Bachelor's Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Circuit Theory

Automatic analysis of speech diadochokinetic task of voiced consonants for the assessment of articulatory deficits in patients with multiple sclerosis

Michaela Měrková

Supervisor: Ing. Michal Novotný, Ph.D.
Subfield: Medical Electronics and Bioinformatics
May 2024

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Měrková Michaela** Personal ID number: **499141**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Circuit Theory**

Study program: **Medical Electronics and Bioinformatics**

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Automatic analysis of speech diadochokinetic task of voiced consonants for the assessment of articulatory deficits in patients with multiple sclerosis

Bachelor’s thesis title in Czech:

Automatická analýza řečové diadochokinetické úlohy znělých konzonant pro hodnocení artikulace pacientů s roztroušenou sklerózou

Guidelines:

Multiple sclerosis (MS) is a debilitating autoimmune neurodegenerative disease resulting in motor and cognitive deficits with a profound impact on the quality of patients' lives. The prevalence of speech disruption termed dysarthria in the MS is 40-50%, which makes it a common sign. Articulatory deficits are an essential part of dysarthria manifestations. The diadochokinetic task based on fast, steady repetition of the consonant-vowel train is one of the most sensitive approaches to assessing articulatory deficits. The analysis based on the evaluation of plosive consonants provides valuable information in differential diagnostics. However, the current acoustic analysis methods are limited to unvoiced consonants, and the effect of articulatory deficits in voiced plosives remains unknown. The aims of the thesis are:
1) Study the topic of dysarthria induced by multiple sclerosis.
2) Search the state-of-the-art methods for the assessment of speech diadochokinetic task.
3) Propose an automatic approach for parametrization of voiced plosives in speech diadochokinetic task
4) Analyze the disruption of voiced plosives in patients with multiple sclerosis

Bibliography / sources:

TYKALOVA, Tereza, et al. Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy, and multiple system atrophy. Brain and language, 2017, 165: 1-9.
NOVOTNÝ, Michal, et al. Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22.9: 1366-1378.
ROZENSTOKS, Kris, et al. 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, 2019, 28.1: 32-41.
DOBSON, Ruth; GIOVANNONI, Gavin. Multiple sclerosis–a review. European journal of neurology, 2019, 26.1: 27-40.

Name and workplace of bachelor’s thesis supervisor:

Ing. Michal Novotný, Ph.D. Department of Circuit Theory FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **06.02.2024** Deadline for bachelor thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

Ing. Michal Novotný, Ph.D. Supervisor’s signature	doc. Ing. Radoslav Bortel, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student’s signature

Acknowledgements

I would like to express gratitude to my supervisor, Ing. Michal Novotný, Ph.D., for his invaluable guidance, insights, and patience throughout the development of my bachelor’s thesis. His support has been instrumental in the successful completion of this work. I would also like to extend my gratitude to my family and friends for their endless support, not only during the final months of working on this thesis but throughout my entire studies.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, 25. May 2024

. . . .

Abstract

Neurodegenerative diseases, such as multiple sclerosis (MS), often impair motor functions, including the articulatory system. This project investigates dysarthria in MS patients by developing an algorithm to detect articulatory events in speech recordings of diadochokinetic (DDK) tasks, focusing on under-studied voiced syllables. The study analysed speech recordings from 120 MS patients and 60 healthy control (HC) individuals, all of whom performed steady repetitions of /ba/-/da/-/ga/ syllables. Four speech events were extracted from these recordings, achieving high detection accuracies of 94.7%, 74.6%, and 79.8% for three of the events with a 5 ms threshold. These events facilitated the calculation of six speech parameters used to differentiate between the HC and MS groups and to train a Support Vector Machine (SVM) classifier.

The comparative analysis revealed minor differences across all calculated features, with the most significant distinctions found in DDK rate and DDK fluctuation. The SVM classifier demonstrated a notable success rate of 70.4% in distinguishing between the two groups. This study underscores the potential of using detailed articulatory event detection and speech feature analysis to improve diagnostic tools for MS-related dysarthria.

Keywords: Speech, Dysarthria, Multiple Sclerosis, Acoustic Analysis, Diadochokinesis

Supervisor: Ing. Michal Novotný, Ph.D.

Abstrakt

Neurodegenerativní onemocnění, jako je roztroušená skleróza (RS), často narušují motorické funkce, včetně hlasového ústrojí. Tento projekt zkoumá dysartrii u pacientů s RS, a to vývojem algoritmu pro detekci artikulačních jevů v řečových záznamech diadochokinetických (DDK) úloh, se zaměřením na dosud málo zkoumané znělé slabiky. Součástí studie byla analýza řečových záznamů od 120 pacientů s RS a 60 zdravých jedinců (HC), kteří opakovaně vyslovovali slabiky /ba/-/da/-/ga/. Z těchto záznamů byly extrahovány čtyři řečové jevy, přičemž tři z nich dosáhly vysoké přesnosti detekce 94,7 %, 74,6 % a 79,8 % při prahu 5 ms. Tyto jevy byly použity k výpočtu řečových parametrů, které sloužily k rozlišení mezi skupinami HC a RS a k natrénování SVM klasifikátoru.

Při porovnání vypočtených parametrů byly zjištěny drobné rozdíly ve všech z nich, přičemž nejvýraznější byly zjištěny v rychlosti DDK a kolísavosti DDK. SVM klasifikátor ukázal význačnou úspěšnost 70,4 % při rozlišování mezi oběma skupinami. Tato studie ukazuje potenciál využití podrobné detekce artikulačních jevů a analýzy řečových rysů ke zlepšení diagnostických nástrojů pro dysartrii související s RS.

Klíčová slova: Řeč, Dysartrie, Roztroušená Skleróza, Akustická Analýza, Diadochokineze

Překlad názvu: Automatická analýza řečové diadochokinetické úlohy znělých konzonant pro hodnocení artikulace pacientů s roztroušenou sklerózou

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Definition and Overview of Multiple Sclerosis

Multiple sclerosis (MS) is a neurodegenerative disease that frequently impacts young adults. This debilitating disease significantly reduces the quality of life of around 2.8 million people globally [1]. Moreover, it has been shown that the incidence of the disease is rising. Yet the cause of the disease remains unresolved, with many causes including genes increasing disease susceptibility or environmental factors such as Epstein–Barr virus, ultraviolet B light (UVB) exposure, obesity or smoking - MS prompts an adverse immune reaction to myelin, disrupting communication between neurons and thus impairing the entire nervous system's function [2].

1.2 Key Signs and Symptoms: Clinical Features

Multiple sclerosis progresses through various phases, including at-risk, preclinical, prodromal, and symptomatic stages. The disease is usually suspected on the basis of the presence of clinically isolated syndrome (CIS) [2]. The CIS can be mono- or poly-symptomatic and with different presentations due to the different eloquent lesion locations.

Historically, the disease was viewed as a two-stage disease, including early inflammation resulting in a relapsing-remitting stage (RRMS) followed after 10 to 15 years by neurodegeneration causing non-relapsing progression, i.e. secondary progressive (SPMM) stage. Disease may also be classified as a primary progressive (PPMS) in the case that the relapsing-remitting phase is not present and only occasional relapses appear [3]. Current state-of-the-art research challenges the traditional two-stage view and suggests that the disease varies across the spectrum between relapsing and progressive stages. Even the RRMS causes persistent damage due to incomplete relapse recovery. The MRI revealed that for every clinically evident attack, approximately ten asymptomatic lesions are present; moreover, on a microscopic level, MS causes large amounts of lesions that are not MRI-visible. The set of symptoms results from the size and location of the MS lesions.[2]

Common symptoms include motor and cognitive impairments, particularly optic neuritis, spinal cord syndrome and brainstem syndrome [2]. Patients may suffer from imbalance, speech and swallowing-related difficulties, depression and anxiety [4]. These challenges not only introduce patients to physical and mental hardships but also influence their surroundings. As of now, a definitive cure for MS remains elusive; hence, treatment predominantly involves symptomatic management and disease-modifying therapies to delay its progression [2].

