMPIÓ una PIERNA cuando iba en la MOTO.
   Translation: Juan BROKE his LEG when he was driving his motorcycle.

9. Estoy muy triste, ayer vi MORIR a un amigo.
   Translation: I am very sad, yesterday I saw a friend DIE.

10. Estoy muy preocupado, cada vez me es más difícil HABLAR.
    Translation: I am very worried, it is getting harder to me to TALK.

A.2 Read text

The patient-doctor dialog as it was used in the read text exercise. "**P**" stands for patients and "**D**" for doctor, respectively.
P: Ayer fui al médico.
D: Qué le pasa? Me preguntó.
P: Yo le dije: Ay doctor! Donde pongo el dedo me duele.
D: Tiene la uña rota?
P: Sí.
D: Pues ya sabemos qué es. Deje su cheque a la salida.

Translation
P: Yesterday I went to the doctor.
D: What happened to you? He asked me.
P: I said him: ah doctor! Where I put my finger it hurts.
D: Do you have the nail broken?
P: Yes.
D: Then we now know what is happening. Leave your check at the exit.

Appendix B

openSMILE features

Here is the list of functionals included in the openSMILE feature set for the2010 INTERSPEECH ComPare challenge. More details can be found in [Eyb13] and the project's website (*https://www.audeering.com/opensmile/*).

The 34 low-level descriptor are the following:

- **pcm loudness** The loudness as the normalised intensity raised to a power of 0.3.
- **mfcc** Mel-Frequency cepstral coeffcients 0 – 14
- **logMelFreqBand** Logarithmic power of Mel-frequency bands 0 – 7 (distributed over a range from 0 to 8 kHz)
- **lspFreq** The 8 line spectral pair frequencies computed from 8 LPC coeffcients.
- **F0fnEnv** The envelope of the smoothed fundamental frequency contour.
- **voicingFinalUnclipped** The voicing probability of the final fundamental frequency candidate. Unclipped means, that it was not set to zero when is falls below the voicing threshold.

The following list shows the 21 functionals:

- **maxPos** The absolute position of the maximum value (in frames)
- **minPos** The absolute position of the minimum value (in frames)
- **amean** The arithmetic mean of the contour
- **linregc1** The slope (m) of a linear approximation of the contour

- **linregc2** The offset (t) of a linear approximation of the contour
- **linregerrA** The linear error computed as the difference of the linear approximation and the actual contour
- **linregerrQ** The quadratic error computed as the difference of the linear approximation and the actual contour
- **stddev** The standard deviation of the values in the contour
- **skewness** The skewness (3rd order moment).
- **kurtosis** The kurtosis (4th order moment).
- **quartile1** The first quartile (25 % percentile)
- **quartile2** The first quartile (50 % percentile)
- **quartile3** The first quartile (75 % percentile)
- **iqr1–2** The inter-quartile range: quartile2 – quartile1
- **iqr2–3** The inter-quartile range: quartile3 – quartile2
- **iqr1–3** The inter-quartile range: quartile3 – quartile1
- **percentile1.0** The outlier-robust minimum value of the contour, represented by the 1 % percentile.
- **percentile99.0** The outlier-robust maximum value of the contour, represented by the 99 % percentile.
- **pctlrange0–1** The outlier robust signal range 'max–min' represented by the range of the 1 % and the 99 % percentile.
- **upleveltime75** The percentage of time the signal is above (75 % · range + min).
- **upleveltime90** The percentage of time the signal is above (90 % · range + min).

The next list contains the four pitch related LLDs and their corresponding delta coeffcients:

- **F0final** The smoothed fundamental frequency contour
- **jitterLocal** The local (frame-to-frame) Jitter (pitch period length deviations)
- **jitterDDP** The differential frame-to-frame Jitter (the “Jitter of the Jitter”)
- **shimmerLocal** The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)

Appendix C

Hyperparameter search for seven-task MTL

Table C.1: The results per fold from the hyperparameter optimization of including all seven tasks into the learning process.

