FILOZOFICKÁ FAKULTA
Univerzita Karlova

MASTER THESIS

Emil Svoboda

Acoustic features of speech in multiple sclerosis

Akustické charakteristiky hlasu při roztroušené skleróze

Institute of Phonetics

Supervisor of the master thesis:	Ing. Tomáš Bořil, Ph.D
Study programme:	Filologie (Philology)
Study branch:	Fonetika (Phonetics)

Prague 2020

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

In Prague on the 25th of May, 2020 signature of the author

The people thanked in this dedication are by no means ordered by the importance of their contribution to this thesis, as attempting to do so would be both disrespectful and futile.

I thank my dear Annie, my girlfriend as of the time of the submission of this thesis, for keeping the practical aspects of my life functional while this thesis was being written.

I thank Tereza Tykalová and Jan Rusz of CTU for letting me use their data and providing me with ideas and knowledge.

I thank my supervisor Tomáš Bořil for helping me flesh out the idea for this thesis, get in touch with the wonderful afomrementioned people of CTU, acquire the necessary data, rally up the annotators and for providing me with various ideas, bits of knowledge and tips on good practice.

I also thank Dominik Škrabal, Tereza Průchová, Andrea Hubertová and Ivan Kartáč for their work in creating and annotating some of the recordings used in this thesis.

Title: Acoustic features of speech in multiple sclerosis

Author: Emil Svoboda

Institute: Institute of Phonetics

Supervisor: Ing. Tomáš Bořil, Ph.D, Institute of Phonetics

Abstract: This thesis analyzes what acoustically sets apart recordings of healthy people from recordings of people afflicted with multiple sclerosis, and how this distinction can be used to automatically detect multiple sclerosis from fairly simple recordings of a subject's voice, potentially discovering early cases of this disease. Chapter 1 includes the theoretical background of the effect of multiple sclerosis on speech and the descriptions of the data, software, hypotheses and assumptions used here. Two sets recordings of read speech were used, a corpus of afflicted speakers and a control corpus of healthy speakers, totalling 250 individuals. A subset of this corpus was manually annotated, resulting in one dataset. Simultaneously, these entire corpora were also annotated automatically, resulting in another dataset, which was created to explore the possibility of detecting multiple sclerosis automatically. Chapter 2 describes the 13 acoustic parameters used in this thesis, their exact hypothesized relationships with the symptoms of multiple sclerosis and the ways they were calculated. Chapter 3 elaborates on the statistical testing of the aforementioned parameters, their interpretation, the success rate of the two machine learning models used to assess their total predictive power, and a potential way to apply the principles of one of these models practically. In the case of the manual dataset, all 13 parameters were used, of which 9 were statistically significant. The machine learning model trained with these data exhibited a 93% raw accuracy with a Cohen's kappa of $\varkappa = 0.63$. The automatically annotated dataset was paremeterized using only 12 parameters, of which 7 were significant. The machine learning model employed on these data exhibited a raw 85% accuracy and a kappa $\varkappa = 0.5$. The thesis then goes on to propose a way to practically apply these findings in a clinical setting.

Keywords: phonetics, dysarthria, multiple sclerosis, neurodegenerative, biometrics

Název: Akustické charakteristiky hlasu při roztroušené skleróze

Autor: Emil Svoboda

Ústav: Fonetický ústav

Vedoucí: Ing. Tomáš Bořil, Ph.D, Institute of Phonetics

Abstrakt: Tato práce se zabývá akustickými odlišnostmi nahrávek zdravých lidí a lidí postižených roztroušenou sklerózou a také tím, jak mohou tyto odlišnosti být použity za účelem automatické detekce roztroušené sklerózy z jednoduchých hlasových nahrávek a z toho vyplývajícího případného brzkého nalezení rozvíjející se choroby. Kapitola 1 obsahuje teoretické pozadí vlivu roztroušené sklerózy na řeč a také popis dat, softwaru, hypotéz a předpokladů v práci užitých. Za tímto účelem byly užity dva soubory nahrávek čteného textu, jeden s nahrávkami nemocných mluvčích a druhý obsahoval zdravé kontrolní mluvčí. Některé z těchto souborů byly ručně anotovány, což představuje první soubor dat. Současně byly celé tyto dva soubory nahrávek anotovány automaticky, čímž byl vytvořen druhý, větší soubor dat. Tento byl vytvořen za účelem přezkoumání možnosti detekovat roztroušenou sklerózu čistě automaticky. Kapitola 2 popisuje 13 akustických parametrů použitých v této práci, jejich předpokládané vztahy se symptomatologií roztroušené sklerózy a metody jejich byly výpočtu. Kapitola 3 se zabývá statistickým testováním parametrů, jejich interpretací, úspěšností dvou machine learningových modelů vytvořených za účelem zhodnocení jejich celkové prediktivní síly a možným praktickým využitím druhého z nich. Z ručního datového souboru bylo vypočteno 13 parametrů, u 9 z nichž byla prokázána statistická významnost. Model na nich založený vykázal 93% hrubou úspěšnost a Cohenovu kappu $\varkappa = 0{,}63$. Automaticky anotovaný soubor dat byl parametrizován pouze 12 z nich, z nichž 7 bylo významných. Model vytvořený pomocí těchto dat vykázal 85% hrubou úspěšnost a kappu $\varkappa = 0{,}5$.

Klíčová slova: fonetika, dysarthrie, roztroušená skleróza, neurodegenerativní, biometrika

Contents

Preface

The primary reason I have decided to write this thesis is trivial – when my mother was diagnosed with multiple sclerosis, it was because I had noticed neurological symptoms in her one year prior, when I was but a 17-year-old boy with absolutely no medical training. About a year ago, this, along with the assumption that the entire situation had not just been pure luck on my part, led me to believe that if neurological red flags can be noticed by a boy, then perhaps these red flags could be theoretically detected by automatic means by a biometric system of sorts.

What inspired this, at least in part, was hearsay of some incredibly successful research going on at ČVUT regarding Parkinson's disease and also, it just so happened that at that time, I was a student of phonetics pondering the topic of his future Master's thesis. I put two and two together and decided to construct a machine learning model that would be able to detect multiple sclerosis automatically.

The original title of this thesis was supposed to be something along the lines of "Automatic detection of multiple sclerosis" from read speech, but my supervisor felt that the selection, assessment and testing of potentially useful acoustic parameters for the purposes of this application was ambitious enough. Trusting his judgment, I hearkened his advice, but deciding not to abandon my hopes of constructing a bona-fide proof of concept of such a detection tool, I regardless attempted to go through with my original concept. I would like to take this opportunity to alert the reader to a matter perhaps obvious but nevertheless critical to mention – *in its current state, the model is absolutely nothing but proof of concept.* It is not practically applicable in any way due to its unwieldiness, inefficiency, overreliance on otherwise irrelevant software and a tedious workflow that nearly defeats its purpose. However, I wish to emphasize this point as well: *all of the tedious labour, inefficiency and unwieldiness can indisputably be automated and streamlined by a team of skilled programmers.* However, such an accomplishment is well beyond both the scope of this thesis and my software engineering skills.

Ultimately, I would say that my personal motivation for choosing this topic is twofold. First, regarding quality of life, the sooner neurological diseases are diagnosed, the better the quality of life of the patient. Since "Patients who begin treatment later do not reap the same benefits as those who begin treatment earlier during the disease course.", as Miller [2004] slightly understates in the abstract of his article, it seems obvious that it would be ultimately highly beneficial to many people to have a cheap, reausable tool that can detect MS early on. As I was fortunate enough to have the opportunity to attempt to do something genuinely *pro bono publico*, I decided to go through with this thesis topic despite the fact that my mathematical and data scientific background can be described as limited at best.

Secondly, I just consider the entire idea of this model to be inherently incredibly appealling. There is something just inherently fascinating about simulating an artificial brain-like structure and training it to acoustically detect and recognize a potentially debilitating disease. I thus knew that I would enjoy working on this, and consequently give my absolute best performance. I am happy to report that I indeed have thoroughly enjoyed working on this model, and thus hopefully have given an exceptional performance.

Please note that I will try my best to refrain from using the term *diagnose*, since I strongly prefer *detect*. This is because a tool like this cannot be used to diagnose MS in the same way that a skilled neurologist equipped with an magnetic resonance imaging (MRI) machine and all the necessary knowledge and intuition can. Any “diagnosis” given by such a model cannot be considered on par with a result from such a physician, and thus this detection tool can only be used to find people who should be referred to a professional. More on this in subsection 3.3.1 on page 45.

1. Introduction

The objective of this thesis is to lay grounds for the construction of a tool that automatically detects multiple sclerosis from the voice recording of a potentially afflicted individual using no input but that individual's voice recordings and their basic biometric data such as age and sex.

This encompasses the statement, theoretical evaluation and statistical testing of relevant acoustical measurements for the purposes of constructing such a model. If possible, the construction of a proof-of-concept machine learning (ML) working model proving the workability of the proposition also falls within the objective of this work, as well as a proposition of a practical application paradigm.

The practical part of this thesis consists of two largely similar parts, one consisting of *manually* annotated data, the other of *automatically* annotated data. Annotation in this context means that in recordings such as used in this thesis, it has to be temporally discriminated *when* relevant phenomena, such as particular phoneme realizations, occur, so that their acoustic characteristics can be measured. This can be done in two ways – it can be done manually or it can be done automatically.

The manual approach is much more reliable in terms of errors and thus produces results which are in and of themselves interpretable and scientifically relevant. Therefore, statistical tests can be performed on thus acquired data and based on them, general statements about dysarthric individuals can be made, conclusions be drawn and the sum of human knowledge expanded.

Annotating data automatically is much cheaper and makes it possible for *full automatic detection* to be potentially applied, but is much less reliable in terms of error rate. Parameters thus acquired are therefore much less suitable for intrepretation, but are reliable enough to train a machine learning model to discriminate healthy recordings from dysarthric ones.

The objective of this thesis is, however, not the assessment of how acoustic features of multiple sclerosis (MS)-afflicted speech correlate with the various subtypes of MS. This work concentrates fully on the possibility of constructing and applying such a model along with the evaluation of possible parameters used therein, without delving too deep into the intricacies of what happens between MS and its associated dysarthrias. This sacrifice is made purely for the sake of focus.

Similarly, for the same reason, no effort will be made to generalize any of these findings beyond the borders of the Czech language. It is not unreasonable to assume that having tweaked parameters and being trained on a different dataset, a similar model could be deployed to detect multiple sclerosis in speakers of languages other than Czech, but this is by no means the objective of this thesis.

1.1 Theoretical background

In this chapter, I will put forward a general theoretical foundation on what multiple sclerosis actually is, along with its symptoms, especially in the context of speech, and why it is important that it is treated as early as possible. This will be done in subsection 1.1.1.

Next, in subsection 1.1.2, I will briefly elaborate on how such biological phenomena are measured acoustically according to the scientific literature in practice. This will be a general overview of sorts; not all techniques and methods described here will ultimately be used for the construction of the ML model.

In section 1.2, all recordings, annotation methods, software and machine learning tools used in this thesis will be described and properly attributed to their respective authors, since they have been aggregated from various sources.

The final section, 1.3, will present all assumptions, theoretical, statistical, and otherwise, that are necessary to draw all the conclusions claimed at the end of this thesis.

1.1.1 Multiple sclerosis

In this subsection, I will give a short introduction on what multiple sclerosis actually is. Following this general description is a deeper explanation of the three main symptoms (or rather clusters of symptoms) most relevant to this work.

Multiple sclerosis, or MS for short, is a neurodegenerative disorder affecting the central nervous system (CNS) of the affected individual. More precisely, for reasons unknown as of the time this thesis is being written, the body's immune system attacks the myelin sheath of the individual's nervous system. The myelin sheath is a protective layer made of a fatty substance that covers one's nervous tissue. If we imagine nerves as being electric wires, the myelin sheath serves as an insulating layer.

When MS develops, the immune system's attack manifests as localized areas of swelling and inflammation in the myelin-rich white matter of the CNS (though grey matter, which contains less myelin, may be affected as well), which, after some time, may develop into areas of hardened scar tissue called lesions.[1] These lesions compromise the myelin sheath's ability to properly insulate the nervous tissue to let it transmit signals around the body properly, resulting in a potentially wide array of neurological symptoms.

Not only that, but these lesions often seep into the nervous tissue itself, disrupting any signals attempting to pass through. ([Amor and Van Noort, 2012, p. 3-8])

Because of the localized nature of the damage, it is critical to understand that the precise neurological difficulties any single patient may experience often vary significantly. Intuitively, this is because these symptoms stem from the exact part of the brain, spine or other location in the nervous system that is affected. This presents a significant challenge for anyone attempting to build a detection system – because of the unusually high (among neurological disorders) symptomatic variability.

Symptoms of MS may include difficulty walking, ataxia and lowered coordination, psychological symptoms such as depression, various cognitive difficulties such as impaired memory, tremors, spastic muscles and cramps, pain, sensory difficulties such as blurred, double or otherwise impaired vision, insufficient bladder function and sphincter functions, general fatigue, dysphagia and dysarthria. (Kister et al. [2013])

Not all of these are necessarily relevant for speech biometrics, but some of them are in an indirect way. For example, impaired vision and memory may contribute to the fact that when reading a text aloud (like in the recordings used in this thesis), subjects may experience increased difficulty understanding and processing the text, which may exhibit in their speech in a measurable way despite not being dysarthria *per se*.

Multiple sclerosis manifests itself in several standardized subtypes depending on author, these being relapsing-remitting, secondary progressive, primary progressive and progressive relapsing. Clinically isolated syndrome is also sometimes mentioned, although that term is reserved for individuals with MS-like symptoms, but who do not otherwise meet necessary criteria to be diagnosed with the disease proper.

The **relapsing-remitting** form is the most common one, constituting about 85% of known cases of multiple sclerosis. (Hurwitz [2009]) Typically, its symptomatology constitutes of so-called attacks, where the patient's condition worsens significantly for a short amount of time spanning roughly days before getting bet-

[1] The word "sclerosis" comes from the Greek σκληρός, meaning hard.

ter again for some time.

This presents another challenge for both the construction of a functional detection model and its application, because the patient might be nearly completely asymptomatic during recording. Relapsing-remitting MS often (but not always) transforms into the **secondary progressive** form, where the attacks cease to appear and the patient's condition instead deteriorates much more gradually.

Similar to this is the **primary progressive** form, which is not preceded by relapsing-remitting MS and typically develops much more steadily. (Hurwitz [2009])

According to Wilkins [2019], the total population prevalence of MS in the USA is about 0.3 % at the time. Despite the fact that MS seems appears more commonly further from the equator, there is no reason to assume that the general prevalence will be significantly different in the Czech republic, at least for detection purposes.

For the purposes of this work and especially this chapter, as well as chapter 2, three relevant symptoms, or rather perhaps three symptom subgroups, have been identified as especially relevant. It is critical that the reader understands that these symptom groups exhibit significant overlap in the manner that they manifest themselves in dysarthric speech, and one symptom may exhibit as several acoustic phenomena, as well as one acoustic abnormality may be caused by symptoms from more than one group. These follow in no particular order.

Spasticity in speech

According to a rigorous definition found in Mclellan [1981], muscular spasticity in MS "is a motor disorder characterised by a velocity dependent increase in tonic stretch reflexes (muscle tone) with exaggerated tendon jerks, resulting from hyperexcitability of the stretch reflex, as one component of the upper motor neurone syndrome."

For the purposes of this thesis, this definition will be loosened somewhat. Various episodes of muscle stiffness, jerkiness, poor muscle control, tremors or similar phenomena all fall into this category.

Assuming the figures found in Barnes et al. [2003] on Northern English patients diagnosed with MS can be extrapolated to the Czech republic with a reasonable degree of accuracy, acoustic correlates of spasticity are of great value to speech-based MS detection, because patients with this family of symptoms constitute 43% of all patients diagnosed with MS. Because the definition used for the purposes of this thesis is looser, acoustical correlates of spasticity logically have a chance of being found in *at least* 43% of positive recordings.