1.3 Management and Treatment: Disease-modifying Therapies

Acknowledging the absence of a definitive cure for MS, the focus shifts to disease-modifying therapies (DMTs), which are key in managing this condition. DMTs are designed to modify the disease course, aiming to reduce the frequency and severity of relapses and slow down the progression of disability. These therapies, while not curative, offer significant benefits in terms of reducing the impact of symptoms and enhancing the overall quality of life for individuals living with MS [5].

To build upon the foundational knowledge of DMTs in MS management, a brief overview of the various therapies is essential for appreciating their role in influencing the disease's progression. Traditionally, MS treatments have fallen into two categories: immunosuppressants, exemplified by drugs like fingolimod and natalizumab, and immunomodulators, including agents such as interferon beta. More recent advances, potentially leading to a definitive cure, have introduced immune reconstitution therapies, such as alemtuzumab, which provide longer-lasting effects. The 'no evidence of disease activity' (NEDA) principle has emerged, concentrating on a combination of clinical assessments, MRI results, and biomarker measures to steer more proactive and robust treatment approaches [2].

In addition to these disease-modifying therapies, symptomatic treatments play a significant role in MS management. This approach comprises of drugs for bladder issues and neuropathic pain, along with MS-specific treatments like sativex for spasticity. Equally important is the focus on managing lifestyle habits, emphasising proper sleep, consistent exercise and a balanced diet. Such strategy targets both the illness and its symptoms, thereby enhancing the quality of patient care overall [2, 5, 3].

1.4 Diagnosis

The diagnosis of MS relies on a set of criteria due to the potential overlap of its symptoms with other conditions. The current standard is the 2017 McDonald Criteria, which combines clinical symptoms with diagnostic tests such as magnetic resonance imaging (MRI) [3]. MS is typically suspected and later identified due to the CIS, which usually occurs at the relapsing-remitting stage. However, in many patients, the MRI showed older inactive lesions, showing that MS begins before CIS occurrence, and the current search suggests that the preclinical phase of the disease may last for decades. Moreover, there is evidence of brain damage in the earliest stages, including brain volume loss in young people with CIS, decreased school performance in children later developing MS, and cognitive impairment in people with radiologically isolated syndrome detected in MRI done due to the unrelated causes [6, 7, 8]. Early diagnosis is critical as treatments are most effective during the initial stages of the disease [2].

Conventional diagnostic methods for multiple sclerosis, such as MRI, while effective, have several drawbacks that limit their accessibility and convenience for patients. Firstly, the cost of an MRI scan can be prohibitive for many, which means that some people may delay or forego essential diagnostic testing. Secondly, MRI machines are not universally available, especially in rural or under-resourced areas. This lack of availability can result in further delaying the diagnosis. Additionally, undergoing an MRI can significantly increase stress levels in patients, especially in those with chronic pain and mobility issues, as the procedure requires the patient to remain still in an enclosed space, which often induces feelings of claustrophobia, discomfort, and anxiety. Given these challenges, there is a growing interest in alternative diagnostic methods that are more accessible, cost-effective, and patient-friendly.

1.5 Speech Impairment in Multiple Sclerosis

Dysarthria in MS is characterised by various speech impairments, including slow articulation, imprecise consonants, and instability in pitch and loudness. Additionally, patients often face prosodic challenges, such as extended pauses and inadequate loudness control [9]. These characteristics are crucial in the conventional evaluation of speech issues in MS, providing a framework for assessment and diagnosis.

Expanding upon these traditional methods, a recent study, "Automated Assessment of Oral Diadochokinesis in Multiple Sclerosis Using a Neural Network Approach: Effect of Different Syllable Repetition Paradigms" by Rozenstoks et al. employed a neural network approach to investigate oral diadochokinesis, offering new insights into the condition. This innovative research focused on analysing different syllable repetition paradigms in MS patients [10]. This study demonstrated significant differences in speech patterns between MS patients and healthy controls, particularly noting slower speech for sequential motion rate tasks and more irregularity in voiced paradigms. These observations underscore the importance of paying closer attention to unvoiced syllables when detecting speech impairments in MS. This approach could potentially lead to more accurate and early diagnosis, thereby improving patient care and treatment strategies.

1.6 Speech Analysis in Multiple Sclerosis

Speech analysis has emerged as a highly promising method for diagnosing neurodegenerative diseases, including multiple sclerosis. This approach is particularly relevant given that such conditions often impair the motor system, leading to distinctive changes in speech patterns, such as disrupted speech loudness, harsh voice quality and imprecise articulation, that can be early indicators of the condition. The importance of speech analysis lies in two primary aspects: firstly, it may serve as an early detection tool to predict disease relapse, and secondly, it can accurately monitor the disease progression. Esteemed for being non-invasive, cost-effective, and convenient for patients, speech analysis is a significantly beneficial tool in medical diagnostics. Its utility extends beyond mere diagnosis; it also offers insights into the effectiveness of treatments and patient responses over time. Previous research has underscored the effectiveness of speech analysis in neurodegenerative diseases, which notably achieved a success rate of up to 88% in accurately distinguishing patients with Parkinson's Disease from healthy individuals [11]. This impressive success rate shows how speech analysis is not only useful for diagnosis but also plays a vital role in ongoing efforts to understand and tackle the challenges of neurodegenerative diseases like MS.

1.7 State-of-the-Art

The exploration of speech analysis in MS, particularly in the context of dysarthria, underscores the importance of previous studies. The work by K. Rozenstoks et al., focusing on the analysis of syllable repetition in MS patients, is a prime example. This research, employing a neural network approach to assess oral diadochokinesis, uncovered significant differences in speech patterns between MS patients and healthy controls, thereby enriching our understanding of this aspect of the disease [10]. However, it's important to highlight that Rozenstoks' study did not explore the articulatory features of voiced syllables, leaving a noticeable gap in the research.

This gap is particularly relevant considering the earlier discussion about the importance of speech analysis in MS and the varied symptoms of dysarthria. While the study by Novotný, Rusz, Cmejla, and Růžička primarily focused on Parkinson's Disease and offers valuable methodologies for speech analysis, its applicability to MS, especially for voiced syllables, remains unexplored. Their work, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease", could provide a framework for future studies in this area [11].

Additionally, the research conducted by J. Rusz provides a comprehensive overview of motor speech phenotypes in multiple sclerosis. This study contributes significantly to the understanding of speech impairments in MS, detailing various speech characteristics and their implications for diagnosis and treatment [12]. However, like the studies before it, this research does not address the analysis of articulatory deficits of voiced syllables in MS.

The absence of in-depth research on voiced syllables in MS highlights a significant opportunity for groundbreaking work in this area. Investigating these syllables could reveal new dimensions of speech impairments in MS, potentially leading to more effective diagnostic and monitoring strategies. This research direction aligns with the need for non-invasive, cost-effective diagnostic tools in neurodegenerative diseases and promises to make a substantial contribution to the field.

