CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Electrical Engineering
Department of Radioelectronics

Analysis of Oral Diadochokinesis in Progressive Neurological Diseases via Automated Acoustic Analysis

Analýza orální diadochokineze u progresivních neurologických onemocnění pomocí automatizované akustické analýzy

Master's thesis

Study programme: Electronics and Communications
Branch of study: Media and Signal Processing

Supervisor: Doc. Ing. Jan Rusz, Ph.D., Department of Circuit Theory, FEE

Bc. Jan Melechovský

Praha 2020

MASTER'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Melechovský Jan** Personal ID number: **435031**

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

Department / Institute: **Department of Radioelectronics**

Study program: **Electronics and Communications**

Branch of study: **Media and Signal Processing**

II. Master's thesis details

Master's thesis title in English:

Analysis of Oral Diadochokinesis in Progressive Neurological Diseases via Automated Acoustic Analysis

Master's thesis title in Czech:

Analýza orální diadochokineze u progresivních neurologických onemocnění pomocí automatizované akustické analýzy

Guidelines:

1. Perform the state-of-the-art in the field of motor speech disorders in progressive neurological diseases and digital speech processing.
2. Based on the available literature, explore the suitable methods for parametrization of the /pa/-/ta/-/ka/ syllable repetition task (so-called oral diadochokinesis) in patients with neurological disorders.
3. Design the algorithm for the automatic evaluation of fundamental aspects of dysarthria based on acoustic analysis of oral diadochokinesis.
4. Using developed methods, perform statistical analyses, and select suitable parameters for differentiation between investigated neurological disorders.

Bibliography / sources:

[1] Novotny M, Rusz J, Cmejla R, Ruzicka E. Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM T Audio Speech Lang Process 2014;22:1366-1378.
[2] Kim Y, Kent RD, Weismer G. An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. J Speech Lang Hear Res 2011;54:417–429.

Name and workplace of master's thesis supervisor:

doc. Ing. Jan Rusz, Ph.D., Department of Circuit Theory, FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **13.09.2019** Deadline for master's thesis submission: **22.05.2020**

Assignment valid until: **19.02.2021**

doc. Ing. Jan Rusz, Ph.D.	doc. Ing. Josef Dobeš, CSc.	prof. Mgr. Petr Páta, Ph.D.
Supervisor's signature	Head of department's signature	Dean's signature

III. Assignment receipt

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

. | Student's signature

Date of assignment receipt

Affidavit

I declare that I completed the presented thesis independently and that all used sources are quoted following the Methodical Guidelines that cover the ethical principles for writing an academic thesis.

Čestné prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Dne 22. května 2020 v Praze

…….

Jan Melechovský

Acknowledgement

I want to thank my supervisor Doc. Ing. Jan Rusz, Ph.D., for providing guidance and being very helpful. I would also like to acknowledge Ing. Michal Novotný, Ph.D., and Ing. Tereza Tykalová, Ph.D., for their willingness to share useful knowledge with me. Last but not least, I want to thank my family for supporting me throughout my whole university studies.

Abstract

This thesis deals with an automated assessment of neurodegenerative diseases by acoustic speech analysis using oral diadochokinesis. Two variants of speech segmentation algorithm necessary for the diadochokinetic task are designed. Their performance is evaluated and compared to previously designed algorithms from the literature. Furthermore, the speech segmentation algorithm is used to extract features from the utterances. The features are evaluated in terms of significance, and a simple classifier is built to distinguish between the neurodegenerative diseases. Finally, we discuss the results and make proposals for future work.

Keywords: oral diadochokinesis, neurodegenerative disease, dysarthria, convolutional neural network, speech segmentation, feature extraction, classification

Abstrakt

Tato práce se zabývá problematikou automatické klasifikace neurodegenerativních onemocnění pomocí akustické analýzy orální diadochokineze. Dvě varianty algoritmu pro segmentaci řeči, který je nedílnou součástí pro vyhodnocení diadochokineze, jsou navrženy. Jejich přesnosti jsou porovnány a ten lepší je porovnán s předešlými algoritmy. S využitím tohoto algoritmu jsou pak z promluv pacientů extrahovány příznaky, jejichž významnost je vyhodnocena. Posledním krokem je klasifikace onemocnění s pomocí jednoduchého klasifikátoru. Práce je zakončena diskuzí nad výsledky a návrhy k budoucí práci.

Klíčová slova: orální diadochokineze, neurodegenerativní onemocnění, dysartrie, konvoluční neuronová síť, segmentace řeči, extrakce příznaků, klasifikace

Contents

List of figures

List of tables

List of abbreviations

AMR	–	Alternate motion rates
ANOVA	–	Analysis of variance
CNN	–	Convolutional neural network
DBS	–	Parkinson's disease (patient) with Deep brain stimulation
DDK task	–	Diadochokinetic task
DEVO	–	Direct EVO (algorithm)
E position	–	Position of an explosive
EVO	–	Explosive, Vowel onset, Occlusion (positions)
HC	–	Healthy control (group)
HD	–	Huntington's disease
IoR	–	Index of Rhythmicity
IS	–	Intensity Slope
MS	–	Multiple sclerosis
MSA	–	Multiple system atrophy
O position	–	Position of vowel occlusion
PD	–	Parkinson's disease
R-CNN	–	Region (proposal) CNN
RA	–	Rhythm Acceleration
RI	–	Rhythm Instability
RIRV	–	Relative Intensity Range Variation
SDEVO	–	Syllable Detection into EVO detection (algorithm)
SMR	–	Sequential motion rates
SR	–	Syllable Rate
sVOT	–	Standard deviation of Voice Onset Time
V position	–	Positon of vowel onset
VL	–	Vowel Length
VOT	–	(mean) Voice Onset Time

1. Introduction

Neurodegenerative progressive diseases and their treatment are a matter of increasing urgency in nowadays world. This has to do with the fact that the population is ageing, and the incidence of neurodegenerative diseases increases with age. Thus, a vast majority of people suffering from the diseases are middle-aged or of old age. With some of the diseases, we still do not know what exactly causes them and whether they could be cured if we knew their source. In connection with these facts, researchers are trying to gather more and more information about the diseases.

One of the markers of a neurodegenerative disease is motor speech impairment, known as dysarthria. This speech impairment is reported to be noticeable even in the early stages of a neurodegenerative process. Due to the fact that speech tests can be performed with relative ease, repeated many times, and can be performed even in distant form, speech tests make for a promising candidate for early disease detection that could lead to helping us discover the source of the diseases. The possibility of distant evaluation may be the most important, as doctors nowadays are very busy, and in low numbers, which makes regular check-ups difficult to make frequent.

Another important factor is that physicians can evaluate speech impairment by perception quite insufficiently; usually, they only assess the severity of patient's dysarthria on a scale of 0 to 4. Additionally, each physician evaluates the speech subjectively, which can lead to inconsistent evaluations. An automated algorithm would solve both these problems by analysing many more features and providing machine-like consistency. The algorithm could then be used as a supportive tool for physicians in clinical practice.

The topic of assessment of neurodegenerative diseases from patient utterances has been dived into dozens of times, although usually only one disease was considered, and compared to a healthy control group. There are just a few works that deal with multiple diseases at once, or multiple types of dysarthria at once.

In this thesis, we will focus on rhythmic speech – a functional task called diadochokinetic task – repeating of prescribed patterns by the patient. For that purpose, a speech segmentation algorithm will be implemented.

In this chapter, we will explain some key terms and introduce the state-of-the-art.

1.1. Dysarthria

Dysarthria is a motor speech disorder characterised by poor articulation. It is a result of neurological injury of the motor component of the motor-speech system. There are multiple types of dysarthria based on the level of brain damage. Patients can suffer from more than one type simultaneously, which is often referred to as mixed dysarthria.

The distinguished types of dysarthria relevant to this thesis are the following [3]:

- Spastic – caused by damage of bilateral upper motor neuron; in speech manifests as harsh, strained-strangled phonation, low-pitch, short phrases, imprecise articulation, slow rate of speech, abnormal prosody with monopitch and monoloudness
- Ataxic – caused by damage to the cerebellum or nerve pathways connecting the cerebellum with rest of the nervous system; manifests as drunken-like speech with

imprecise articulation and irregular articulatory breakdowns, distorted vowels, prolonged phonemes, slow rate, abnormal prosody, and harsh vocal quality

- Hypokinetic – caused by damage to basal ganglia control circuits; primary characteristics are muscle rigidity and reduced range and force of movements; in speech manifests as reduced loudness, rapid speech, short rushes of speech, abnormal prosody, reduced facial expression, mumbling
- Hyperkinetic – caused by damage to basal ganglia control circuits; primary characteristics involve involuntary, abnormal movements; in speech manifests as unexpected inhalations and exhalations, irregular articulatory breakdowns, harsh voice, voice tremor, excess loudness, hypernasality, spasms

1.2. Neurodegenerative diseases

Neurodegenerative diseases are diseases that gradually affect the function of the nervous system of specific populations of neurons [4]. This happens in response to various conditions, including inflammation. It is not yet clear whether inflammatory changes within the brain tissue are a consequence or a primary cause of brain damage. There are multiple neurodegenerative diseases, but we will only list out the ones considered in this thesis.

1.2.1. Parkinson's disease

Parkinson's disease (PD) is an idiopathic, progressive neurodegenerative disease affecting the central nervous system and is characterised by a decrease of dopaminergic neurons in substantia nigra [5]. This leads to dysfunction of basal ganglia. Thus, patients suffer mainly from, but not restricted to, motor system disorders, such as rest tremor, muscular rigidity, postural instability, and bradykinesia (slowed movement). Besides motor impairment, patients can also experience disorders of mood, behaviour, and cognition. A large number of patients also show some form of vocal impairment (70–90 %) [6]. The common dysarthria type of PD is the hypokinetic dysarthria.

Parkinson's disease can be treated with dopaminergic agents, such as levodopa, to suppress the symptoms, i.e. help with movement, speech, tremor, etc. However, in time the efficiency of such treatment lowers until it is no longer helpful at all. Patients in these later stages of the disease can then undergo a so-called deep brain stimulation (DBS), which is brain surgery, in which electrodes are implemented into the patient's brain. These are connected to a stimulator, also known as a pacemaker, which is placed under the skin of the chest. Even though the DBS helps patients with movement and tremor, the patients often experience worse speech ability than before the surgery. Their previous hypokinetic dysarthria can be mixed with spastic, ataxic or hyperkinetic [7].

1.2.2. Multiple system atrophy

Multiple system atrophy is a rapidly progressive neurodegenerative disease characterised by parkinsonism, cerebellar, dysautonomic and pyramidal features in any combination [8]. Speech impairment develops in all MSA patients, often in early disease stages. Dysarthria type in MSA tends to be mixed with ataxic, spastic, hypokinetic components.

1.2.3. Huntington's disease

Huntington's disease (HD) is another progressive neurodegenerative disease, which is just as Parkinson's disease mainly related to basal ganglia dysfunction, but on the contrary, is of a genetic origin. The death of cells in the striatum leads to under-stimulation of the motor cortex in the direct pathway and to over-stimulation in the indirect pathway. The first mentioned

manifests as slow speed of movement in persons with HD, the latter manifests as irregular, sudden, jerky movements [9]. Patients with HD suffer from hyperkinetic dysarthria.

1.2.4. Multiple sclerosis

Multiple sclerosis is the most common demyelinating disease of the central nervous system. The patient suffers from widespread atrophy of white and grey matter, which results in the disrupted ability of the nervous system to transmit signals, manifesting as motor, sensory and cognitive impairments [10]. Patients can also experience some form of vocal impairment. The dysarthria types associated with MS are ataxic, spastic or mixed.

1.3. Diadochokinesis and Diadochokinetic task (DDK task)

Diadochokinesis in general is the ability to repeat antagonistic movements in quick succession, e.g. alternately bringing a limb into opposite positions or pronating and supinating hands. In speech pathology, diadochokinetic tasks (DDK tasks) are performed to measure the speed necessary to stop a determined motor impulse and start another of the opposite nature [11]. They are referred to as alternate motion rates (AMR) or sequential motor rates (SMR), the latter being more complicated than the first. In AMR patient is required to repeat one syllable, while in SMR they are required to repeat a sequence of syllables.

A very common SMR is to repeat syllables /pa/-/ta/-/ka/ in alternating rapid sequence as constant and as long as possible on one breath [12]. This syllable sequence consists of so-called plosive consonants, also known as stop consonants or plosives, and vowels.

Plosives are a group of consonants in which the vocal tract is blocked so that the airflow stops. The stopping can be done by the tongue tip or blade in /t/ and /d/, tongue body in /k/ and /g/, and lips in /p/ and /b/. A subcategory of plosives is ejectives, also known as explosives, which include /p/, /t/ and /k/. The fact that each of these stop consonants is quite the opposite in the way they are produced requires good muscle coordination and might prove as a difficult task for patients with neurodegenerative progressive diseases.

A common mistake can occur when instructing patients for this DDK task is not emphasising on the fact that there should be no pauses not only inside the /pa/-/ta/-/ka/ syllable train but also in between these trains. A correct utterance should be a constant stream of syllables that alternate in the predetermined /pa/-/ta/-/ka/ pattern, e.g. /pa/-/ta/-/ka/-/pa/-/ta/-/ka/-/pa/, etc., but not /pa/-/ta/-/ka/-/pause/-/pa/-/ta/-/ka/.

In the DDK task evaluation, we are not only interested in the syllable positions alone, but in each of the syllables, we detect so-called EVO positions. The abbreviation EVO stands for explosive (plosive consonant), vowel (vowel onset) and occlusion (the end of a vowel).

In [13], an overview of speech characteristics among motor speech disorders is given based on dysarthria type. For both AMRs and SMRs, four important characteristics are listed. Slow and regular AMRs are a prominent marker of spastic dysarthria and less prominent, yet possible marker of hypokinetic dysarthria. Irregular AMRs are significant in ataxic and hyperkinetic dysarthrias. Rapid, blurred AMRs are prominent for hypokinetic dysarthria. Finally, slow and irregular AMRs are typical in hyperkinetic dysarthria and less prominent but possibly expected in ataxic dysarthria.

1.4. Automated evaluation of oral diadochokinesis in dysarthrias

A couple of works have tried to describe the relations between neurodegenerative diseases, dysarthria types, speech severity and acoustic features extracted from the speech. In [14], a study of oral diadochokinesis in 4 different motor disorders was performed. They report that the features used (syllable rate, syllable duration, intratrain variation coefficient of syllable durations, and percentage of incomplete stop consonant occlusions) provide a sensitive measure of orofacial motor impairment, e.g. syllable rate distinguished healthy controls from three of the disorders significantly. A more recent study by [2] compared acoustic features in relation to neurological diseases, dysarthria type and dysarthria severity. In that study the speech recordings comprised of words and sentences, which does not allow for a good evaluation of rhythm features. Nevertheless, they achieved classification accuracies of 68.6 %, 54.9 %, and 31.7 % for aetiology, dysarthria type, and dysarthria severity, respectively. Significant contributors in terms of features were, e.g. articulation rate, fundamental frequency range, or voiceless interval duration.

In this study, we will focus on the DDK task, a rhythmic task, therefore we need to have a speech segmentation algorithm that will detect syllable positions to allow for feature extraction. Such algorithms were implemented in [1, 15, 16, 17, and 18]. The algorithm in [1] is based on traditional signal processing, and it was designed for PD patients and HC (Healthy Control) group. It is able to detect not only syllables but also the EVO positions. The algorithm in [15] is an improved version of the one in [16] and is also based on traditional signal processing. It detects both syllables and EVO positions and its design is not focused on any group, on the contrary to [1]. The algorithm in [17] uses bandpass filtering to detect syllables but does not detect the EVO positions. The algorithm in [18] is based on deep learning, namely a Faster R-CNN structure based on RESNET-101. It detects syllables from spectrogram images and was tested on MS patients and HC group only.

The limitations of the aforementioned algorithms are the small datasets used with mostly mild dysarthria and one specific type of the disease, except for [15] which uses the very same dataset that we will use. In terms of performance, the algorithm in [17] works fine but detects only syllables. The algorithm in [18] showed very high syllable detection accuracy, but it cannot detect EVO positions either. Finally, the algorithms in [1] and [15] are able to detect the EVO positions, but with lower syllable detection accuracy than [18] reported.

1.5. Goal

The goal of this thesis is to design an algorithm for the automated assessment of neurodegenerative diseases. This comprises of implementing a speech segmentation algorithm that will try to overcome existing algorithms, mentioned in the previous subchapter, in terms of syllable detection accuracy and the EVO positions detection accuracy. We will design features for automated assessment of the diseases and use the aforementioned speech segmentation algorithm to extract them. Finally, we will try to classify the diseases based on these features using a classifier.

2. Methodology

In this part, we will introduce methods used to achieve the goal of the analysis of oral diadochokinesis and automated assessment of neurodegenerative diseases. To distinguish between different diseases, we have to use certain features. We will use some previously invented features as well as some new features to perform this task.

In the first part of this chapter, we will talk about the dataset. Then we will introduce our approach for speech segmentation, which is an important part of the assessment algorithm and describe other algorithms that will be compared to ours. Finally, we will provide an overview of the features and statistical methods used in this thesis.

2.1. Dataset

The dataset consists of 6 groups, namely Healthy Control (HC), Huntington's Disease (HD), Multiple Sclerosis (MS), Multiple System Atrophy (MSA), Parkinson's Disease (PD), and PD with Deep Brain Stimulation (DBS) groups. The dataset details can be seen in Table 1.

Table 1 – Dataset details, n = number of recordings in a group, m = number of speakers in a group

Group (g. size) (no. s.)	Gender balance (F/M) (%)	Age (years) Mean/SD (range)	Disease duration (years) Mean/SD (range)	Disease severity Mean/SD (range)*	Speech severity Mean/SD (range)
HC (n=162) (m=71)	43.7/56.3	52.97/15.27 (21–75)	N/A	N/A	N/A
HD (n=50) (m=25)	46.1/53.9	44.08/12.28 (23–66)	N/A	27.08/11.74 (3–54)	0.73/0.45 (0–1)
MS (n=50) (m=25)	56.0/44.0	41.36/8.84 (25–61)	12.92/7.27 (2–31)	3.86/1.19 (1.5–6.5)	0.72/0.98 (0–3)
MSA (n=26) (m=13)	53.9/46.1	61.21/5.01 (55–72)	3.75/1.28 (2–6)	77.14/19.94 (46–123)	1.86/0.66 (1–3)
PD (n=16) (m=8)	54.5/45.5	64.41/9.65 (48–82)	9.27/5.55 (1–24)	15.86/7.60 (6–34)	0.72/0.70 (0–2)
DBS (n=258) (m=32)	17.9/82.1	60.51/6.33 (49–72)	N/A	N/A	N/A

*Scores on the performance section of the Unified Huntington's Disease Rating Scale (UHDRS) motor subscore for HD (ranging from 0 to 124), the Expanded Disability Status Scale (EDSS) for MS (ranging from 0 to 10), the Natural history and neuroprotection in Parkinson plus syndromes–Parkinson plus scale (NNIPPS) for MSA (ranging from 0 to 332), the Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS III) for PD (ranging from 0 to 132). Higher scores indicate more severe disability in all scales, N/A = not available

Speech recordings were performed in a quiet room with a low ambient noise level using a head-mounted condenser microphone (Beyerdynamic Opus 55, Heilbronn, Germany) situated approximately 5 cm from the mouth of each subject. Speech signals were sampled at 48 kHz with 16-bit resolution. Each participant was instructed to repeat the syllables /pa/-/ta/-/ka/ as quickly and accurately as possible at least seven times per one breath. Each of the syllable repetition tasks was performed twice.

2.2. EVO labels

The EVO positions in the recordings were manually labelled according to rules mentioned in [1].

