Computers in Biology and Medicine 148 (2022) 105853

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Assessing clinical utility of machine learning and artificial intelligence approaches to analyze speech recordings in multiple sclerosis: A pilot study

E. Svoboda [a,b], T. Bořil [b], J. Rusz [c,d,e], T. Tykalová [c], D. Horáková [d], C.R.G. Guttmann [j], K.B. Blagoev [i], H. Hatabu [g], V.I. Valtchinov [f,g,h,*]

[a] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[b] Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czech Republic
[c] Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[d] Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic
[e] Department of Neurology & ARTORG Center, Inselspital, Bern University Hospital, University of Bern, Switzerland
[f] Center for Evidence-Based Imaging, USA
[g] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[h] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[i] Department of Biophysics, Johns Hopkins University, Baltimore, MD, 21218, USA
[j] Center for Neurological Imaging, Brigham & Women's Hospital and Harvard Medical School, USA

ARTICLE INFO

Keywords:
Multiple sclerosis
Dysarthria
Machine learning
Speech acoustics
Phonetics
Technology assessment
Biomedical

ABSTRACT

Background: An early diagnosis together with an accurate disease progression monitoring of multiple sclerosis is an important component of successful disease management. Prior studies have established that multiple sclerosis is correlated with speech discrepancies. Early research using objective acoustic measurements has discovered measurable dysarthria.

Method: The objective was to determine the potential clinical utility of machine learning and deep learning/AI approaches for the aiding of diagnosis, biomarker extraction and progression monitoring of multiple sclerosis using speech recordings. A corpus of 65 MS-positive and 66 healthy individuals reading the same text aloud was used for targeted acoustic feature extraction utilizing automatic phoneme segmentation. A series of binary classification models was trained, tuned, and evaluated regarding their Accuracy and area-under-the-curve.

Results: The Random Forest model performed best, achieving an Accuracy of 0.82 on the validation dataset and an area-under-the-curve of 0.76 across 5 k-fold cycles on the training dataset. 5 out of 7 acoustic features were statistically significant.

Conclusion: Machine learning and artificial intelligence in automatic analyses of voice recordings for aiding multiple sclerosis diagnosis and progression tracking seems promising. Further clinical validation of these methods and their mapping onto multiple sclerosis progression is needed, as well as a validating utility for English-speaking populations.

1. Introduction

It is almost universally accepted that Multiple sclerosis (MS), a chronic inflammatory autoimmune neurological disease of the central nervous system (CNS), is due to the changes of the CNS myelinated axons, creating inflammatory plaques that effectively cause demyelination with axonal transection [1,2].

The clinical development and manifestation course of MS is highly varied and unpredictable. In most MS patients, episodes of reversible neurological deficits is often followed by a progressive neurological deterioration [2]. Epidemiologically, MS affects at least 14,908 people in the Czech Republic and an estimated 620,000–720,000 people in the US [3]. It typically presents in young adults (mean age of onset, 20–30 years); it is expected that up to one-half of subjects will need physical help to walk within 15 years after the onset of the disease [2,4]. Additionally, it has been discovered that risk of MS increases 32-fold after infection by the Epstein-Barr virus [5]. Additionally, it has been shown in mouse models that the administration of an mRNA vaccine can

* Corresponding author. Center for Evidence Based Imaging (CEBI), One Brigham Circle, 1620 Tremont Street, 3rd Floor, Suite 3010, Boston, MA, 02120, USA.
E-mail address: vvaltchinov@bwh.harvard.edu (V.I. Valtchinov).

https://doi.org/10.1016/j.compbiomed.2022.105853
Received 28 December 2021; Received in revised form 9 April 2022; Accepted 23 May 2022
Available online 15 July 2022

suppress MS symptoms without introducing overt symptoms of general immune suppression [6].

There is no single diagnostic test for MS. Diagnosis is made based on clinical evidence from multiple testing procedures, some of which are quite invasive: a combination of presentation signs and symptoms, diagnostic imaging findings (for example, magnetic resonance imaging (MRI) T2 lesions), and laboratory findings (ie cerebrospinal fluid (CSF)–specific oligoclonal bands), which are components of the 2017 McDonald Criteria [7].

Recently, speech patterns have been shown to be a good indicator for the presence of neurological disorders, specifically in the case of MS. An early study in 1987 by Gerald et al. [8] was the first to describe the effects of this disease not only on speech, but also on linguistic capabilities in general. In a small sample of 23 individuals, they established that multiple sclerosis has noticeable effects on how afflicted individuals communicate. This study was extended by Rusz et al., in 2018 [9,10], where for the first time they introduced objective acoustic criteria showing that MS-afflicted speech differs significantly from normal speech. According to Hartelius [11], at least some vocal impairment is perceptually present in 51% of all MS patients.