1.8 Aim of This Project

The primary aim of this project is to delve into the relatively unexplored area of voiced syllable analysis in multiple sclerosis patients. While previous studies have provided valuable insights into unvoiced syllables, this project seeks to extend that knowledge by focusing on how MS affects the articulation of voiced syllables, specifically the syllables /ba/-/da/-/ga/. By analysing these specific syllable types, the study aims to uncover potentially distinct speech patterns associated with MS, which could lead to a more accurate and early diagnosis of the disease. Additionally, this research could contribute to the development of improved monitoring techniques for tracking the progression of MS, ultimately aiding in the optimisation of treatment plans and enhancing the quality of life for patients. This project stands to fill a significant gap in current MS research and could pave the way for more comprehensive speech analysis methods in the diagnosis and management of neurodegenerative diseases.

Chapter 2

Methods

2.1 Participants and Recordings

Speech recordings were obtained from a group consisting of 120 individuals diagnosed with MS (31 males, 89 females) and a control group of 60 healthy subjects (16 males, 64 females). Both groups comprised individuals aged between 18 and 74 years, with the MS group having a mean age of 43.9 ± 10.9 years and the control group 43.9 ± 12 years. Within the MS group, the duration of disease varied from 2 to 37 years with an average of 14.5 ± 7.6 years. Disease severity within the MS patients was assessed using the Expanded Disability Status Scale (EDSS), developed by Kurtzke. This scale ranges from 0 to 10, where 0 indicates no disability, and 10 signifies death due to MS. Scores from 1 to 4.5 refer to patients who are fully ambulatory despite increasing disability, while scores above 5.0 indicate assistance requirements for walking. The participants' scores involved in this project varied from 1 to 6, with a mean of 3.4 ± 1.5, which signifies mild to moderate disability [13].

Recordings were obtained in a quiet room to minimise ambient noise. Each participant was instructed to quickly and consistently repeat the syllable sequence /ba/-/da/-/ga/, which is pronounced as voiced in the Czech language. Each subject completed two recordings of this sequence, resulting in 360 recordings in total. The recordings were captured at a sampling rate of 48 kHz and a 16-bit resolution to ensure high-quality audio, clarity, and detail.

2.2 Reference Labels

To assess the performance of the algorithm, 20 recordings from each group were previously labelled by a speech expert with experience in the field. These labels provided timestamps for key speech events such as the beginning of the voicing segment, initial vowel burst, vowel onset, and occlusion. The labels, which followed previously established guidelines [11], were provided with the recordings. This setup allowed assessing the algorithm's accuracy in detecting these events, ensuring that it could identify speech patterns effectively and reduce the time and potential bias associated with manual marking.

2.3 Algorithm for Signal Analysis

2.3.1 Pre-Processing

The initial phase of signal processing involves resampling the audio data from 48 kHz to 20 kHz. This reduction in sampling rate decreases memory requirements and reduces computational complexity in later processing stages. Importantly, a sampling frequency of 20 kHz preserves all essential speech information, preventing the loss of essential data during the transformation. After resampling, any existing DC offset, which is the mean amplitude shift from zero, was removed from the signal. The signal was then normalised to standardise the maximum amplitude from -1 to 1 across all recordings to ensure easier processing and application of thresholds during the following syllable segmentation.

2.3.2 Automatic Syllable Segmentation

The diadochokinetic task requires participants to rapidly repeat a sequence of syllables in a single breath, resulting in a variable number of syllables within each recording. To analyse and identify pronunciation events within each syllable, it was essential to segment the signal and individually extract each syllable for further investigation. Each syllable in the /ba/-/da/-/ga/ sequence consists of a consonant followed by a vowel, with the vowel typically displaying higher energy than the consonant. This difference in energy served as the primary criterion for segmenting the signal.

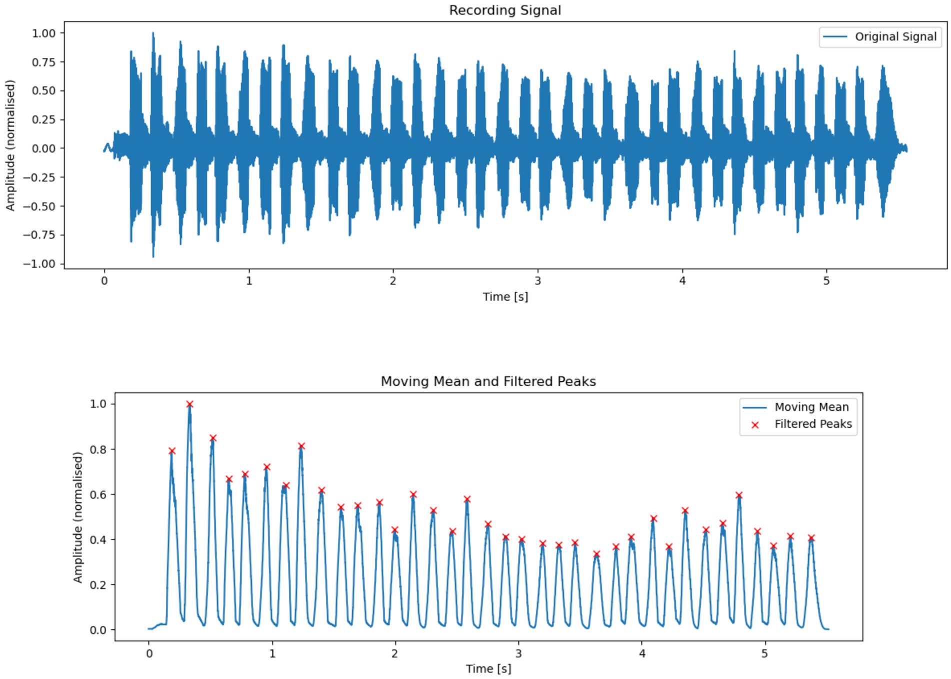

Figure 2.1: The upper plot in this figure displays the original audio signal of a diadochokinetic task, capturing the rapid repetition of syllables. The lower plot presents the smoothed signal after applying a moving average, with highlighted peaks indicating the high-energy vowel components of each syllable.

The segmentation process began with applying a moving average to smooth the signal. This method averaged data points over a specified window length set as 800, or 1/25 of the sampling frequency, reducing noise, clarifying underlying trends, and enabling more accurate segmentation. The next step was to identify peaks in the smoothed signal, which correspond to vowels due to their higher energy. To ensure accurate detection and minimise false positives, the following filters were implemented:

1. Minor peak elimination: Peaks significantly lower in amplitude than neighbouring peaks were discarded as they often represent minor stammers or noise.

2. Adjacent peak analysis: Peaks without a significant drop in amplitude relative to adjacent peaks were likely components of the same vowel sound and were, therefore, removed.

3. Initial peak evaluation: The initial peak was excluded from analysis if its amplitude was significantly lower than the subsequent peaks, typically indicating it resulted from the initial breath rather than a vowel.

These processing steps are visualised in Figure 2.1, which displays two plots: the upper plot shows the original audio signal and the lower plot presents the smoothed signal with highlighted peaks, marking the locations identified as significant for analysis.