The release of the airflow in a stop consonant causes a burst of energy, which can be visible in the time domain, but even clearer as a broadband burst in the spectrogram. We will designate the position of the initial burst as the explosion (E) position. In case there are multiple bursts visible, we will take the first one as the E position.

The vowel onset position (V) is the position where a periodic signal with the highest acoustic energy caused by vocal fold vibration starts (in our case, the vowel /a/). In the frequency domain, it is sought as the onset position of fundamental (F0) and first formant (F1, F2, and F3) frequencies. In the time domain, this is the position of highest (amplitude) contrast.

Finally, the occlusion position (O) marks the end of vowel and is marked at the position of F0, F1, F2, and F3 weakening, with F2 being considered the best indicator. In the time domain, a decrease in energy can be seen but is not considered a very reliable indicator as occlusion weakening can manifest in some patient's speech.

An example of EVO position labelling is given in Figure 1.

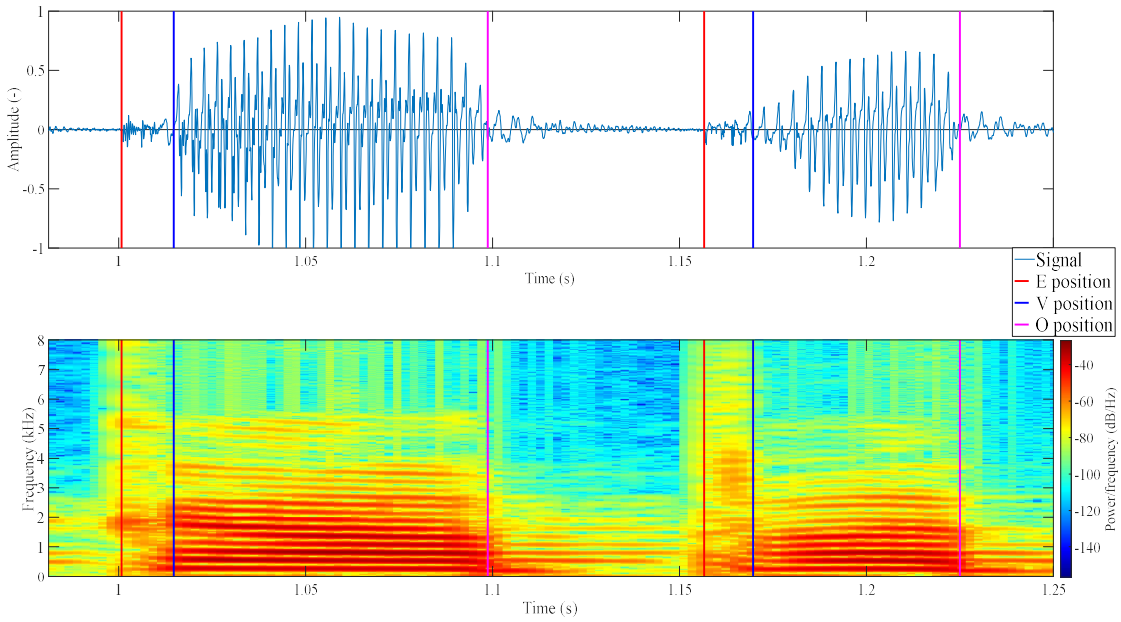


Figure 1 – EVO positions in syllables – the time domain and spectrogram

2.3. Speech segmentation

A necessary part for the feature extraction is a speech segmentation algorithm. The goal of this algorithm is to detect syllables in speech, distinguishing them from pauses. In addition, the algorithm will label each syllable with three important positions in the DDK task related to articulation, which are the EVO positions.

Multiple options are feasible to create such an algorithm, but not all of them are able to achieve the same accuracy. In this thesis, two different approaches for EVO detection were considered, both based on a convolutional neural network (CNN) and a bit of post-processing. The first idea was to create a syllable detection algorithm to detect syllables first and then apply another algorithm to detect EVO positions within each of the detected syllables. The second idea was an

evolution of the first proposal, and it combines the two separate algorithms for syllable detection and EVO detection into one.

2.3.1. Syllable detection into EVO detection (SDEVO)

The first proposed speech segmentation algorithm, SDEVO, works in two steps. The first step is syllable detection, and the second step is EVO detection in the previously detected syllables. This idea of double-stepped detection comes from the way a human researcher would label the speech signals themself. First, they see the syllable in the signal because of the amplitude pattern and clear vowel related periodic waveform. Then they focus on the EVO positions individually.

Using Tensorflow 2.0, we designed a neural network (Figure 2) to detect syllables. First, the data were split into training and test sets in 80 to 20 ratio speaker-wise, so that all the recordings from one speaker are always in one set or another but never split. The training set is further split into training and validation sets in 80 to 20 ratio speaker-wise as well. Even though all the sets are unbalanced, for the purpose of training (only), we made the training set balanced by setting the number of windows (mentioned further) for each class equal to the number of windows in the smallest class (in our case PD). The test and validation sets remain unbalanced, so the class distribution ratios accord to the number of recordings in each class as listed in Table 1.

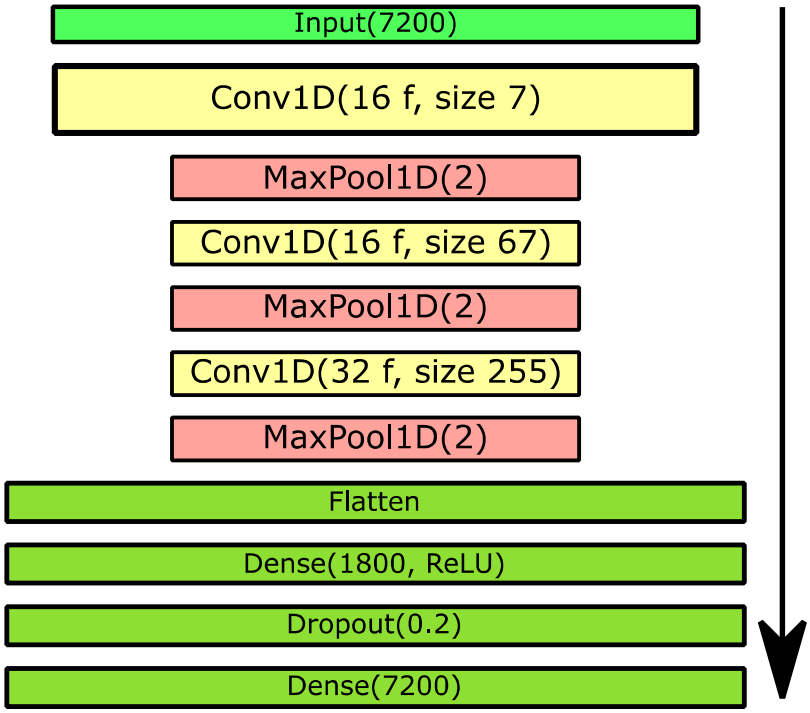


Figure 2 – Syllable detection network architecture for SDEVO, arrow shows the propagation way, # f means number # of filters, size is the size of filters in the layer

In the training phase, speech signals are first downsampled to 8 kHz, and the syllable positions are marked for the network output using a welch window ranging from each syllables E to O position (Figure 3), which creates an output vector for each window. Then the signals are cut into windows of 7200 samples (0.9 seconds) with a shift of 720 samples (0.09 seconds). These windows are fed into the network for training using MSE loss and ADAM optimiser with learning rate 10^{-4}. The network is initially trained using only HC group and after converging the network is trained using all the groups until convergence again.

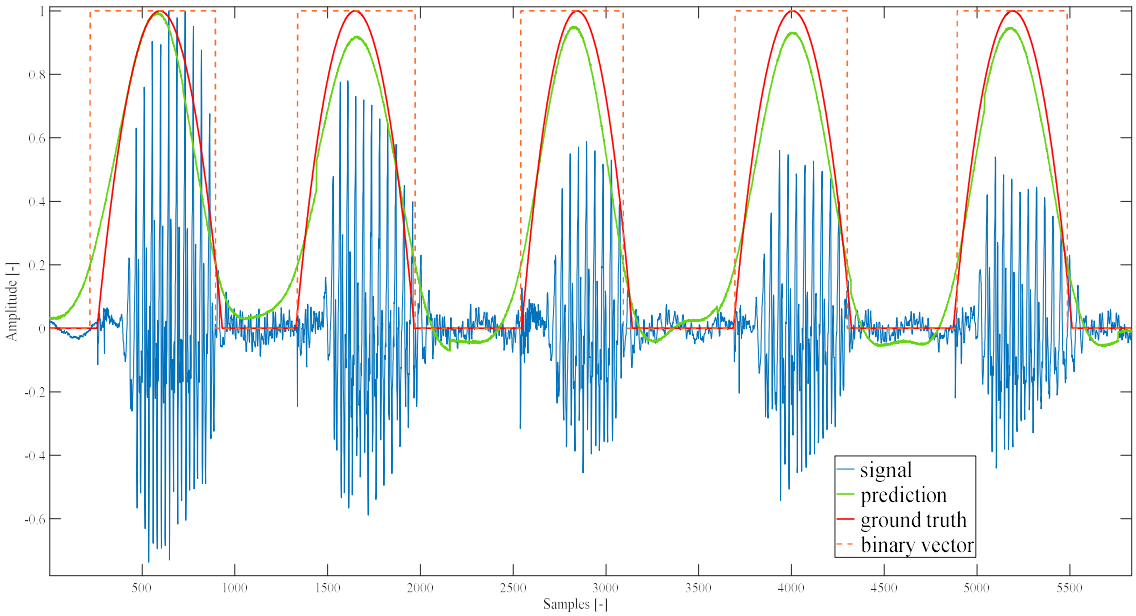


Figure 3 – Example of syllable detection part of SDEVO. We can see the welch window labels for each syllable along with the syllable prediction and thresholded prediction – binary vector.

In the testing phase, signals are downsampled to 8 kHz and zero-padded by a window of 7200 both in front and rear of the signals. Each recording is then proceeded through the net, resulting in a series of 7200 length output windows, each representing syllable location prediction for their corresponding input window, which are then summed up using OLA (overlap and add) method, and finally, the padding is removed. Thus, we get an output vector of syllable location prediction (Figure 3) with the same length as the input signal.

The prediction vector is normalised and made into a binary vector using a threshold equal to the mean of the amplitude of the prediction vector. Syllable beginnings and endings are determined by rising and falling edges, respectively. If a detected syllable is less than 50 samples long, it is marked as false detection and removed, as this can be an artefact of OLA summing. Start, centre and end locations of syllables are then passed on to the EVO detection part.

The EVO detection part comprises of a very similar CNN structure as seen in Figure 4.

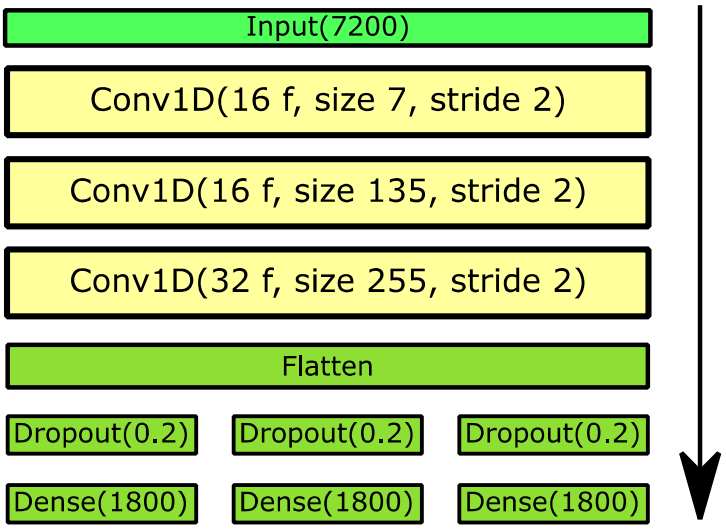


Figure 4 – EVO detection network architecture in SDEVO and DEVO, arrow shows the propagation way, # f means number # of filters, size is the size of filters in the layer and stride # is the filter stride in the layer

In the training phase, signals are downsampled to 16 kHz, and each labelled syllable is taken in a separate window of 7200 samples length (0.45 seconds). To ensure there is only one syllable in each window the rest of the signal until point 1 and from point 2 onwards is zeroed, where point 1 is the middle between current syllable's E position and previous syllable's O position and point 2 is the middle between current syllable's O position and next syllable's E position. Initially, the position of each syllable in its corresponding window is centred, but to prevent overfitting, a simple augmentation of data was introduced resulting in shifting the position of each syllable inside their corresponding windows by a random amount for each iteration. The ground truth positions of EVO are marked by a Gaussian window of 72 samples length (Figure 5). The used training loss was MSE, and optimiser was ADAM. The net was first trained on non-augmented HC group for a few epochs, and then other groups along with augmentation were added.

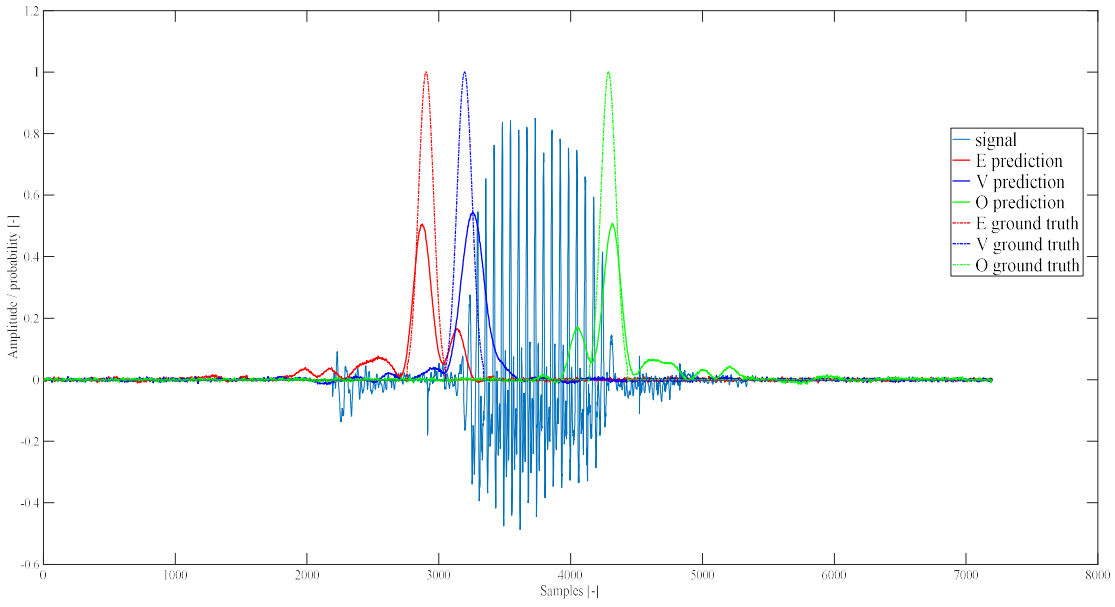


Figure 5 – Example of EVO detection in a single syllable. We can see the gaussian window EVO labels as well as EVO predictions given by the neural network part of SDEVO.

In the testing phase, the signals are downsampled to 16 kHz, and their corresponding syllable predictions obtained from the first part are added. Syllables are picked one by one for the EVO prediction. As in the training phase, to ensure only one syllable is taken as input, the signal is zeroed from start to point 1 and from point 2 onwards, where point 1 is the middle between previous syllable's end prediction and current syllable's start prediction and point 2 is the middle between current syllable's end position and next syllable's start position. The output of the net is an array of 3x1800 dimension representing EVO for each syllable window. The maximum of EVO for each syllable is taken, and the EVO prediction for the whole speech recording is reconstructed.

2.3.2. Direct EVO detection (DEVO)

Our second proposed algorithm, DEVO, differs from SDEVO mainly in using only one network - one step. It outputs EVO positions straight from the signal without any need for syllable detection. The architecture is the same as for SDEVO part 2 (Figure 4) (implemented using Tensorflow 2.0), but the input differs because there is no zeroing inside windows here, resulting in multiple possible syllables in one window. The dataset was split in the same manner as for SDEVO, and for the training we made the number of windows balanced just like for SDEVO. Validation and test sets still remain unbalanced.

In the training phase, signals are downsampled to 16 kHz, and EVO ground truth positions are marked by Gaussian windows of 72 samples length. Afterwards, windows of 7200 samples with a step of 720 samples are taken throughout the whole signal as training input data. First, the net is trained with the HC group only until convergence, then other groups are added too until another convergence.

In the testing phase, signals are downsampled to 16 kHz and padded with 7200 zeros on both sides. Then they are cut into windows of 7200 samples with 720 sample shift and put into the network. Outputs are then summed up using OLA, and the padding is removed. This results in a 3xN output (Figure 6), where N stands for signal length. To determine the final positions of EVO, we decided to use peaks of V as default indicators of syllables, because their reliability is the best among the three. Major peaks of V are sought using MATLAB's findpeaks function with a peak distance of 500 and a peak height of 0.5. Then, the time difference between consecutive peaks is calculated and sorted. We take the median of the lowest fourth of these differences and multiply by 0.75 factor, giving us peak distance for further search. Using this new peak distance and a peak height of 0.09 other peaks of V as well as peaks of E and O are sought. Now that we have all EVO peaks, it is important that they fit together, i.e. each syllable has all 3 positions and that there are no false detections of either E, V or O.

For this purpose, a logical build-up part had to be implemented. First, the EVO positions are given numbers 1, 2 and 3 respectively and are sorted based on time appearance. Then imaginary positions of 123 are added in front and behind this vector (as a kind of padding) and a search for the 123 pattern is initialised. If the pattern 123 is found in the first three positions, pattern shifts to 231 and the search index increments too. Another successful search leads to pattern shifting to 312 and another index increment, etc.

However, if the pattern is not found, the positions have to be corrected. Whenever this happens, we will designate the newest scanned position a "sinner". We will use the position information along with the prominence of peaks information, which is obtained from the findpeaks function, and apply the following correction rules:

- Pattern 123 is not detected
  - If sinner is 1 (that means we detected 121 instead), find a position 3 to be inserted
  - If the sinner is 2, look at the prominence of the two peaks before sinner, sinner and one after sinner - if they are prominent enough, find a position 3 to insert
  - Else remove the sinner as false detection
- Pattern 231 is not detected
  - If sinner is 2 and after sinner comes 3, find a 1 to be inserted
  - If sinner is 2 and is a prominent peak, find a 1 to be inserted
  - If sinner is 3 and is a prominent peak, find both 1 and 2 to be inserted
  - Else remove the sinner as false detection
- Pattern 312 is not detected
  - If sinner is 3 and after sinner comes 1, find a 2 to be inserted
  - If sinner is 3 and is a prominent peak, find a 2 to be inserted

The prominence peak threshold was set to 0.7 of mean peak prominence for single peak prominence check and to 0.6 of mean peak prominence for double peak prominence check. The mean peak prominence was gathered from the initially detected peaks.

After the correction, the search pattern returns to 123 and search index returns to 1 again. Once the whole correction process is finished, the 123 padding in front and behind the vector is removed. One final correction comprises of removing predicted EVO positions that are located at the very same sample, e.g. E is at sample 5702, V is at 5702 and O is also at 5702. This is a very rare occurrence, but it may happen due to the rules set previously and is a problem for the feature evaluation algorithm.