Acoustic abnormalities hypothesized to be found in dysarthric speech with spastic components in this study include irregularites in amplitude due to impaired control of the breathing muscles – spasms, even mild ones, can be presumed to produce abnormal **fluctuations in amplitude**. It is also hypothesized that spastic individuals would have a **larger number of glottal stops** present in their speech due to them involuntarily closing the glottis inappropriately. It can also be assumed that articulating any full oral aperture, spastic individuals will have trouble reopening that same aperture due to muscle stiffness, resulting in **longer stop consonants**.

Longer recordings due to articulatory muscle control difficulties and frequent mistakes (and subsequent corrections) stemming thereof, although this parameter is presumably related to all three of the main symptom groups roughly equally, albeit for different reasons. It should be noted here that all data used in this thesis are recordings of the same text read by various individuals.

Fatigue in speech

Fatigue is one of the less outwardly salient symptoms of MS. It appears in roughly 78% of MS patients (applying the same assumptions and caveats as in 1.1.1). (Freal et al. [1984]) In this thesis, the term “fatigue” refers to what one intuitively expects it to – a feeling of tiredness, be it cognitive, psychological, physical or otherwise.

Acoustical abnormalities hypothesized to be found in dysarthric speech with components of fatigue are **longer vowels** and **longer pauses**. This relies on the presumption that a fatigued individual can be expected to try and find moments of respite during speech, which is an activity involving a non-negligible amount of physical strain, a fairly high cognitive strain and an exceedingly high coordinative strain.

Because of this, fatigued individuals may attempt to rest by prolonging vowels, since this takes off coordinative strain for a moment, reduces physical strain and buys them time to think or read slightly ahead, or they may rest by prolonging pauses, since those take off all of the aforementioned types of strain.

Additionally, as previously mentioned, **longer recordings** are to be expected, due to fatigued individuals gradually lowering their articulation rates as they speak.

Ataxia in speech

Ataxia (adj.: atactic) is a neurological symptom describing a compromised ability to voluntarily control muscle movements. Outwardly, it often exhibits as a sort of perceived clumsiness, since it often affects gait, which may become visibly unsteady. This is arguably the most obvious symptom observable by a layman in the disease. It is assumed to occur in roughly 80% of all cases of MS, which makes it a potentially invaluable biomarker for early detection. (Wilkins [2017])

Acoustical components of hypothesized to be found in dysarthric speech with components of ataxia is **general articulatory decay**. The main challenge in measuring this phenomenon is the fact that atactic individuals are unlikely to mispronounce anything in particular in a *consistent* manner, so a method that takes general articulatory inconsistency is required. Prime candidates for such measurements are **vowel quality** and **duration**, due to their ease of measurement, and **realizations of the phoneme /s/**, due to its difficulty of pronunciation.

Longer recording lengths are to be expected as well as in previous cases, due to atactic individuals having a tendency to mispronounce words and correct themselves much more often, leading to longer recording times.

1.1.2 Speech in neurological disorders

In this subsection, various methods already used in the past in regards to multiple sclerosis will be described. Especially important to this thesis is Jan Rusz's work *Characteristics of motor speech phenotypes in multiple sclerosis* (Rusz et al. [2018]), which will be mentioned here as well on multiple occasions throughout this thesis.

Speech processing itself is based on signal processing. What this means is critical – it means that by using recordings of speech as data, we can extract incredible amounts of information out of them thanks to the likes of Claude Shannon and Jean-Baptiste Fourier, as well as many others. This poses a significant advantage over say data in the form of videos of subjects ambulating, because analysis of movement (and especially from video) lacks such a robust, established and relatively accessible mathematical background.

Furthermore, as mentioned in section 1.1.1, speech is physiologically incredibly complex, meaning that it can be reasonably expected that many neurological abnormalities appear therein, often before they manifest themselves elsewhere or are noticed by the afflicted individual.

We call speech afflicted like this *dysarthric*. Using tools including but not limited to spectral analysis, we can then define parameters that we deem systematically influenced by the disease we are studying, measure them in the recordings that we have acquired, and then perform statistical tests which confirm or falsify our conjectures.

Pioneering in the study of MS is a study by Gerald et al. [1987], which, while revealing interesting findings, had the drawback of being perception-based and thus insufficiently objective, in addition to a rather small sample of 23 indiviudals. The authors primarily mention deviations in vocal quality, imprecisions in consonant production and amplitude control, which are mostly objectively measurable. It did, however, delve into topics which are a lot more difficult to measure objectively, much less automatically, such as deficits in sentence construction and comprehension of logico-grammatical constructions.

Critical to the existence of this thesis and the study of MS-related dysarthria is the work of Jan Rusz and his colleagues, whose studies will be mentioned along others here briefly in chronological order.

The most relevant study this thesis is called *Characteristics of motor speech phenotypes in multiple sclerosis.* Rusz et al. [2018] Here, it was discovered that in a sample of 141 MS patients, *at least* 56% exhibited some degree of dysarthria, typically with both spastic and atactic components, as measured using objective acoustic measurements. This dysarthria typically exhibited low pitch variation, articulation problems, amplitude fluctuations and a slow articulatory rate. Because of this, this thesis attempts to cover these symptoms as uniformly as possible, as described in subsection 1.1.1. Most importantly, this study, along with Orozco-Arroyave et al. [2016] and Rusz [2018], opened up the possibility that automatic detection of MS is a feasible idea. The results of the first of these studies will be compared in-depth to the results of this thesis in subsection 3.3.2 on page 47.

Worth mentioning is also Rusz et al. [2019], where the authors attempt to find connections between specific acoustic abnormalities and MRI-based volume measurements. In other words, they try to find out how specific forms of dysarthria relate to which parts of the brain are attacked by the disease. Since this thesis conflates all MS subtypes into one group of simply positive patients, this particular study is not immediately relevant. It is, however, potentially highly in the context of the possibility of a construction of a different, more advanced algorithm and testing paradigm, which could output prognoses or discern between MS subtypes.

What is much more relevant is Rozenstoks et al. [2020], because the method described in this article it may potentially be used in conjunction with the model described here. On the details of this possibility, please refer to subsection 3.3.1 on 45. This article presents a method on how to detect syllables in repetition paradigms with more than reasonable accuracy for the purposes of temporal measurements. The study discovers that MS patients have an impaired ability to perform alternating, repeated articulatory movements, resulting in both slower articulation rates and temporal irregularities.

Lastly, Hlavnička et al. [2020] is relevant to this thesis in that it states that objectively measurable vocal tremor is only present in 8% of all MS patients. Realizing this, the decision to not try to cover vocal tremor with the parameters used in this thesis becomes obvious.

1.2 Data and software

In this section, all data and software will be properly described, mentioned, evaluated and attributed to all the incredibly kind people who have made the usage of their tools and data available, knowingly or unknowingly, for the purposes of creating this thesis and the accompanying ML model.

1.2.1 Recordings

In this thesis, two sets of recordings of various individuals reading the same text aloud were used. Various subsets of these two corpora were then used in the two experiments that comprise the practical segment of this thesis, meaning that some recordings were dropped for various reasons. Some recordings were not technically usable, sometimes critical personal data (such as age or gender) on a given individual were missing, not enough recordings were manually annotated for the purposes of a given experiment or some speakers were discarded because the data had to satisfy certain statistical criteria.

The text the subjects were reading is an excerpt from Karel Čapek's short story *Měl jsem psa a kočku.* It contains 230 syllables, takes about just under a minute to read and goes as follows:

> I na tom, že člověk si opatří psa, aby nebyl sám, je mnoho pravdy. Pes opravdu nechce být sám. Jen jednou jsem nechal Mindu o samotě v předsíni; na znamení protestu sežrala všechno, co našla, a bylo jí pak poněkud nedobře. Podruhé jsem ji zavřel do sklepa s tím výsledkem,

> že rozkousala dveře. Od té doby nezůstala sama ani po jedinou minutu. Když píši, chce, abych si s ní hrál. Když si lehnu, považuje to za znamení, že si mně smí lehnout na prsa a kousat mě do nosu. Přesně o půlnoci s ní musím provádět Velikou Hru, při níž se s velikým hlukem honíme, koušeme a kutálíme po zemi. Když se uřítí, jde si lehnout; pak si smím lehnout i já, ovšem s tou podmínkou, že nechám dveře do ložnice otevřené, aby se Mindě nestýskalo. (Čapek [1939])

It is not particularly cognitively demanding, but upon being read aloud forces one to realize all Standard Czech phonemes.

Healthy individuals

A healthy corpus of recordings of 133 female and 104 male native Czech speakers, totalling 237 individuals, was kindly provided by the Institute of Phonetics, Faculty of Arts, Charles University. This corpus had originally been recorded for the purposes of studying dysarthric speech in neurological disorders; thus, it was not recorded in a studio. The participants were given a small fee for their efforts and were recorded in a quiet furnished room with a high-quality condenser microphone at the Institute of Phonetics. The median age of the subjects at the time of recording is 55 years with a standard deviation of about 19.

Affected individuals

Two sets of recordings of MS patients were kindly provided by the Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague. These were recorded using a professional handheld voice recorder set to uncompressed mode (48 kHz sampling rate and 16 bit resolution) in a quiet room equipped with a *Bayerdynamic Opus 55* head-mounted microphone.

The first set contains 26 female and 23 male native Czech speakers, with an age median of 42 years and a standard deviation of about 13. These recordings had previously all been labelled as perceptually dysarthric by trained speech therapists and thus consist a more salient, but also biased (for the purposes of recognition and ML model evaluation) sample of MS speakers. The second set contains 17 females and 13 males, with an age median of 42 years and a standard deviation of about 10. Unlike the previous set of recordings, these were selected at random with no regard as to their dysarthric status as evaluated by ear.

Thus, 65 females and 36 males form the backbone MS-afflicted corpus of this thesis, median age of 42 with a standard deviation of about 12.

1.2.2 Software

In this subsection, all software tools used in this thesis will be properly mentioned and cited.

The TextGrids and some preliminary perceptual and spectrographic analyses were made with the help of Praat (Boersma and Weenink [2009]), a free computer software package intended for use in phonetics.

The R programming language (R Core Team [2019]) and its integrated development environment (IDE), RStudio (RStudio Team [2015]), were used for data wrangling, statistical analysis and some signal processing. R packages used include *tuneR* (Ligges et al. [2018]), *rPraat* (Bořil and Skarnitzl [2016]), *tidyverse* (Wickham [2017]) and *stringr* (Wickham [2019]).

The Python programming language (Van Rossum and Drake [2009]) and its IDE, PyCharm (JetBrains) were used for the construction of an ML algorithm along with its *scikit-learn* (Pedregosa et al. [2011]) and *pandas* (McKinney et al. [2010]) packages.

Unless stated otherwise, the *default* settings and parameters of all functions of all software used in these thesis were used.

1.2.3 Annotation

As mentioned in 1, apart from the recordings, text files that delimit where phonetically relevant phenomena begin and end in a given recording are necessary for proper statistical analysis of some parameters. The process of generating these text files is called annotation, as the name of this subsection suggestis, and have been dubbed TextGrids by Paul Boersma, the author of Praat.

This can be done in two ways, either manually, which requires some amount of workforce at least semi-competent in phonetics, or automatically, which at least *theoretically* requires nothing but computing power.

Manual annotation

Manual annotation requires a considerable amount of tedious and time-consuming labour, which is typically provided by students of phonetics and/or linguistics. Due to them being human, this typically requires some form of compensation, be it academical or financial in nature.

Luckily, all recordings from the healthy corpus already came with manually created TextGrids, so for the purposes of this thesis, nothing in particular had to be done regarding annotation.

The MS-afflicted recordings, however, came with no TextGrids to accompany them, which meant these had to be made by hand. Several students including the author of this thesis were thus rallied to provide their services as annotators for either a small amount of money, credits, or both.

These had originally been being made from scratch until it was discovered that it was in fact practical to use automatically generated TextGrids as templates to save time. As no difference in quality was found between the results of the two approaches, both of these are considered to be manually annotated for the purposes of this thesis.

All TextGrids used in the thesis were annotated in a consistent way, using explicit guidelines described in *Fonetická segmentace hlásek.* (Machač and Skarnitzl [2010])

Similar files containing the base frequency (f_0) within given time windows called PitchTiers of the recordings had to be created using Praat. Since it measures vocal fold pulses (the acoustic correlate of f_0) using autocorrelation, meaning it correlates slices of the signal with itself. Because vocal folds often pulsate irregularly (especially if the associated muscles are impaired in some way), this leads to so-called octave jumps, which occur when the algorithm correlates the slices not by period, or by half-period or double-period. These need to be corrected by hand in a process that is surprisingly difficult to automate.

In the case of manually annotated data, this was simply done by whoever annotated the recording. In the case of automatically annotated data, this was not done at all.

Automatic annotation

As already mentioned, a tool exists that allows for the automatic creation of TextGrids, called Prague Labeller. It was created by a doctoral student from the Czech Technical University, as commissioned by the Institute of Phonetics. Pollák et al. [2007]

It is a ML-based algorithm implemented in Perl and the *HTK Toolbox* (and thus exhibiting certain peculiarities) that takes TextGrids with the orthographic transcript of the recording as input, and outputs phonetically annotated TextGrids. This is why it was mentioned earlier that *theoretically*, this requires only computing power – there *is* a non-zero amount of busywork involved with the generation of these. However, all of this busywork is definitely *automatable*. This point becomes important in subsection 3.3.1 on page 45.

It is critical to point that out in order for the tool to work properly, the orthographic transcript of the recording must exist. This is the main reason why the potential MS detection tool would probably need recordings of the same text.

In the case of automatically annotated data, octave jump correction was not performed at all.

1.3 Assumptions

In this section, all statistical and non-statistical assumptions that are required for all of the conclusions presented by this thesis to be logically valid will be presented.

The first assumption regularly relied on throughout this thesis is the presupposition that given an extreme enough sample from a certain distribution under a null hypothesis, the null hypothesis can be safely rejected. The metric used to assess the extremeness of a sample will be the p value, with its arbitrary threshold being 0.05, as convention dictates. Furthermore, *no data* in this thesis are assumed to be normally distributed. The reasons for this decision are further elaborated upon in chapter 3 on page 24.

Next, it is assumed that multiple sclerosis influences speech acoustically in a way that can be reasonably modelled as linear. To be more exact, it is assumed that the acoustic abnormalities associated with MS do not generally discontinuously appear or disappear at any point as the disease develops, but rather appear in a gradual way as the patient's condition worsens and get more and more salient along with the disease's progress.

It is also assumed that the success rate and general performance of the machine learning models presented in subsections 3.1.2 and 3.2.1 on pages 39 and 40 respectively do not owe their success to random chance. This *can* be made arbitrarily improbable by repeated testing, but is impossible to *completely* rule out using a finite amount of tests. Details on the methods used to mitigate this possibility are to be in found in the aforementioned subsection of this thesis.

Regarding the data used in this work, the acoustical differences between the two sets of recordings used are assumed to be made negligible by the parametrization methods described in chapter 2. These differences stem from the fact that the recordings were made using different equipment in different acoustical conditions. This presupposition mainly rests upon the fact that the parametrization methods used in this thesis are robust against such differences.

It is also assumed that these models will be reasonably reliable in predicting recordings of multiple sclerosis cases which are generally significantly less developed (and thus less acoustically abnormal) than the ones in the trainings datasets. In other words, it is assumed that the model's relatively high reliability in predicting strongly progressed MS cases (if achieved) can be extrapolated to reasonable reliability in predicting weakly progressed MS cases. This is by far the weakest presupposition, but practicality dictates that it is logically unavoidable, since the entire point of these models is to catch multiple sclerosis cases *before* they can be reasonably detected by other means. If these cases cannot be detected by conventional means, then there is no way to reliably produce data on them, and thus there is no way to train a model using them. Thus any such model logically has to be trained on more developed cases than it is designed to detect, at least in the case of detection paradigms similar to the one described in subsection 3.3.1 on page 45.