In 2021, Noffs et al. [12] demonstrated that some objective acoustic measurements of speech correlate with disability scores in MS-afflicted patients even when there is no perceivable dysarthria present, specifically intensity decay and decreased frequency variability. The sensitivity of speech disorders towards MS has not insofar been assessed using fully automated methods based on individual phoneme segmentation, nor has the potential of such a vocal fingerprint to detect and track the progression of MS been tested using machine learning methods.

Machine learning has been demonstrated to show potential in a wide array of healthcare applications. For example, it has been shown that the technique, especially in the form of convolutional neural networks, can be used to detect Coronavirus based on chest X-Rays and computer tomography [13]. Similarly, detection of a wider range of ailments, such as Alzheimer's disease, glaucoma, arrhythmia, diabetes and brain tumors, based on various types of available data, such as PET or MRI scans [14,15].

Our aim is to demonstrate the value and predictive power of automatic phoneme segmentation in the context of neurological disease detection, as well as to assess of the potential of machine learning methods in the detection of multiple sclerosis from speech recordings.

To assess the utility of automated methods for speech analysis in MS patients in this study we undertook the following aims: a) create a set of acoustic parameters able to discern recordings of speakers with MS from healthy controls (HC); b) find out how strong are the differences between the MS-afflicted and the healthy speakers with respect to each of these parameters separately; c) test how well these parameters discriminate between these speakers using machine learning; and d) assess the feasibility of creating an automated tool for diagnosis and disease progression monitoring based on these (and potentially additional) parameters.

2. Methods

2.1. Study setting and human subjects approval

All MS patients were diagnosed with a neurologically-confirmed diagnosis of MS according to the revised McDonald Criteria [16]. All patients were relapse-free for at least 30 days prior to testing. Each patient was ranked according to the Expanded Disability Status Scale (EDSS) [17]. For neuropsychological assessment, the patients were tested with the Paced Auditory Serial Addition Test-3 (PASAT-3) [18] and the Symbol Digit Modalities Test (SDMT) [19].

In addition, a healthy control group free of neurological or communication disorders was included. All participants were native speakers of the Central Bohemian dialect of Czech.

2.2. Speech recordings

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) placed approximately 5 cm from the subject's mouth [20]. Speech signals were sampled at 48 kHz with 16-bit resolution. Each subject was recorded during a single session with a speech specialist.

All of these individuals were recorded reading out loud the same excerpt from Karel Čapek 's *Měl jsem psa a kočku* in the original Czech [21]. This text has the benefit of being cognitively and articulatorily little to moderately demanding while also utilizing the entirety of the Czech phonemic inventory. It is 230 syllables long and a healthy native speaker of Czech can be expected to read it aloud within 40–50 s.

2.3. Feature selection

The acoustic features measured on the recordings were deliberately chosen such that they would map onto the symptomatology of MS as meaningfully as possible. We identified three primary symptom clusters of MS, namely (*muscular) spasticity*, *ataxia*, and (*overall) fatigue*. We then hypothesized as to how these might manifest themselves acoustically by means of abnormal prosody, phoneme articulation and other speech production phenomena, and based the feature selection thereon.

For example, we predicted that individuals suffering from ataxia would have trouble appropriately modulating the amplitude of their speech due to reduced control of their breathing muscles, leading to pressure and by extension intensity spikes or drops throughout MS-afflicted recordings. We therefore included a relevant feature, CSI of intensity (Cumulative Slope Index – see the Appendix for exact definition and formula). Detailed descriptions of each of the features and their hypothesized mapping onto the symptom clusters can be found in the Appendix.

2.4. Annotation and feature extraction

Prague Labeller, a *HTK*-based implementation of the Hidden Markov Model algorithm originally intended for use in phonetics, was used to automatically delimit boundaries of phoneme realizations within each of the recordings [22]. In a recording of the Czech word *Minda*/minda/, for example, this tool finds the beginning and end of the articulation of the/m/phoneme, considering that a short silence may precede the word and thus delimiting silences as well as phonemic boundaries.

Additionally, *Praat* [23] was used alongside *Prague Labeller* to extract the intensity and fundamental frequency and formant curves for each of the recordings.

2.5. Validation of the automatic algorithm for phoneme extraction

To validate the accuracy of *Prague Labeller*, the features extracted using the tool were correlated against features extracted from the same speakers using human experts according to rules strictly defined in *Fonetická segmentace hlásek* [24]. In the case of one MS-speaker recording, it was discovered that *Prague Labeller* had placed the last boundary of the annotation midway through the recording, creating an artificial outlier. This recording was discarded, alongside its matched Control counterpart. Pearson's r correlation coefficient was used to cross-correlate the automatically extracted features with their counterparts extracted by human experts.