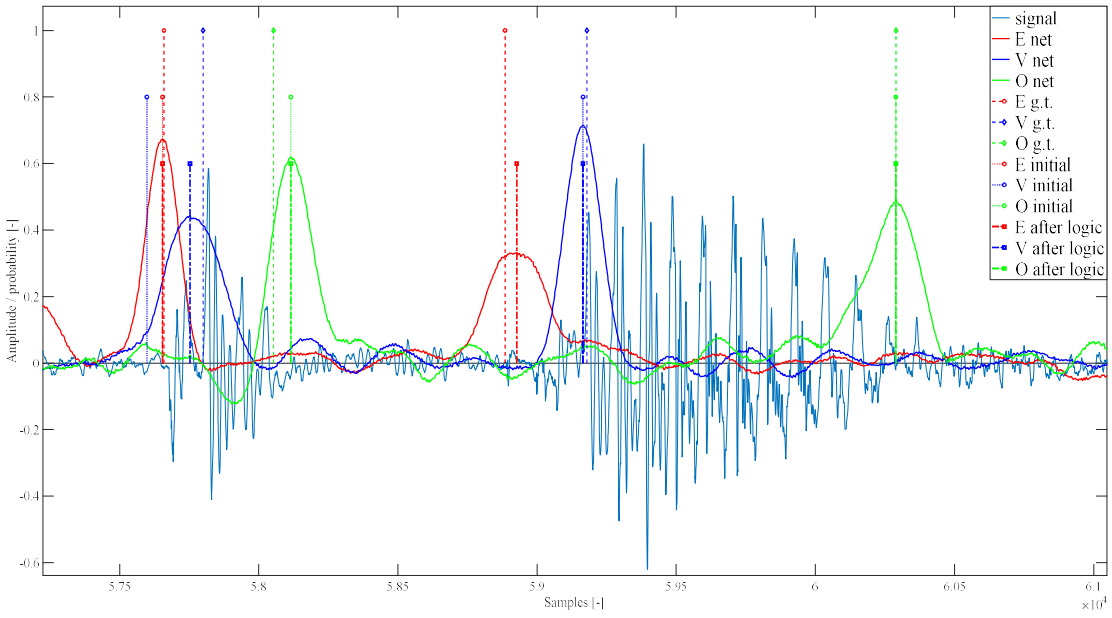


Figure 6 – Example of DEVO output. We can see the network EVO predictions, EVO ground truth, EVO initial positions taken from findpeaks function and EVO positions after the logical build-up. Here a V position was found before the E position in the first syllable and had to be removed. A new one was assigned between the E and O position. In the second syllable, the E position was missing and had to be added using the logical build-up.

2.3.3. Evaluating the algorithms

To evaluate syllable accuracy, we looked at each syllable's centre position determined as the mean of its E and O position. If this centre fits into any truth interval of E and O ground truth

pair, the syllable is considered as a successful prediction, and the syllable pair is removed from the evaluation loop (and considered as matched pair). Once the loop is finished, we look at how many predicted and ground truth syllables are left and determine the percentage of missing predictions as

$$syl_{miss} = \frac{m_{GT} - m_{pair}}{m_{GT}} \cdot 100\ [\%]\,, \tag{1}$$

where m_{GT} is the number of unmatched ground truth syllables, and m_{pair} is the number of matched syllables.

Then we determine the percentage of excess syllable predictions as

$$syl_{excess} = \frac{m_{pred} - m_{pair}}{m_{GT}} \cdot 100\ [\%]\,, \tag{2}$$

where m_{pred} is the number of unmatched predicted syllables, and m_{pair} is the number of matched syllables.

The accuracy of EVO is measured by the percentage of EVO predictions that fit into a certain tolerance interval of EVO ground truth. This can be elegantly plotted as a cumulative distribution function (CDF).

2.3.4. Comparison to other algorithms

Once we choose the better algorithm of the two proposed (SDEVO, and DEVO), we will compare the performance with previously implemented algorithms in [1] and [15], which are based on traditional signal processing. In the algorithm from [1], the syllables are first detected using filtering and energy peak detection. Then the E position is sought through spectrogram filtering. To detect the V position, the Bayesian Step Changepoint Detector was implemented and finally, to detect the O position, polynomial thresholding was used. The algorithm in [15] works similarly to the one from [1]. It first detects syllables and the O position using Mel-frequency cepstral coefficients (MFCC). Then it detects the V position using filtering and k-means clustering. The E position is detected using the phase information of the burst. For comparison of the algorithms, we will use the same dataset for all the algorithms, i.e. our test set.

2.4. Features

For feature extraction, we did not use the whole speech signals but instead focused only on the part between 4th and 33rd syllable inclusive, i.e. we skip the first syllable train because of occasional inequalities in volume and/or rhythm (so to say before the speakers start “going”). Additionally, if the signal has less than 34 syllables, we always omit the last one, e.g. for a signal with 26 syllables, we would use syllables 4 to 25. Described below are the used features.

2.4.1. Relative Intensity Range Variation (RIRV)

Taken from [12]; the RIRV feature describes loudness in the utterance. It is calculated as the standard deviation of the intensity envelope.

2.4.2. Intensity Slope (IS)

This feature describes the decrease in volume in time. We compute the intensity envelope of the signal, construct a regression line and then take its slope value.

2.4.3. Syllable Rate (SR)

Syllable Rate represents the pace (speed) at which the patient speaks. It is defined as the number of syllables pronounced per time interval, computed as

$$SR = \frac{m_{end} - m_{start}}{E_{end} - E_{start}}, \tag{3}$$

where m_{end} is the number of the last considered syllable, E_{end} is this syllable's E position, m_{start} is the number of the first considered syllable, and E_{start} is the starting syllable's E position.

2.4.4. Vowel Length (VL)

This feature describes the ratio of vowel vocalisation length to syllable length. We take the time difference of V and O and divide it by time difference of E and O in each syllable and then take the average, that is

$$VL = \frac{1}{m}\sum_{i=1}^{m}\frac{O_i - V_i}{O_i - E_i}, \tag{4}$$

where m is the number of considered syllables, O_i is the O position of ith considered syllable, V_i is the V position of ith considered syllable, and E_i is the E position of ith considered syllable.

2.4.5. Rhythm Acceleration (RA)

As the name suggests, this feature describes speeding up or slowing down in rhythm. First, we take the time differences of each consecutive pairs of syllables (we consider V positions) and construct a linear regression line. Then we take the slope of this line as the RA feature.

2.4.6. Voice Onset Time (VOT)

Voice Onset Time usually refers to the temporal coordination between the articulation of a stop consonant and the laryngeal mechanism required to produce periodic vibration of the vocal folds [12]. As features, we will consider the mean (VOT) and standard deviation (sVOT) of the time difference between E and V positions of each syllable.

2.4.7. Rhythm Instability (RI)

Also known as DDK regularity [12], Rhythm Instability is the time position variance of syllable vocalisations. We calculate the time difference of each consecutive pair of syllables and then take the variance.

2.4.8. Index of Rhythmicity (IoR)

This feature was designed as an improvement to previous rhythmicity measures that struggle when the patient does not understand the DDK task properly, i.e. when the patient repeats the syllable train /pa-ta-ka/ with distinctive pauses between the syllable trains, making it a repetition of triplets rather than a continuous stream of single alternating syllables.

To extract this feature, we first filter the signal to get its intensity envelope and then construct a correlation function. We further take the maximum of the first correlation peak and divide it by the mean of maxima of the next 3 peaks. This should ensure that both streams of triplets or single syllables are perceived as rhythmical.

Rhythmicity index feature can be an important marker of cerebellar damage as patients with damaged cerebellum are likely to have dysdiadochokinesia (impaired ability to perform rapid alternating motions).

2.5. Statistics

In order to increase the performance of classification and feature evaluation, the extracted features for each of the patients were averaged across their two utterances. To test for normality, we used Kolmogorov–Smirnov test. The extracted features were analysed using one-way analysis of variance (ANOVA) afterwards to determine their significance. To evaluate the ability to distinguish between groups, Tukey's HSD post hoc test was performed on each of the features. Additionally, Pearson's correlation coefficients were calculated to see whether the features are highly correlated or not.

2.6. Classification

For the purpose of automated distinguishing between the diseases, a simple classifier was built. We decided to use a small neural network, as seen in Figure 7 because of its ability to learn nonlinearities, output multiple class probabilities and ease of use and training. The input of the classifier comprises of the extracted features and output is the probability of class (disease).

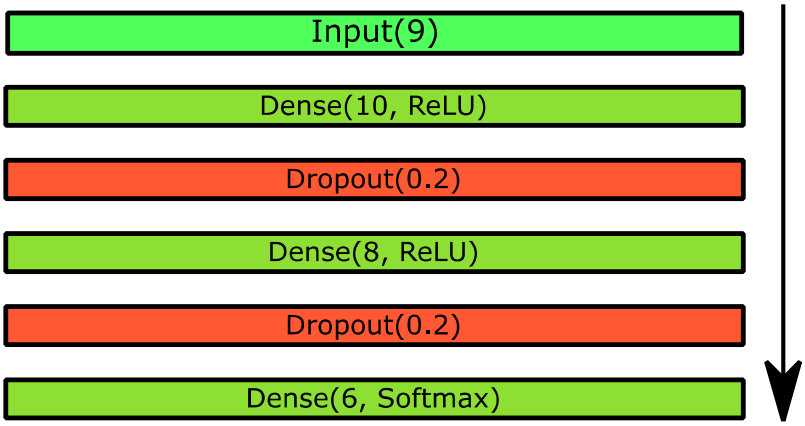


Figure 7 – Architecture of the neural network classifier

The data for this purpose were split into train, and test sets in train/test ratio of 10:6 and further the train set was split into train/validation set in ratio 8:2. The network was trained using a modified loss function that set the higher weight of importance to less numerous classes so that when calculating the loss of bad prediction for these classes it would be higher than for more numerous classes. This had to be done due to the fact of our dataset being highly imbalanced, as without this modification, the classifier would be biased towards the larger classes. The weights were set as the inverse of each class' size and then multiplied by modifiers that rose from observing the performance. The final modifiers were (9, 9, 8, 8, 10, 10) assigned to classes in order (DBS, HC, HD, MS, MSA, PD). We split the data and trained the model, i.e. cross-validated the model 10 times.

3. Results

In this chapter, we will show the results of both SDEVO and DEVO as speech segmentation algorithms and then the results of feature evaluation and disease classification.

3.1. Speech segmentation algorithm

In this subchapter, we will give results of SDEVO and DEVO performance in terms of syllable detection accuracy and EVO positions detection accuracy. Then we will take the better performing algorithm and compare it to the algorithms from [1], and [15], and also evaluate the performance preceding feature evaluation, i.e. performance on shortened utterances as described in Chapter 2.4.

First, we measured the accuracy of syllable detection of both SDEVO and DEVO, as seen in Table 2. In terms of syllable detection accuracy, SDEVO seems to perform slightly better than DEVO.

Table 2 - Comparison of syllable detection accuracy of SDEVO and DEVO

Algorithm, set	Missing syllables (%)	Excess syllables (%)
SDEVO, test	0.079	0.700
SDEVO, val	0.647	0.953
SDEVO, train	0.440	3.428
DEVO, test	0.146	1.198
DEVO, val	0.689	1.161
DEVO, train	1.090	2.308

3.1.1. SDEVO

In this section, we will show the EVO detection performance results of SDEVO. These results were obtained only from the successfully detected syllables, i.e. we omit the unsuccessfully detected syllables (whose rate of occurrence is depicted in Table 2) and keep only the matched syllables, as described in chapter 2.3.3.

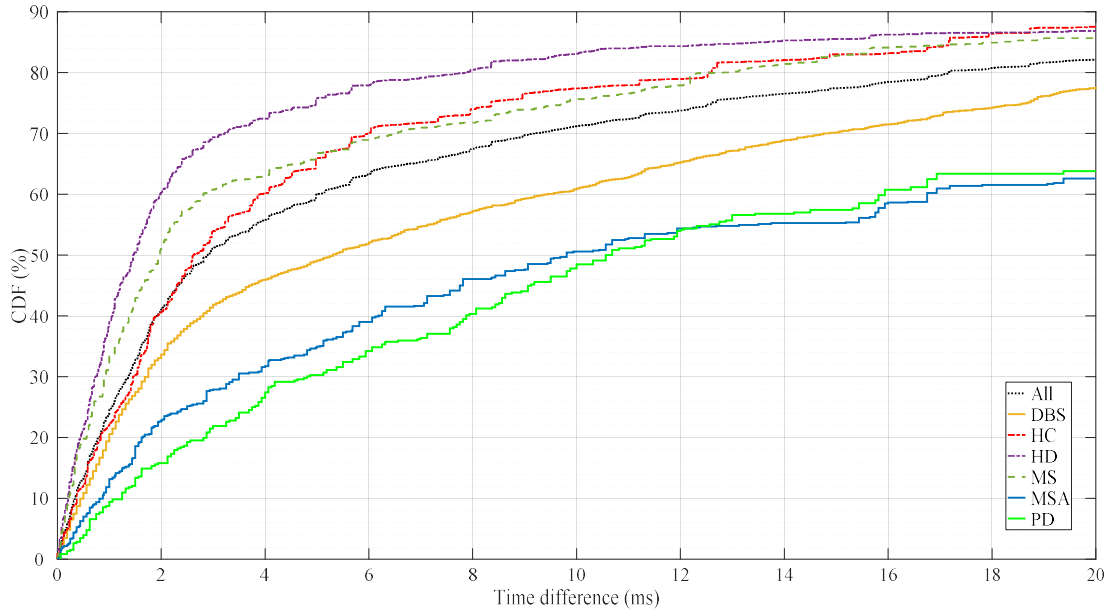


Figure 8 – Performance of SDEVO at E position, test dataset

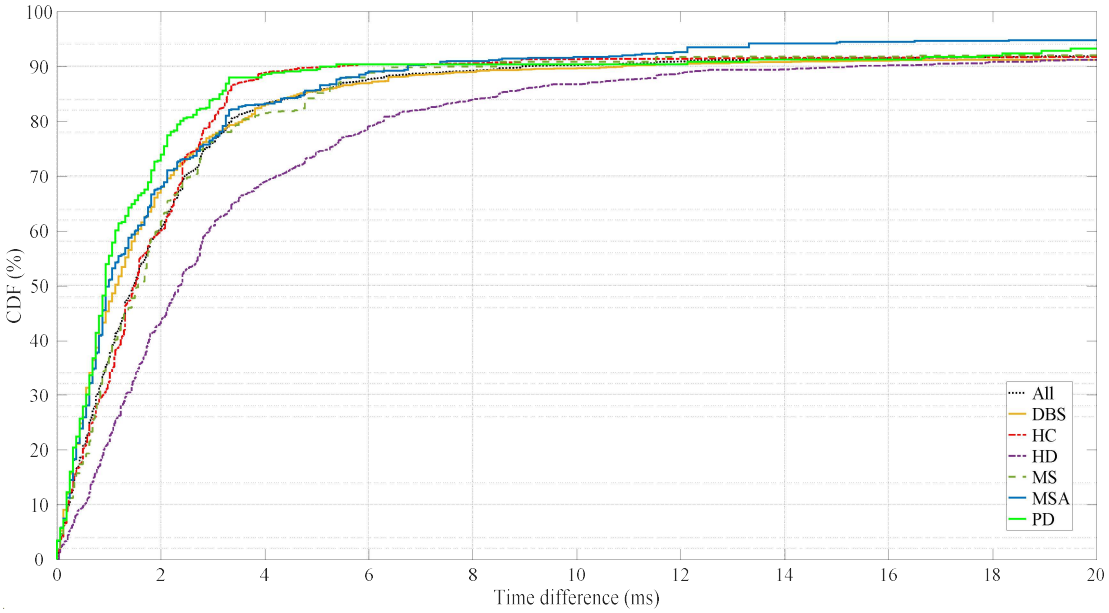


Figure 9 – Performance of SDEVO at V position, test dataset

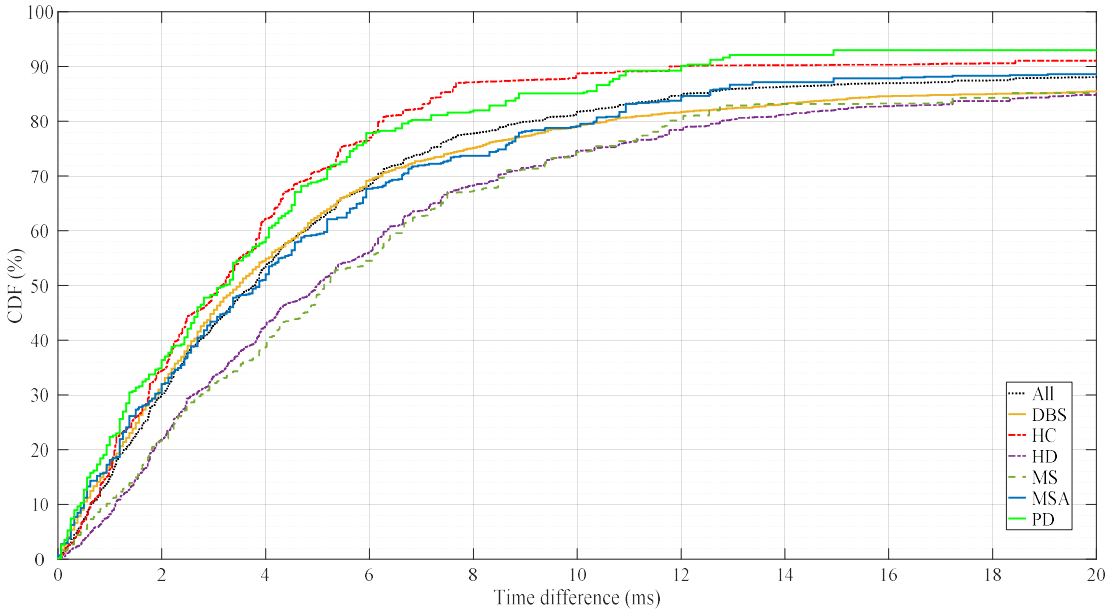


Figure 10 – Performance of SDEVO at O position, test dataset

It can be seen that the performance differs with class, especially in E position detection as seen in Figure 8, where the algorithm is underperforming for MSA, PD and DBS classes. However, in Figure 9 and Figure 10, we can see these two classes are among the better-performing ones. A summary of performance at selected tolerance key points for all the classes (average of syllables across all classes) is given in Table 3, Table 4, and Table 5 for the test, validation, and training sets respectively.

Table 3 – Performance of SDEVO on the test set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	57.03	85.62	62.22
10 ms	68.46	90.05	81.08
20 ms	80.94	91.69	87.44

Table 4 – Performance of SDEVO on the validation set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	54.41	79.33	55.01
10 ms	66.78	85.87	74.34
20 ms	83.36	87.52	82.98

Table 5 – Performance of SDEVO on the training set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	44.58	73.10	51.97
10 ms	55.98	78.72	74.99
20 ms	75.27	86.20	83.83

3.1.2. DEVO

In this section, we show the results of EVO detection performance using DEVO algorithm. Just as in SDEVO evaluation, these results were obtained only from the successfully detected syllables as described in chapter 2.3.3.