Next, it is assumed that all individuals labelled as healthy do not suffer from any neurological disorders that might interfere with statistical tests and ML models. This is assumed because none of these subjects reported any such difficulties.

Lastly, it is assumed that the acoustic speech parameters most affected by any slowly progressing neurological disease will be the ones that are relatively difficult to detect by (untrained) ear alone. This is because it can reasonably be assumed that the human brain, and by extension the entire nervous system, continually adapts to everything including neurological difficulties. As the functionality of the nervous system slowly decays, the brain can be presumed to focus on trying to maintain speech functions more relevant to communication at the expense of those less relevant. It is critical to note that this assumption, unlike the others, is *not* related to the *validity* of any of the conclusions presented in this work, but is rather the driving force behind the *choice* of the presented parameters.

2. Method

The method of how all data used in this thesis were evaluated and tested will presented in this chapter for the purposes of reproducibility.

2.1 Hand-annotated data

The data used for this section of the thesis were obtained using TextGrids that were manually annotated, as described in 1.2.3.

This means that a *subset* of the individuals from the afflicted corpus were used. 16 recordings of individuals (13 females and 3 males) with MS were used here, since those are the TextGrids that we had the time to annotate. All of these were taken from the dataset where dysarthria was confirmed by a speech therapist. The median of their age was 41 with a standard deviation of about 7.

By contrast, 145 individuals (110 females and 35 females) from the healthy data were used. The ones left out from the larger dataset mentioned in 1.2.1 were removed for two reasons:

- Some of the recordings were incomplete, corrupted or were technologically unusable for other reasons,
- some males were removed so that the male-to-female ratios would be roughly equal in both the MS and healthy datasets.

These individuals had a median age of 59 with a standard deviation of about 18.

What follows is a list of what methods were employed to objectively measure the various hypothesized effects of multiple sclerosis on speech, as described in 1.1.1 on page 5.

2.1.1 Parameters mostly related to spasticity

The methods that have been used in this thesis to measure the effect of spastic muscles on speech are described here. Brief descriptions of the computation method and the meaning of the cumulative slope index (CSI) are also included.

Glottal stop rate

It was hypothesized in 1.1.1 that individuals with spastic muscles would have more difficulty keeping control of the size of their glottis. Assuming that this would manifest as their closing or constricting the glottis when inappropriate, instances of this happening were simply measured by counting the occurrences of the glottal stop [ʔ] and dividing them by the recording's durations.

This rather simple and elegant measurement carries a significant drawback by not being practical using automatic annotation, for reasons described in section 3.2 on page 40.

Cumulative Slope Index of intensity

Similarly, spastic individuals can be presumed to have poor control of their breathing muscles. Because spasticity, at least according to the ad-hoc definition presented in this thesis (see 1.1.1) involves both **involuntary contraction** and **rigidity**, the effect on the way this can have on the *amplitude* development throughout a given person's recording can be quite diverse. The absolute value of intensity across the whole recording is useless as a parameter, because measured sound intensity decreases with the square of the distance from the source according to the inverse-square law described by the equation:

$$I \propto \frac{1}{r^2}$$

where I is the measured sound intensity, r is the distance from the source, and $\propto$ denotes proportionality.

Because the r is squared, measured intensity is much more sensitive to distance from the source than the intensity of the source itself. Thus, if has their microphone further away from their face than another subject, the difference in measured intensity between the two subjects has much more to do with that and much less to do with the behaviour of their breathing muscles or anything relevant.

A much better choice is, then, to calculate the *net change* of the measured intensity across the recording, as spastic activity in muscles presumably introduces spikes to the intensity contour, while stiffness could lead to an unusually flat intensity curve. CSI is a good formula to apply in these cases. It is the sum of the absolute values of differences between each point:

$$CSI = \sum_{n}^{N-1} |x[n+1] - n[1]]|$$

where x is the vector of values (be it intensity at certain points in time or something else), n represents the index of a certain value in that vector and N is the length of vector x, as introduced by Volín et al. [2017].

Because the number of syllables is constant (or *should be* in unaffected speech), whenever CSI will be used an normalized in this thesis, it will be normalized by the duration of the recording. CSI was implemented in R using the following code:

```
csi <- function(x) {
    sum(abs(diff(x)))
}
```

Average unvoiced stop duration

As already mentioned, MS patients often have difficulty performing voluntary antagonistic movements in quick succession.

Prime candidates for measuring this phenomenon are stop consonants, because they require a full oral aperture to be made and then quickly reopened (see 1.1.1 on page 7, which is a good example of oral diadochokinesis.

To measure the effect of MS on this, the durations of all detected unvoiced stop consonants were measured and the mean of those was taken for each individual. The advantage of the mean in this case is its sensitivity to outliers; this should properly capture a stronger statistical tendency to overextend the timing of oral apertures in individuals with MS, if it exists.

2.1.2 Parameters mostly related to fatigue

The methods that have been used in this thesis to measure the effect of various kinds of fatigue on speech are described here. Recording duration is somewhat arbitrarily included here, as it has been deemed that duration the relationship between it and fatigue is the most obvious of the three symptom clusters.

Please note that none of these parameters need take larger values for affected indiviuals. Fatigue may result in frustration in subjects, leading to hastiness.

Silence percentage

It can be assumed that speakers who feel tired may try and find moments of rest whenever possible. Since not speaking is less tiring on just about every level than speaking, it stands to reason that fatigued individuals will pause more often, make longer pauses, or both.

All three of these possibilities can be measured by taking the sum of silent seconds in the recording and dividing them by the duration of the recording, resulting in the percentage of the recording which constitutes silence.

Despite the fact that all recordings have been trimmed, the first and last silent segment of each recording (before and after each speaker reads their text out loud) is not included in this metric.

Vowel percentage

Similarly to the previous parameter, vowels can present a short moment of articulatory respite for a speaker, since no precise movements have to be made and vowels are generally less straining. For this reason, this parameter might also be influenced by patients who have problems with spasticity. As mentioned in chapter 1 on 4, precisely mapping symptom clusters to individuals measurements does not fall into the scope of this thesis.

Analogically, the total time the speaker spends producing vowels will be added up and divided by the recording duration.

Recording duration

Fatigued speakers may take longer to read a text out loud. This is for a variety of reasons – they might spend more time being silent or producing vowels, as described earlier, or they might simply slow down their articulation rate, spend more time fiddling around, they might have difficulty reading certain words if their vision is affected, or perhaps problems processing the unfamiliar name *Minda* and so on.

For this reason, this is a seemingly simple, but in reality deep and complex parameter.

CSI of f_0

The principle here is described here is the formula described in 2.1.1. The purpose of this is to measure the intonation range of the individual across the whole recording, since it can be assumed that fatigue may distract the reader from intonating properly, focusing more on (in their view) important aspects of the exercise, such as not misreading words. Unlike in the case of intensity, base frequency (or f_0) absolute values *are* meaningful in and of themselves, though we are still more interested in the rate of change across the whole recording.

The CSI of f_0 the whole recording is recorded for each individual, having been converted from Hertz to semitones to account for differences between males and females.

Quantile difference of f_0

What we are measuring with the quantile difference is similar to the Cumulative Slope Index of f_0. The reason for this seemingly reduntant measurement is the fact that CSI measures total rate of change across time, while quantile difference measures the *span* of all values.

The 10th quantile of the entire PitchTier is taken and subtracted from the 90th quantile after the PitchTier is converted from Hertz to semitones.

2.1.3 Parameters mostly related to ataxia

The methods that have been used to measure the effect of various kinds of ataxia on speech are described here. Speech is an incredibly motorically complex activity and thus atactic symptoms should presumably reliably produce acoustic phenomena.

Formant value standard deviation

Formants can be described as local maxima in the spectrum of a speech signal in a given interval. The first two, first formant (f_1) and second formant (f_2), correlate to jaw angle and tongue blade position respectively (Skarnitzl and Volín [2012]), while third formant (f_3) correlates mainly with vocal fold laxness and tightness, which makes f_3 also relevant to spasticity. (Sawyer [2013]) However, f_3 may sometimes also correlate with tongue tip positioning in certain contexts. (Monahan and Idsardi [2010])

Because atactic individuals presumably exhibit worsened with speech organ precision, it can be presumed that the range that their formant values take across all vowels will be larger than in healthy subjects. Thus, the standard deviations of f_1, f_2 and f_3 will be taken, with the express presupposition that these value will be larger for affected individuals.

Sibilant spectral centroid standard deviation

Sibilants are valuable for articulatory precision measurement because they present a challenging sound to produce. To produce a sibilant, one must create a groove in their tongue while maintaining proper tongue positioning to produce a sharp sound. If either of these is incorrect, a sound of the incorrect perceptual **brightness** *for the given language* is produced. (Reidy [2016]) The acoustic correlate of **brightness** (and thus tongue positioning and groove depth) is called the spectral centroid and is given by the following formula:

$$Centroid = \frac{\sum_{n=0}^{N-1} f[n]x[n]]}{\sum_{n=0}^{N-1} x[1]]}$$

where $x[n]$ is the magnitude of bin number n, and $f[n]$ represents the central frequency of the spectral bin, as described by Grey and Gordon [1978].

Praat was not used, the formula was instead implemented in R and was performed on the raw *.wav* recordings using the *tuneR* package using the following code:

```
rfft <- function(x) {
  spec <- fft(x)
  real_spec <- spec[1:(round((length(spektrum)/2))+1)]
  return(real_spec)
}

fft_freq_real <- function(n, fs) {
  vec <- 0:length(n)
  vec <- vec*(fs/length(n))
  return(vec[1:(round((length(n)/2))+1)])
}

spektralni_teziste <- function(x, fs) {
  magnitudes <- abs(rfft(x))
  length <- length(x)
  freqs <- fft_freq_real(x, fs)
  return(sum(magnitudes*freqs) / sum(magnitudes))
}
```

where rfft() is a function that extracts the positive frequency terms of a given spectrum, fft_freq_real() determines the central frequencies of its spectral bins and spektralni_teziste() is simply an implementation of the aforementioned formula.

Because of the assumption in section 1.3 on page 15 regarding linguistically relevant phenomena, it can be assumed that the central value of the sibilants remains unchanged in afflicted subjects, otherwise they would have a perceptible lisp (or similar), which they do not, at least according to a random informal listening test performed by the author. Thus, similarly to the previous measurement method, the *standard deviation* of the spectral centroids of their [s] consonants was taken.

CSI of vowel duration

Atactic individuals may have trouble keeping rhythm in speech, because proper rhythm requires a complex interplay of speech organs such as breathing muscles. Apart from that, as already described in 2.1.2, fatigue may also contribute to abnormal rhythm.

To measure this, CSI was simply applied to individual vowel durations across the individuals' recordings.

3. Results and discussion

In this section, statistical findings of all the aforementioned measurements will be described and discussed as to their potential scientific impact and practical application.

3.1 Results

All hypotheses were using the Kolmogor-Smirnov test due to the fact normality was not proved for the afflicted dataset in any parameter. This was tested visually using a Q-Q plot of the dataset and a randomly generated normal distribution of the same parameters and the one-sample Kolmogorov-Smirnoff test.

Unless stated otherwise, the *two-tailed* alternative hypothesis was used. This was because it is quite difficult to predict how exactly the disease will affect a certain parameter, as stated numerous times in chapter 2.

Please note that in figures, "**MS**" stands for "multiple sclerosis" and "**H**" stands for "healthy".

3.1.1 Manually annotated data

The reader may notice quite a significant age disparity between the studied group and the healthy control group, as described in section 2.1. There had originally been an intention to balance things out by removing some of the older individuals from the control group. However, this disparity was later deemed to serve as a failsafe to make sure that the physiologically more aged voice of older outliers did not skew the statistical results in favour of the hypotheses proposed here.

Recording duration

Recording duration was found to be nearly statistically significant under the alternative hypothesis that the cumulative distribution function of the afflicted individuals lies below that of the healthy ones, with a p value $= 0.056$. The reasoning for the decision is the simple assumption that afflicted individuals take longer to read their text for various reasons, as described in chapter 2. The turning point of this measurements might be the aforementioned age disparity.

As we can see in Figure 3.1, the medians do not differ very much at all. It is however clear that there are many more outliers with longer recordings, resulting in a much more skewed distribution on part of the afflicted. It is thus

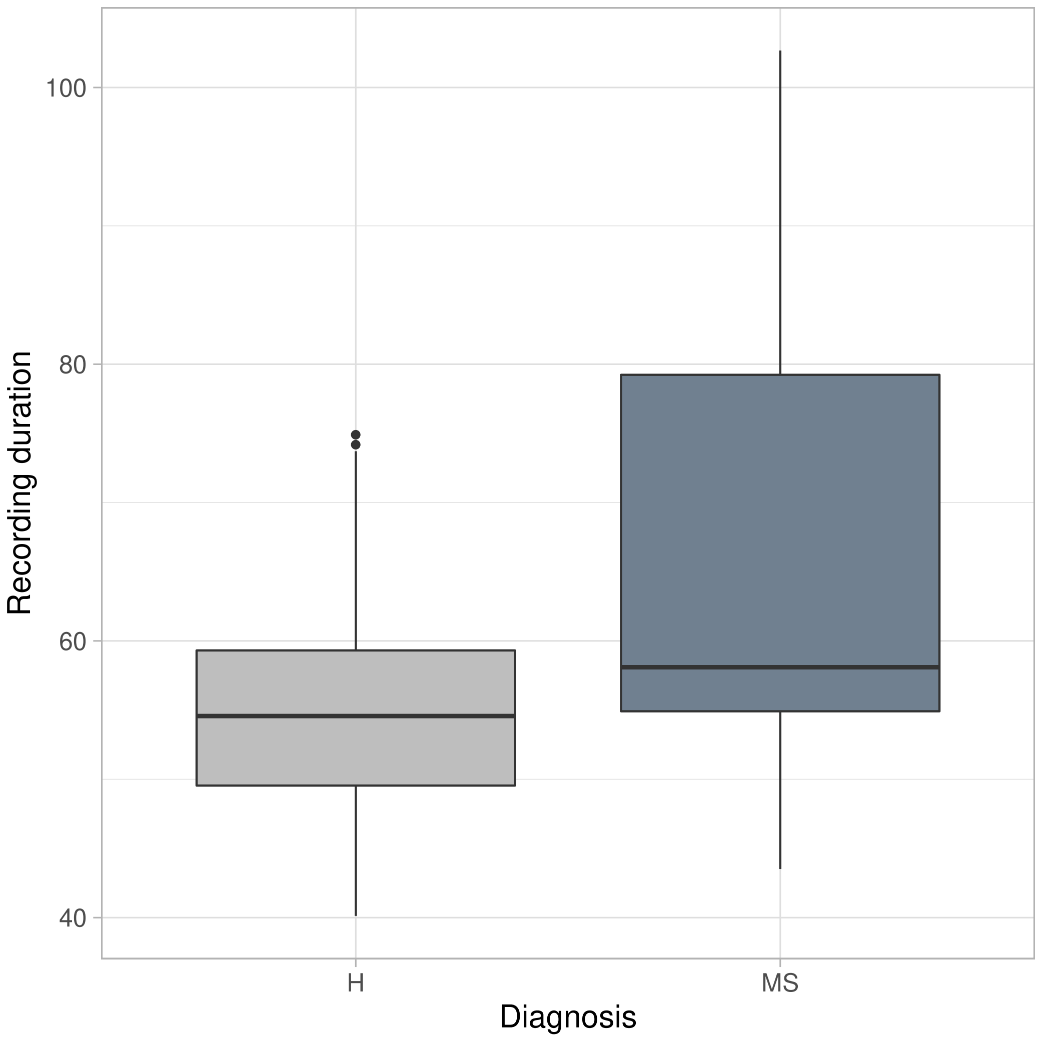

Figure 3.1: Boxplot of the recording durations of afflicted and healthy individuals.

not the difference in medians the carries the near significance, but rather the entire skewness of the distribution.