Fold	LR	Layer	Dropout	TW PD vs. HC	UPDRS	m-FDA	Exercise	Ac. cond.	Gender	Age
1	0.0001	5	0.0	0.04	0.39	0.04	0.39	0.04	0.04	0.04
2	0.1	3	0.0	0.03	0.29	0.03	0.29	0.03	0.29	0.03
3	0.0001	5	0.0	0.06	0.60	0.06	0.06	0.06	0.06	0.06
4	0.0001	5	0.0	0.29	0.29	0.03	0.03	0.03	0.29	0.03
5	0.0001	5	0.0	0.04	0.39	0.04	0.39	0.04	0.04	0.04
6	0.0001	5	0.0	0.39	0.39	0.04	0.04	0.04	0.04	0.04
7	0.0001	4	0.0	0.03	0.29	0.03	0.29	0.03	0.29	0.03
8	0.0001	4	0.0	0.29	0.29	0.03	0.03	0.03	0.29	0.03
9	0.0001	4	0.0	0.03	0.29	0.03	0.39	0.03	0.29	0.03
10	0.0001	5	0.0	0.29	0.29	0.03	0.03	0.29	0.03	0.03

Appendix D

Confusion matrices

D.1 Adaboost baseline

Table D.1: Confusion matrices of the m-FDA task for the Adaboost baseline.

		Prediction			
		1	2	3	4
Reference	1	866	383	159	121
	2	363	564	347	284
	3	188	485	309	236
	4	133	291	274	142

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.57	0.25	0.10	0.08
	2	0.23	0.36	0.22	0.18
	3	0.15	0.40	0.25	0.19
	4	0.16	0.35	0.33	0.17

(b) Normalized confusion matrix

D.2 Single-task neural network

Table D.2: Confusion matrices of the PD vs HC task single task neural network.

		Prediction	
		PD	HC
Reference	PD	2610	970
	HC	480	1085

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.73	0.27
	HC	0.31	0.69

(b) Normalized confusion matrix

Table D.3: Confusion matrices of the m-FDA task single task neural network.

Reference \ Prediction	1	2	3	4
1	827	460	124	118
2	419	625	269	245
3	255	510	242	211
4	130	343	166	201

(a) Confusion matrix

Reference \ Prediction	1	2	3	4
1	0.54	0.30	0.08	0.08
2	0.27	0.40	0.17	0.16
3	0.21	0.42	0.20	0.17
4	0.15	0.41	0.20	0.24

(b) Normalized confusion matrix

D.3 Multitask neural network

D.3.1 Focus on PD vs. HC

Table D.4: Confusion matrices of the PD vs HC task for focusing on PD vs. HC.

Reference \ Prediction	PD	HC
PD	2656	924
HC	457	1108

(a) Confusion matrix

Reference \ Prediction	PD	HC
PD	0.74	0.26
HC	0.29	0.71

(b) Normalized confusion matrix

Table D.5: Confusion matrices of the m-FDA task for focusing on PD vs. HC. The values from 1 to 4 represent the respective classes.

Reference \ Prediction	1	2	3	4
1	900	194	277	215
2	338	248	322	257
3	356	154	412	490
4	136	171	316	359

(a) Confusion matrix

Reference \ Prediction	1	2	3	4
1	0.57	0.12	0.17	0.14
2	0.29	0.21	0.28	0.22
3	0.25	0.11	0.29	0.35
4	0.14	0.17	0.32	0.37

(b) Normalized confusion matrix

D.3.2 Focus on MDS-UPDRS-III

Table D.6: Confusion matrices of the PD vs HC task for the MTL focusing on MDS-UPDRS-III.

		Prediction	
		PD	HC
Reference	PD	2336	1244
	HC	572	993

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.65	0.35
	HC	0.37	0.63

(b) Normalized confusion matrix

Table D.7: Confusion matrices of the MDS-UPDRS-III task for the MTL focusing on MDS-UPDRS-III.