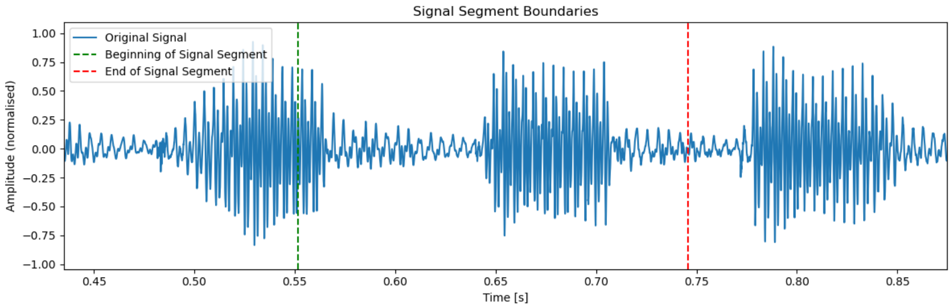

Figure 2.2: A segment of a speech signal with roughly established boundaries that define each syllable. The end of the previous vowel and the start of the next voicing segment are included to provide general comparisons and validate detections.

Lastly, the total count of valid peaks was calculated. As each complete /ba/-/da/-/ga/ cycle should contain three syllables, the presence of a total number of peaks not divisible by three indicates possible misdetections. Signals that failed to meet this criterion were later olny used for label comparison, however, not for the SVM training, as it was not possible to confirm whether individual detected syllables corresponded accurately to the intended syllables in the /ba/-/da/-/ga/ sequence.

After identifying each vowel in the signal, the borders between individual syllables were approximated at the midpoint between two adjacent peaks. Since the peak of each vowel might not be perfectly centred, each segment was expanded by 1.25 times in both directions. This adjustment ensured that both the beginning of the consonant and the end of the vowel were included within the segment, as visible in Figure 2.2. In the signal segment, the end of the previous vowel was also present, allowing for comparison between the occlusion of the previous vowel and the beginning of the voicing segment of the current syllable. This comparison helped to identify and eliminate any invalid detections. This method effectively divided the signal into overlapping segments, allowing for individual analysis of each one.

2.3.3 Speech Rate

The analysis of intervals between detected vowel peaks was crucial in evaluating the participants' speech rate and the consistency of their syllable repetition. Initially, the speech rate was calculated from these peak intervals to determine the number of syllables pronounced per second. Following that, variability in speech rate was measured with higher variability, often suggesting a lack of control over speech muscles, while more consistent intervals indicated stable motor control.

2.3.4 Detection of Initial Burst

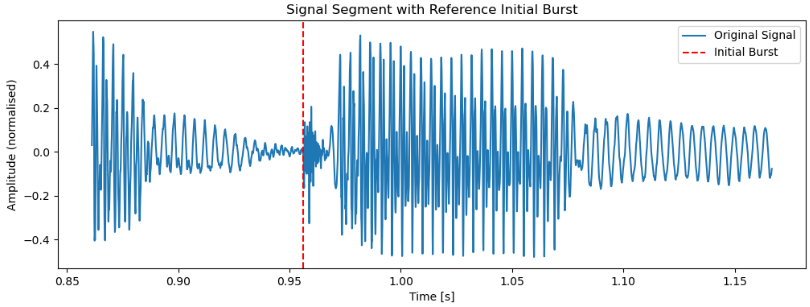

Figure 2.3: A signal segment showing an initial burst of a vowel within a syllable sequence. The initial burst is characterised by a distinct high-frequency spike which occurs due to the release of built-up pressure as the articulators position themselves to produce the vowel.

The initial burst, the first event detected in the analysis, is marked by a distinct, high-frequency spike that occurs just before the full onset of each vowel in the sequence and is marked in Figure 2.3. Each burst arises from the release of built-up pressure as the articulators move into position to produce the next vowel sound, resulting in a brief, intense burst of sound energy. These bursts serve as indicators that provide valuable insights into the timing and precision of syllable articulation within the speech sequence.

To effectively analyse the initial bursts preceding each vowel in the /ba/-/da/-/ga/ sequence, the signal first underwent a high-pass filtering process. This was essential to emphasise the high-frequency components indicative of the bursts and eliminate lower-frequency noise. A Butterworth filter with a 1500 Hz cutoff was used due to its efficiency in isolating these frequencies and its minimal phase distortion, ensuring that the crucial characteristics of the bursts were preserved.

After filtering, the signal was converted into the frequency domain using a spectrogram, which provides a visual representation of the spectrum of frequencies as they vary with time, as illustrated in Figure 2.4. The spectrogram itself was then subjected to further processing using a thresholding technique. Specifically, a threshold matrix was applied such that any values in the spectrogram falling below 80% of the mean value of each frequency bin were set to zero [11]. This step effectively enhanced the visibility of higher-frequency components by suppressing lower amplitude frequencies, thereby emphasising the significant features relevant for further analysis.

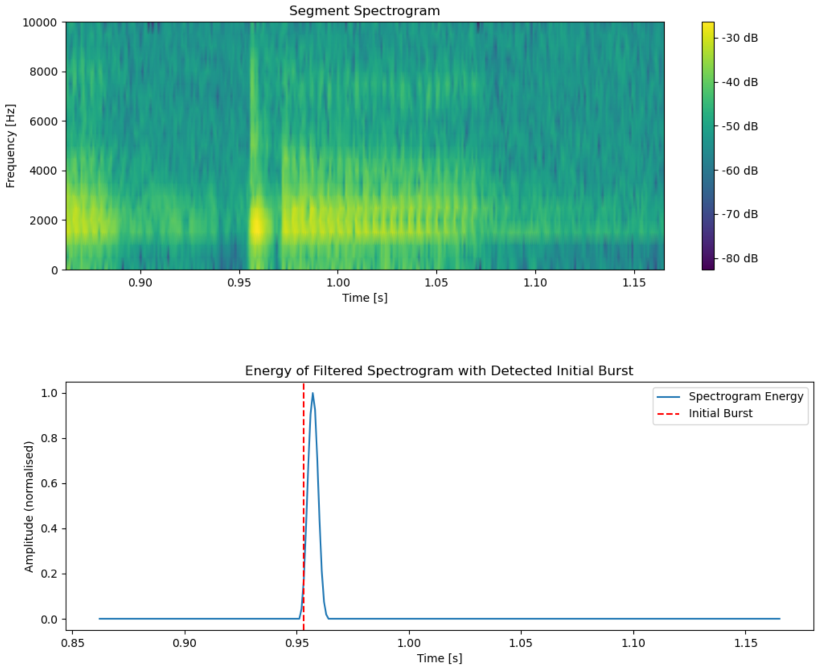

Figure 2.4: Analysis of the Initial Burst in a Speech Signal Segment: The first plot displays the spectrogram of a signal segment, highlighting frequency variations over time, while the second one shows the energy plot of a filtered spectrogram, pinpointing the beginning of a significant peak that marks the initial burst.

The signal was reconstructed from the pictogram by summing the upper half, which contains frequencies high enough to capture the initial bursts. This approach helped emphasise peaks in the signal that correspond to these bursts. After this, the signal was smoothed with a moving average and normalised to even out the amplitude, making the data clearer and more consistent for analysis.

From the reconstructed signal, the peak corresponding to the initial vowel burst could then be detected. Since this burst occurs before the vowel onset, the search focused on peaks preceding the vowel peak within the specific signal segment. Additionally, because high-frequency peaks from the previous vowel can also be present at the beginning of the segment, the search for the initial burst started at the midpoint between the two vowel peaks. All peaks from the midpoint and before the vowel peak time of the current syllable were identified and examined. The first peak detected in this sequence was identified as the initial burst. To confirm the accuracy of this detection, it was essential to later compare this identified burst with the voice onset and occlusion times of the previous syllable, ensuring that the correct peak was pinpointed.