Silence percentage

Silence duration was found to be moderately significant with a p value $= 0.02$ under the two-sided alternative hypothesis.

As seen in Figure 3.2, this significance comes from two main factors. One, the standard deviation of the MS is larger, and two, the group contains two massive outliers, with one of whom spends nearly half of the reading being silent.

Vowel percentage

Vowel percentage was not found significant at all with a p value $= 0.74$ under the two-sided alternative hypothesis. This is surprising due to the fact that

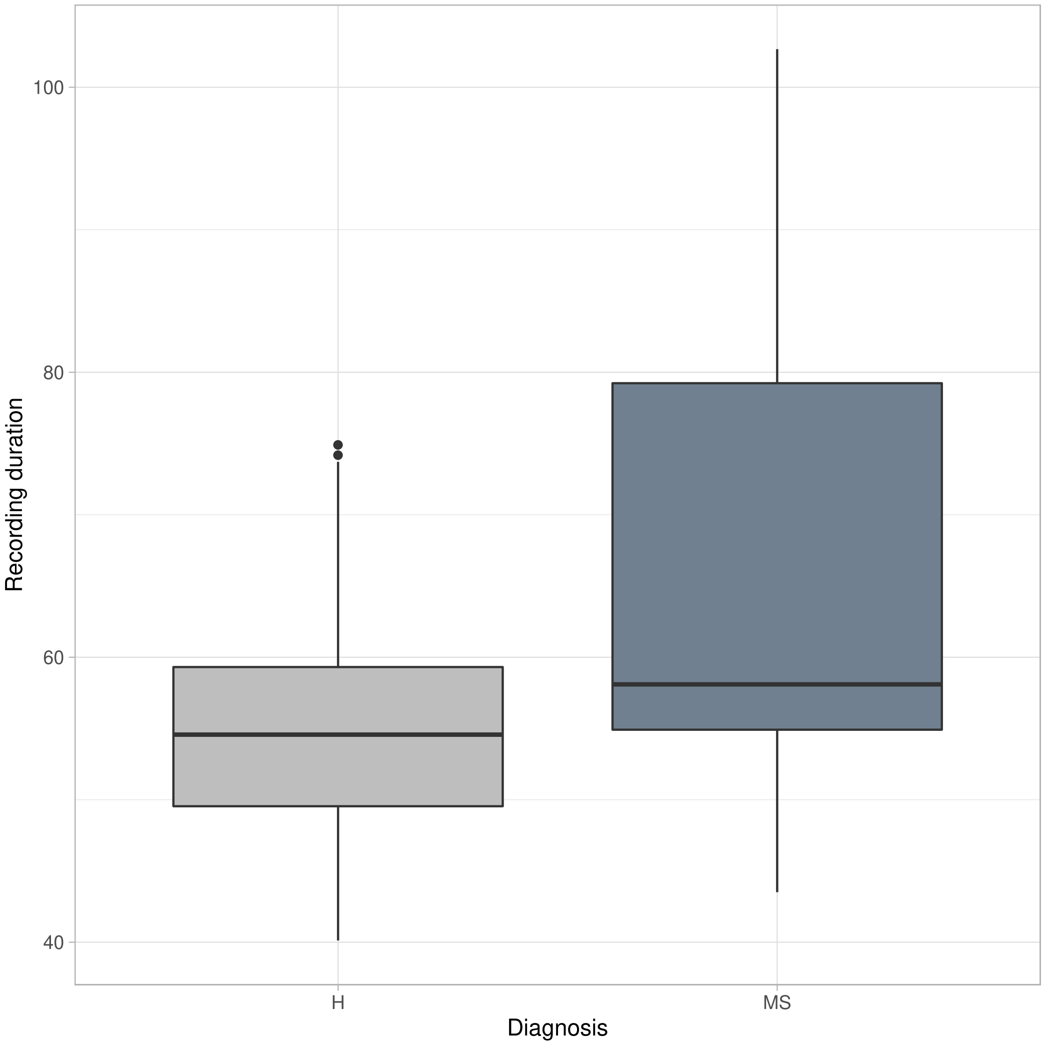

Figure 3.2: Boxplot of the silence percentages of afflicted and healthy individuals.

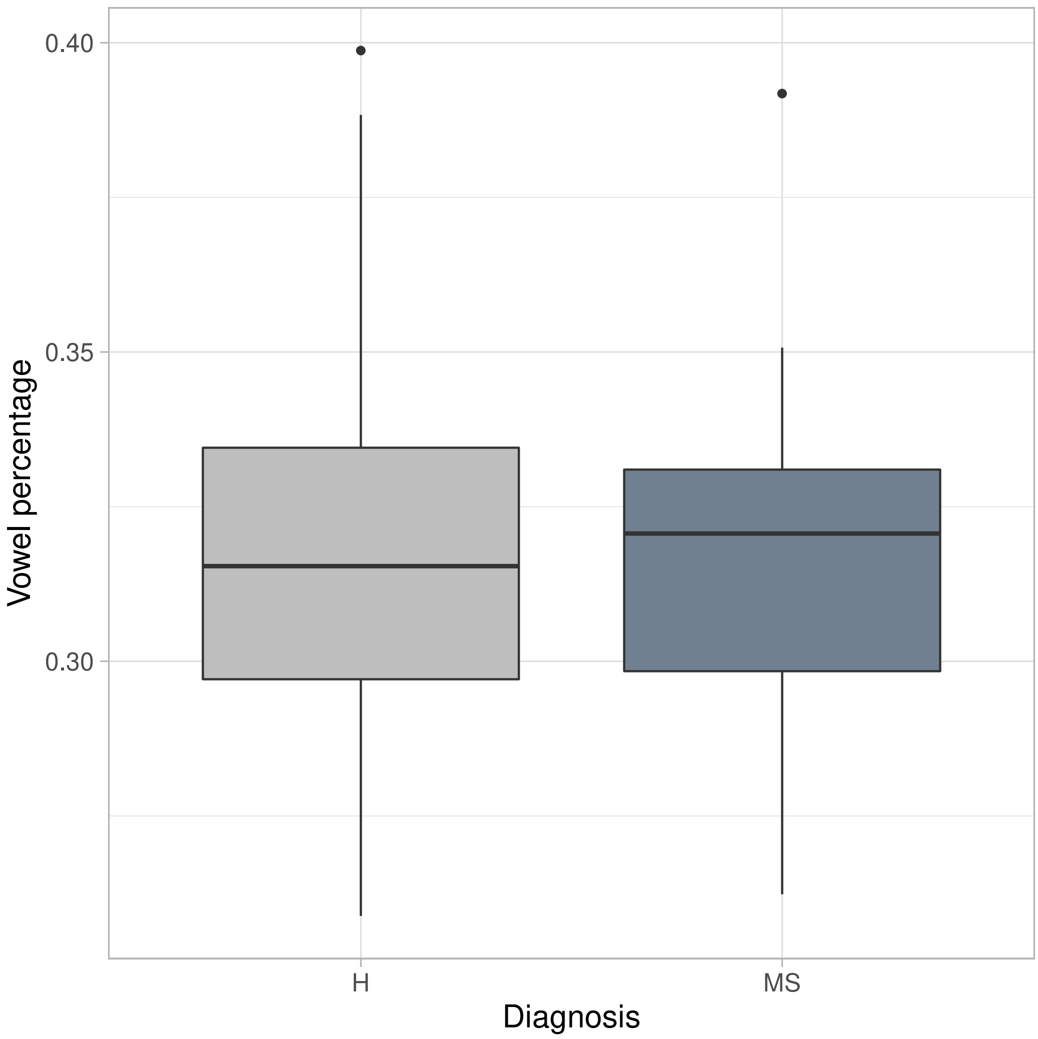

Figure 3.3: Boxplot of the vowel percentages of afflicted and healthy individuals.

informally, vowel percentage has recently become a sort of joke among researchers at the Institute of Phonetics for being typically *incredibly* reliable in predicting speech difficulties of essentially any kind ranging from voice aging to inebriation.

A brief look at Figure 3.3 reveals that the respective boxplots look very similar, thus the high p value.

CSI of vowel duration

CSI of vowel duration, a measurement of rhythm, was found to be relatively strongly significant under the two-sided alternative hypothesis with a p value = 0.005.

Looking at Figure 3.4, it is fairly obvious as to why this is. The distributions have exhibit a significant difference in their medians, along with a substantial difference in standard deviation. This suggests that not only do vowel durations

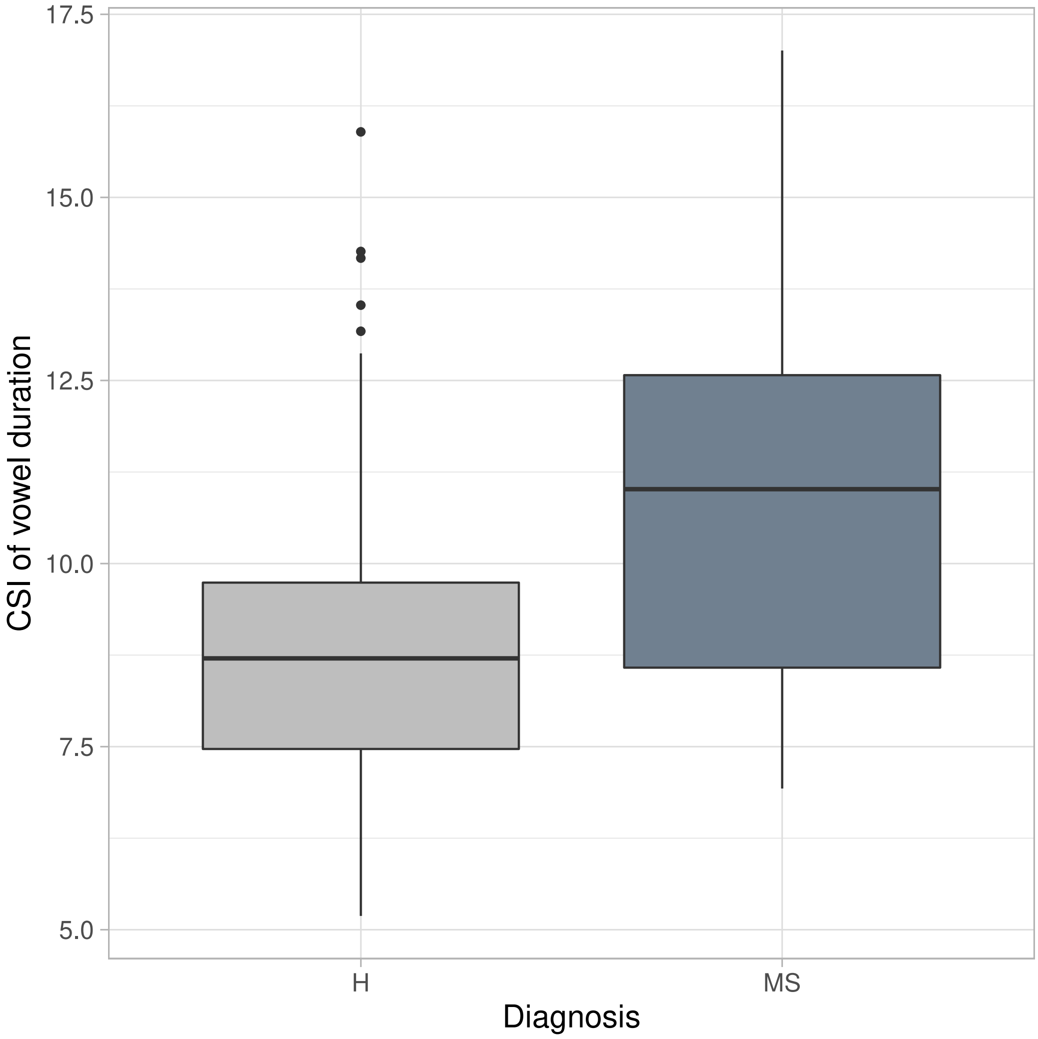

Figure 3.4: Boxplot of the CSIs of vowel durations of afflicted and healthy individuals.

fluctuate more in afflicted individuals, but also that these fluctuations vary individually. Not uninteresting are the several outliers in the top part of the healthy group – it illustrates the high individual variability that we see in phonetics even among healthy individuals.

CSI of f_0

CSI of f_0 normalized by recording duration was not found to be significant under the two-sided hypothesis with a p value = 0.2284. This is surprising, since Rusz et al. [2018] has found monopitch to be one of the manifestations of MS in speech.

As Figure 3.5 shows, we can assess that the two boxplots look fairly similar, though a difference can be seen. What sets the figure apart from the other more is the relative symmetricity of the MS group, though some skewness can be

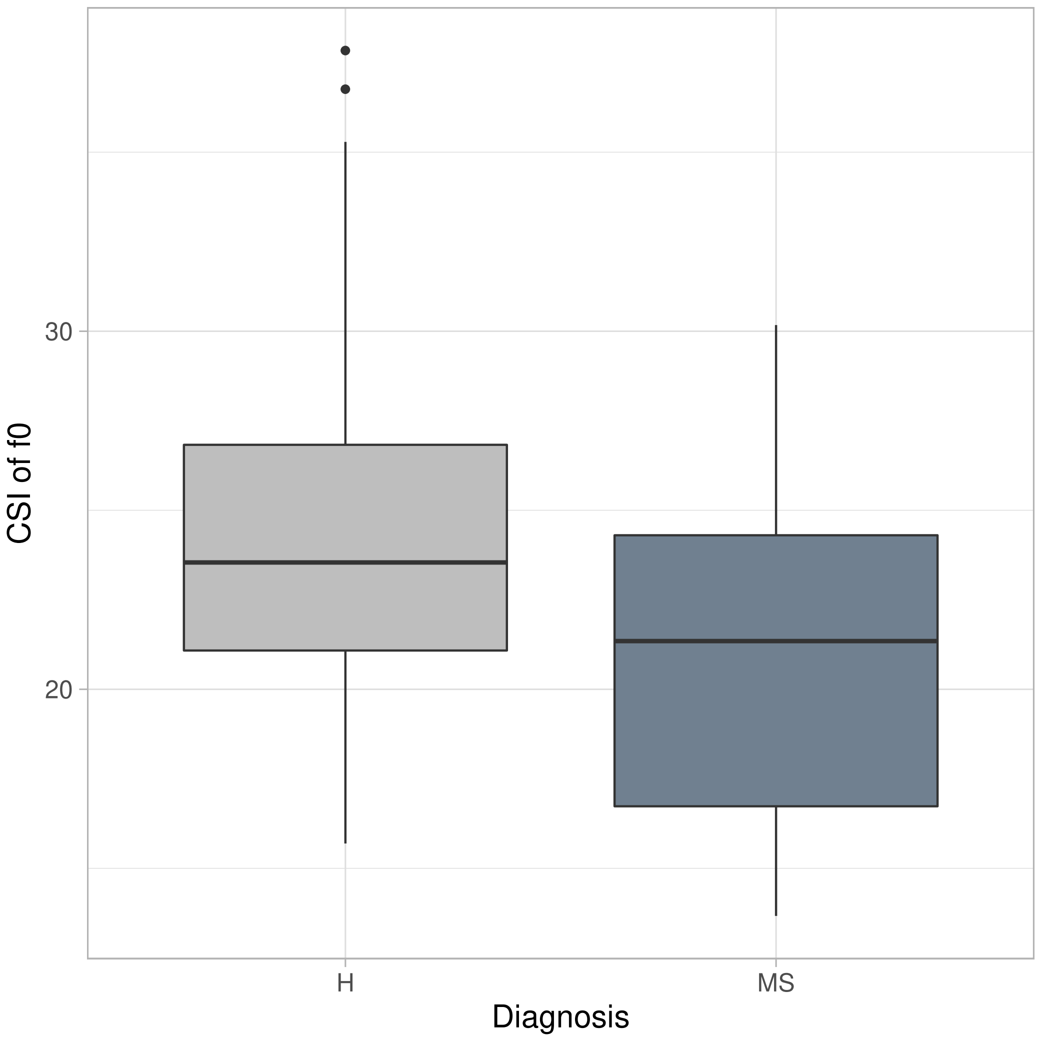

Figure 3.5: Boxplot of the CSIs of afflicted and healthy individuals.