		Prediction			
		1	2	3	4
Reference	1	783	769	216	344
	2	152	182	64	144
	3	164	176	79	130
	4	91	120	71	208

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.37	0.36	0.10	0.16
	2	0.28	0.34	0.12	0.27
	3	0.30	0.32	0.14	0.24
	4	0.19	0.24	0.14	0.42

(b) Normalized confusion matrix

Table D.8: Confusion matrices of the m-FDA task for the MTL focusing on MDS-UPDRS-III

		Prediction			
		1	2	3	4
Reference	1	674	485	168	202
	2	384	617	281	276
	3	199	580	182	257
	4	102	333	185	220

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.44	0.32	0.11	0.13
	2	0.25	0.40	0.18	0.18
	3	0.16	0.48	0.15	0.21
	4	0.12	0.40	0.22	0.26

(b) Normalized confusion matrix

D.3.3 Focus on m-FDA

Table D.9: Confusion matrices of the PD vs HC task for focusing on m-FDA.

		Prediction	
		PD	HC
Reference	PD	2391	1189
	HC	540	1025

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.67	0.33
	HC	0.35	0.65

(b) Normalized confusion matrix

Table D.10: Confusion matrices of the MDS-UPDRS-III task for focusing on m-FDA.

		Prediction			
		1	2	3	4
Reference	1	957	693	122	340
	2	230	162	42	108
	3	176	192	56	125
	4	180	147	35	128

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.45	0.33	0.06	0.16
	2	0.42	0.30	0.08	0.20
	3	0.32	0.35	0.10	0.23
	4	0.37	0.30	0.07	0.26

(b) Normalized confusion matrix

Table D.11: Confusion matrices of the m-FDA task for focusing on m-FDA.

		Prediction			
		1	2	3	4
Reference	1	787	552	60	130
	2	370	791	170	227
	3	229	572	170	247
	4	141	441	78	180

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.51	0.36	0.04	0.09
	2	0.24	0.51	0.11	0.15
	3	0.19	0.47	0.14	0.20
	4	0.17	0.53	0.09	0.21

(b) Normalized confusion matrix

D.3.4 Learn loss function weights

Table D.12: Confusion matrices of the PD vs HC task for learned loss function weights.

		Prediction	
		PD	HC
Reference	PD	1956	1624
	HC	409	1156

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.55	0.45
	HC	0.26	0.74

(b) Normalized confusion matrix

Table D.13: Confusion matrices of the MDS-UPDRS-III task for learned loss function weights.

		Prediction			
		1	2	3	4
Reference	1	1017	741	250	104
	2	231	233	46	32
	3	177	253	57	62
	4	175	175	73	67

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.48	0.35	0.12	0.05
	2	0.43	0.43	0.08	0.06
	3	0.32	0.46	0.10	0.11
	4	0.36	0.36	0.15	0.14

(b) Normalized confusion matrix

Table D.14: Confusion matrices of the m-FDA task for learned loss function weights.

		Prediction			
		1	2	3	4
Reference	1	589	546	283	111
	2	329	586	430	213
	3	154	572	262	230
	4	117	351	178	194

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.39	0.36	0.19	0.07
	2	0.21	0.38	0.28	0.14
	3	0.13	0.47	0.22	0.19
	4	0.14	0.42	0.21	0.23

(b) Normalized confusion matrix

D.3.5 MTL with seven tasks

Table D.15: Confusion matrices of the PD vs HC task for MTL with seven tasks.

		Prediction	
		PD	HC
Reference	PD	2549	1031
	HC	613	952

(a) Confusion matrix

		Prediction	
		PD	HC
Reference	PD	0.71	0.29
	HC	0.39	0.61

(b) Normalized confusion matrix

Table D.16: Confusion matrices of the MDS-UPDRS-III task for MTL with seven tasks.

		Prediction			
		1	2	3	4
Reference	1	1139	565	176	232
	2	224	154	69	95
	3	203	215	46	85
	4	269	126	21	74

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.54	0.27	0.08	0.11
	2	0.41	0.28	0.13	0.18
	3	0.37	0.39	0.08	0.15
	4	0.55	0.26	0.04	0.15

(b) Normalized confusion matrix

Table D.17: Confusion matrices of the m-FDA task for MTL with seven tasks.