2.6. Algorithms for predictive risk modelling

These vectors were combined with information about the speakers' age and gender at birth, on which a binary classification array of models was trained using the R programming language and the package *caret* [25].

We evaluated the performance of 7 ML (machine learning) algorithms in building a predictive MS risk model with acoustic features and demographic variables as independent predictors. A common simulation and evaluation framework was set up using utilities from the caret R package [26]. The following algorithms were implemented and assessed: eXtreme Gradient Boosting, gbm (Generalized regression Boosting Model), GLMnet (Generalized Linear Model), KNN (k-Nearest Neighbors), Multi-Layer Perceptron Neural Network, Random Forrest (RF) and Support Vector machine with a radial kernel (SVM). We used a 5-fold cross-validation (CV) with an 80-20% balanced split for training and validation datasets. All acoustic and demographic variables were used in the model building and validation, regardless of the amount of variability in the dataset they explain, or any variable-selection procedures. The system performance was measured by the accuracy (the proportion of times the model's predictions agree with the labels of the data) using the validation set and the mean Area Under Curve (AUC) for the Receiver Operator Characteristic (ROC) across the individual CV runs on the training set.

2.7. Univariate statistical analyses

In addition, to assess the statistical significance of each variable, analysis of the individual features was performed. As the dataset was found not to be normally distributed, the Kolmogorov-Smirnov two-sample test was used.

2.8. Outcome measures

The primary outcome metrics of our analyses were the Accuracy on the holdout and the AUC for the binary classification models. Secondary outcomes were the values of the 7 acoustic and 2 demographic features extracted from the voice recording and used to construct a vector space to quantify and predict the risk of MS.

3. Results

3.1. Study cohort

A total set of 131 matched recordings of healthy and MS individuals was recorded. A total of 65 MS patients (41 females, 24 males), with a mean age of 43.9 (standard deviation [SD] 10.5) years, mean disease duration of 14.7 (SD 8.3) years, and mean EDSS of 3.9 (SD 1.4) were recruited. The patients exhibited a mean SDMT of 51.7 (SD 13.4) and a mean PASAT of 44.6 (SD 16.4). Fifty-one patients were diagnosed with relapsing remitting MS, 7 with secondary progressive MS, 2 with clinically isolated syndrome and 5 with primary-progressive MS.

According to the consensus judgment of two speech-language pathologists, the dysarthria was imperceptible in 30 MS patients. The perceptible dysarthria in remaining 35 MS patients mainly featured a combination of spastic and ataxic components with primary signs of slow rate, irregular speech timing, imprecise articulation, strained-strangled voiced and unnatural word stress expression.

In addition, 66 individuals comprised the controls cohort (41 females, 24 males) with a mean age of 45.5 (SD = 11.2) years.

The overall workflow implemented for collecting, storing, extracting the 7 speech features (plus Age and Gender) and systematically building and evaluating the predictive models is depicted in Fig. 1.