recognized (this may be attributed to small sample size, though). Thus, it can be assumed that this parameter is of non-negligible importance for the detection of MS

Glottal stop rate

Glottal stop rate was found to be slightly significant under the two-sided hypothesis with a p value = 0.038. A caveat has to be expressed here – some of the MS group annotators were expressly asked to specifically pay attention to glottal stops, so the amount may be overestimated in favour of the afflicted glottalizing. Nothing of the sort was explicitly asked of the *annotators* of the healthy dataset, although they were, of course, tasked to annotate glottal stops as well – they just were not told to pay special attention to them. A short informal check revealed no sloppiness regarding glottal stop annotation in the healthy TextGrids, but this

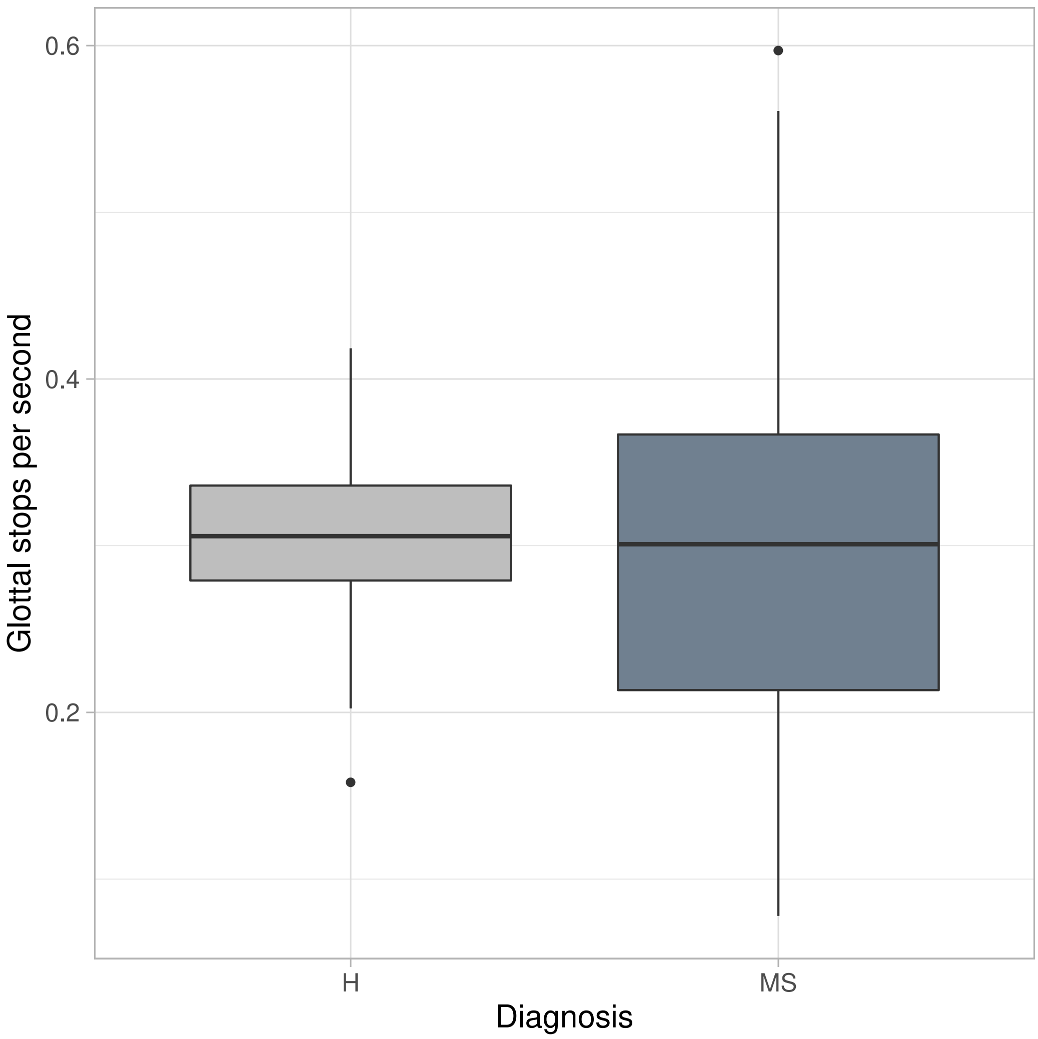

Figure 3.6: Boxplot of glottal stops per second as realized by afflicted and healthy individuals.

was not investigated thoroughly.

Please refer to Figure 3.6 to reveal a surprising fact – contrary to the hypothesis mentioned in chapter 2, on average, MS patients do not glottalize any *less* – they just have a much larger standard deviation, suggesting that they have as much trouble holding the glottis closed when appropriate as they have holding it open.

Quantile difference of f_0

The quantile difference of f_0 was found to be slightly significant under the two-sided hypothesis with a p value of 0.02. This is surprising, as this measurement is close related to 3.1.1. This means that while the relative rate of change of f_0 is generally similar in both healthy and afflicted speakers, their intonation span

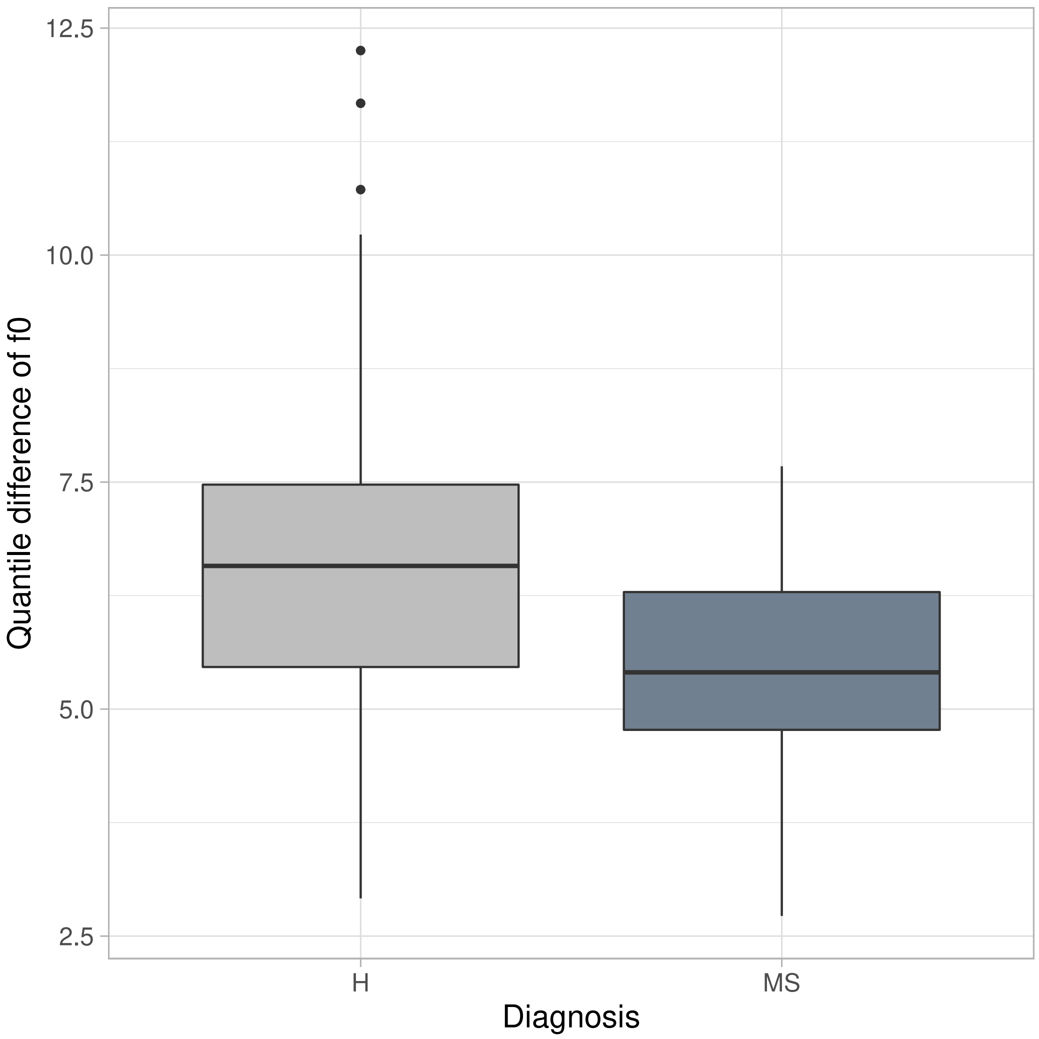

Figure 3.7: Boxplot of quantile differences of f0 of afflicted and healthy individuals.

differs significantly.

As posited in Rusz et al. [2018], we can see from Figure 3.7 that MS speakers do indeed exhibit monopitch, as basically every parameter of their sample is smaller than those of healthy speakers.

Average unvoiced stop duration

Average unvoiced stop duration exhibits strong significance under the two-sided hypothesis: p value $= 0.005$.

Visible from Figure 3.8 is the general tendency of MS patients to drag out their unvoiced stops in accordance with the hypothesis posited in subsection 2.1.1. The outliers visible in the upper part of the healthy group are important – they again highlight the importance of using a large amount of parameters when au-

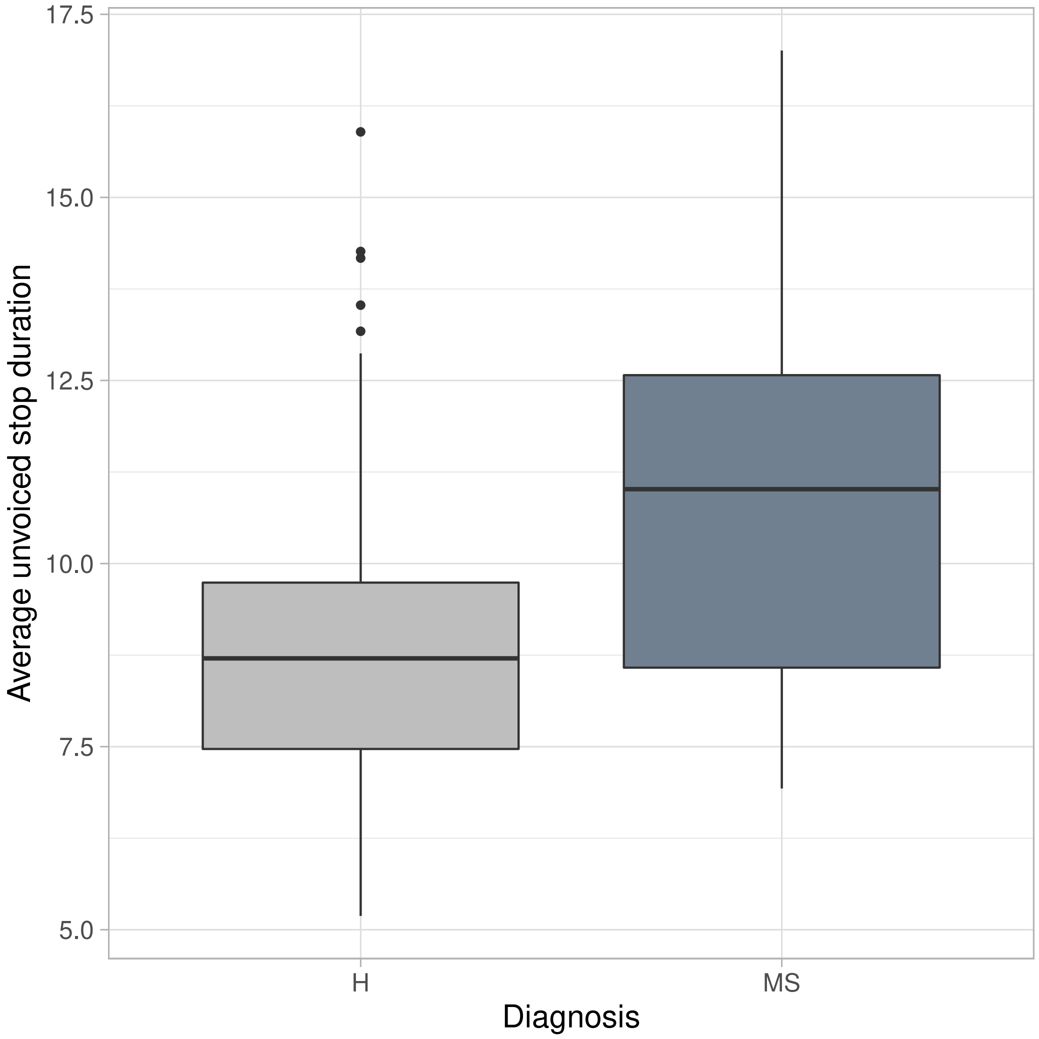

Figure 3.8: Boxplot of average unvoiced stop durations of afflicted and healthy individuals.

tomatically detecting MS.

CSI of intensity

CSI of intensity has failed to yield a significant result under the two-sided hypothesis with a p value of 0.38.

Despite this, there are two clear presumably spastic (according to the hypothesis in subsection 2.1.1 on page 17) inviduals in the top part of the MS group graph in Figure 3.9. These individuals illustrate the need to have diverse parameters in this ML model even when they are not significant, because they can serve as an auxiliary parameters for discrimination.

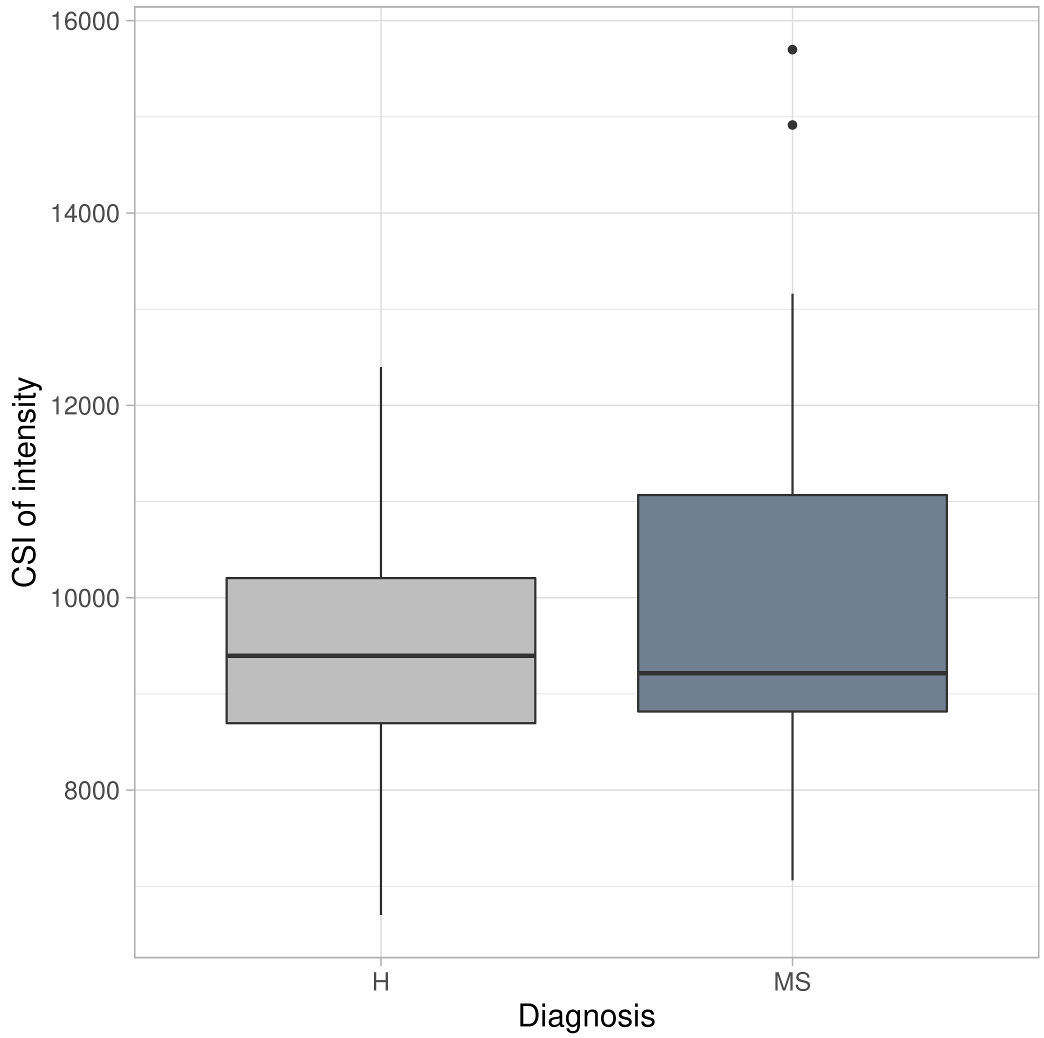

Figure 3.9: Boxplot of the CSI of the intensity of afflicted and healthy individuals.