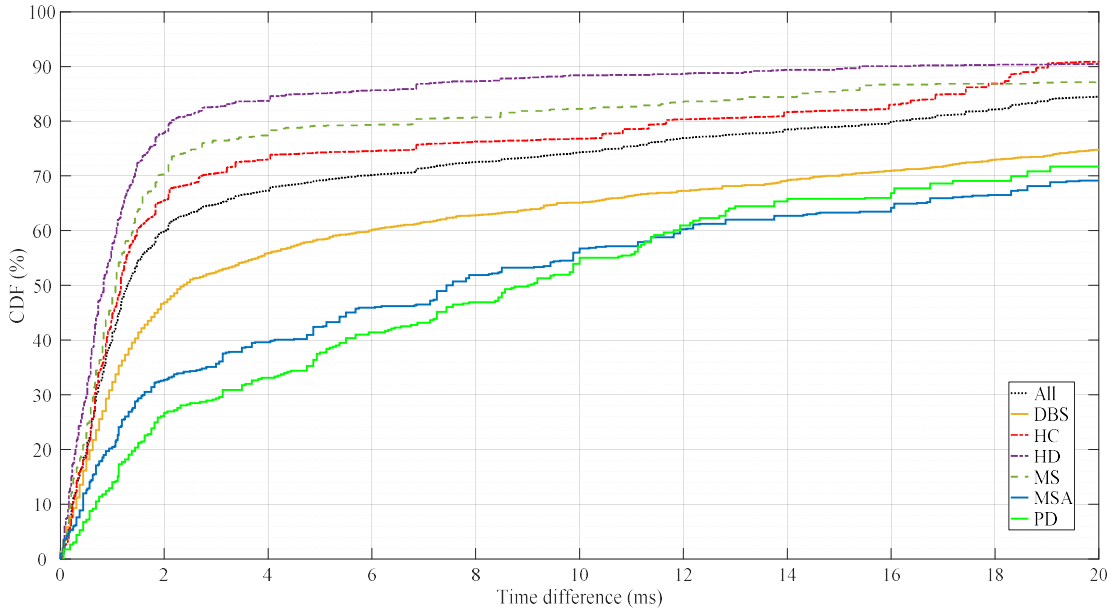


Figure 11 – Performance of DEVO at E position, test dataset

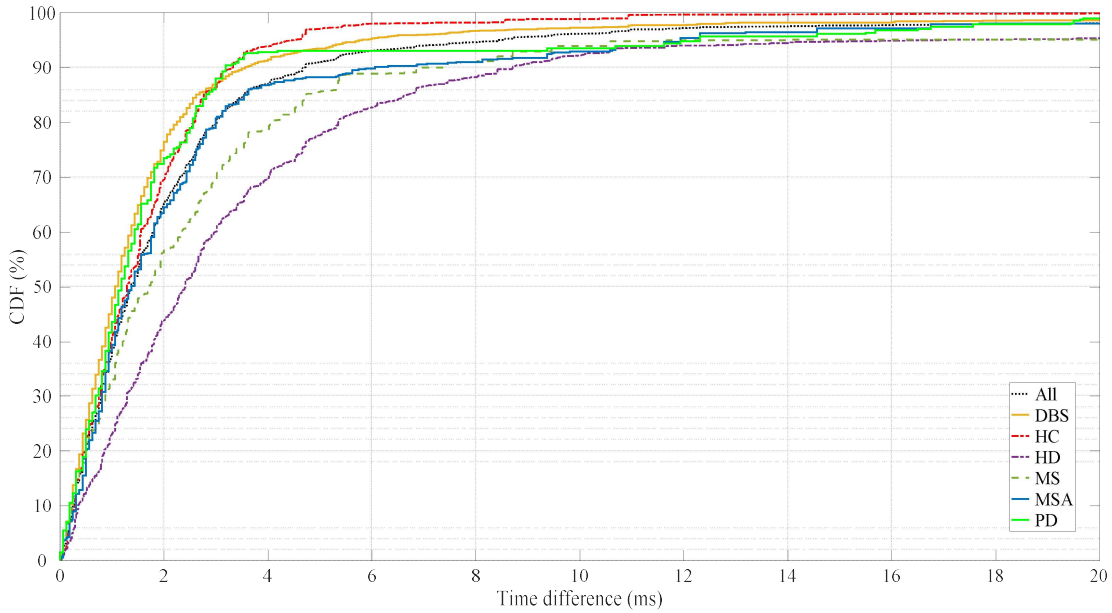


Figure 12 – Performance of DEVO at V position, test dataset

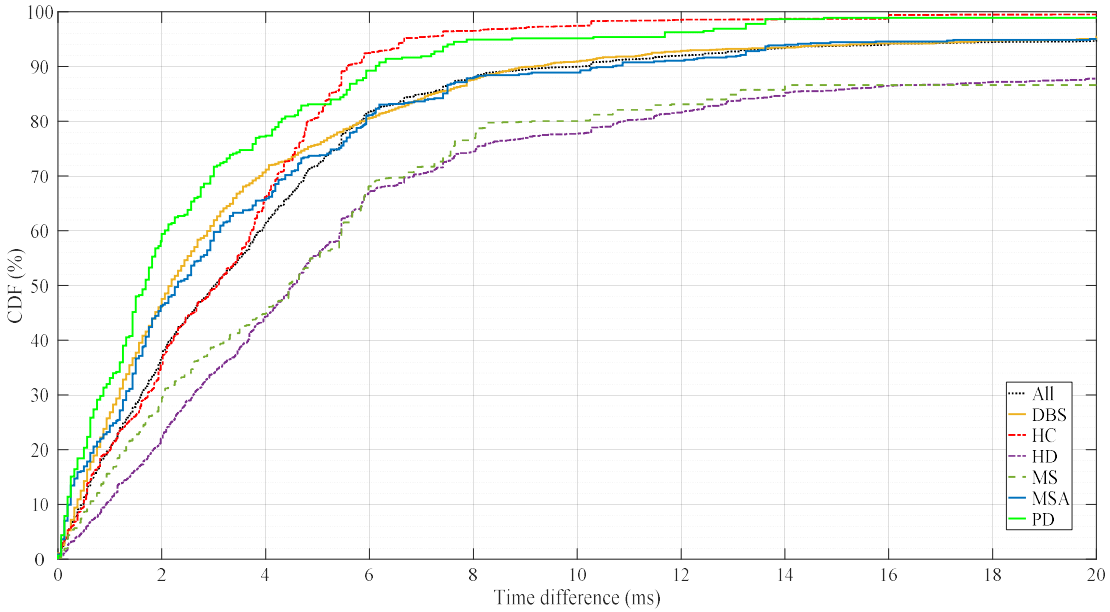


Figure 13 – Performance of DEVO at O position, test dataset

In Figure 11, Figure 12, and Figure 13, we can see that similarly to SDEVO, DEVO performance differs with class, especially in E position. A summary of DEVO performance at selected tolerance key points for all the classes (average of syllables across all classes) is given in Table 6, Table 7, and Table 8 for the test, validation, and training sets respectively.

Table 6 – Performance of DEVO on the test set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	66.37	91.62	73.01
10 ms	71.92	96.40	90.28
20 ms	81.95	98.18	94.79

Table 7 – Performance of DEVO on the validation set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	67.11	84.98	66.15
10 ms	76.32	92.61	81.83
20 ms	85.03	94.18	91.52

Table 8 – Performance of DEVO on the training set at selected tolerance points for all classes

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	52.68	78.84	59.11
10 ms	61.80	85.76	77.58
20 ms	78.88	93.27	88.38

Based on the results shown in Table 3, Table 6, and Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, and Figure 13, DEVO seems to perform better than SDEVO at all stages, and even

though syllable detection accuracy as reported in Table 2 is in favour of SDEVO, it is not by a large margin. Therefore, we chose DEVO as the main speech segmentation algorithm for all the following experiments.

3.1.3. DEVO in comparison

We compared the performance of DEVO to previously implemented algorithms in [1] and in [15]. For comparison, we ran all the algorithms on the same set, which is the same test set as mentioned and used before. The comparison was made for all the syllables and among each class, just like in the previous performance evaluations. The comparison with [1] is depicted in Figure 14 – Figure 19, and the comparison with [15] is depicted in Figure 20 – Figure 25.

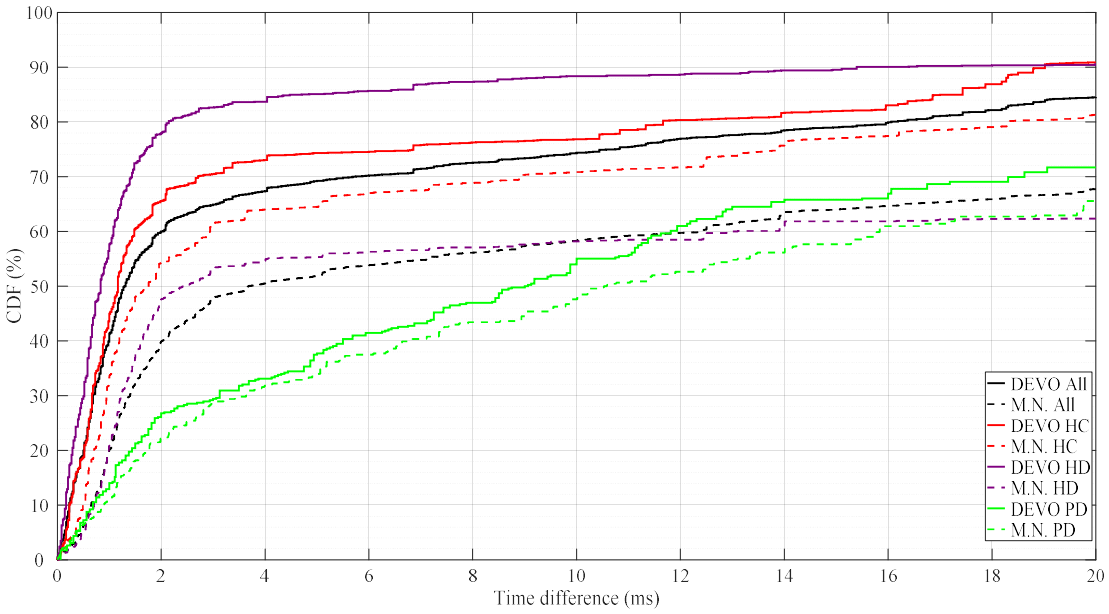


Figure 14 – Comparison of DEVO and algorithm by Michal Novotny [1], E position, part 1

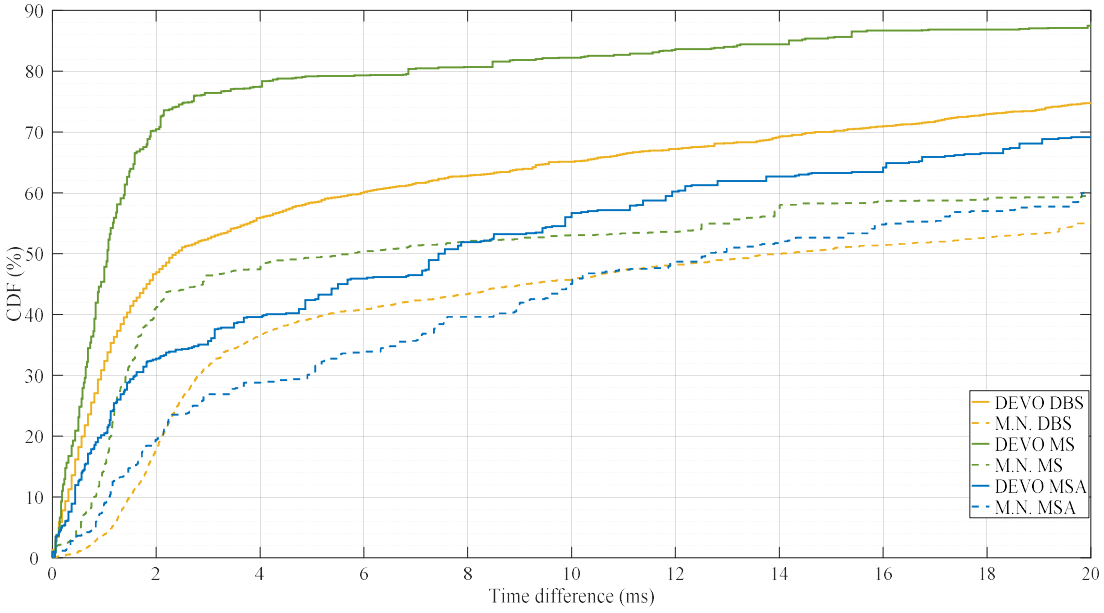


Figure 15 – Comparison of DEVO and algorithm by Michal Novotny [1], E position, part 2

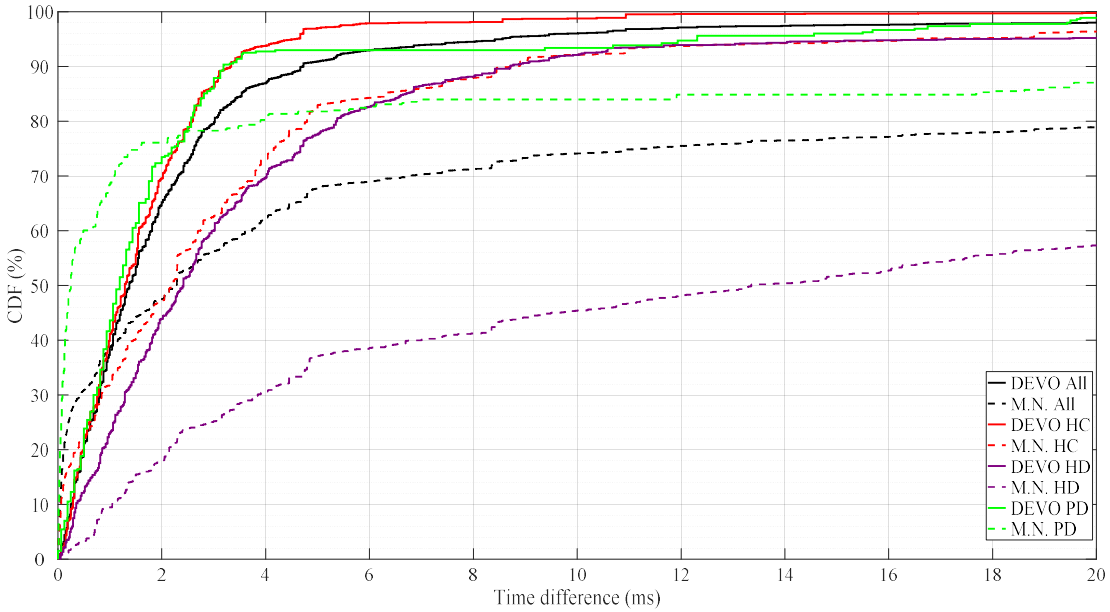


Figure 16 – Comparison of DEVO and algorithm by Michal Novotny [1], V position, part 1

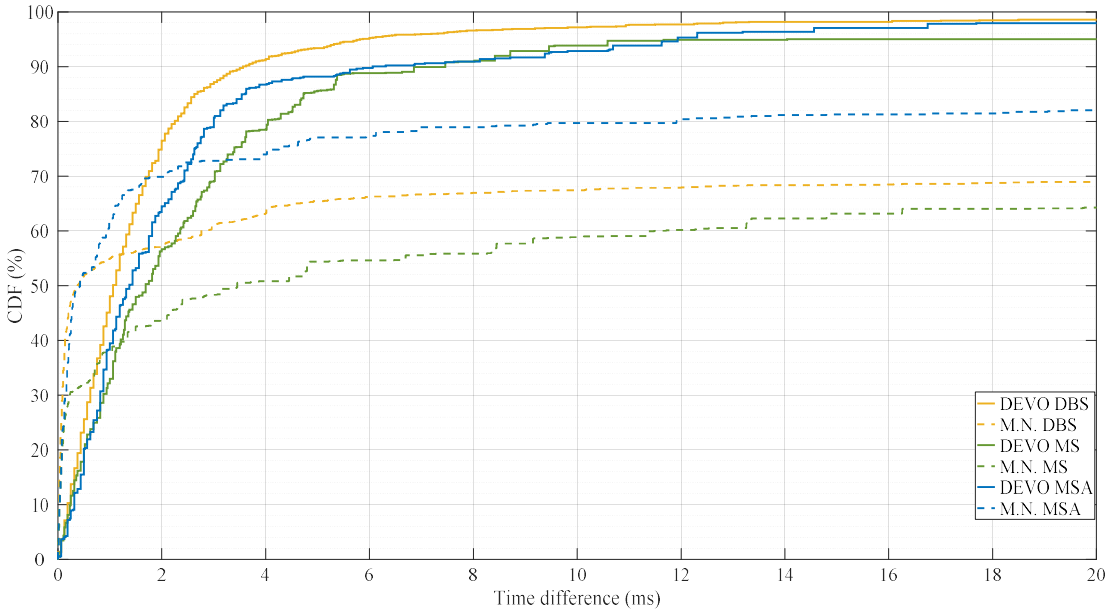


Figure 17 – Comparison of DEVO and algorithm by Michal Novotny [1], V position, part 2

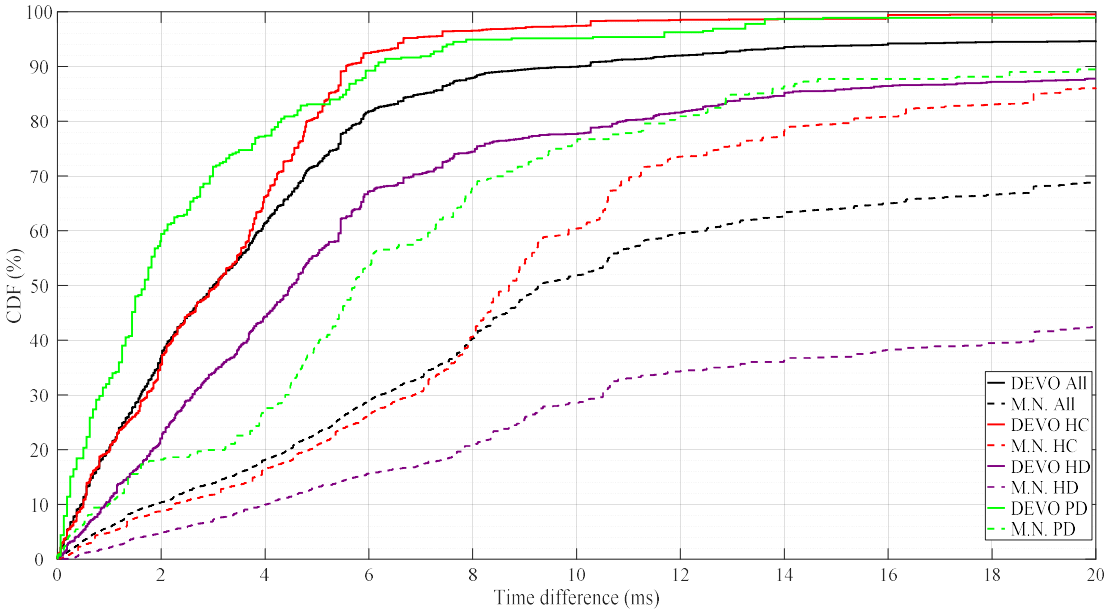


Figure 18 – Comparison of DEVO and algorithm by Michal Novotny [1], O position, part 1

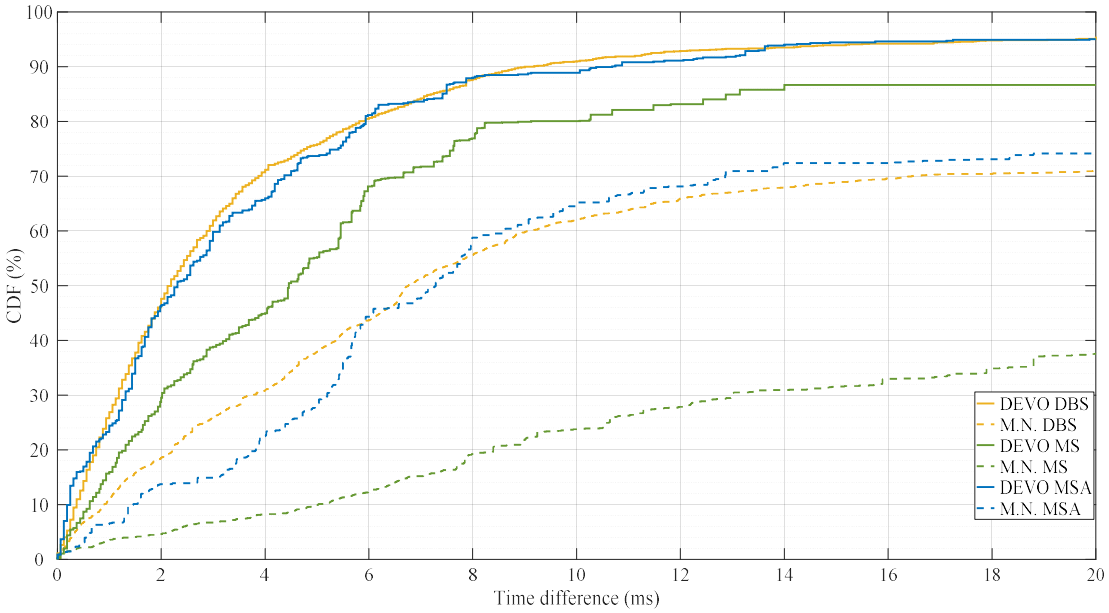


Figure 19 – Comparison of DEVO and algorithm by Michal Novotny [1], O position, part 2