2.3.5 Detection of Voice Onset

To accurately detect the onset of a vowel in the signal, the analysis primarily focused on regions preceding the vowel's peak energy. Given that vowels typically exhibit higher energy compared to consonants, pinpointing the onset required careful examination of the signal leading up to this peak. The detection process commenced by computing the energy envelope of the signal with a window length of 5 ms. This specific window length was chosen to preserve the visibility of individual glottal pulses, which are crucial for setting an accurate threshold for detecting vowel onset.

Following this initial calculation, the segment of the signal was normalised, and peaks corresponding to the glottal pulses were identified. These steps were essential as they enhanced the signal's characteristics, making the bursts of energy that mark the vowel onset more discernible. Detecting these peaks accurately is critical because they reflect the precise moments when the vocal folds start to vibrate, marking the beginning of vowel articulation.

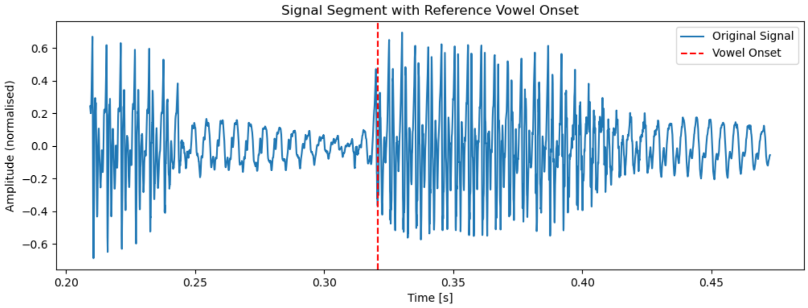

Figure 2.5: The original speech signal with the indicated vowel onset. This speech event is characterised by a noticeable increase in energy, marking the transition where vocal fold vibration begins, leading to the production of vowel sounds.

The peaks in the signal segment leading up to the previously identified vowel centre were analysed to determine the vowel onset. This analysis involved reviewing all peaks before the vowel centre, focusing on their amplitudes. A peak exceeding half of the maximum vowel amplitude was initially identified as the potential start of the onset.

However, to reduce the influence of random spikes in energy, a verification process was implemented. It was confirmed that the marked onset was preceded by two consecutive glottal pulses, each with an amplitude less than 50% of the maximum, and immediately followed by two consecutive glottal pulses, each at or above 50% of the maximum amplitude. The vowel onset was then definitively marked at the first peak within this higher amplitude region, ensuring accurate detection of the true beginning of the vowel sound. This method helps to ensure that the detected onset is not merely a random fluctuation but a consistent increase in vocal activity, signalling the start of vowel articulation.

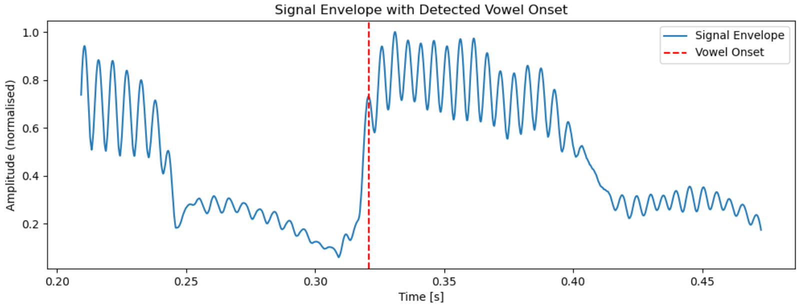

Figure 2.6: The signal envelope of a speech segment, marking the vowel onset. Yhis event is identified by a sharp rise in amplitude, illustrating the transition from lower to higher energy levels, which indicates the start of a vowel sound.

2.3.6 Detection of Occlusion

The occlusion, marking the end of a vowel, is characterised by a gradual decrease in signal energy. This is in contrast to the sharper changes seen at the vowel's onset. Due to this subtler shift, detecting the occlusion requires a more targeted approach than simply applying a universal threshold. To address this, techniques such as dynamic thresholding were used to highlight and track these gradual changes, ensuring that the precise moment of occlusion is accurately identified in the analysis.

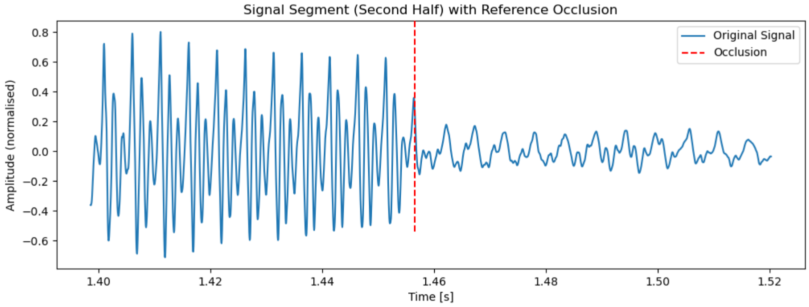

Figure 2.7: The second half of a speech signal segment with the marked occlusion. This event, indicating a moment where the articulators close to block the airflow, is evident from the abrupt change in the signal amplitude.

The first step in detecting the occlusion involved filtering the signal to highlight the frequency components that are key for accurately identifying the vocal activity. Unlike the analysis of initial vowel bursts, which targeted higher frequencies, the detection of the occlusion primarily focused on lower frequencies. To achieve this, a low-pass Finite Impulse Response (FIR) filter was employed. This filter was designed using a Hamming window to minimise the effects of spectral leakage, which helps in maintaining the integrity of the frequency components near the cutoff edge. The cutoff frequency was set at 1500 Hz, effectively suppressing frequencies higher than this threshold while preserving those lower, which are significant for capturing the subtle changes that indicate occlusion. This selective filtering ensured that only the essential lower-frequency components were retained for further analysis, enhancing the accuracy of detecting the onset within the vocal sequence.

Following the application of the low-pass filter, the next step involved squaring the filtered signal. This transformation amplified variation in the signal's amplitude, particularly emphasising the decreasing energy typical for an occlusion. By squaring the signal, subtle fluctuations in amplitude became significantly more pronounced, providing easier identification of the gradual decline that marks the end of a vowel.

Once the signal has been squared, it underwent polynomial fitting. A ninth-order polynomial was fit to the squared signal, as it was previously shown to be a sufficient approximation [11], starting from the vowel peak time identified earlier. This polynomial fitting was used to smooth out any noise and irregularities, providing a clearer view of the underlying trends in the signal. Additionally, it helped to set a dynamic threshold for detecting the occlusion, which adjusted better to the complexities of vocal fading compared to a fixed threshold.

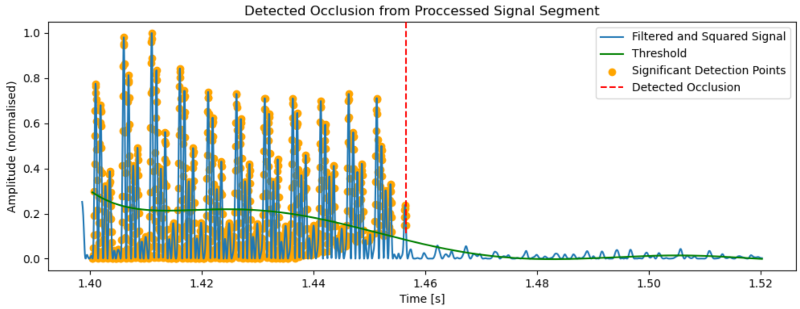

Figure 2.8: Filtered and squared signal used to identify vocal occlusion. The green curve represents a dynamic threshold adapted to changes in the signal, enhancing the detection accuracy. Yellow points represent points compared against the threshold, and the red dashed line marks the detected occlusion, pinpointing where the examined points drop below this threshold for a longer period of time, indicating the end of vocal activity.