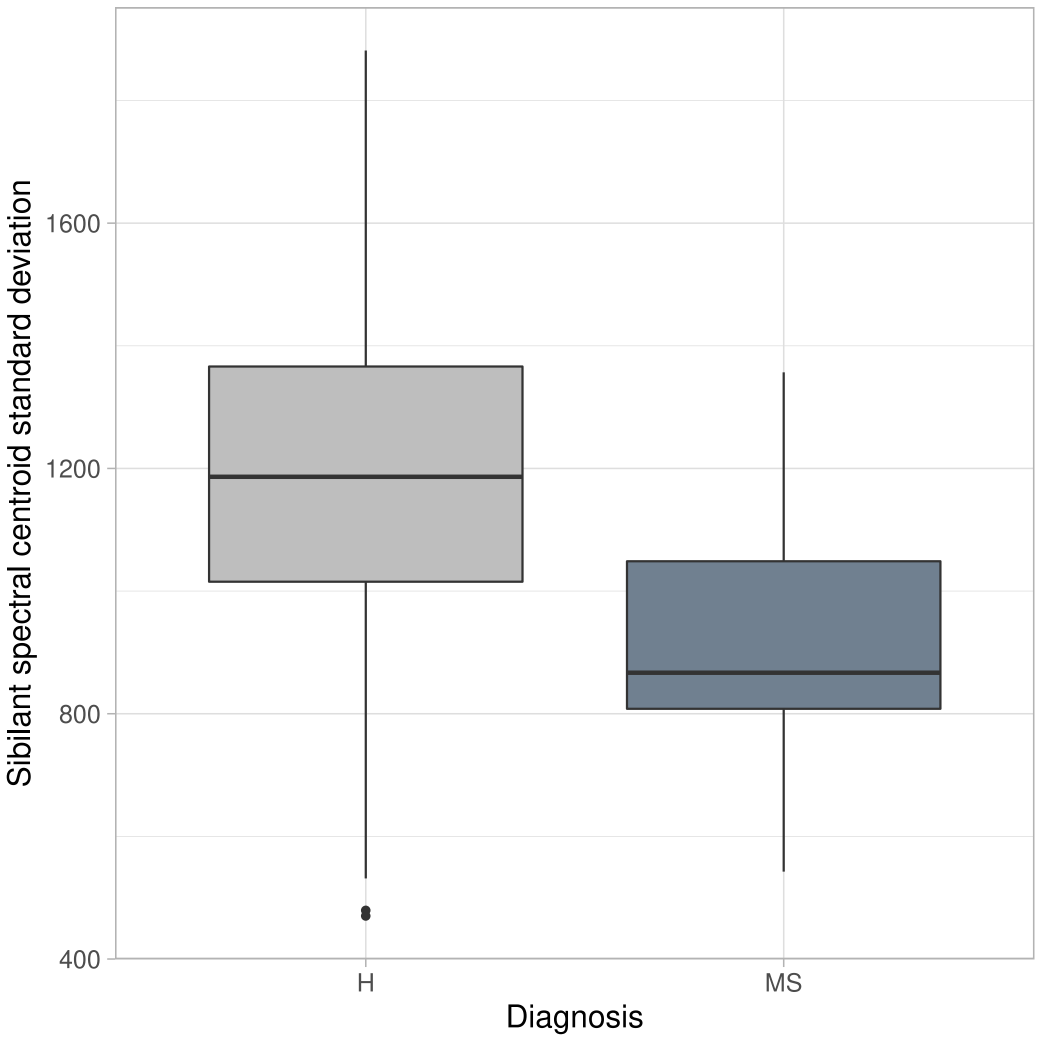

Figure 3.10: Boxplot of the CSI of the standard deviation of the spectral centroid of [s] of afflicted and healthy individuals.

Sibilant spectral centroid standard deviation

The standard deviation of the spectral centroid of /s/ realizations turned out to be strongly significant with a *p* value of 0.0003 under the two-sided hypothesis. This is surprisingly significant, despite the fact that both spastic and atactic individuals can be presumed to struggle with sibilants due to their difficulty of pronunciation.

Not following the prediction in subsection 2.1.1, the afflicted individuals seem to have a smaller median than the healthy ones according to Figure 3.10. This seems to suggest that the [s] phone is much more influenced by spasticity, because the lower median suggests a smaller range of motion of the tongue.

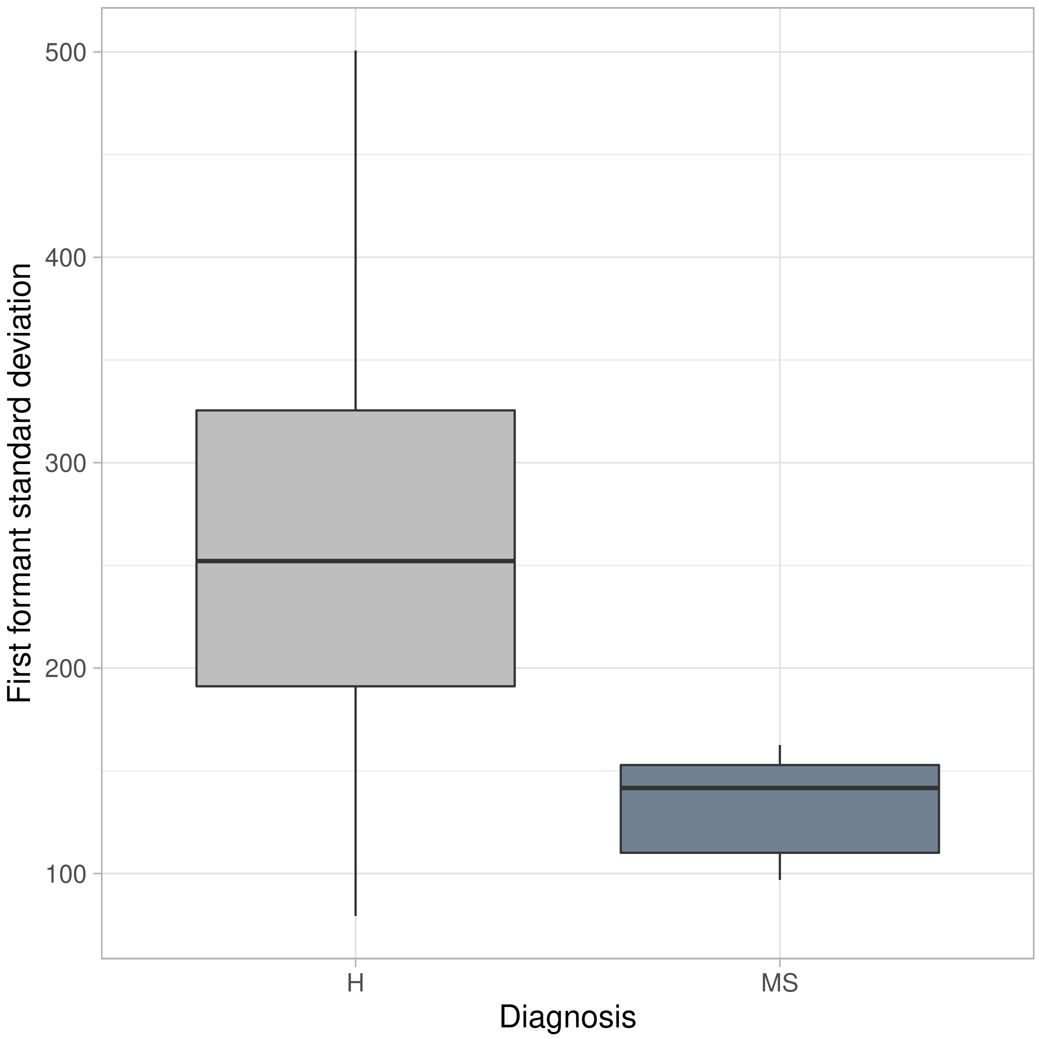

Figure 3.11: Boxplot of the standard deviation of f_1 across all vowels of afflicted and healthy individuals.

Formant standard deviations

The standard deviations of the of formants f_1, f_2 and f_3 of all vowels combined turned out to be statistically significant, with f_1 being extremely significant with a p value of 1.4×10^{-11}, f_2 less so with a p value of 0.01 and f_3 being also extremely significant with a p value of 9.9×10^{-8}.

All of these can be attributed to impaired movement of the tongue and jaw. Moving the jaw especially requires a relatively high amount of kinetic energy (due to the organ's mass), which is difficult in afflicted individuals to do, which can clearly be seen in how the two groups almost do not overlap at all in Figure 3.11.

Contrary to that, it seems that the tongue is easier to move in comparison, which makes sense due to its smaller mass and more ease of movement. Despite this, a significant difference can be seen in the formant value standard deviation

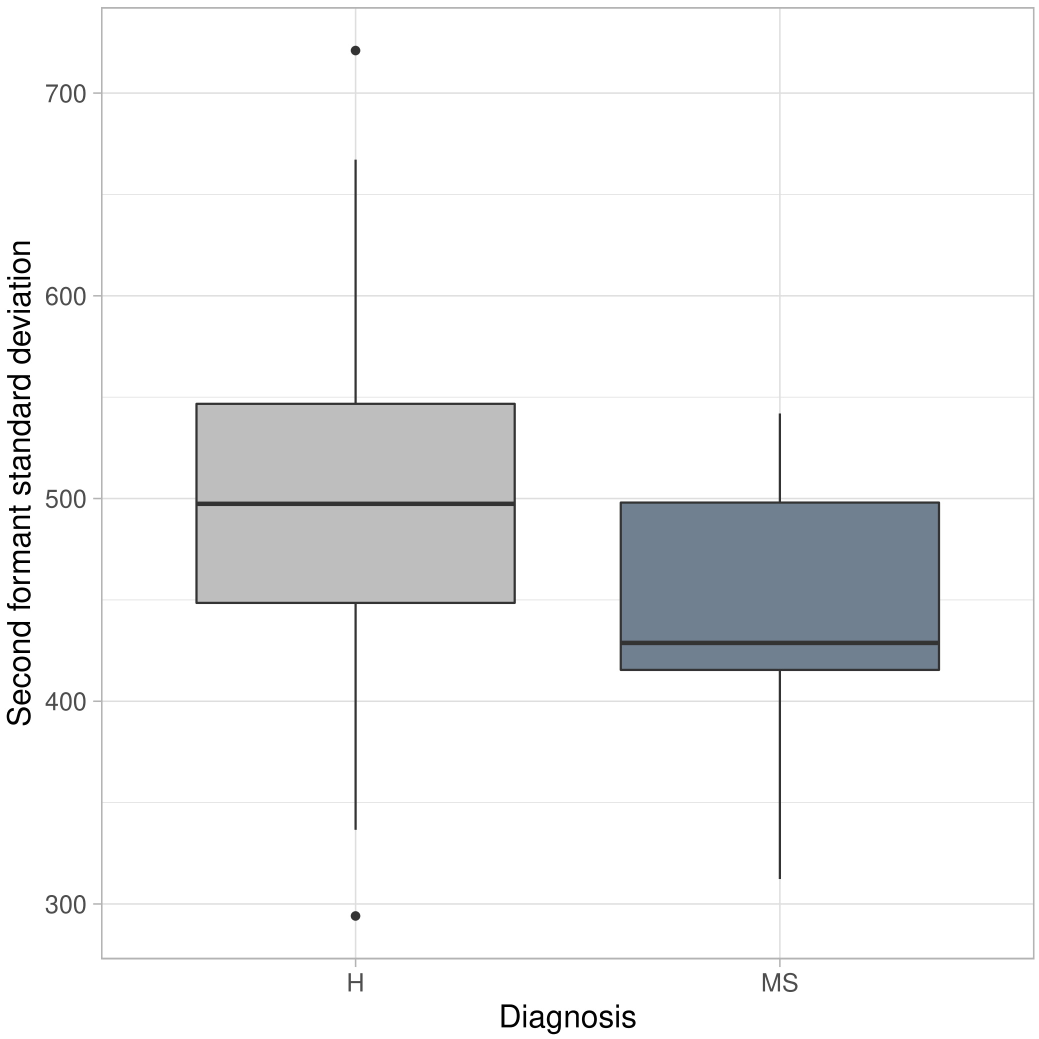

Figure 3.12: Boxplot of the standard deviation of f_2 across all vowels of afflicted and healthy individuals.

of f_2 in both groups in Figure 3.12.

Finally, the MS group seems to have the control of their vocal folds impaired both ways, meaning that their glottis tends to be too open or closed at inappropriate times. This is obvious from Figure 3.13, where the standard deviation of the MS is much smaller and thus suggests afflicted individuals have a harder time keeping their glottis under control.

Overall, it has turned out that even though these results violate the presupposition posited in subsection 2.1.3 on page 21, they are still incredibly informative and useful.