		Prediction			
		1	2	3	4
Reference	1	705	306	292	226
	2	434	353	454	317
	3	253	253	366	346
	4	150	249	211	230

(a) Confusion matrix

		Prediction			
		1	2	3	4
Reference	1	0.46	0.20	0.19	0.15
	2	0.28	0.23	0.29	0.20
	3	0.21	0.21	0.30	0.28
	4	0.18	0.30	0.25	0.27

(b) Normalized confusion matrix

Table D.18: Confusion matrices of the Exercise task for MTL with seven tasks. The numbers 1 to 4 represent DDK, monologue, read text and sentence exercises.

Reference \ Prediction	1	2	3	4
1	1481	3	28	421
2	30	246	64	35
3	4	3	275	96
4	28	2	12	2471

(a) Confusion matrix

Reference \ Prediction	1	2	3	4
1	0.77	0.00	0.01	0.22
2	0.08	0.66	0.17	0.09
3	0.01	0.01	0.73	0.25
4	0.01	0.00	0.00	0.98

(b) Normalized confusion matrix

Table D.19: Confusion matrices of the Acoustic condition task for MTL with seven tasks. The numbers 1 to 4 represent the sounproof booth, portable soundproof both, headset and at-home conditions.

Reference \ Prediction	1	2	3	4
1	781	185	537	32
2	47	204	257	50
3	102	226	1625	228
4	51	95	465	260

(a) Confusion matrix

Reference \ Prediction	1	2	3	4
1	0.51	0.12	0.35	0.02
2	0.08	0.37	0.46	0.09
3	0.05	0.10	0.75	0.10
4	0.06	0.11	0.53	0.30

(b) Normalized confusion matrix

Table D.20: Confusion matrices of the Gender task for MTL with seven tasks. M stands for male and F for female.

Reference \ Prediction	M	F
M	2190	405
F	666	1884

(a) Confusion matrix

Reference \ Prediction	M	F
M	0.84	0.16
F	0.26	0.74

(b) Normalized confusion matrix

Table D.21: Confusion matrices of the Age task for MTL with seven tasks.

Reference \ Prediction	1	2	3	4
1	170	293	275	508
2	157	412	280	718
3	196	181	430	446
4	158	229	245	447

(a) Confusion matrix

Reference \ Prediction	1	2	3	4
1	0.14	0.24	0.22	0.41
2	0.10	0.26	0.18	0.46
3	0.16	0.14	0.34	0.36
4	0.15	0.21	0.23	0.41

(b) Normalized confusion matrix

Appendix E

Exercise confidence scores

These figures show the exercise confidence scores for the DDK (see Figure E.1) and read text (see Figure E.2) exercises.