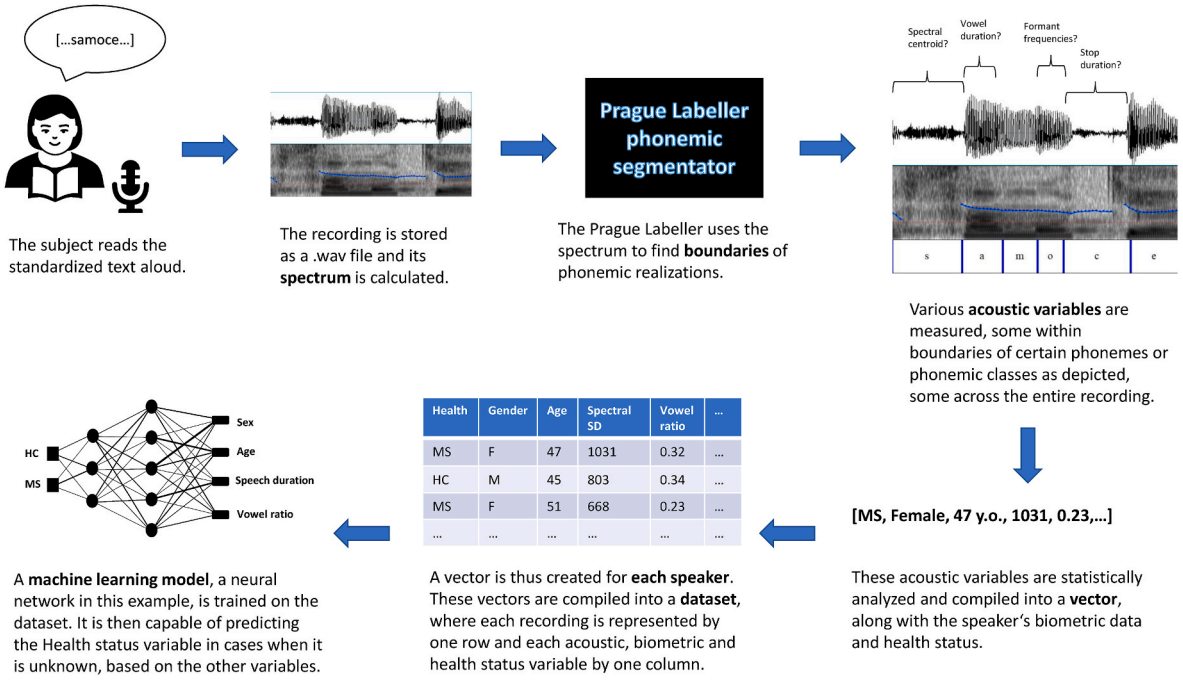

Fig. 1. A diagram of the feature extraction process workflow. The Prague Labeller tool – shown as a black box in the diagram – is used to find points in the recordings when one individual speech sound ends and another begins. In the case of the Czech word *samotě*/samoce/, for instance, Prague Labeller uses the fact that/s/ represents a noisy sound as opposed to/a/, which represents a tonal sound, to find the boundary between these two at a certain point in time. This is called a *phonemic boundary* and comes attached with a label of its respective phoneme.
Some acoustic features are then measured within these boundaries, like the spectral centroid of/s/, which simultaneously roughly corresponds to the perceived "sharpness" of the sound and the configuration of the tongue while it is being pronounced. Similarly, acoustic features are measured across the *entire* recording, such as the CSI of f_0, which represents the total melody fluctuation across the recording. These measured acoustic features, along with the given subject's sex and age, are compiled into vectors, which themselves are compiled into a dataset where each row represents one speaker with their corresponding acoustic features. One of the machine learning algorithms presented in Table 1, represented in the diagram using a neural network as an example, is then trained to predict the health status based on the biometric and acoustic data. Thus, it is possible to train a model to predict the neurological health status of an individual using nothing but a.wav recording of them reading a predetermined text out loud.

3.2. Feature extraction and validation

A resulting set of 12 acoustic features, listed in Table 1 and defined in the Appendix, were extracted from the delimited phonemes. These variables were used to distinguish between healthy and MS-afflicted individuals, as we have hypothesized that changes in them are correlated with articulatory and perhaps cognitive impairments. Fig. 2 includes some descriptive statistics for the 12-feature set extracted from the voice files.

To validate the algorithm from the *Labeller* utility, we used human expert annotation that were available for the entire HC cohort and a subset of the MS cohort, totaling 18 speakers. 7 of these features were found to be significantly correlated with expert annotation, as shown in Table 1.

The features whose automatically extracted values were not demonstrated to be correlated with human annotation were deemed unreliable and were not fed into the ML models or used in any other way.

Each recording was thus represented by a feature vector of length 9, where 7 positions represent a diagnostically relevant acoustic feature of a speaker's recording, and 2 represent the age and gender of the speaker.

3.3. Univariate and multivariate statistical analyses of variable significance

Next, we present the results of the univariate statistical significance testing using the Kolmogorov-Smirnoff (K–S) statistics, see Table 2. Only the 7 validated variables were included in the K–S testing, with five variables (listed in bold) were statistically significant or borderline significant (defined here as $p < \sim 0.1$).

Table 3 lists the results of the multi-variate statistically significant analyses after adjustment with all validated acoustic characteristics (variables) and gender and age, using a generalized linear regression model. Of note is that after adjustment, only two variables (CSI of vowel duration and Quantile difference of fundamental frequency) were borderline significant (see Table 4).

Table 1
A table of Pearson correlation coefficients and associated p values obtained by running a correlation test ofthe acoustic features extracted using *Prague Labeller* against features extracted using annotation by human experts on the same recordings, under the hypothesis that true correlation is greater than 0. Parameters found to be significant under $p < 0.05$ for both groups are in **bold**. "f_0" is a shorthand for fundamental frequency.

Parameter	Cases		Controls	
	p value	corr. coeff.	p value	corr. coeff.
Speech duration	**1.9×10^{-10}**	**0.99**	**8.9×10^{-84}**	**0.99**
Silence-to-speech ratio	0.41	0.06	0.005	0.31
Vowel-to-speech ratio	**2.5×10^{-5}**	**0.78**	**3.9×10^{-26}**	**0.91**
***CSI* of vowel duration**	**0.01**	**0.59**	**3.3×10^{-22}**	**0.87**
CSI of f_0	0.12	0.32	4.9×10^{-06}	0.51
Quantile difference of f_0	**0.001**	**0.70**	**2.1×10^{-16}**	**0.81**
Unvoiced stop mean duration	**1.4×10^{-7}**	**0.93**	**1.9×10^{-39}**	**0.97**
***CSI* of intensity**	**2.4×10^{-27}**	**0.99**	**2.5×10^{-84}**	**0.99**
Spectral centroid of/s/, SD	**5.3×10^{-7}**	**0.83**	**1.3×10^{-15}**	**0.79**
Vowel F1, SD	0.40	0.07	1.1×10^{-4}	0.44
Vowel F2, SD	0.41	0.04	0.42	0.02
Vowel F3, SD	0.53	−0.02	1.1×10^{-15}	0.79