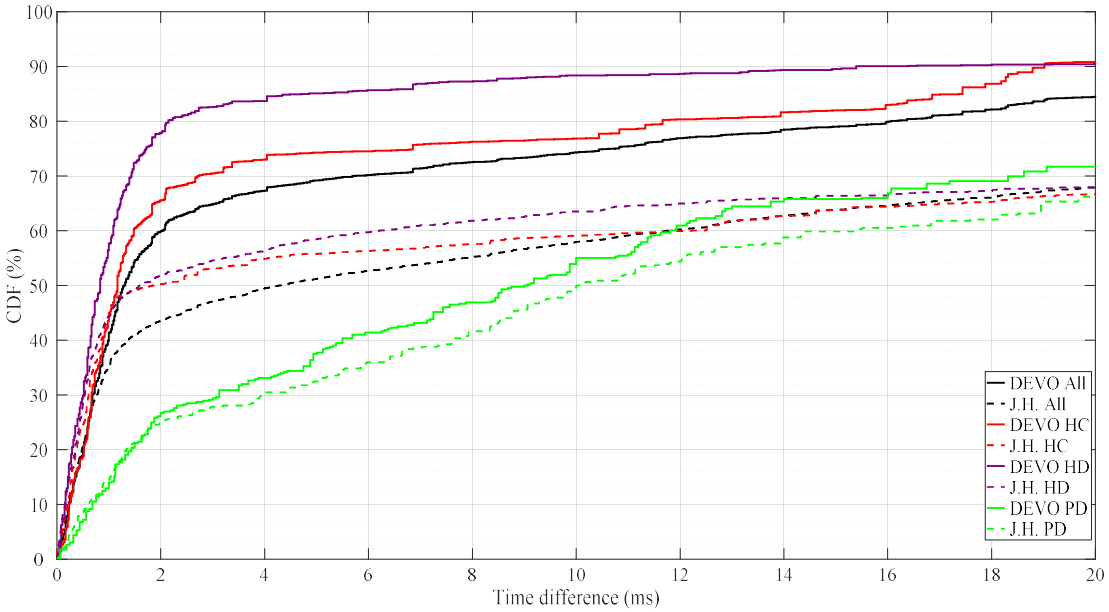


Figure 20 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], E position, part 1

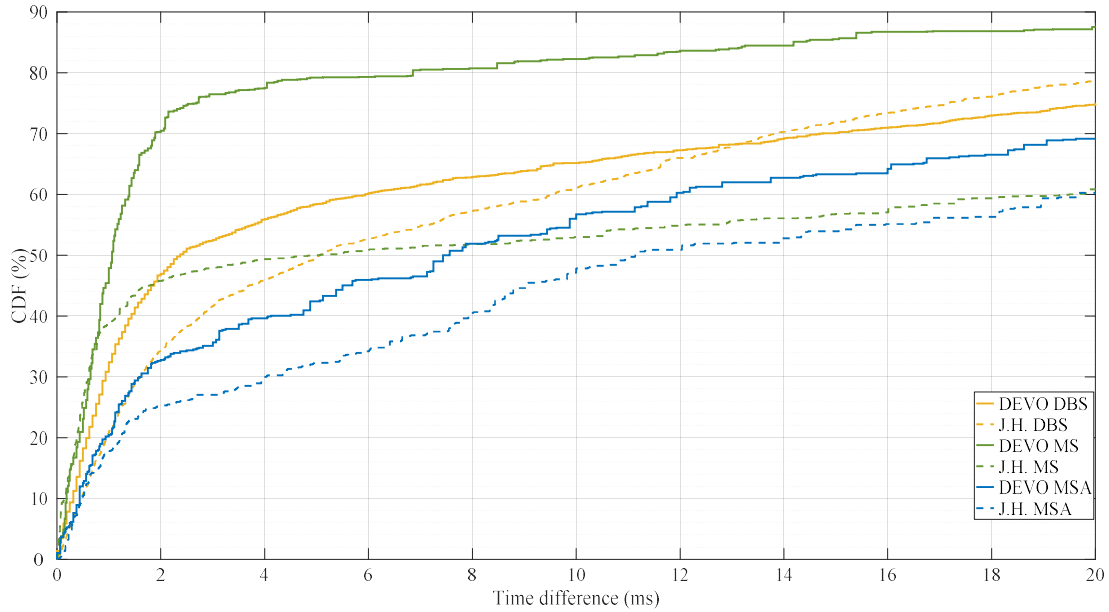


Figure 21 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], E position, part 2

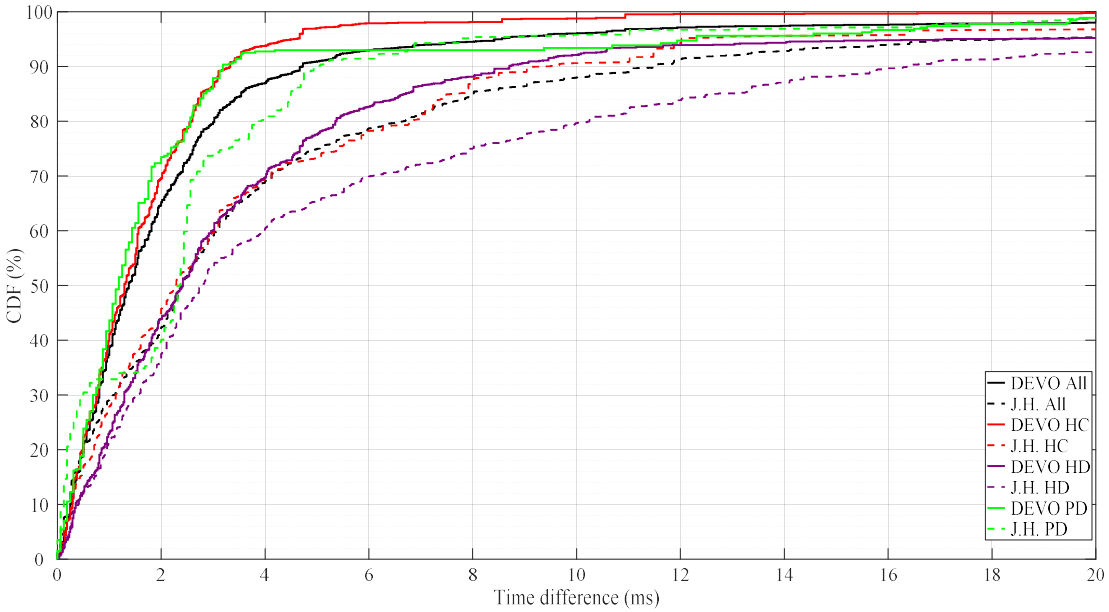


Figure 22 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], V position, part 1

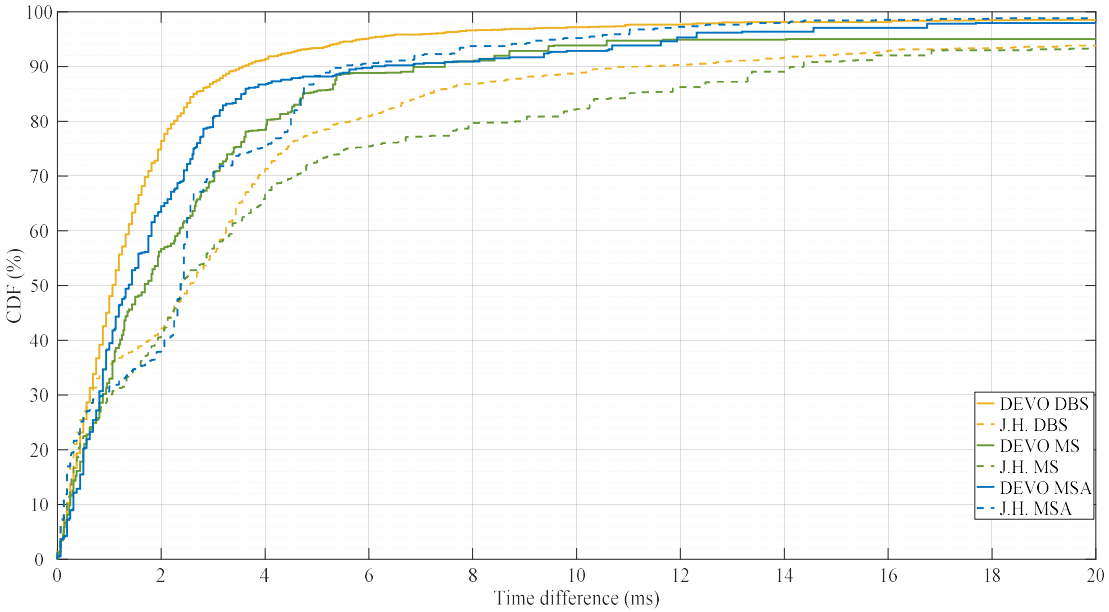


Figure 23 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], V position, part 2

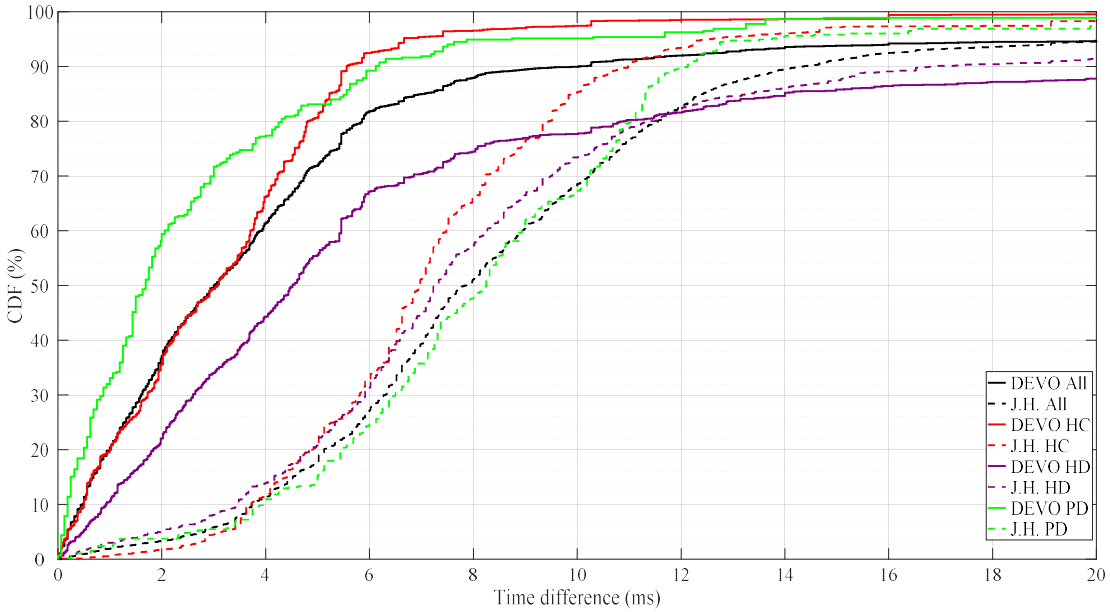


Figure 24 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], O position, part 1

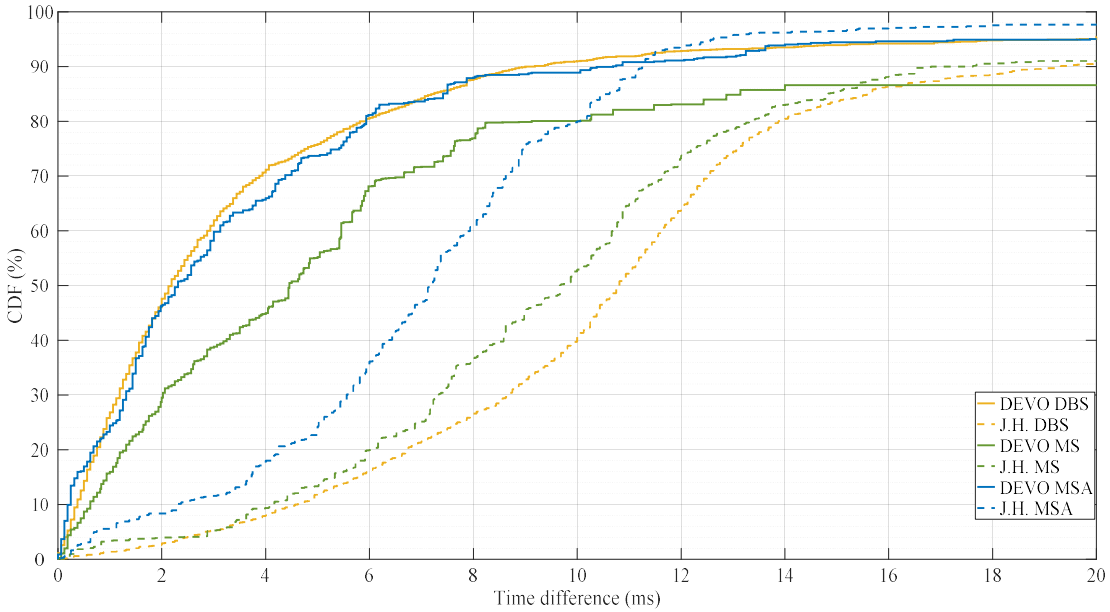


Figure 25 - Comparison of DEVO and algorithm by Jan Hlavnicka [15], O position, part 2

For a simpler comparison, we calculated the EVO performance at selected key points, as seen in Table 9 and excess and missing syllables, as seen in Table 10.

Table 9 – Comparison of DEVO, algorithm by Michal Novotny [1], and algorithm by Jan Hlavnicka [15] in performance at selected EVO key points for the whole test set

	Position	E			V			O		
	Algorithm	DEVO	JH	MN	DEVO	JH	MN	DEVO	JH	MN
Time tolerance / accuracy (%)	5 ms	66.37	50.71	48.39	91.62	75.86	67.09	73.01	16.03	27.37
	10 ms	71.92	58.95	54.9	96.4	88.18	72.29	90.28	60.18	54.82
	20 ms	81.95	71.12	64.31	98.18	94.92	76.22	94.79	93.49	69.55

Table 10 – Comparison of DEVO, algorithm by Michal Novotny [1], and algorithm by Jan Hlavnicka [15] in excess and missing syllable percentage gathered from the test set

	Missing syllables			Excess syllables		
Algorithm	DEVO	JH	MN	DEVO	JH	MN
Miss / excess (%)	0.146	1.179	3.049	1.198	0.463	0.402

The performance of DEVO seems to be better than both algorithms in almost all cases. The algorithm from [1] performs better at V position detection at time tolerance of 0 to 1 ms (Figure 16, Figure 17). It has a lower number of excess syllables detected (Table 10), but it is worse in all other metrics. The algorithm from [15] is performing better than [1], but not yet as good as DEVO. It is slightly better in V position detection for the MSA class (Figure 23), and O position for the MSA, HD, DBS and MS classes (Figure 24, Figure 25) at the higher time tolerances and has a lower number of excess syllables detected (Table 10), but is beaten by DEVO everywhere else.

A short summary of comparison is given by Figure 26, Figure 27, and Figure 28, where we show the performance of the algorithms for the whole test set only (all classes together). From these, it is clear that DEVO outperforms both the previous algorithms quite significantly.

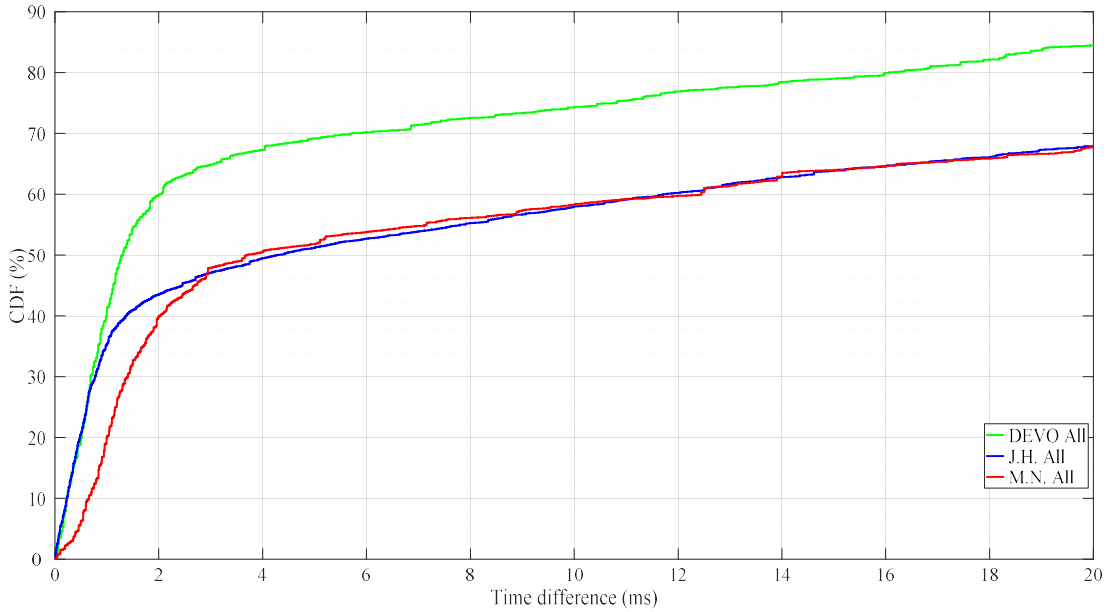


Figure 26 – Comparison of DEVO, [1] algorithm, and [15] algorithm in E position performance for the whole test set (all classes)

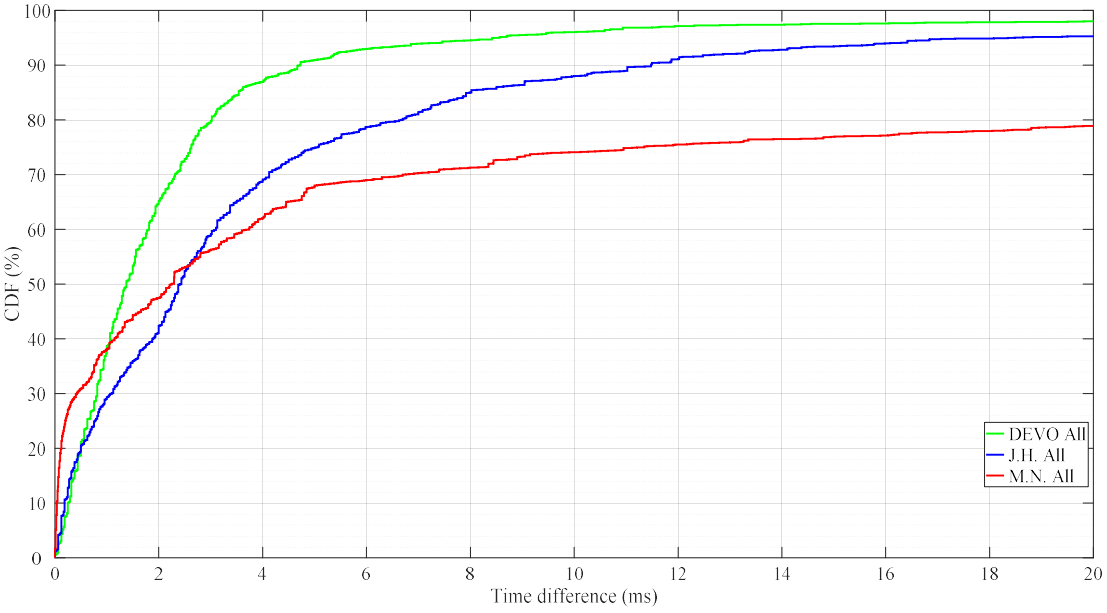


Figure 27 – Comparison of DEVO, [1] algorithm, and [15] algorithm in V position performance for the whole test set (all classes)