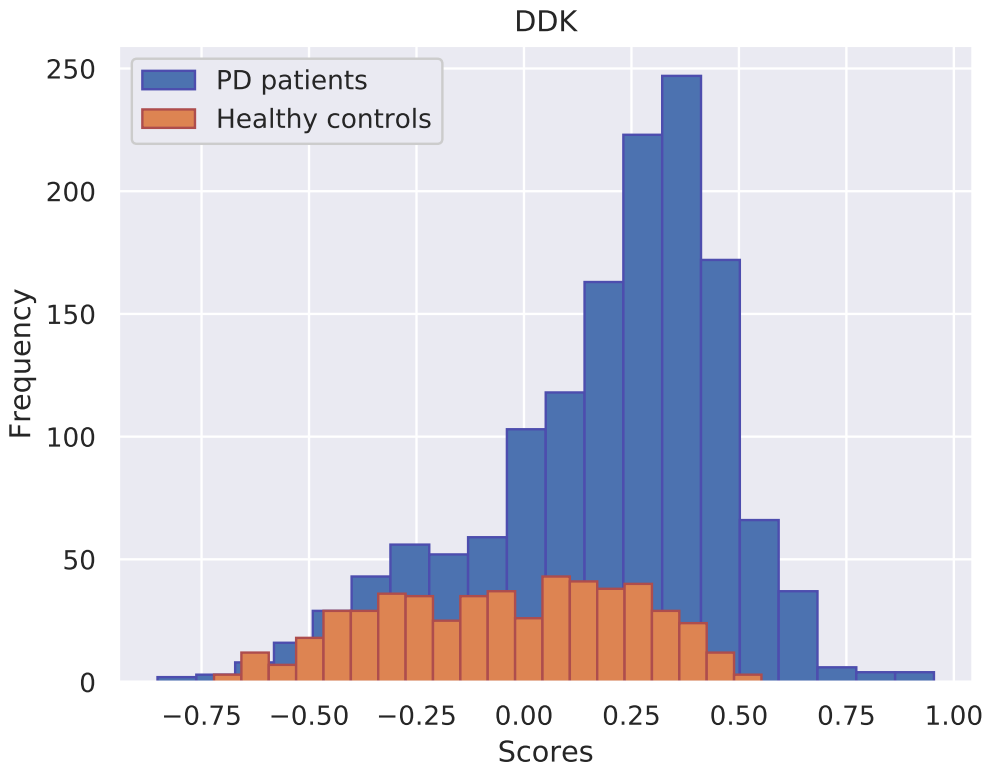

Figure E.1: Score distribution according to the PD vs. HC. classifications in the DDK exercises. Similarly to the sentences it can be seen that PD is confidently classified with this exercise.

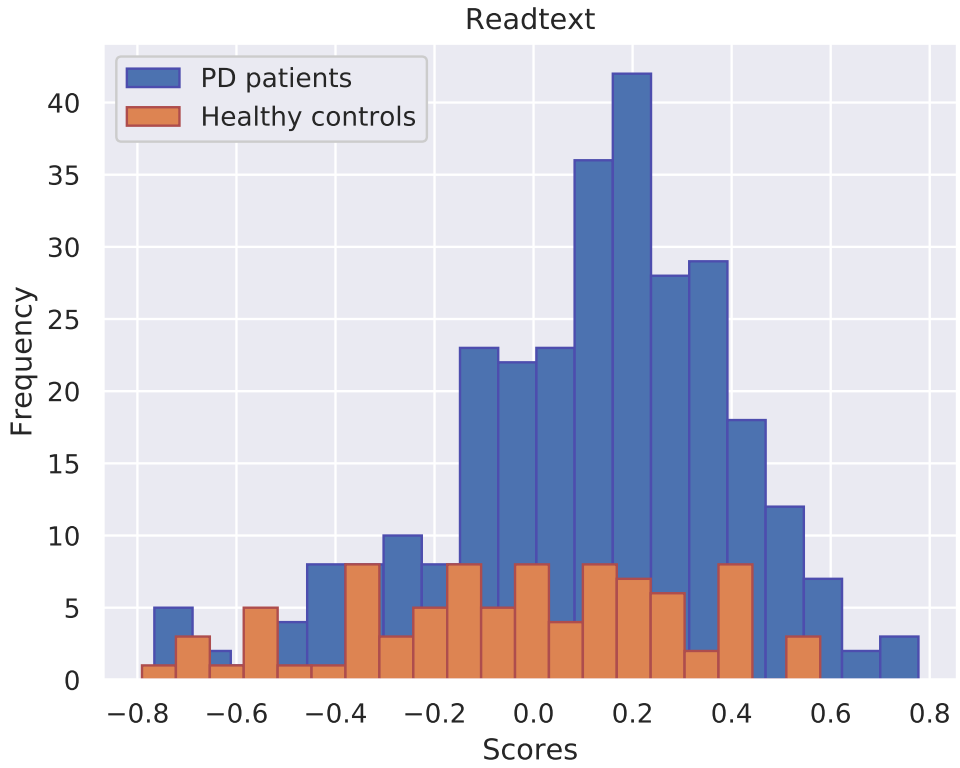

Figure E.2: Score distribution according to the PD vs. HC. classifications in the read text exercise. While PD classification tends to the right direction, HC discrimination does not seem to benefit from this exercise.