3.4. Classification models

The results of the systematic model building, and evaluation are presented in Table 2. There were several models whose AUC achieved similar performance: eXGB, Gradient Boosting Machine, Random Forest and, to a degree, the Neural Network model. We chose to select the best-performing model based on a combination of both accuracy and the AUC measures (we used the rank of each model in each metric): it appears the best performing model was a Random Forest, which achieved an accuracy of 0.82 on a holdout validation dataset, and an AUC of 0.76 as measured across 5 train/test cycles on the training data.

3.5. Discussion

In this study, we present a fully automated, quantitative assessment methodology to detect and objectively measure the footprint of the disease in the speech of MS patients. This opens the door for a scalable and unbiased diagnostic, disease progression and treatment response assessment of MS that can in principle aid the current standard clinical diagnostic assessment for MS (i.e., review of clinical history and examination, brain and spinal cord MRI, CSF analysis etc.; the McDonald Criteria [7]). Using only speech recordings and basic demographic factors, we have been successful in achieving an accuracy of 0.82 (AUC = 0.76) using fully automated algorithmic methods. This points to the fact that MS carries a possible distinct vocal fingerprint which can in principle be utilized to diagnose or at least pre-diagnose the disease using methods orthogonal to the current standard-of care set of tests and procedures.

Currently, most available acoustic methods to capture motor speech disorders are designed for highly functional vocal paradigms such as sustained phonation or syllable repetition (for the most recent review, please see Ref. [27]). In fact, this is the first study to quantify speech disorder in MS based on a fully automated approach allowing precise segmentation of individual phonemes from connected speech, leading to the detection of a wide variety of natural speech abnormalities associated with MS that might not be seen using simpler functional vocal paradigms. Indeed, despite the fact that our classification results are based on a single paradigm of reading a short passage, our classification accuracy of 0.82 is superior to the accuracy of 0.79 previously reported using three types of speaking tasks including sustained vowel, fast/-pa-/ta/-/ka/syllable repetition and reading passage [10]. In addition, our findings appear to be superior to different voiced and unvoiced syllable repetition paradigms, with reported classification accuracy between MS and controls ranging between 0.68 and 0.74 [28]. To the best of our knowledge, only these two studies [10,28] have been previously demonstrated discrimination accuracy between MS and controls based on machine learning or deep learning/AI approach, and are therefore the only studies whose results are directly comparable to ours.

Indirectly, however, our study can be compared to Vavougios et al. [29], who were able to produce a discriminant function equation able to reach an accuracy of 100% in discriminating MS patients from healthy controls using. The experiment was, however, performed using electroglottographic (EGG) data, as opposed to acoustic data (as in our experiment). Similarly, Noffs et al. [30] built a unified acoustic speech score significantly correlated with cerebellar white matter volume and quality of life. This speech score was able to predict abnormal 9-hole peg test (9HPT) results with 85% accuracy. The predictive power of this speech score with regards to discriminating healthy controls from MS speakers has, however, not been measured in the study.

Our study also corroborates some of the results of the landmark study published by Gerald et al., in 1987 [8], in which MS-related dysarthria was assessed auditorily only. Because the results in our study have been measured objectively, they are much more rigorous; Gerald's study however covers other areas of affected linguistic capabilities, such as impaired grammar, which are much more non-equivocal to assess automatically. Therefore, finding a way to objectively assess such