The dynamic threshold was calculated as the negative value of the polynomial evaluation, adjusted by twice the mean of the squared signal post-peak. This adjustment allowed the threshold to adapt to the signal's specific characteristics at the given moment, thus providing a tailored baseline against which the signal's decline could be compared. When the squared signal remained consistently above this dynamic threshold, it indicated ongoing vocal activity.

In contrast, a drop below this threshold for a longer period of time, set as 40 milliseconds, signified occlusion, marking the termination of vocalisation.

The exact point of occlusion was determined by identifying the first significant gap in indices where the squared signal values dropped below the dynamic threshold. This gap, defined as being greater than 2% of the sampling rate, signified a substantial decrease in signal energy, marking the end of vocal activity. This method provided a precise and reliable means of detecting occlusion, which is crucial for speech assessment in the DDK task.

2.3.7 Detection of Voicing

The final event detected in the analysis was the onset of the voicing segment, visible in Figure 2.9, that precedes each consonant in the sequence of /ba/-/da/-/ga/. This characteristic differs from tasks involving unvoiced syllables such as /pa/-/ta/-/ka/, as discussed in the study by Novotný et al.[11]. Identifying the precise onset of voicing is especially challenging in sequences like /ba/-/da/-/ga/ because the voiced segments directly follow vowels. This continuous alternation between vowels and consonants with minimal interruption complicates the detection of where exactly the consonant voicing begins.

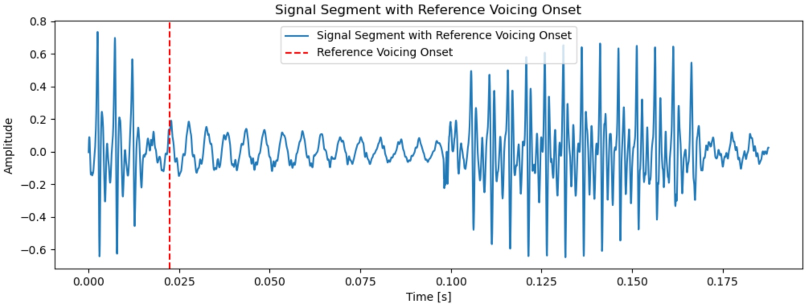

Figure 2.9: The dashed red line indicates the onset of the voicing segment, directly following vowels and preceding the consonants in the /ba/-/da/-/ga/ syllables sequence.

To address the challenge of detecting voicing onset in the sequence of /ba/-/da/-/ga/, the focus turned on the analysis of two indicators: zero-crossing rate (ZCR) and energy. The ZCR, which counts how frequently the signal crosses the baseline, helps to pinpoint moments of transition between vowels and consonants. This measure is particularly useful when combined with the calculation of short-time energy, which assesses the power within short segments of the audio signal and highlights areas of significant vocal activity.

After initial calculations, both the ZCR and short-time energy were smoothed with a Gaussian filter to reduce noise and minimise temporary fluctuations. The smoothed signals were then normalised, enabling the application of consistent threshold levels.

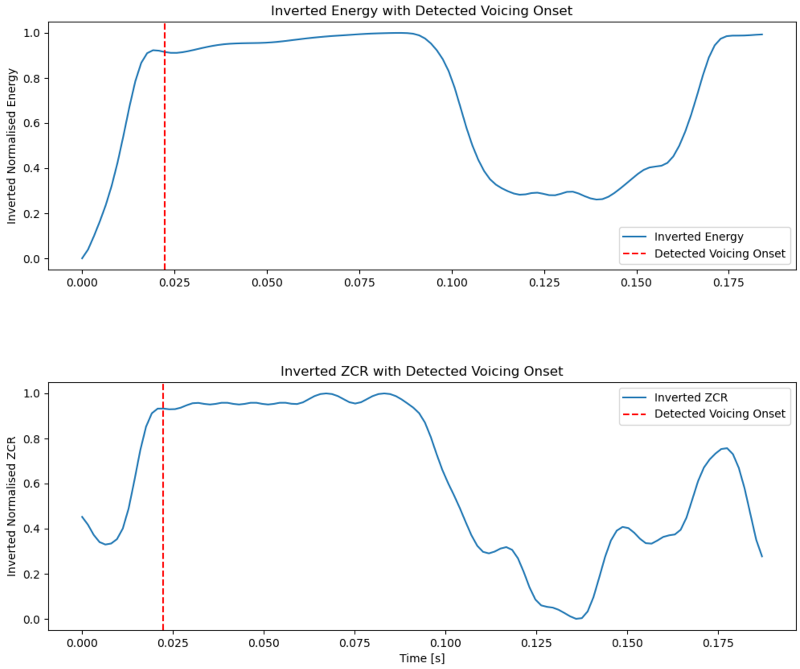

Figure 2.10: Inverted zero-crossing rate (ZCR) and inverted energy of the signal segment to determine voicing onset in the signal segment. The ZCR identifies transitions between vowels and consonants, while energy highlights significant vocal activity. The dashed red lines in both plots mark the detected voicing onset.

To accurately determine the onset of the voicing segment, both the energy and ZCR signals were analysed after smoothing and normalisation. Peaks were identified in these inverted signals, representing the troughs of the original measurements where the voicing typically starts with lower energy and ZCR values compared to the preceding vowel. The initial peaks detected below an elevated threshold in both signals were compared, and the later of these peaks was marked as the onset of voicing, as displayed in Figure 2.10. This approach ensured that the voicing was not detected prematurely, given that both energy and ZCR can momentarily fall below the threshold during the occlusion. If a detected voicing onset occurred before the occlusion of the preceding vowel, it was considered an invalid detection, ensuring only valid voicing onsets were recognised.

2.4 Data Evaluation

2.4.1 Algorithm Accuracy Evaluation

The initial phase of the data evaluation focused on comparing the timestamps of reference labels with those of detected events, specifically the initial vowel burst, vowel onset, occlusion, and the onset of voicing in the consonant. All timestamps for these labels were compared to the reference, and the average differences for each event were calculated. This analysis offered a detailed assessment of the algorithm's performance in event detection, enabling the identification of which events were accurately detected and could be reliably used in the later classification experiment.

Events that were successfully detected by the algorithm were the only ones compared with the reference. If an event was not detected, its timestamp was marked as invalid and excluded from the comparison. Across the 20 labelled recordings from each group, a total of 859 timestamps in the MS group and 1,094 in the HC group were evaluated.

2.4.2 Calculation of Speech Features

After verifying the accurate detection of events, these were utilised to calculate parameters for comparison between the two groups. These parameters included voice onset time (VOT) and vowel onset to occlusion time (VO-OT). Additionally, the diadochokinetic rate (DDT rate) and diadochokinetic fluctuation (DDK fluctuation) were derived from the timestamps of vowel peaks. Each of these calculated parameters was saved in a separate file corresponding to its respective speech recording to be further utilised in the classification segment of the project.

2.4.3 Statistics

Firstly, the data corresponding to each recording was utilised to calculate the average of all speech features for each individual. This means that if a subject had two valid recordings, the data from those recordings were averaged; however, if there was only one valid recording, then that single recording was utilised. If there was no valid data in any of the recordings, the individual was excluded from the analysis. This approach ensured that each subject had the same weight in the final statistical analysis.