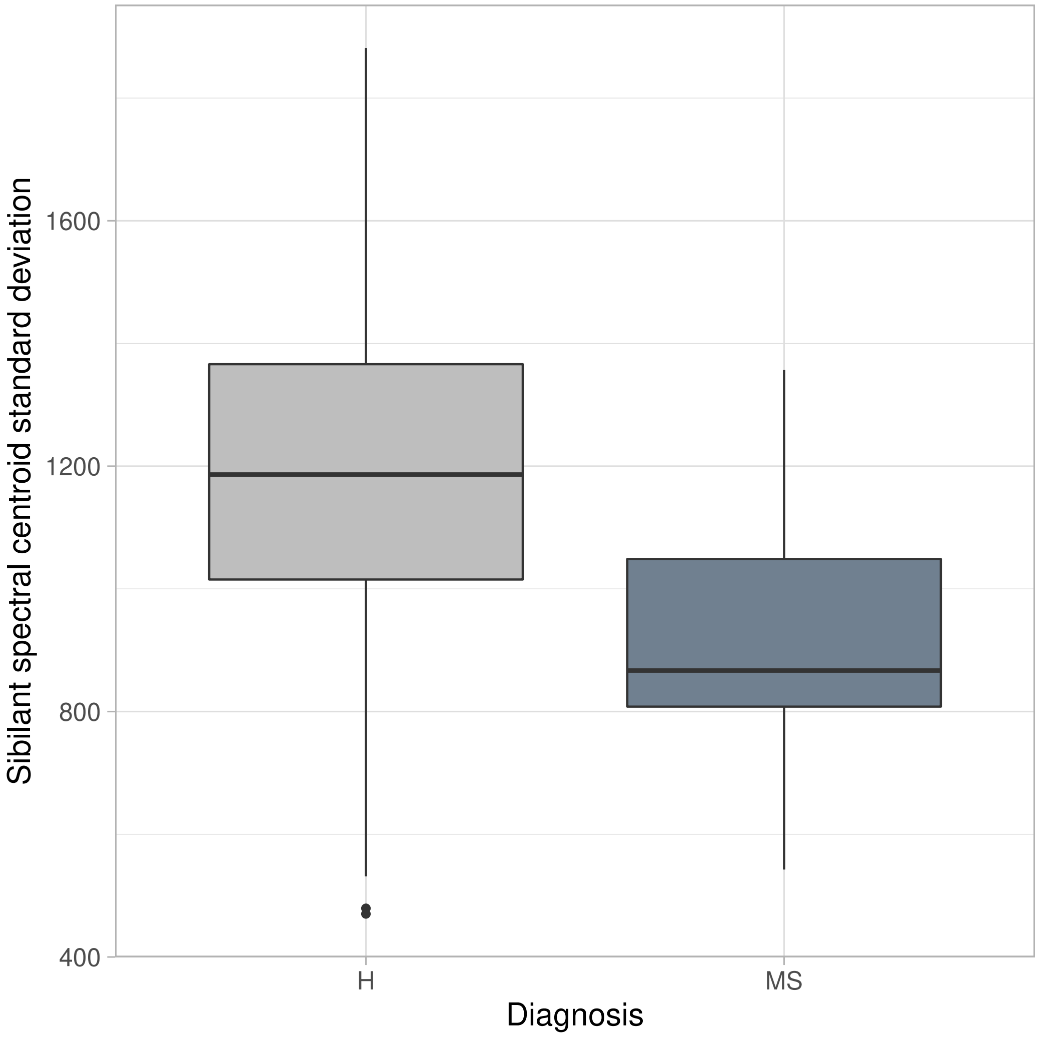

Figure 3.13: Boxplot of the standard deviation of f_3 across all vowels of afflicted and healthy individuals.

Parameter	p value	hypothesis
Recording duration	0.056	Less
Silence percentage	0.02	Two-sided
Vowel percentage	0.74	Two-sided
CSI of vowel duration	0.005	Two-sided
CSI of f_0	0.2284	Two-sided
Glottal stop rate	0.038	Two-sided
Quantile difference of f_0	0.02	Two-sided
Average unvoiced stop duration	0.005	Two-sided
CSI of intensity	0.38	Two-sided
Sibilant spectral centroid standard deviation	0.0003	Two-sided
Standard deviation of f_1	1.4×10^{-11}	Two-sided
Standard deviation of f_2	0.01	Two-sided
Standard deviation of f_3	9.9×10^{-8}	Two-sided

Table 3.1: A table of the p values calculated from parameters measured using manual labelling.

Summary and figures

For a concise summary of the measurements of parameters extracted using manual annotation of data, please refer to Table 3.1, where all the parameters, p values and hypotheses used are respectively listed.

3.1.2 Machine learning model

Overall, significant results were mostly found. There is an important caveat here, however – the fact that some of the results were not statistically significant does not mean that they are irrelevant, as already stated.

Because MS is such a symptomatologically complex disease, it was decided that a neural network model would be trained for the manual annotation datasets as well, as such ML models are good at capturing complex parameters. If it is possible to train such a model using these parameters, we can assert that they capture the MS in a good way *as a whole*. In this case, it can be assumed that the whole is more than just a sum of its parts.

Because of this, it has been decided that a ML model will be trained for the manual dataset as well, despite the fact that this is presumably not practically applicable. Such a model would need someone competent to laboriously annotate TextGrids and correct PitchTiers for an estimation to be made, which would be expensive and impractical.

Instead, this model will in this case serve as a benchmark for how all the parameters described can work *in conjunction* to reliably determine multiple sclerosis on a theoretical level.

In other words, this model will be used to test the hypothesis that MS, as a disease, creates a characteristic acoustic fingerprint, and that by assessing the parameters put forth in this thesis, this fingerprint has been captured reasonably well. This capturing ability of the parameters can be tested *holistically* as consistent performance of a ML model trained using these paremeters. As mentioned in section 1.3 on page 15, the validity of this claim can only be assured by repeated training and testing of the model.

A monotone multi-layer perceptron was initiated using the *caret* package of R as a binary discriminator between healthy and afflicted individuals. It was mainly left with the default settings, the only exceptions being that all data were decorrelated and standardized before being fed into the network and the bootstrapping train control technique was specified. The data was first split into a training and testing set. The entirety of the training and bootstrapping process took place over the training set and the testing set was only used as the last to verify that the model had not overtrained itself. In other words, the model's accuracy was measured only on data the model had never seen.

This entire process was repeated 5 times and the model yielded a median accuracy of 93%, with median Cohen's kappa (which we might call effective accuracy) being 0.63. This metric is normalized for a situation when a model decides completely randomly, yet has a non-zero success rate due to random chance. Generally, this can be assessed as more than satisfactory, as according to Landis and Koch [1977], any result over 0.6 is **substantial**. The reader should be reminded here that the MS patients in this particular dataset all had perceptually detectable dysarthria.

3.2 Automatically annotated data

A dataset of 250 individuals was used, of that 66 MS-afflicted ones. Unlike in the case of manually annotated data, the automatically annotated parameters are much less individually meaningful, because they are mostly heavily biased by the shortcomings of the automatic annotation. For example, the algorithm tends to assume that any silence preceding an unvoiced stop is part of the unvoiced stop. They make up for this by being much more telling in whether or not an automatic detection tool is feasible. For this reason, only Table 3.2 of all p values is included. It is clear that those parameters that are unburduned by the bias of annotation tend to carry over their significance. These are CSI of intensity, Quantile difference of f_0 and Recording duration.

Please note that one of the parameters, glottal stop rate, has been excluded. The reasoning for this decision is the simple fact that the Prague Labeller tool does not annotate glottal stops. Even if it did, it would presumably not be very reliable anyway – it is not very good at detecting elisions and tends to annotate phenomena that are not present in the speech signal if it expects them being there based on the input transcription.

3.2.1 Machine learning model

Unlike previously, this time the Python module *scikit-learn* was used to construct a monotone multi-layer perceptron binary classification model using the class *MLPClassifier*. Its performance was tested much more thoroughly. The maximum epoch amount (how many times the entire dataset is run through during training) was set to 10000 cycles and the learning rate was set to 0.001. The data were normalized before being fed in but were not decorrelated. All other settings were left in their default state.

A similar train/test split was made as in the previous model, but this time, the entire cycle was carried out 50 times and the performance of the model was measured therefrom. The model had a median accuracy of 85% and a Cohen's

Parameter	p value	hypothesis
Recording duration	0.009	Two-sided
Silence percentage	0.11	Two-sided
Vowel percentage	0.72	Two-sided
CSI of vowel duration	0.0002	Two-sided
CSI of f_0	0.13	Two-sided
Quantile difference of f_0	0.004	Two-sided
Average unvoiced stop duration	0.034	Two-sided
CSI of intensity	0.09625	Two-sided
Sibilant spectral centroid standard deviation	0.35	Two-sided
Standard deviation of f_1	2.5×10^{-5}	Two-sided
Standard deviation of f_2	1.3×10^{-5}	Two-sided
Standard deviation of f_3	0.79	Two-sided

Table 3.2: A table of the p values calculated from parameters measured using automatic labelling.

kappa of 0.5. This is by Landis and Koch [1977] to be interpreted as a *moderate* success rate.

3.3 Discussion

It has been proven beyond reasonable doubt that MS leaves a specific, objectively measurable acoustic fingerprint on read speech. It is a much deeper and more complicated question whether or not this fingerprint is substantial enough to make a detection paradigm based on its automatic detection feasible.

Looking at the statistics of the measured recordings, we can see that some hypotheses originally posited in chapter 1 were false. It seems that most of the differences between the two datasets boil down to overall reduced muscle mobility. This does not counterpoint the usefulness or validity of these findings, since two-sided hypotheses were used in the Kolmogorov-Smirnoff statistical tests as a failsafe mechanism for the unpredictability of MS even at the cost of smaller decision power.

It seems from the data, specifically the significance of pause percentage and vowel percentage, but the non-significance of vowel percentage that afflicted individuals generally have a tendency to speak in short laborious bursts with frequent articulatory mistakes.[1]

Poor glottis control due to affected muscle coordination is also obvious, which is backed by the massive significance of the standard deviation of f_3. Similar to this is worsened jaw coordination, evident from the standard deviation of f_1.

Similarly, MS speakers also seem to not use the prosodic and intonational potential of their language, because their intensity and f_0 patterns seem to be consistently quite flat, as seen in figures 3.9 and 3.7.

As for the total predictive power of all parameters combined, the first machine learning model speaks for itself, with an unbalanced accuracy of 93% and a Cohen's kappa of 0.63. This is even with the fact that the model did not train using the age parameter, which would normally be required to account for natural aging processes. The model's practicality is another matter entirely, though.

The constraint of manually annotating data before parametrization renders it practically unusable, because there is no way to practically implement a paradigm that contains annotators labelling data in a cost- and time-efficient way. The negligible operating costs (both financial and temporal) are in fact the number one advantage of such an algorithm, because it is in fact what complements the incredibly high expenses of neurologists and MRI machines.

Thus the other model, the purely automatic one, will be now discussed in-depth. Its main advantage is of course its practically non-existent operating costs, which would presumably boil down to some non-zero but nearly negligible amount of electricity, comparable to a mobile phone.

Its investment costs would depend on its practical implementation and range from zero in the form of a free smartphone app to roughly the price of a new laptop if implemented as a standalone device. This will be further elaborated upon in subsection 3.3.1.

With an accuracy of 85% and a Cohen's kappa of 0.5, the model seems to be fairly good as far as such models go in general. However, this is way below the performance necessary for practical implementation, because during the performed cycles of training, validation and testing, false negative and false positive errors appeared with roughly equal frequency. This model would thus terrify healthy people into visiting neurologists for no reason while simultaneously sometimes falsely reassuring people with possible neurological symptoms.

The good news is that chances are in favour of the presupposition that the model's exhibited performance is largely suboptimal when compared to its po-

[1] This observation is further supported by informal listenings performed by the author. These are not part of this thesis *per se*.

tential. The main reasoning points for this assumption are presented as follow, in presumed order of importance:

- The dataset is very small, which means the model will not generalize very well. 250 observations might even be the bare minimum to construct a semi-workable model with real-life data like this. Obtaining more recordings would presumably improve the model's performance significantly.

- The parametes used in this thesis constitute little more than a pilot set based mostly on theoretical hypotheses with some inspiration from Rusz's research. Many more useful parameters can surely be found, perhaps based on comparing different parts of the recording to one another.

- The recordings that were available were not created with this *particular* application in mind. Thus, only recordings of subjects reading a fixed text were analyzed. No extra exercises like syllable repetition, prolonged vowel production, spontaneous dialogue or elicited monologuing were thus used.

- The author of this thesis is a phoneticist, not a speech technologist. While this offers an advantage in the form of a strong intuition regarding the linguistic relevance of certain phenomena and thus the presumed human behaviour in relation to them, it presents a drawback that the signal processing techniques used for parameterization in this thesis are rather basic.

- The author of this thesis is a phoneticist, not a data scientist, statistician or mathematician. This means the model selection and hyperparameter tuning process can be described as crude at best. Related to this is the fact that all models exhibited a 100% success rate when tested on the entire *training* set on each iteration, suggesting some amount of overfitting.

- Due to the scientific nature of this thesis, the data have been overly parametrized quite heavily. For example, it would probably be beneficial to have separate parameters for each realization of the /s/ phoneme as opposed to just taking the standard deviation of all of them, since it would allow the model to fit to deeper patterns present in the data.

- Binary classification might not be the optimal solution. It may be more practical to use a regression algorithm that outputs an estimate expanded status disability scale (EDSS) score, which is a kind of numeric representation of the patient's disability due to MS.

There are some caveats keep in mind, however. While the model's accuracy seems promising, there are some reasons to think they might be somewhat overexaggerated. These are again presented in the order of their presumed importance.

- There is a slight bias in the dataset in that about half of the MS patients reliably do have hearable dysarthria, while in the other half the ratio is unknown. This is mitigated by the fact that what is mostly measured here are parameters that are difficult to detect by ear, such as standard deviations of long-term trends.
- The dataset has an age bias. It was deemed to be necessary to include the subject age parameter for model training and assessment, because otherwise the model would not be able to map natural aging processes to the ages of the subjects, reducing its accuracy by an unrealistic amount – since in any practical setting, the age of any subject is known *a priori*. This might have skewed the results somewhat.
- Despite the fact that all parameters used in this thesis are robust against noise, it cannot be mindlessly assumed they are not influenced by it at all. Since the recordings were recorded in different rooms, there is a chance that the differing noise levels and acoustical setting skewed the results in favour of a favourable outcome.
- The ratio of healthy and afflicted individuals in the dataset is not the same as in the general population. That is, the model would probably produce a large amount of false positives upon being deployed on a truly random sample of individuals.[2]

Since there is no way to address any of these as of now, further research into this are is necessary. While this statement is true by default (so much so that its negation approaches absurdity) and a staple of the conclusion of every thesis discussion, the liberty has been taken to rephrase it slightly to make it less obvious and more informative:

Further research is necessary if the lives of undiagnosed MS patients are to be improved.

The next part describes how this might be achieved.

[2] Assuming that the set of people who would end up being assessed in a practical scenario by the model would truly be a random sample, which in turn might not be true. People who suspect neurological symptoms would presumably have a stronger tendency to submit to the test, skewing the numbers.

3.3.1 Detection paradigm proposition

Assuming a model exhbiting sufficient accuracy is constructed, the question of its practical implementation arises.

It is important to remind ourselves that no such model can probably ever replace a proper neurological examination performed by a specialist physician using an MRI machine and a spinal tap, as pointed in out in chapter 1. However, such a model can be implemented as a catch-early mechanism for patients exhibiting subclinical symptoms of MRI to refer suspectedly dysarthric individuals to a specialist. As is mentioned in the very preface of this thesis, beginning MS therapy as early as possible is crucial in MS. So how does one apply this model to catch MS cases earlier?

One approach would be to create a smartphone app, which in the author's personal opinion has several significant drawbacks. These are listed here in no particular order:

- A smartphone app severely limits the potential for any financial revenue stemming from such a project, making it much more difficult to acquire funding;
- With a smartphone, one has very limited control over the acoustical properties of the measuring device, specifically its microphone, which a) may be of low quality and b) varies significantly smartphone to smartphone;
- Test subjects may find apps in general to be untrustworthy and may refrain from trusting their output;
- There is no one available apart from the app test subjects themselves to properly interpret the output of the app, which in some implementations may be more complicated than a simple yes/no answer;
- There is no one to assure that the test subject is doing everything properly.

All of these drawbacks can be solved in one fell swoop by implementing the model not as an app, but by a specialized physical device intended for use in a practical physician's office. This would be physically and conceptually similar to a device called an ESR Sed Screener used to measure erythrocyte sedimentation. (Bio-One) Such a device would essentially be just a voice recorder (possibly equipped with a condenser microphone) with a small attached computer with the model and an interface loaded in, which would automatically analyze acoustic input similarly to the aforementioned app. A small speaker can also be included, should it be necessary for the speaker to imitate a pre-recorded task.

Such a device is much more attractive to potential investors than an app because there is a clear potential buyer and thus source of revenue – specifically, physicians' offices looking to improve their services. In an app, this is much less obvious. Since it is a physical object rather than a piece of sotware, a non-zero price is always justified, offering the opportunity for a profit margin, thus making the project interesting to investors.