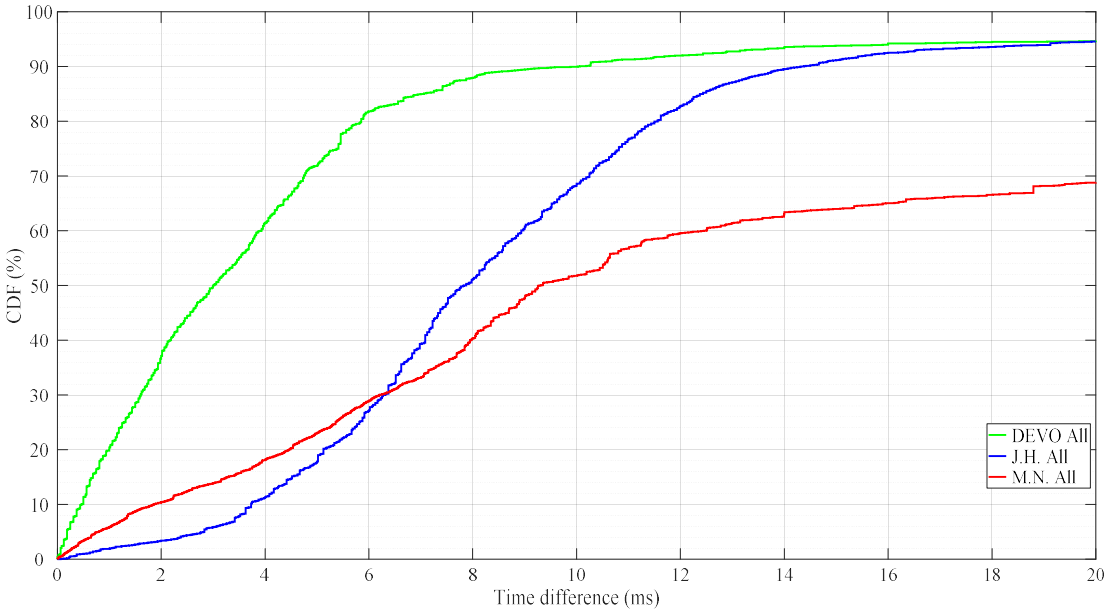


Figure 28 – Comparison of DEVO, [1] algorithm, and [15] algorithm in O position performance for the whole test set (all classes)

3.1.4. DEVO performance for feature evaluation

In this section we will present DEVO performance results for shortened utterances, i.e. we only pick predicted syllables number 4 to 33 as described in chapter 2.4, which is 10 /pa/-/ta/-/ka/ syllable trains (or less), and test the accuracy the same way as before.

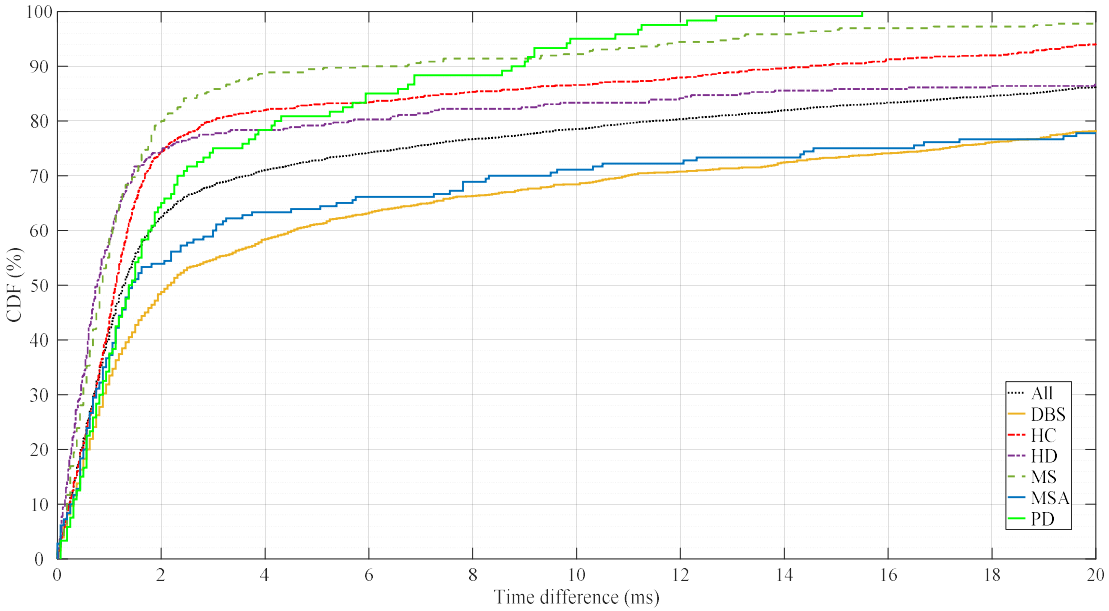


Figure 29 – DEVO performance on shortened utterances, E position, test set

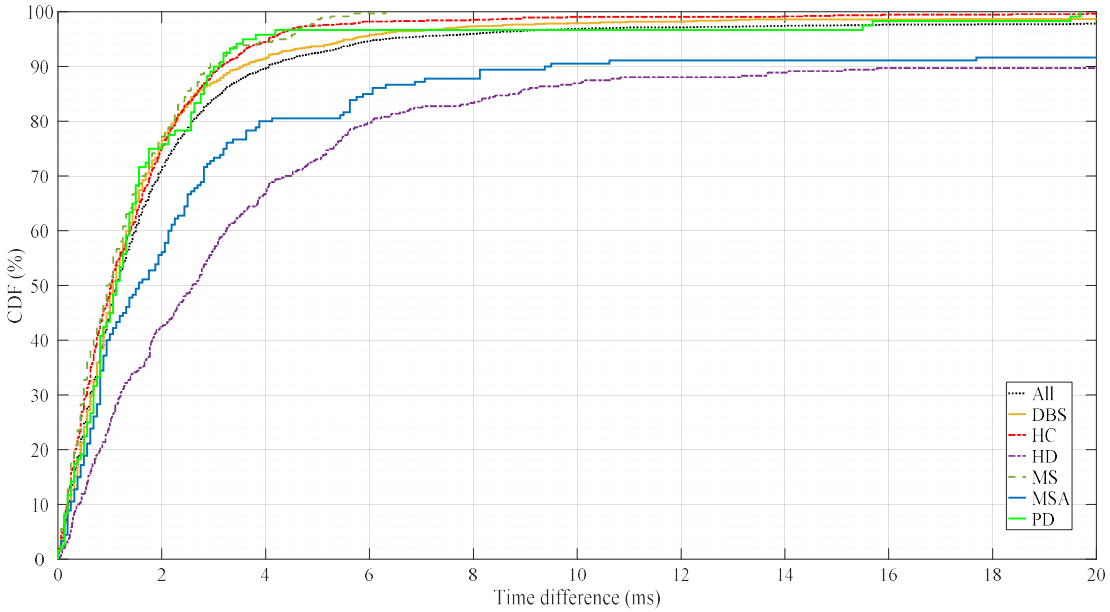


Figure 30 – DEVO performance on shortened utterances, V position, test set

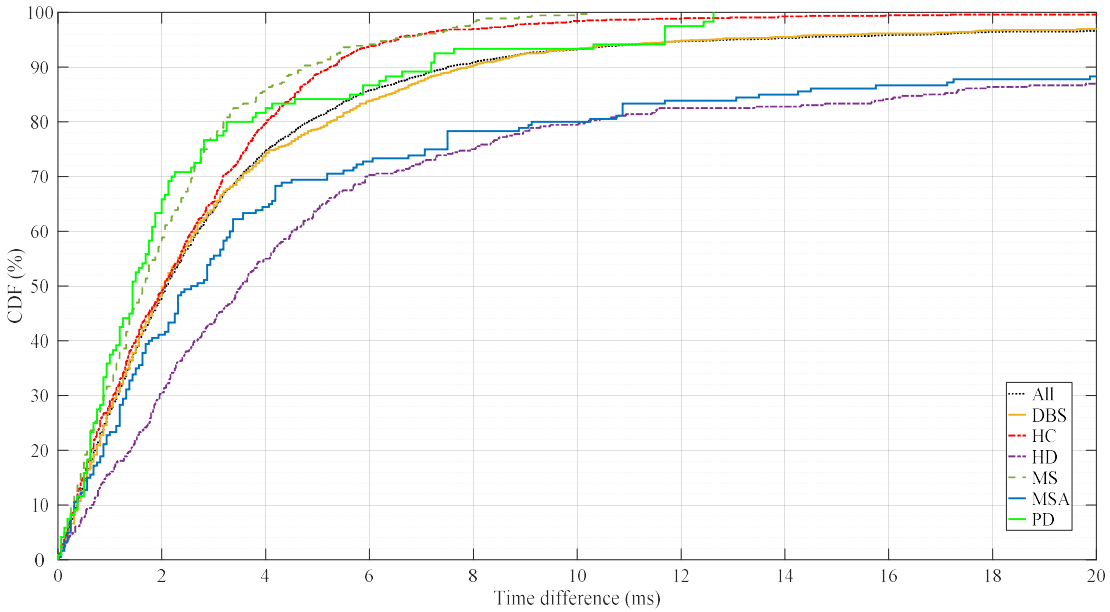


Figure 31 – DEVO performance on shortened utterances, O position, test set

On the test set, the performance of DEVO on shortened utterances, as seen in Figure 29, Figure 30, and Figure 31, is better than on the whole utterances, as seen in Figure 11, Figure 12, and Figure 13. In Table 11, the performance of DEVO on the test set of shortened utterances is summarised at selected key points.

Table 11 – DEVO performance on shortened utterances at selected key points, the test set only

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	72.78	92.49	80.82
10 ms	78.50	96.90	93.24
20 ms	86.16	97.84	96.63

We also calculated the performance of DEVO on the whole dataset for shortened utterances, as seen in Figure 32, Figure 33, and Figure 34. This reflects the final accuracy of EVO position detection with which we will approach the feature evaluation in the following step.

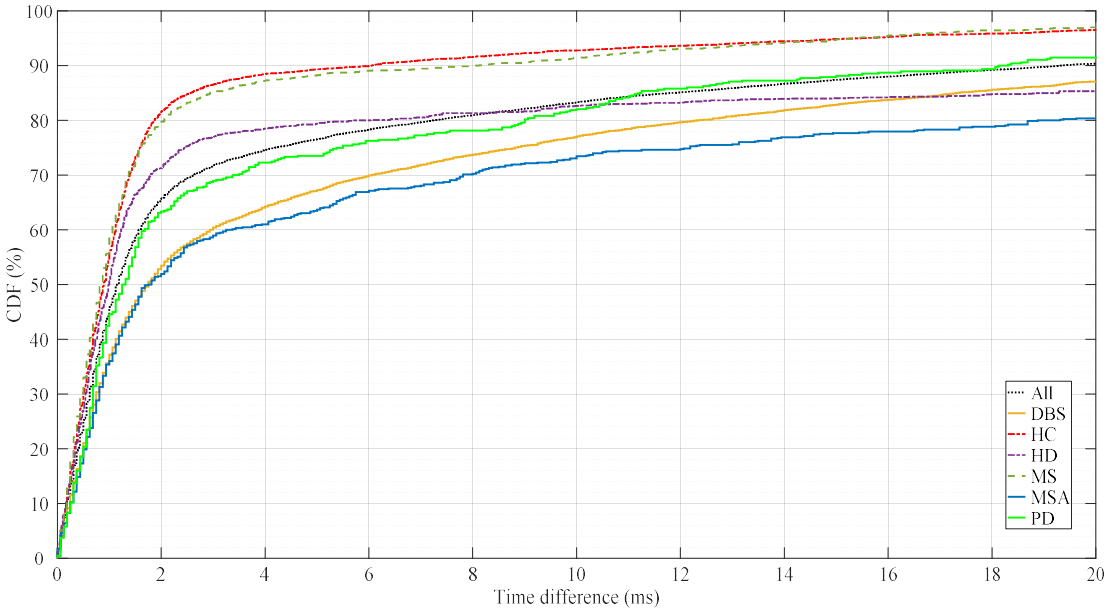


Figure 32 – DEVO performance on shortened utterances, E position, the whole dataset

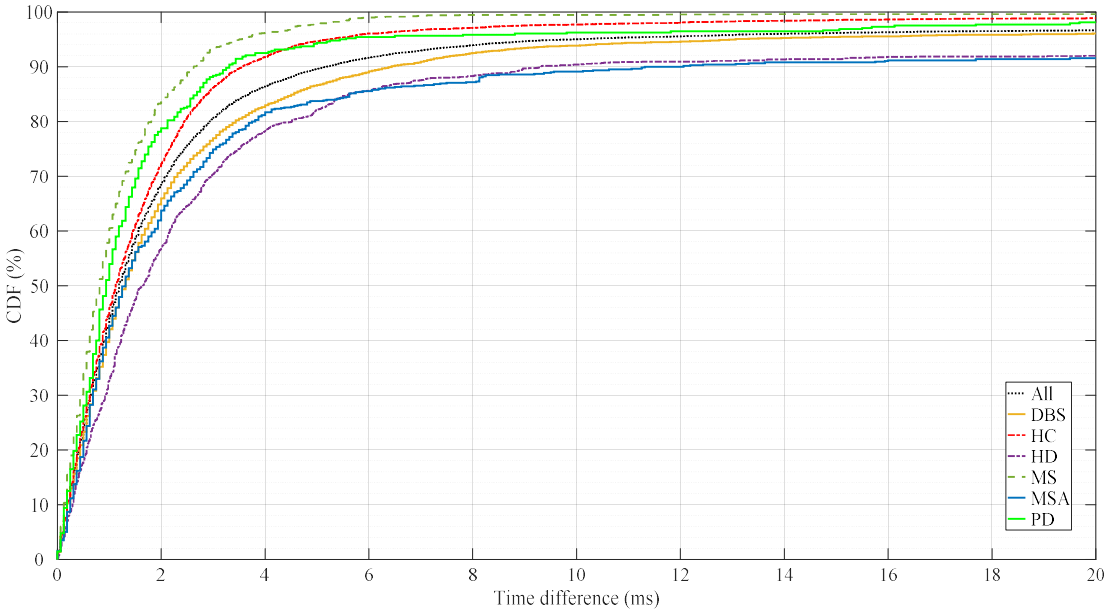


Figure 33 – DEVO performance on shortened utterances, V position, the whole dataset

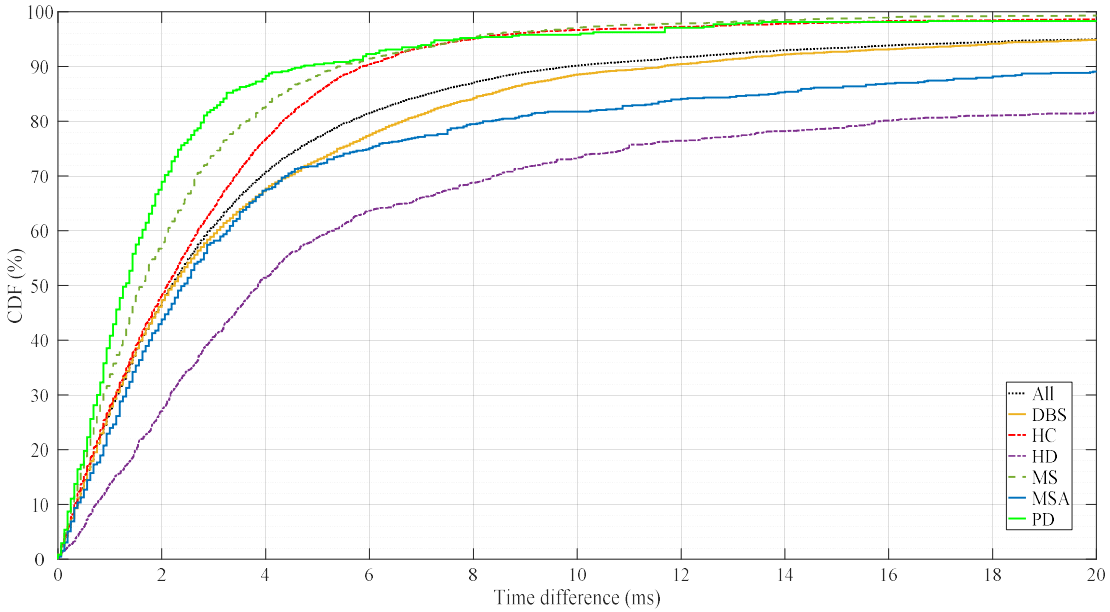


Figure 34 – DEVO performance on shortened utterances, O position, the whole dataset

A summary of the performance of DEVO for the whole shortened dataset is given in Table 12. Compared to the performance on the test, validation and training sets unshortened (Table 6, Table 7, Table 8), the performance on shortened utterances seems to be overall better.