Specifically, for each subject, the following features were extracted: VOT for the three syllables /ba/-/da/-/ga/, VO-OT, DDK rate, and DDK fluctuations. If there was insufficient data for one of these features, the rest could still be utilised in the final statistics, resulting in a slight variance in the number of participants contributing to the calculation of each feature. Once the data were calculated for each individual, it was averaged across the entire group for the final result.

Once the mean values for both groups were calculated and compared, statistical tests were employed to establish the significance of differences. Specifically, a two-sample t-test was used to compare the groups' means, with results including the t-statistic, measuring the difference size relative to sample variation, and the p-value, indicating the probability of the difference occurring by chance. Additionally, the number of subjects contributing to each feature's statistical test was reported, ensuring clarity in the sample size for comparison.

2.5 SVM training

The data extracted from each recording were used to train a Support Vector Machine (SVM) classifier. This classifier was designed to differentiate between the HC and MS groups based on key speech features. These features include the VOT for syllables /ba/, /da/, and /ga/, as well as the VO-OT, DDK rate, and DDK fluctuation. These features were chosen because the events used to calculate them were detected with high accuracy. Additionally, they have the potential to uncover significant differences in speech between the HC and MS groups, providing insights into the neurological impact of MS on speech.

For this analysis, each participant's file was checked to exclude any with missing data, ensuring the model was trained on complete and accurate information. The features extracted from these files were standardised to ensure uniform treatment across all data points. To ensure fair representation of both groups and to reduce model bias toward the more prevalent MS group, the dataset was manually balanced beforehand by duplicating all entries for the HC group, given that there was approximately twice as much data for the MS group.

The dataset was randomly divided into training and testing subsets, with 90% allocated for training and 10% for testing. This process was repeated 10 times to ensure robustness. Additionally, different cross-validation methods, specifically 3-fold, 5-fold, and 7-fold cross-validation, were used to test the model's stability and efficiency.

Chapter 3

Results

3.1 Algorithm Performance

The average differences between manually labelled events and those detected by the algorithm are presented in Table 3.1.

Event	HC (ms)	MS (ms)
Beginning of Voicing	10 ± 14	20 ± 64
Initial Burst	2 ± 2	2 ± 2
Vowel Onset	3 ± 2	4 ± 3
Occlusion	4 ± 10	4 ± 5

Table 3.1: Average differences between reference labels and detected labels by the algorithm.

The algorithm effectively detected the initial burst, vowel onset, and occlusion in both the HC and MS groups, with success rates of 94.8%, 77.8%, and 88.0% for the HC group, and 94.7%, 74.6%, and 79.8% for the MS group respectively, when using a 5 ms accuracy threshold. However, the detection of the beginning of the voicing segment was considerably less accurate in both groups. Specifically, the MS group's success rate was markedly low at 38.9%, and in the HC group, it was 39.8%. The number of compared voicing segments was also considerably lower than any other event due to the high number of invalid detections. Given the inadequate performance in voicing segment detection, this component was not included in the further analysis across both groups.

3.2 Articulatory features comparison

Feature	HC (ms)	MS (ms)	t-value	p-value
VOT /ba/	12.3 ± 3.2	14.8 ± 6.1	-3.1	<0.005
VOT /da/	15.1 ± 6.5	17.6 ± 8.0	-2.0	=0.05
VOT /ga/	21.5 ± 7.0	22.8 ± 8.8	-0.9	=0.4
VO-OT /all/	82.7 ± 15.5	89.7 ± 14.7	-2.5	=0.01
Performance Metrics	**HC**	**MS**	**t-value**	**p-value**
DDK Rate (syll/s)	6.54 ± 0.63	6.01 ± 0.77	4.4	<0.001
DDK Fluctuation (%)	20.3 ± 8.5	26.5 ± 9.4	-3.9	<0.001

Table 3.2: Articulatory features calculated for HC and MS groups with t-values and p-values.

The articulatory features extracted from the data analysis are detailed in Table 3.2. For this analysis, only signals where the number of syllables was divisible by three were used, as this configuration was more likely to ensure that all syllables were correctly assigned to the corresponding syllable from the /ba/-/da/-/ga/ sequence. Notable differences were observed across all calculated parameters between the groups, with the VOT being consistently longer in the MS group for all analysed syllables, indicating a slight delay in the activation of vocal mechanisms compared to the HC group. Specifically, the average VOT for the syllable /ba/ was longer by 2.5 ms (t(132)=-3.1, p<0.005), for /da/ by 2.5 ms (t(136)=-2.0, p=0.05), and for /ga/ by 1.3 ms (t(135)=-0.9, p=0.4). The table also shows that the standard deviations are more pronounced for the MS group, particularly for the /da/ and /ga/ syllables, with standard deviations of 8.0 ms and 8.8 ms, respectively, indicating greater variability among patients with multiple sclerosis. The VO-OT measurements in the MS group were also slightly elevated (t(138)=-2.5, p=0.01), suggesting prolonged durations in the transition phases between syllables.

Further analysis of the performance metrics showed that the HC group could repeat the syllable sequence faster than the MS group. On average, the HC group repeated 6.54 syllables per second, while the MS group managed slightly fewer, averaging 6.01 syllables per second, meaning the average difference was 0.53 syllables per second (t(139)=4.4, p<0.001). Additionally, the HC group showed slightly less variability in the intervals between individual syllables (t(139)=-3.9, p<0.001), indicating a more stable and consistent speech production pattern.

3.3 SVM Classification

The SVM classifier demonstrated consistent performance across 10 random training-testing splits and various cross-validation methods. The average accuracy rates were as follows: $70.5 \pm 4.4\%$ for 3-fold, $70.6 \pm 7.6\%$ for 5-fold, and $70.7 \pm 7.8\%$ for 7-fold cross-validation. The best-performing model from each split was then evaluated on the test data, achieving an average accuracy of $70.4 \pm 7.8\%$.

Chapter 4

Discussion

This project introduced an automated approach to detect articulatory events in diadochokinetic tasks among patients with multiple sclerosis, with a specific focus on voiced syllables—a shift from past studies that primarily concentrated on unvoiced syllables. This shift to focusing on voiced syllables introduced unique challenges, particularly in detecting the voicing segment that precedes the vowel. Despite these complexities, the SVM classification achieved an impressive success rate of 70.4%, which can be compared to the 70.6% success rate reported by Rozenstoks et al. in their study, which utilised a CNN-based approach to detect voiced syllables in the DDK task and used calculated repetition paradigms to classify multiple sclerosis patients. [10].

This success rate in distinguishing between the MS and HC groups is particularly remarkable given the complex nature of multiple sclerosis, which is characterised by its variable stages, including RRMS and progressive forms, which lead to fluctuating symptoms over time. The patients involved in this project were all free from relapses for at least 30 days before their recordings were taken, meaning that their symptoms were not in their most severe or noticeable phase, making the task of distinguishing them based on speech patterns more challenging. The ability of the algorithm to accurately differentiate in this context highlights its sensitivity and potential usefulness in clinical settings.