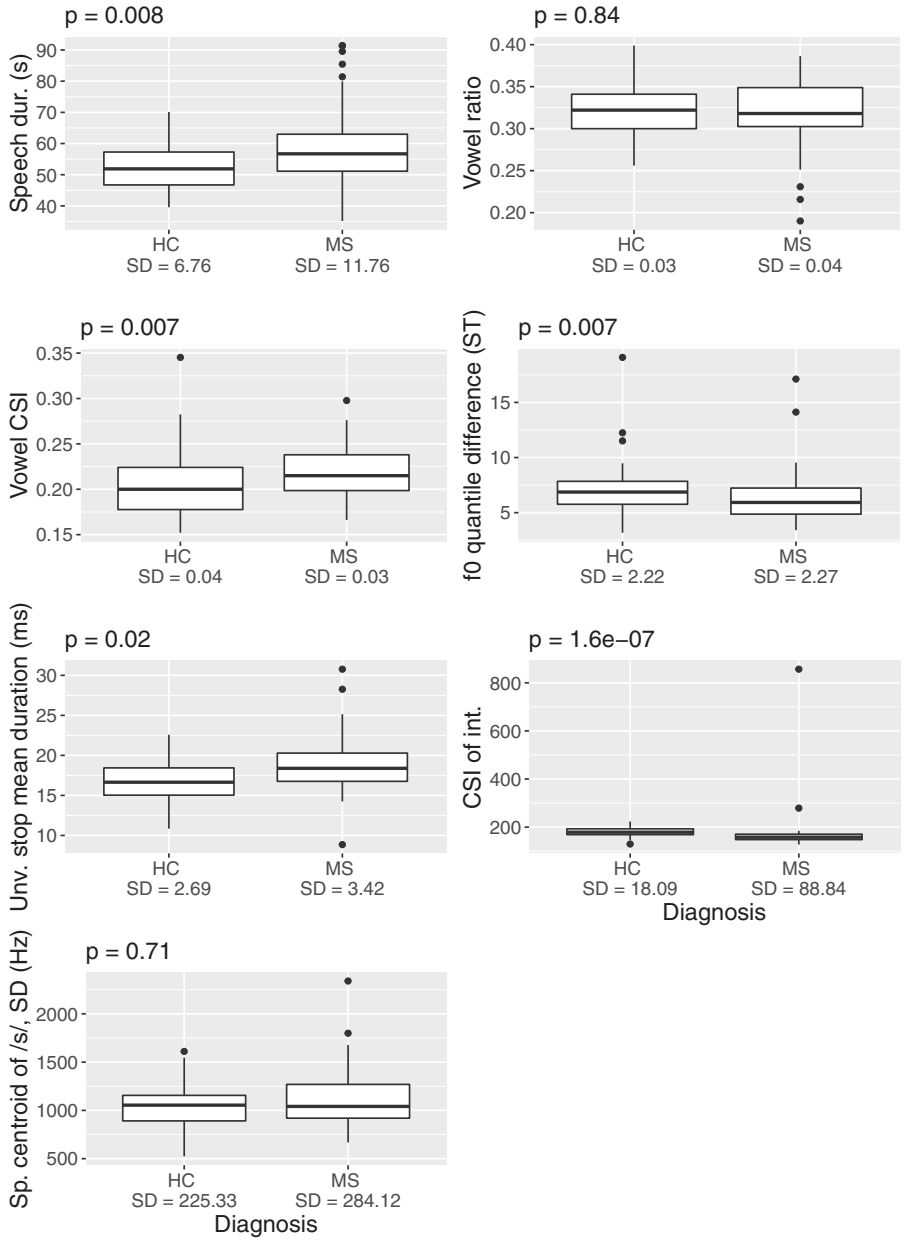

Fig. 2. Boxplots of the 7 acoustic measurements that were used as features. Note that recording and unvoiced durations are measured in milliseconds, the fundamental frequency measurements ("f_0") are in semitones (because fundamental frequency corresponds to articulatory phenomena logarithmically) and the formant measurements are listed in Hertz.

Table 2
A table of the univariate statistics *p*-values of MS against HC calculated using the Kolmogorov-Smirnov test for the 7 variables extracted by the *Labeller* tool andvalidated against human expert annotations, see Methods and also Table 1. Parameters deemed significant or borderline statistically significant are in **bold**. "f_0" is a shorthand for fundamental frequency.

Parameter	*p* value
Speech duration	0.008
Vowel-to-recording ratio	0.84
CSI of vowel duration	0.007
Quantile difference of f_0	0.007
Unvoiced stop mean duration	0.02
CSI of intensity	1.6×10^{-7}
Spectral centroid of/s/, SD	0.71

MS-related linguistic impairment beyond articulatory difficulties may present another set of indicators to help preliminary diagnosis of the disease.

Recently, a number of studies have attempted to use "real world" data (normally, clinical records data) to assess the risk of MS patient trajectories that transition from various established states in the MS disease progression, i.e. in the general disease course [31], or for example the initial Relapsing-Remitting (RR) to the Secondary Progressive (SP) form of the disease [32]. It is of interest to see how the alternative set of variables extracted from the voice patterns as shown in this study could be added to these types of clinical predictive approaches to potentially enhance the accuracy of the resulting models.

This study brings to the forefront another interesting research subject, e.g.the use of acoustic feature characterization from voice as a proxy measure when trying to find anatomical correlates of the symptoms in brain magnetic resonance imaging (e. g. lesion-symptom

Table 3
A table of the p values individual features (variables) after adjustment with the 7 validated acoustic features (see Table 1) and gender and age included, as calculated using a generalized linear model, MS against HC. The variables in **bold** are borderline significant. "f_0" is a shorthand for fundamental frequency.