Table 12 – DEVO performance on shortened utterances at selected key points, the whole dataset

Position	E	V	O
Time tol. / accuracy	(%)	(%)	(%)
5 ms	76.50	89.54	76.91
10 ms	83.26	95.00	90.14
20 ms	90.36	96.66	95.00

3.2. Feature evaluation

After applying DEVO to detect EVO positions, all the mentioned features from chapter 2.4 were extracted from the whole dataset, and the methods from chapter 2.5 were applied. First, the Kolmogorov–Smirnov test of normality was performed. The vast majority of the features agreed with the null hypothesis of normality. Next, the features were compared using one-way ANOVA and Tukey's HSD post hoc test. The results of feature significance are shown in Table 13, and in the following boxplot figures (Figure 35 – Figure 43), where the asterisk * denotes the significance as follows: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 13 – Significance of extracted features from one way ANOVA

Feature	F score	p significance
Syllable rate	37.16	p<0.001
Vowel length	6.13	p<0.001
RA	1.54	p=0.176
IoR	15.53	p<0.001
RI	18.58	p<0.001
VOT	19.23	p<0.001
sVOT	14.08	p<0.001
RIRV	13.87	p<0.001
IS	5.82	p<0.001

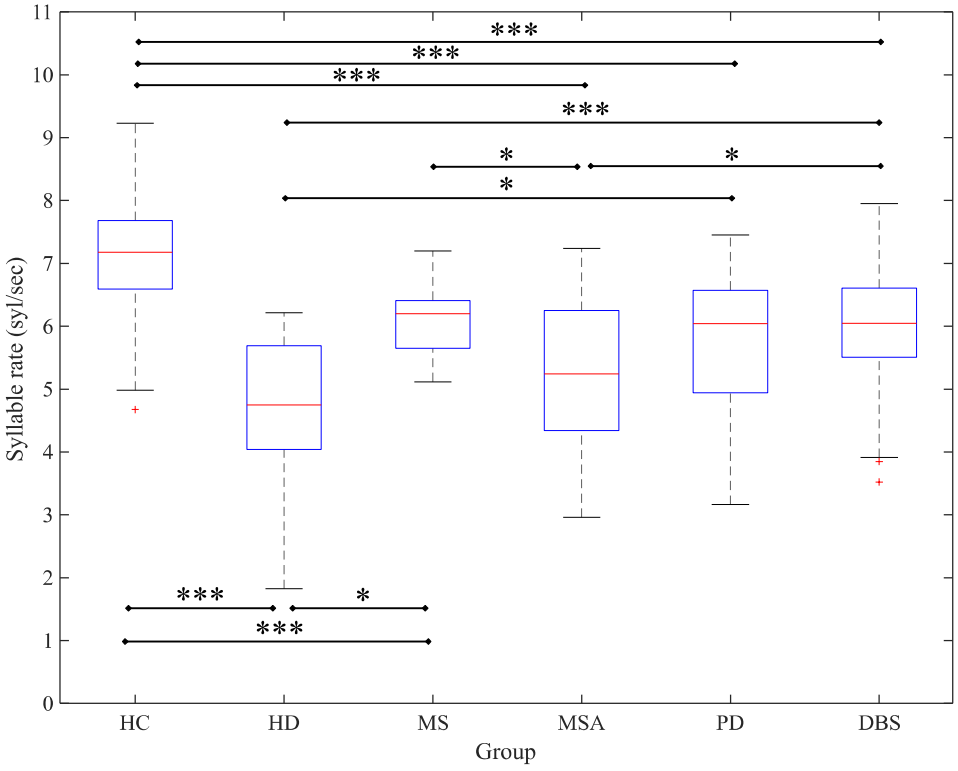


Figure 35 – Feature evaluation – Tukey's HSD on Syllable Rate (SR)

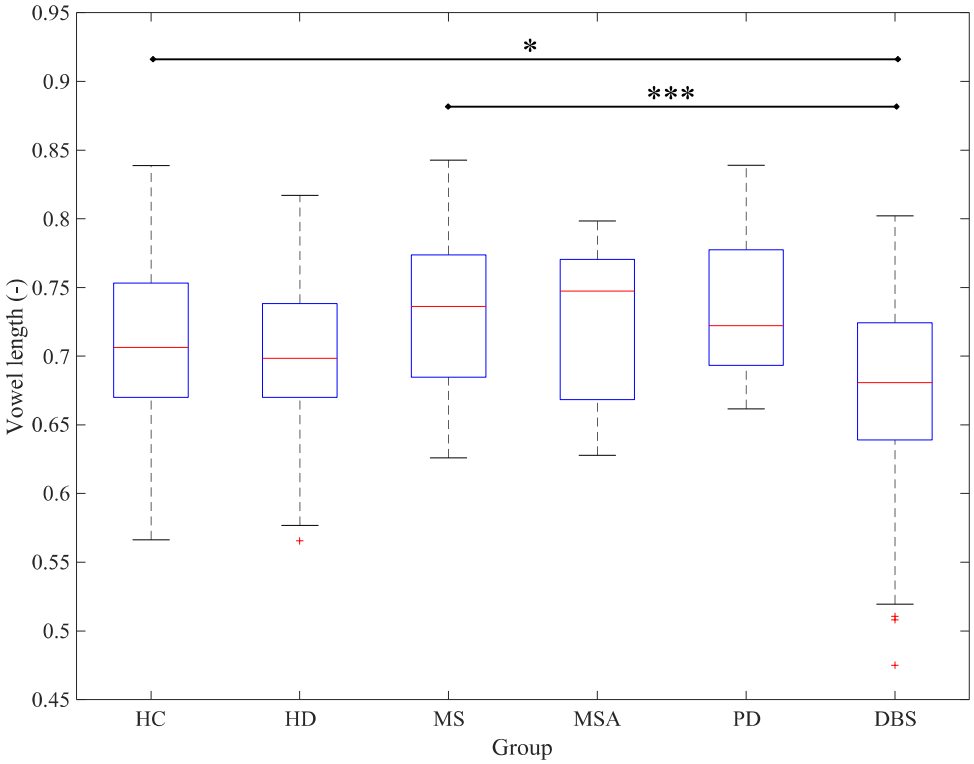


Figure 36 – Feature evaluation – Tukey's HSD on Vowel Length (VL)

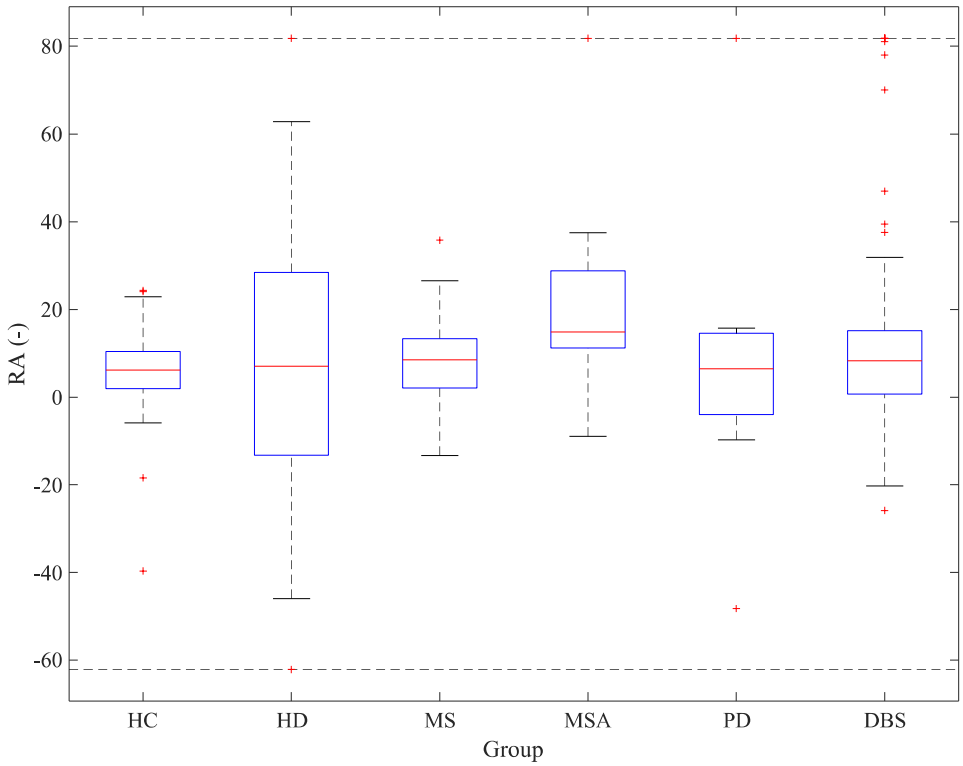


Figure 37 – Feature evaluation – Tukey's HSD on Rhythm Acceleration (RA)

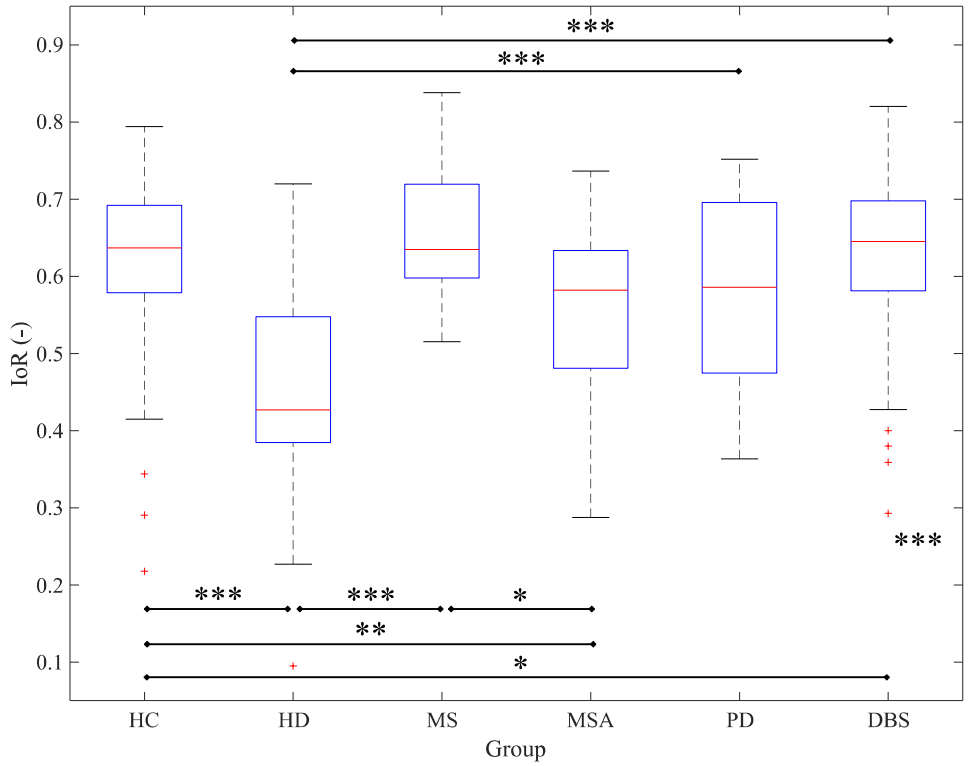


Figure 38 – Feature evaluation – Tukey's HSD on Index of Rhythmicity (IoR)

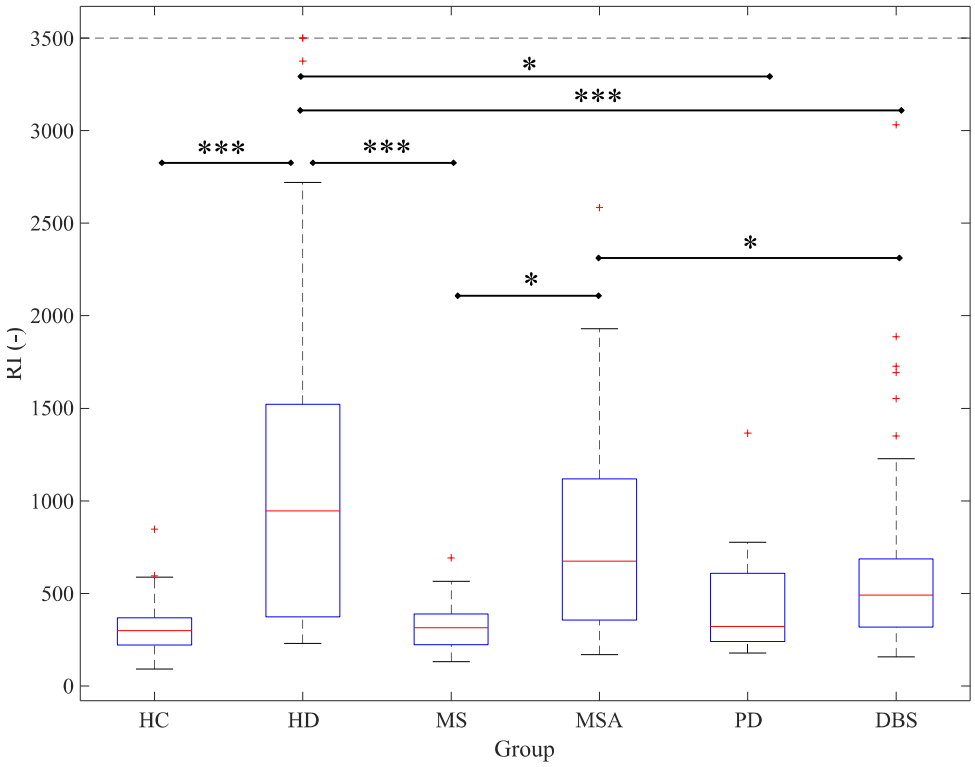


Figure 39 – Feature evaluation – Tukey's HSD on Rhythm Instability (RI)

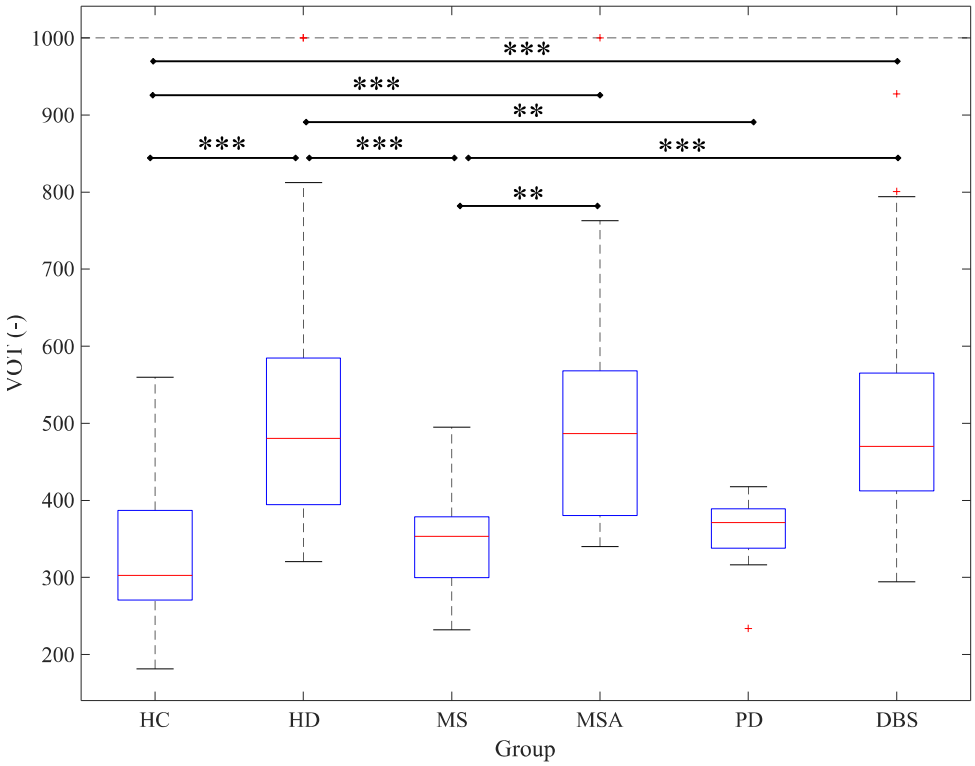


Figure 40 – Feature evaluation – Tukey's HSD on mean Voice Onset Time (VOT)

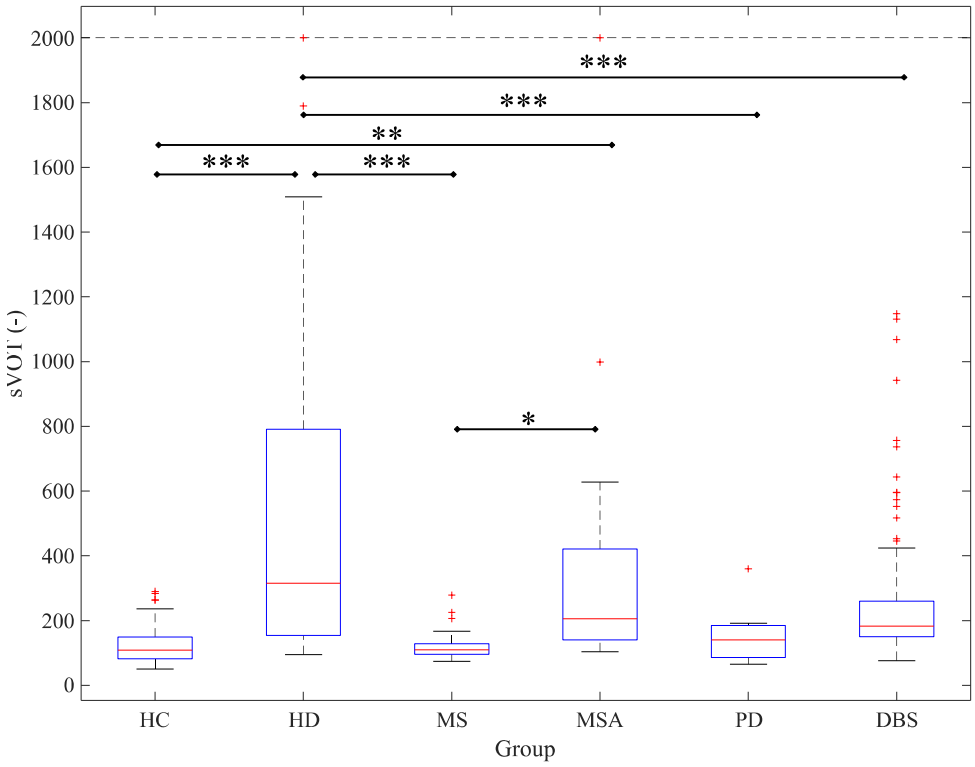


Figure 41 – Feature evaluation – Tukey's HSD on the standard deviation of Voice Onset Time (sVOT)

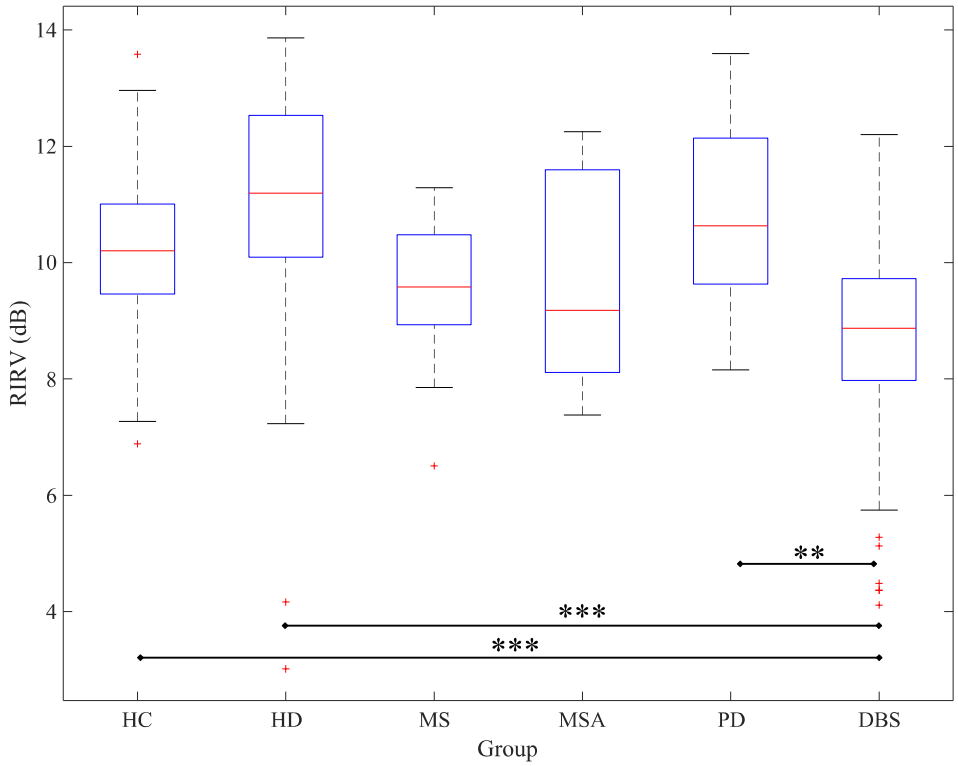


Figure 42 – Feature evaluation – Tukey's HSD on Relative Intensity Range Variation (RIRV)

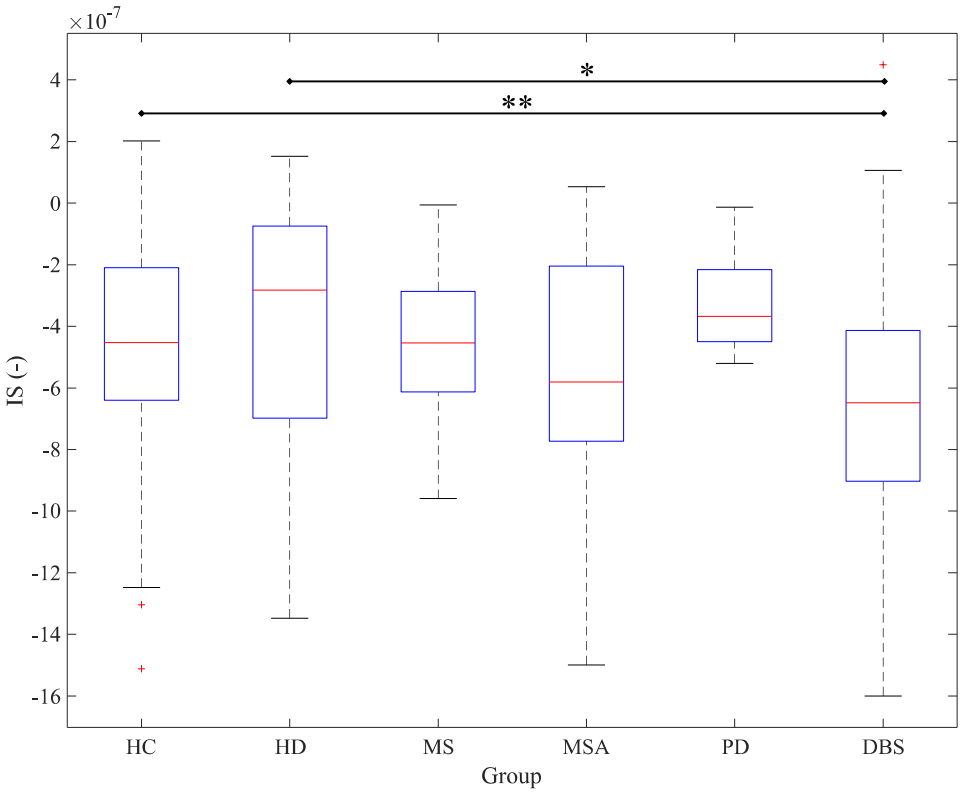


Figure 43 – Feature evaluation – Tukey's HSD on Intensity Slope (IS)

We performed a correlation test (Pearson's correlation coefficient) to see how the features are correlated. The results are shown in Table 14.