Statistical analyses showed significant differences between HC and MS groups in several features. Notably, the DDK rate ($t = 4.4$, $p < 0.001$), DDK fluctuation ($t = -3.1$, $p < 0.001$), and the VOT for /ba/ ($t = -3.1$, $p < 0.005$) suggested a potential alteration in articulatory timing in MS patients. The VOT for /da/ ($t = -2.0$, $p = 0.05$) and VO-OT ($t = -2.5$, $p = 0.01$) also suggested some impairment, though less definitively. However, the VOT for /ga/ ($t = -0.9$, $p = 0.4$) was not significantly different, indicating variability in sensitivity. These results suggest that while DDK measures and some

VOT parameters could be reliable indicators of speech dysfunction in MS, a comprehensive assessment is necessary.

The event detection algorithm demonstrated considerable effectiveness, achieving an average success rate of 94.7% in the less-performing MS group for detecting the initial burst, 74.6% for vowel onset, and 79.8% for occlusion, with a threshold of 5 ms. It performed particularly well in identifying the initial burst from the signal spectrogram, which is typically very distinct, allowing precise localisation due to a pronounced shift in frequencies. Interestingly, the success rate for detecting the vowel onset was lower than for the occlusion, despite the onset of a vowel typically showing a clearer and quicker change in energy. This lower performance might be due to the more complex detection techniques used for occlusion, where squaring the signal to make energy changes more visible and applying a dynamic threshold were key to successful detection. For future analysis, applying similar techniques to vowel onset detection could likely enhance its accuracy.

On the other hand, the detection of the beginning of the voicing segment posed significant challenges, with a notably lower overall accuracy of 39.4%. This lower accuracy can be attributed to the method and threshold settings selected for this detection. Occasionally, the energy and the ZCR would temporarily drop below the set threshold too soon, causing the event to be detected before the actual occlusion of the preceding vowel. This timing difference led to a higher rate of invalid detections compared to other detected events. Moreover, the ZCR frequently fluctuated at the onset of the voicing segment, which often resulted in detections being significantly delayed. To enhance detection accuracy, implementing a dynamic threshold that considers both the energy level and the zero-crossing rate might provide a more reliable identification of the voicing segment. Given these challenges, if improvements were made to accurately detect the onset of voicing, it could then become an important parameter that might potentially enhance the SVM classification process, thereby improving the model's overall predictive accuracy.

It is also important to acknowledge certain limitations of the project. Firstly, the detection of voicing onset was shown to be unreliable, which prevented further analysis of features related to this event, such as the length of the voicing segment before the vowel onset. Accurate detection of the voicing segment's length could add another parameter to the classification task, potentially improving the results if differences are found between the HC and MS groups. Additionally, better voicing onset detection could allow for a more thorough analysis of speech patterns, uncovering subtle variations that help differentiate these groups. This could lead to more accurate diagnostics and better-targeted treatments. Furthermore, as all participants were Czech speakers, the findings may not generalise to other languages regarding the speech features of voiced consonants between healthy individuals and those with multiple sclerosis.

The following goals were addressed in this project:

1. The primary aim was to explore the less-studied area of dysarthria in MS patients, focusing on voiced syllables. This was achieved by reviewing the literature and understanding the specific speech challenges faced by MS patients.

2. The project included a review of current methods for assessing speech in DDK tasks. These techniques were identified and compared, providing a foundation for using existing methods in a new context.

3. An automatic approach was proposed and implemented for event detection in voiced plosives in DDK tasks. In the less-performing MS group, the event detection algorithm had high accuracy: 94.7% for initial burst detection, 74.6% for vowel onset, and 79.8% for occlusion, with the success rates being even higher for the HC group. Challenges in detecting the voicing segment were noted, with suggestions for future improvements.

4. The disruption of voiced plosives in MS patients was analysed. An SVM-based classification algorithm was developed, demonstrating significant effectiveness with a 70.4% success rate. Additionally, all calculated parameters were compared between the HC and MS groups using statistical tests. The most significant differences were observed in the DDK rate ($t = 4.4$, $p < 0.001$), DDK fluctuation ($t = -3.9$, $p < 0.001$), and the VOT for the syllable /ba/ ($t = -3.1$, $p < 0.005$).

Chapter 5

Conclusion

This study explored the challenges of detecting dysarthria in MS patients by introducing an automated approach for identifying articulatory events in DDK tasks. The proposed event-detection algorithm demonstrated high reliability in pinpointing speech events in voiced syllables, achieving significant accuracy rates. Additionally, an SVM-based classification experiment was conducted, resulting in an overall success rate of 70.4% in distinguishing between MS patients and healthy controls. The speech features calculated from these detected events showed great potential for assessing and monitoring the progression of MS. These findings demonstrate the algorithm's clinical relevance and potential usefulness in medical settings. This project provides a foundation for further advancements in speech analysis for better diagnosis of neurodegenerative diseases.

Bibliography

[1] Walton, C., King, R., Rechtman, L., et al. *Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition.* Multiple Sclerosis (Houndmills, Basingstoke, England). 2020 Dec;26(14):1816-1821.

[2] Dobson, R., Giovannoni, G. *Multiple sclerosis - a review.* Eur J Neurol. 2019 Jan;26(1):27-40.

[3] McGinley, M.P., Goldschmidt, C.H., Rae-Grant, A.D. *Diagnosis and Treatment of Multiple Sclerosis: A Review.* JAMA. 2021;325(8):765–779.

[4] Hosseini, Z., Homayuni, A. Etemadifar, M. *Barriers to quality of life in patients with multiple sclerosis: a qualitative study.* BMC Neurol 22, 174 (2022).

[5] Robertson, D., Moreo, N. *Disease-Modifying Therapies in Multiple Sclerosis: Overview and Treatment Considerations.* Fed Pract. 2016 Jun;33(6):28-34.

[6] Aubert-Broche, B., Fonov, V., Narayanan, S., et al. *Onset of multiple sclerosis before adulthood leads to failure of age-expected brain growth.* Neurology. 2014;83:2140–2146.

[7] Sinay, V., Perez Akly, M., Zanga, G., Ciardi, C., Racosta, J.M. *School performance as a marker of cognitive decline prior to diagnosis of multiple sclerosis.* Mult Scler. 2015 Jun;21(7):945-52. doi: 10.1177/1352458514554054. Epub 2014 Oct 24. PMID: 25344372.

[8] Amato, M.P., Hakiki, B., Goretti, B., et al. *Association of MRI metrics and cognitive impairment in radiologically isolated syndromes.* Neurology. 2012;78:309–314.

[9] Noffs, G., Perera, T., Kolbe, S.C., Shanahan, C.J., Boonstra, F.M.C., Evans, A., Butzkueven, H., van der Walt, A., Vogel, A.P. *What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis.* Autoimmun Rev. 2018 Dec;17(12):1202-1209.

[10] Rozenstoks, K., Novotny, M., Horakova, D., Rusz, J. *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. 2020 Jan;28(1):32-41.

[11] Novotný, M., Rusz, J., Cmejla, R., Růžička, E. *Automatic Evaluation of Articulatory Disorders in Parkinson's Disease.* IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014;22:1366-1378.

[12] Rusz, J., Benova, B., Ruzickova, H., Novotny, M., Tykalova, T., Hlavnicka, J., Uher, T., Vaneckova, M., Andelova, M., Novotna, K., Kadrnozkova, L., Horakova, D. *Characteristics of motor speech phenotypes in multiple sclerosis.* Mult Scler Relat Disord. 2018 Jan;19:62-69. doi: 10.1016/j.msard.2017.11.007. Epub 2017 Nov 8. PMID: 29149697.

[13] Kurtzke, J.F. *Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS).* Neurology. 1983;33(11):1444-1452.