Feature	p value
Speech duration	0.40
Vowel-to-recording ratio	0.82
CSI of vowel duration	**0.10**
Quantile difference of f_0	**0.09**
Unvoiced stop mean duration	0.64
CSI of intensity	0.53
Spectral centroid of/s/, SD	0.77
Age	0.42
Gender	0.61

Table 4
A comparison of the Accuracy and AUC scores of each model as calculated on the validation dataset, and the mean area under the ROC curve as calculated across 5 resamples of cross-validation on the training set. The best performing model is in **bold**.

Model	Accuracy	Sensitivity	Specificity	AUC
eXtreme Gradient Boosting	0.70	0.73	0.67	0.79
Gradient Boosting Machine	0.77	0.77	0.77	0.75
Neural Network	0.69	0.73	0.67	0.66
Random Forest	**0.82**	**0.90**	**0.75**	**0.76**
k-Nearest Neighbors	0.58	0.67	0.56	0.66
Support Vector Machine	0.46	0.45	0.47	0.66

mapping experiments) [33,34]. For example, Rusz et al. [35] correlated specific articulatory difficulties with particular brain volume changes, and in 2020, Rozenstoks et al. [28] demonstrated that MS patients exhibit significant difficulty in performing certain tasks involving alternating syllables.

Analyzing Fig. 2, MS speakers show greater variability than the healthy controls regarding speech duration. This is likely a consequence of pyramidal involvement due to widespread grey and white matter reductions in speakers with spastic dysarthria, leading to slower reading, which is consistent with Clark et al. [36]. The generally lower pitch range of MS speakers, as shown by the fundamental frequency quantile difference, probably reflects low range of vibration frequencies resulting in monotonic speech. The overall greater mean durations of unvoiced stops of MS patients shows a tendency to hold full articulatory closures for an abnormally long time, indicating muscular spasticity of the tongue, which is consistent with the findings of Tykalová et al. [37]. The two individuals exhibiting extreme high variability of intensity, as reflected in their f0 quantile difference values, pointing towards spasticity of the breathing muscles. Finally, articulation of the phoneme/s/requires producing a shallow groove along the center of the tongue, which requires a high degree of coordination. The greater standard deviation of the phoneme's spectral centroid of MS speakers seems to reflect ataxia due to their compromised ability to consistently articulate the difficult phoneme.

Lastly, the procedure correlating the values of the automatically obtained features against the feature values obtained with the help of manual annotation constituted a validation step ensuring that the information fed into the models is meaningful and not simply a collection of artifacts caused by *Prague Labeller* or *Praat*, ensuring the validity of our models.

3.6. Implications

This study describes a fully automated, quantitative assessment methodology to detect and objectively measure the footprint of the disease in the speech of MS patients. This opens the door for a scalable and unbiased diagnostic, disease progression and treatment response assessment of MS that can in principle be used in conjunction with current methods and aid the current standard clinical diagnostic assessment for multiple sclerosis.

3.7. Limitations

Our study is not without limitations. First and foremost, it is possible that for a significant proportion of MS patients, there might be no dysarthria present at all. Theoretically, these zero-dysarthria patients should, however be relatively rare (at least as detected by acoustic analysis [19]), because speech motor control is governed by a high number of different areas of the CNS [8,35]. Related to this lack of diagnostic specificity to MS, this initial work has not established a precise map between MS disease progression stages and the speech patterns (variables) that are predictive of the disease status. This latter correspondence and its clinical validation need additional studies.

Additionally, the subjects have had MS for an average of 15 years, which is a rather long time.

Furthermore, the set of features and ML-models trained thereon used in this study has been limited to a specific language, Czech. Theoretically, there is nothing overly language-specific about this set of features and the workflow utilized, but additional studies and validation, including a clinical validation, are needed for English-speaking populations such as in the US, as well as other languages. We expect our results to extend to other languages, because it has been demonstrated that motor speech disorders can be acoustically measured cross-linguistically [38].

Error analysis of the predictions of the Random Forest model on the validation set reveals that the model's predictions of MS status seem much more reliable under the condition that the given speaker also has perceptible dysarthria. Specifically, as described in Table 5, while all MS speakers with perceptible dysarthria were correctly classified, only 50% of MS without perceptible dysarthria were correctly classified as MS. This suggests that the classifier's performance was mainly limited by the presence or non-presence of dysarthria in a given speakers speech rather than the fine-tuning of the models or feature selection. The validation dataset is relatively small, however, so further research is necessary to confirm this (see Table 6).

Finally, our dataset is of rather limited size. Larger cohorts are desirable to validate and extend the findings reported here.