Table 14 – Correlation matrix of extracted features

Features	SR	VL	RA	IoR	RI	VOT	sVOT	RIRV	IS
SR	1.00	-0.15	-0.09	0.30	-0.61	-0.58	-0.59	-0.08	-0.07
VL	-0.15	1.00	-0.06	0.19	0.01	-0.45	-0.09	0.13	0.00
RA	-0.09	-0.06	1.00	-0.08	0.31	0.22	0.24	0.07	0.06
IoR	0.30	0.19	-0.08	1.00	-0.47	-0.27	-0.41	-0.29	-0.45
RI	-0.61	0.01	0.31	-0.47	1.00	0.71	0.89	0.14	0.13
VOT	-0.58	-0.45	0.22	-0.27	0.71	1.00	0.82	-0.05	0.00
sVOT	-0.59	-0.09	0.24	-0.41	0.89	0.82	1.00	0.13	0.16
RIRV	-0.08	0.13	0.07	-0.29	0.14	-0.05	0.13	1.00	0.21
IS	-0.07	0.00	0.06	-0.45	0.13	0.00	0.16	0.21	1.00

3.3. Disease classification

A classifier, as described in chapter 2.6 was built, and a 10 fold cross-validation was performed to obtain the confusion matrix in Table 15.

Table 15 – Confusion matrix of NN classifier, rows represent ground truth, columns predictions

Predictions (%)	DBS	HC	HD	MS	MSA	PD
DBS	58.78	7.14	4.49	6.53	14.49	8.57
HC	9.35	59.35	0.97	11.61	3.87	14.84
HD	18.89	4.44	52.22	3.33	10.00	11.11
MS	15.56	22.22	0.00	34.44	3.33	24.44
MSA	36.00	2.00	30.00	6.00	16.00	10.00
PD	0.00	26.67	23.33	13.33	6.67	30.00

The classifier's accuracy averaged across all testing data was 53.48 ± 4.62 % and 41.80 ± 5.66 % when averaged across classes, i.e. average of average accuracies as seen in Table 15, so to respect the class imbalance.

4. Discussion

In this chapter, we will explain and discuss the results and their meaning. First, we evaluate the performance of our speech segmentation algorithm, then we look at the feature evaluation and classification and last we illustrate the direction of future work in this topic.

4.1. Speech segmentation algorithm

In this thesis, we proposed two algorithms for speech segmentation and EVO position detection, SDEVO and DEVO, and tested their performance. In both the algorithms the E position detection (Figure 8, Figure 11) is the worst-performing among EVO, followed by the O position (Figure 10, Figure 13), which is performing a bit poorly at 5 ms tolerance, but makes a big jump in accuracy when considering 10 ms tolerance window. The best performing position is the V position (Figure 9, Figure 12), which is showing exquisite results, compared to the other two positions. Both the algorithms performed very well, and while SDEVO showed better syllable detection accuracy (Table 2), DEVO seemed to be better in terms of EVO detection accuracy (Table 3, Table 6). Therefore, we chose DEVO as the main speech segmentation algorithm for all the following experiments.

We compared the performance of DEVO to algorithms from [1] and [15]. Our DEVO algorithm beat the algorithm in [1] at almost all fronts. The only two better results of [1] are the excess syllable metric (Table 10), where DEVO seems to make up more of nonexistent syllables most probably due to the rules set in chapter 2.3.2, and the accuracy of the V position detection at 0 to 1 ms time tolerance (Figure 16, Figure 17), which is still an impressive fact about the algorithm in [1]. We think that this shows that traditional signal processing can achieve greater accuracy at very precise detection but lacks the robustness of the neural network approach to deal with more severe dysarthria.

As for the comparison with [15], our DEVO algorithm is still much better overall. Even though the performance of V and O position detection is not much different at 20 ms tolerance, it differs a lot for tighter tolerance interval, as seen in Figure 27, and Figure 28, and it differs in E position significantly. The algorithm from [15] produces fewer excess syllables (Table 10), but DEVO is still much better in terms of the number of missing syllables.

All the three algorithms showed to be class dependent in the performance, e.g. in Figure 14, Figure 15, Figure 20, and Figure 21 we can see that the performance for MSA and DBS classes is poor by all the algorithms. The fact that it occurs similarly in all the three algorithms performances suggests it is rather due to the nature of the dataset, not the algorithms. However, for some classes and some positions, the algorithms perform differently, e.g. in Figure 25, DEVO performs the best on the DBS class while the algorithm from [15] performs the worst on the DBS class.

Furthermore, we evaluated the performance of DEVO on shortened utterances (Figure 29 – Figure 34), i.e. considering only detected syllables between the 4^{th} and the 33^{rd} syllable inclusive, which shows the accuracy of EVO detection that was used for further feature extraction. Overall, the performance of DEVO on these shortened utterances was better than on full-length utterances. This is most likely due to the fact that the last syllables in utterances tend to be worse in terms of intelligibility and intensity as the patient runs out of breath and muscles may be getting exhausted. The very first syllables in an utterance can also be tricky, as the patient is trying to get “into rhythm” and is using perceptual feedback to adjust his speech. Therefore, as expected, removing these problematic syllables raised the EVO accuracy.

All in all, we showed that the neural network approach for speech segmentation is working very well and is a way forward as it is an improvement to the previous methods.

4.2. Feature evaluation

After using DEVO on the dataset, we extracted features and evaluated their significance (Table 13, Figure 35 – Figure 43). The worst performing feature was Rhythm Acceleration (RA), which was not able to significantly separate investigated groups. The best distinguishing feature was Syllable Rate (SR), followed by mean Voice Onset Time (VOT), Rhythm Instability (RI), Index of Rhythmicity (IoR), the standard deviation of VOT (sVOT), and Relative Intensity Range Variation (RIRV). Quite poorly, but still significantly, performed the Intensity Slope (IS) and Vowel Length (VL) features.

Even though SR is the most significant feature, its significance comes mainly from the difference between HC group and other groups. The differences among the rest are not that significant. As for RI and IoR, most of its significance value comes from HD comparison to other groups. On the other hand, VOT seems to have split groups into two new groups, one comprising of HC, MS, and PD, the other of HD, MSA, and DBS. However, this is more likely due to the severity of dysarthria present in those groups as MS and PD patients in our database do not have that severe dysarthria as patients from HD, MSA or DBS groups. In fact, a similar hypothesis was stated in [19] where they observed that the VOT feature reflected dysarthria severity in patients. A similar feature to VOT is sVOT, which shows similar results to VOT, but in sVOT the HD group is much more significant than in VOT, while DBS group is not, which makes for a highly significant difference ($p<0.001$) between the two. The remaining features of IS, RIRV and VL show significance only between DBS and some of the other groups.

As expected, there is no universal feature, so a combination of features is always required for better assessment of the diseases. What's worse, some of the groups are not distinguishable in even a single of the presented features, e.g. MS and PD groups. Therefore, we have to rely on a combination of features to give us some significance between the two. Looking at the groups that are somewhat distinguishable in some of the features, a decision tree approach could be taken to first remove these highly distinguishable groups from the decision based on the significant dividing feature.

We implemented a simple neural network classifier to try to distinguish the diseases using the extracted features. It reached a satisfactory accuracy (way better than guessing), but not high enough to be much useful. While DBS, HC and HD groups were classified fine, the classifier struggled to predict MS, PD and most of all MSA patients. However, as said before, MS patients in this dataset have only mildly severe dysarthria, which explains confusing them with mild PD and HC groups. As for MSA, their dysarthria is mixed with spastic, ataxic and hypokinetic components, which is quite similar to the DBS group, therefore explaining the confusion between MSA and DBS. The fact that our dataset had many more DBS than MSA recordings could contribute to this result as well, even though we weighted the loss function to take the imbalance into account.

4.3. Future work

The features used in this thesis are not yet optimal as the results present. In Table 14, we can see that some of them are highly correlated, which might not contribute to the classification positively. Usually, if the patient's dysarthria is severe, it will be reflected in all the features as

worse performance, which explains the correlation. This leads us to consider distinction by dysarthria severity in the future.

Besides dysarthria severity, our main goal in the future is to perform an assessment by dysarthria type as we believe that types could be distinguished better using the given acoustic (mainly rhythm) features. It is mainly due to the fact that dysarthria is a motor speech disorder that should manifest with certain perceptual features. In [2] they mention that highly trained clinicians or people with a small amount of intense training can make reliable distinctions between dysarthria types by perception, which makes us believe there definitely is some acoustic information that would allow for an automated distinction between the dysarthria types.

Regarding the speech segmentation algorithm, there is room for improvement. We believe that a net with more convolution filters could perform better. Another improvement of the net could be in the structure. The branch leading to E position detection might not need to have a large receptive field, while the V and O position branches could use larger receptive field, given that the layers get more filters to work with too. In this thesis, we were limited by GPU memory as all the computations were done on a PC.

References

[1] M. Novotný, J. Rusz, R. Čmejla and E. Růžička, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1366-1378, Sept. 2014, doi: 10.1109/TASLP.2014.2329734

[2] Y. Kim, R. D. Kent, and G. Weismer, "An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, pp. 417–429, 2011, doi: 10.1044/1092-4388(2010/10-0020)

[3] C. Roth, "Dysarthria," *Encyclopedia of Clinical Neuropsychology*, Springer, New York, pp. 905–908, 2011. doi:10.1007/978-0-387-79948-3_880

[4] M. D. Binder, N. Hirokawa, U. Windhorst, *Encyclopedia of Neuroscience*, Springer, Berlin, 2009, ISBN: 978-3-540-29678-2

[5] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," *Natural Reviews Disease Primers*, vol. 23, no. 3., art. 17013, 2017, doi: 10.1038/nrdp.2017.13

[6] A. K. Ho, R. Iansek, C. Marigliani, J. Bradshaw, and S. Gates, "Speech impairment in large sample of patients with Parkinson's disease," *Behavioural neurology*, vol.11, pp. 131–137, 1998, doi: 10.1155/1999/327643

[7] T. Tsuboi, H. Watanabe, Y. Tanaka, R. Ohdake, N. Yoneyama, K. Hara, et al., "Distinct phenotypes of speech and voice disorders in Parkinson's disease after subthalamic nucleus deep brain stimulation," *Journal of neurology, neurosurgery, and psychiatry*, vol. 86, no. 8, pp. 856-864, 2015, doi: 10.1136/jnnp-2014-308043

[9] I. S. Robles, "Huntington's disease," HOPES Huntington's Disease Information, *HOPES: Huntington's Disease Outreach Project for Education at Stanford* [online]. Copyright © 2019 HOPES Stanford University [quoted 18.04.2020]. Available at: https://hopes.stanford.edu/glossary/huntingtons-disease/

[8] G. K. Wenning, and A. Fanciulli, *Multiple System Atrophy,* Springer, Vienna, 2014, ISBN: 978-3-7091-0686-0

[10] B.G. Weinshenker, "The natural history of multiple sclerosis," *Neurologic Clinics*, vol. 13, pp. 119–146, 1995, doi: 10.1016/S0733-8619(18)30064-1

[11] F. S. Juste, S. Rondon, F. C. Sassi, A. P. Ritto, C. A. Colalto, and C. R. Andrade, "Acoustic analyses of diadochokinesis in fluent and stuttering children," *Clinics (Sao Paulo, Brazil)*, vol. 67, no. 5, pp. 409–414, 2012. doi: 10.6061/clinics/2012(05)01

[12] J. Rusz, R. Čmejla, H. Růžičková, and E. Růžička, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *Journal of the Acoustic Society of America*, vol. 129, no. 1, pp. 350–367, 2011, doi: 10.1121/1.3514381

[13] J. R. Duffy, "*Motor Speech Disorders: Substrates, Differential Diagnosis and Management*," 3rd ed., Mosby, St. Louis, p. 363, 2013, ISBN: 9780323087605

[14] H. Ackermann, I. Hertich, and T. Herh, “Oral diadochokinesis in neurological dysarthrias,” *Folia Phoniatrica et Logopaedica*, vol. 47, pp. 15-23, 1995, doi: 10.1159/000266338

[15] J. Hlavnička, “Automated analysis of speech disorders in neurodegenerative diseases,” *Doctoral thesis*, CTU FEE, 2019

[16] J. Rusz, J. Hlavnička, R. Čmejla, and E. Růžička, “Automatic evaluation of speech rhythm instability and acceleration in dysarthrias associated with basal ganglia dysfunction,” *Frontiers in Bioengineering and Biotechnology*, vol. 3, art. 104, July 2015, doi: 10.3389/fbioe.2015.00104

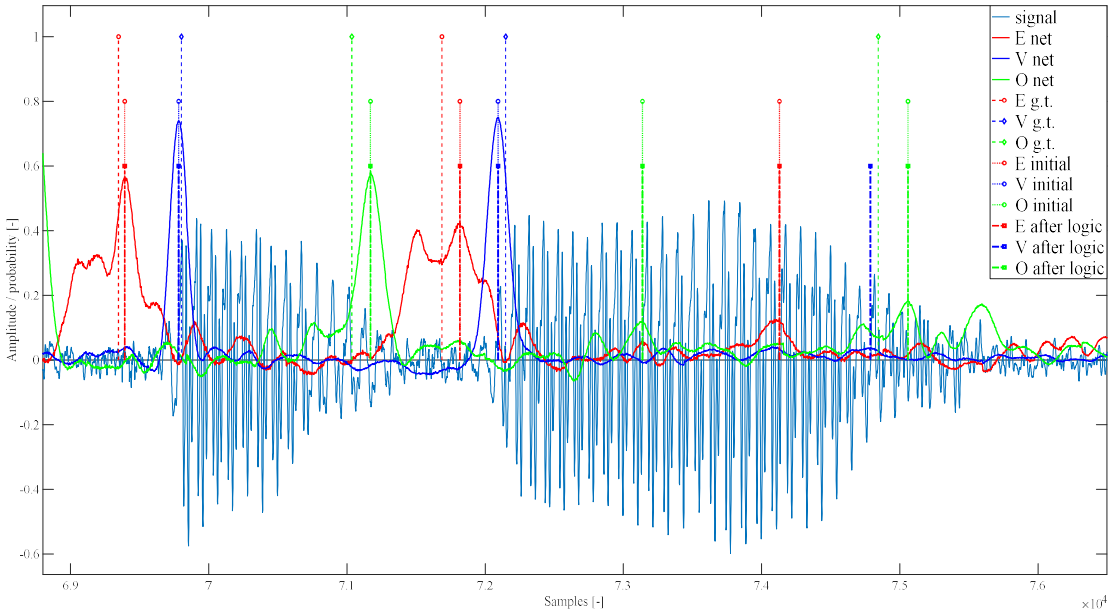
[17] P. Rong**, “**Automated Acoustic Analysis of Oral Diadochokinesis to Assess Bulbar Motor Involvement in Amyotrophic Lateral Sclerosis,” *Journal of Speech, Language, and Hearing Research,* vol. 63*,* pp. 59–73, 2020, doi: 10.1044/2019_JSLHR-19-00178

[18] K. Rozenstoks, M. Novotny, D. Horakova and J. Rusz, “Automated Assessment of Oral Diadochokinesis in Multiple Sclerosis Using a Neural Network Approach: Effect of Different Syllable Repetition Paradigms,” in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 32-41, Jan. 2020, doi: 10.1109/TNSRE.2019.2943064

[19] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, E. Ruzicka, “Distinct patterns of consonant articulation among Parkinson’s disease, progressive supranuclear palsy and multiple system atrophy,” *Brain and Language*, vol. 165, pp. 1–9, 2017, doi: 10.1016/j.bandl.2016.11.005

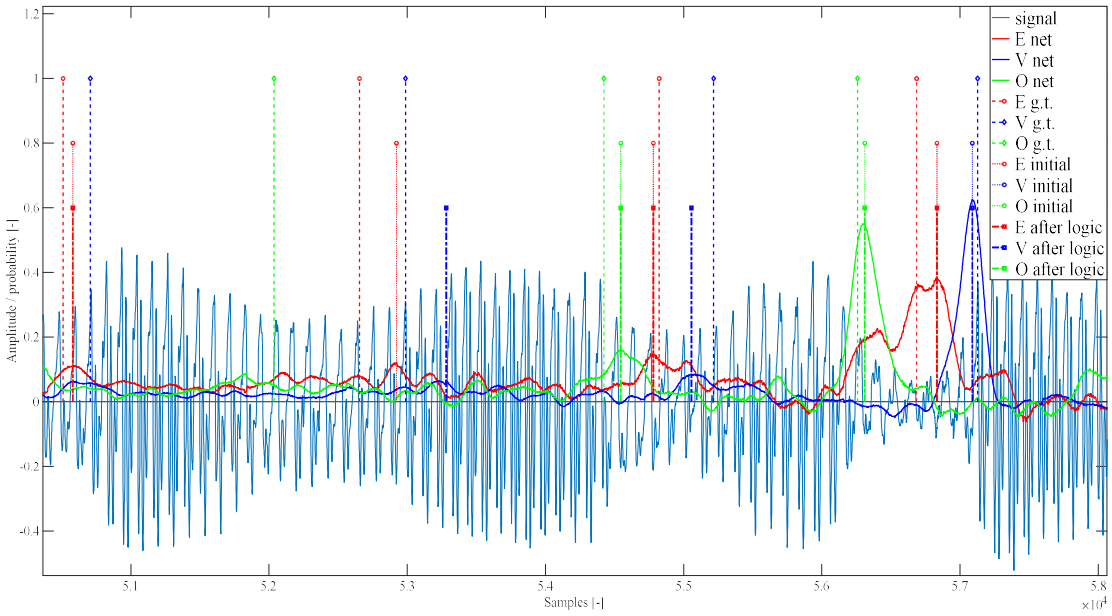
Appendix A – EVO detection examples

In Appendix Figure 1, we can see that the first syllable was detected fine, but in the second, additional O and E positions were detected which resulted in the logic to add a V position to complement, creating an excess syllable



Appendix Figure 1 – An example of an excess syllable detected

In Appendix Figure 2, we can see that the second ground truth syllable was not detected correctly. Initially, only the E position was detected, which was then deleted by the logic. In the first syllable, the V position was not found either. This results in one long detected syllable, which is actually two ground truth syllables melted into one.



Appendix Figure 2 – Missing syllable example

Appendix B – DVD with codes