An app can of course be offered for money, but then it might get used only sporadically, because people presumably rarely pay money for an app that they plan on using once. Additionally, people suspecting that they might have debilitating disease may tend to put off the decision to find out if they have one. A price tag on the app may stimulate this behaviour further, defeating the original purpose of the app. An app can also be offered for free, but then there is little space for financial revenue, placing the entire project in danger of being underfunded and thus not reaching its full potential or not being realized at all.

A physical device also gives the manufacturer full control over the quality of the measuring device, specifically the microphone, while also (and this is nothing but the author's conjecture) feeling much more trustworthy to potential patients. This effect is strengthened by the device's placements in a physician's office, who can also be used to interpret the output of the physical device should it be more complex than just a recommendation to see a neurologist.

If then during check-up, a patient mentions difficulties that could potentially be caused by MS, has a history of neurological diseases in the family, is in some other way at a high risk of developing the disease or simply wishes to check if everything is alright, the physician then assesses the patient's eligibility to take the test.

Non-native Czech speakers, for example, would have to be excluded (at least until such time as similar models are implemented for their particular language), as would patients with already existing speech disorders, because this could very easily skew the results.

Next, the patient is guided through the entire testing process, which will probably include more exercises than the ones presented in this thesis. The device automatically outputs a result, which is then interpreted by the physician and the patient either is or is not referred to a neurologist by the physician. If the results look severe, the device's output can be used to prioritize the patient in question so that they get treatment as soon as possible.

Similarly, the device can be used to perform screening tests in schools, for example, because it is in young adults that the disease typically develops. (Murray [2006]) A physician or other competent person would visit a school, taking recordings of individuals who wish to find out if they may be developing MS. Since the test would take no more than a few minutes, individuals with suspected dysarthria could then be referred to a specialist almost immediately.

Implementing the model as a piece of software intended for use on personal computers that physicians have in their offices is another option that is in between the app/dedicated device question. It has its own limitations in that physicians would be required to buy specific recording equipment, possibly having to set it up, which a) lowers the possible revenue potential of the project by decreasing the profit margin, making the project less attractive to investors, and b) making matters much more complicated for the physicians, discouraging them from using it.

3.3.2 Comparison with previous research

This thesis bears some amount resemblance to some previous works, be it by topic, method of study or scope.

The oldest work this thesis can be compared to is Merson and Rolnick [1998], who used perceptual analysis to lay the ground for following research. Because his parameters were perceptual, it obviously follows that it was not possible to construct a working automatic detection model. This was not the goal of the study, however. The purpose, unlike that of this thesis, was to map general speech and language impediments found in MS patients for mainly for the purposes of direct therapy. The research was not limited to strictly phonetic phenomena and included elaborations on cognitive defects as they manifest linguistically.

Next, worth mentioning is definitely Rusz et al. [2019], where the goal was to produce a mapping the set various dysarthric manifestations of MS onto parts of the brain affected by the disease's lesions. This study likewise does not attempt to predict MS.

All of these studies are similar to this thesis in their topic, but differ in that their general purpose of attempting to expand the knowledge of how MS influences speech, not to assess the overall predictive power of the parameters associated with the MS vocal fingerprint as a whole, extending that to the applicability of these findings to a possible detection device.

A similar study that does attempt to do this is Rusz et al. [2018]. In it, the authors use monopitch and articulatory decay to reach an accuracy 78%

in discriminating healthy speakers from asymptomatic MS speakers, despite the fact that they were only able to detect dysarthria in 56% of *all* MS speakers. This study used three speech tasks, as opposed to this thesis measuring one, and measured a larger amount of parameters. The study unfortunately does not mention Cohen's kappa or any other metric of the model apart from its raw accuracy, which makes comparison with the model in this thesis impossible. By raw accuracy, the model in this thesis performs slightly better (+7%), but this number is unreliable in unbalanced datasets and the model presented in Rusz et al. [2018] might actually still be better.

That said, presumably the most successful model would presumably be constructed by combining the approaches used in both works. Further research is thus necessary.

Conclusion

This thesis has shown that MS conclusively does create a robust acoustic fingerprint that can be objectively measured and evaluated and which furthermore has good predictive value. Aside from that, there is a high chance that these acoustic characteristics can be used to construct a physical device which can then be deployed as early warning mechanisms either in practical physicians' offices, or as bulk screening tests in schools.

The thesis combined data extracted from manually annotated recordings of healthy and afflicted individuals reading a standardized text with data extracted from automatically annotated recordings. The former had the advantage of being much more accurate and interpretable *per se*, while the second approach had the advantage of more closely modelling a practical implementation while also working with more data overall.

13 acoustic parameters variously relating to spasticity, fatigue and ataxia were measured on the smaller manually annotated dataset. Of these, 9 were found to be statistically significant when comparing the afflicted and control groups with Kolmogorov-Smirnov tests. A multi-layer perceptron model implemented in R's *caret* library was employed to try and assess the ability of these parameters to discriminate between the two groups, using separate training and testing datasets. The model had an accuracy of 93% with Cohen's kappa (a measure of model accuracy adjusted for random successful guesses) of 0.63, which indicates that the parameters as a set are robust and have high minimum theoretical predictive power.

To extended this to a more practical scenario, a larger dataset was used, this time annotated automatically using Praat and the Prague Labeller tool. 12 parameters were used this time, of which 7 were found to be significant. A similar model was employed using Python's *scikit* library, with an accuracy of 85% and a Cohen's kappa of roughly 0.5 after repeated testing using split training/testing datasets. This can be interpreted as the automatic parameters having a moderate minimum practical predictive power.

Since it can be assumed that a much better success rate can be achieved by using a better tuned model with a larger dataset, along with more parameters and exercises, the question of practical application then arises. In the author's opinion, an optimal way to do this would be to create a dedicated physical device intended for use by practical physicians either in their offices or by performing screening tests in schools. Those individuals who would get a positive result would get referred to a professional, thus potentially increasing their quality of life significantly due to early therapy, which is of utmost importance in MS.

Further research is necessary to confirm that more accurate models are indeed possible to create, but if that is the case, there is nothing stopping the development of a practical solution.

Bibliography

Sandra Amor and Hans Van Noort. *Multiple Sclerosis*, volume 1st ed of *The Facts*. OUP Oxford, Oxford, 2012. ISBN 978-0-19-965257-0. URL `http://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=475459&lang=cs&site=ehost-live`.

M. P. Barnes, R. M. Kent, J. K. Semlyen, and K. M. McMullen. Spasticity in Multiple Sclerosis. *Neurorehabilitation and Neural Repair*, 17(1):66–70, March 2003. ISSN 1545-9683. doi: 10.1177/0888439002250449. URL `https://doi.org/10.1177/0888439002250449`. Publisher: SAGE Publications Inc STM.

Greiner Bio-One. ESR Instruments. URL `https://shop.gbo.com/en/row/products/preanalytics/instruments/esr-instruments/`.

Paul Boersma and David Weenink. *Praat: doing phonetics by computer (Version 5.1.13)*. 2009. URL `http://www.praat.org`.

Tomáš Bořil and Radek Skarnitzl. Tools rPraat and mPraat. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, pages 367–374, Cham, 2016. Springer International Publishing. ISBN 978-3-319-45510-5. doi: 10.1007/978-3-319-45510-5_42. URL `http://dx.doi.org/10.1007/978-3-319-45510-5_42`.

J. E. Freal, G. H. Kraft, and J. K. Coryell. Symptomatic fatigue in multiple sclerosis. *Archives of physical medicine and rehabilitation*, 65(3):135–138, 1984.

Fiona J. Fitz Gerald, Bruce E. Murdoch, and Helen J. Chenery. Multiple Sclerosis: Associated Speech and Language Disorders. *Australian Journal of Human Communication Disorders*, 15(2):15–35, December 1987. ISSN 0310-6853. doi: 10.3109/asl2.1987.15.issue-2.02. URL `https://doi.org/10.3109/asl2.1987.15.issue-2.02`. Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/asl2.1987.15.issue-2.02.

John M. Grey and John W. Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5): 1493–1500, May 1978. ISSN 0001-4966. doi: 10.1121/1.381843. URL `https://asa.scitation.org/doi/10.1121/1.381843`. Publisher: Acoustical Society of America.

Jan Hlavnička, Tereza Tykalová, Olga Ulmanová, Petr Dušek, Dana Horáková, Evžen Růžička, Jiří Klempíř, and Jan Rusz. Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 131(5):1155–1165, May 2020. ISSN 1872-8952. doi: 10.1016/j.clinph.2020.02.005.

Barrie J. Hurwitz. The diagnosis of multiple sclerosis and the clinical subtypes. *Annals of Indian Academy of Neurology*, 12(4):226–230, 2009. ISSN 0972-2327. doi: 10.4103/0972-2327.58276. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824949/`.

JetBrains. PyCharm. URL `https://www.jetbrains.com/pycharm/`. Library Catalog: www.jetbrains.com.

Ilya Kister, Tamar E. Bacon, Eric Chamot, Amber R. Salter, Gary R. Cutter, Jennifer T. Kalina, and Joseph Herbert. Natural History of Multiple Sclerosis Symptoms. *International Journal of MS Care*, 15(3):146–156, October 2013. ISSN 1537-2073. doi: 10.7224/1537-2073.2012-053. URL `https://www.ijmsc.org/doi/full/10.7224/1537-2073.2012-053`. Publisher: Consortium of Multiple Sclerosis Centers.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006-341X. doi: 10.2307/2529310. URL `https://www.jstor.org/stable/2529310`. Publisher: [Wiley, International Biometric Society].

Uwe Ligges, Sebastian Krey, Olaf Mersmann, and Sarah Schnackenberg. *tuneR: Analysis of Music and Speech.* 2018. URL `https://CRAN.R-project.org/package=tuneR`.

Pavel Machač and Radek Skarnitzl. *Fonetická segmentace hlásek.* Epocha, 2010. ISBN 978-80-7425-031-6. URL `https://search.ebscohost.com/login.aspx?authtype=shib&custid=s1240919&profile=eds`.

Wes McKinney et al. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

D. L. Mclellan. Spasticity: disorder motor control. *Journal of Neurology, Neurosurgery, and Psychiatry*, 44(10):961, October 1981. ISSN 0022-3050. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC491190/`.

Richard M. Merson and Michael I. Rolnick. Speech-language Pathology and Dysphagia in Multiple Sclerosis. *Physical Medicine and Rehabilitation Clinics of North America*, 9(3):631–641, August 1998. ISSN 1047-9651. doi: 10.1016/S1047-9651(18)30254-7. URL `http://www.sciencedirect.com/science/article/pii/S1047965118302547`.

James R. Miller. The importance of early diagnosis of multiple sclerosis. *Journal of managed care pharmacy: JMCP*, 10(3 Suppl B):S4–11, June 2004. ISSN 1083-4087.

Philip J. Monahan and William J. Idsardi. Auditory Sensitivity to Formant Ratios:Toward an Account of Vowel Normalization. *Language and cognitive processes*, 25(6):808–839, July 2010. ISSN 0169-0965. doi: 10.1080/01690965.2010.490047. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893733/`.

TJ Murray. Diagnosis and treatment of multiple sclerosis. *Bmj*, 332(7540):525–527, 2006. Publisher: British Medical Journal Publishing Group.

J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth. Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500, January 2016. ISSN 0001-4966. doi: 10.1121/1.4939739. URL `https://asa.scitation.org/doi/10.1121/1.4939739`. Publisher: Acoustical Society of America.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

P Pollák, J Volín, and R Skarnitzl. HMM-based phonetic segmentation in Praat environment. In *Proceedings of the VII th International Conference "Speech and Computer–SPECOM*, volume 1, pages 537–541, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

Patrick F. Reidy. Spectral dynamics of sibilant fricatives are contrastive and language specific. *The Journal of the Acoustical Society of America*, 140(4):

2518–2529, October 2016. ISSN 0001-4966. doi: 10.1121/1.4964510. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5132428/.

Kris Rozenstoks, Michal Novotný, Dana Horáková, and Jan Rusz. Automated Assessment of Oral Diadochokinesis in Multiple Sclerosis Using a Neural Network Approach: Effect of Different Syllable Repetition Paradigms. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 28(1):32–41, 2020. ISSN 1558-0210. doi: 10.1109/TNSRE.2019.2943064.

RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Jan Rusz. Detecting speech disorders in early Parkinson's disease by acoustic analysis. 2018.

Jan Rusz, Barbora Benová, Hana Růžičková, Michal Novotný, Tereza Tykalová, Jan Hlavnicka, Tomas Uher, Manuela Vaneckova, Michaela Andelova, Klara Novotna, Lucie Kadrnozkova, and Dana Horakova. Characteristics of motor speech phenotypes in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 19:62–69, January 2018. ISSN 2211-0356. doi: 10.1016/j.msard.2017.11.007.

Jan Rusz, Manuela Vaněčková, Barbora Benová, Tereza Tykalová, Michal Novotný, Hana Ruzickova, Tomas Uher, Michaela Andelova, Klara Novotna, Lucie Friedova, Jiri Motyl, Karolina Kucerova, Jan Krasensky, and Dana Horakova. Brain volumetric correlates of dysarthria in multiple sclerosis. *Brain and Language*, 194:58–64, 2019. ISSN 1090-2155. doi: 10.1016/j.bandl.2019.04.009.

Jean Sawyer. The acoustic properties of vowels, 2013. URL http://my.ilstu.edu/~jsawyer/consonantsvowels3/consonantsvowels24.html.

Radek Skarnitzl and Jan Volín. Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny [Reference Values of Vowel Formants for Young Adult Speakers of Standard Czech]. *Akustické listy*, 18:7–11, 2012.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009. ISBN 1-4414-1269-7.

Jan Volín, Tereza Tykalová, and Tomas Boril. Stability of Prosodic Characteristics Across Age and Gender Groups. pages 3902–3906, 2017. doi: 10.21437/Interspeech.2017-1503.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*. 2017. URL `https://CRAN.R-project.org/package=tidyverse`.

Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. 2019. URL `https://CRAN.R-project.org/package=stringr`.

Alastair Wilkins. Cerebellar Dysfunction in Multiple Sclerosis. *Frontiers in Neurology*, 8, June 2017. ISSN 1664-2295. doi: 10.3389/fneur.2017.00312. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5487391/`.

Lippincott Williams & Wilkins. The prevalence of MS in the United States: A population-based estimate using health claims data. *Neurology*, 93(15): 688–688, October 2019. ISSN 0028-3878, 1526-632X. doi: 10.1212/WNL.0000000000007915. URL `https://n.neurology.org/content/93/15/688.2`. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Correction.

Karel Čapek. *Měl jsem psa a kočku*. Městská knihovna v Praze, 1939. URL `https://search.mlp.cz/cz/titul/mel-jsem-psa-a-kocku/3347549/`. Library Catalog: search.mlp.cz.

List of Figures

List of Tables

Glossary

MRI magnetic resonance imaging. 3, 10, 42, 45

ML machine learning. 4, 5, 11, 12, 13, 14, 16, 32, 39

MS multiple sclerosis. 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 19, 24, 25, 28, 29, 30, 31, 32, 36, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

CNS central nervous system. 5

IDE integrated development environment. 13

$\boldsymbol{f_0}$ base frequency. 14, 21, 28, 30, 38, 40, 41, 42

CSI cumulative slope index. 17, 18, 20, 21, 23, 27, 28, 29, 32, 33, 34, 38, 40, 41, 56

$\boldsymbol{f_1}$ first formant. 21, 35, 38, 41, 42, 56

$\boldsymbol{f_2}$ second formant. 21, 35, 36, 38, 41, 56

$\boldsymbol{f_3}$ third formant. 21, 35, 37, 38, 41, 42, 56

EDSS expanded status disability scale. 43