4. Conclusions

The use of machine learning and artificial intelligence in automated analyses of voice recording for aiding diagnosis, disease and treatment

Table 5
The performance of the Random Forest model compared with the presence of perceptible dysarthria as evaluated by the consensus of two speech experts. Please note that this table only shows the MS-positive subset of the validation dataset.

Health status	Prediction	Age	Disease duration	Perceptible dysarthria
MS	MS	25	11	0
MS	MS	46	19	0
MS	H	42	7	0
MS	MS	54	3	0
MS	H	40	21	0
MS	MS	41	3	0
MS	H	57	23	0
MS	H	47	15	0
MS	MS	49	14	1
MS	MS	40	23	1
MS	MS	47	6	1
MS	MS	40	16	1
MS	MS	39	17	1

Table 6
Comparison of the results of our study compared to recent similar studies. An asterisk denotes a study that differs from ours either in the type of data used or in the predictive task, and should therefore not be compared to ours directly, but is nevertheless worth mentioning. Please refer to the Discussion section for further details.

Study	Data	Prediction	Accuracy
Our study	Acoustic	MS vs. HC	0.82
Rusz et al.	Acoustic	MS vs. HC	0.79
Rozenstoks et al.	Acoustic	MS vs. HC	0.74
Vavougios et al.*	EGG	MS vs. HC	1.00
Noffs et al.*	Acoustic	Normal vs. abnormal 9HPT	0.85

progression of multiple sclerosis holds promise. Further clinical validation of the specificity and the mapping to the MS disease progression phases is needed and validating utility for other languages.

In the future, we would like to apply deep learning techniques such as convolutional neural networks to try and either increase the performance of the detection process, or confirm what the error analysis seems to point toward – that the model fails to detect MS in some patients due to their objective lack of dysarthria as opposed to problems such as suboptimal feature selection, model tuning, or training dataset size.

It would also be advisable to transfer the techniques applied in this paper on speakers of English.

Ethics approval and consent to participate

The study was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All participants provided written, informed consent to the neurological examination and recording procedure.

Consent for publication

Not Applicable.

Funding

This study was supported by the Czech Ministry of Education: National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107) - Funded by the European Union – Next Generation EU, and by the Cooperatio Program, research area Neuroscience.

Authors' contributions

All authors have read and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not Applicable.

Appendix

Speech duration

The total duration of the recording in seconds, beginning with the first word and ending with the last, as delimited by either *Prague Labeller* or the human experts. Motivated by the assumption that trouble with *ataxia* would lead to articulatory difficulty, leading to longer recording times.

Silence-to-speech ratio

The total time spent by the speakers being silent divided by Speech duration.

Motivated by the assumption that speakers struggling with muscular *spasticity* would speak in short, labored bursts, and that speaker struggling with *fatigue* would pause frequently to rest.

Vowel-to-speech ratio

The total time spent by the speakers articulating vowels divided by Speech duration.

Motivated by the assumption that because vowels perceptually require less articulatory effort per unit of time to pronounce than consonants, speakers struggling with *fatigue* would spend more time articulating them so as to give themselves a moment of perceived respite.

CSI of vowel duration

Cumulative Slope Index is the absolute value of the sum of differences between each two consecutive elements of a vector of values, in this case the vector of durations of each vowel pronunciation across a given recording. It can be interpreted as a scalar describing the total rate of change of a variable over a series of steps in time. It is given by the formula

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

where x is the vector of values in question, n is the index of each element of that vector and N is the total length of the vector, according to Volín et al. [39].

Additionally, the entire sum can be divided by Speech duration, which is referred to as *normalized CSI*. Every *CSI* measurement used in this study has been normalized.

Motivated by the fact that abnormally high or low normalized *CSI* of vowel duration suggests abnormal speech rhythm, possibly pointing toward

either of *spasticity* and *ataxia* or even higher-level cognitive difficulties.

CSI of fundamental frequency

CSI calculated from the *fundamental frequency* contour of the recording, or in other words, the vector of zeroth formant frequency values calculated at each time step, measured in semitones with a reference frequency of 100 Hz. The reason semitones were chosen is the fact that the logarithmic nature of the unit compensates for gender-wise differences in average fundamental frequencies between genders. It was measured using the auto-correlation method and octave jumps were not manually corrected. This may informally be interpreted as the *total rate of change in voice pitch* across the recording.

The time duration of these time steps were determined by *Praat*'s default setting, which applies to all *CSI* and formant measurements used in this paper [40].Motivated by the fact that poor control of the laryngeal muscles related to *ataxia* may contribute to an abnormal base frequency *CSI* value.

Quantile difference of fundamental frequency

The difference between the third quartile value of the base frequency vector and the first quartile value of the base frequency vector. Measured the same way as in the previous parameter. May be interpreted informally as the *voice pitch range* of the recording.

Motivated by the fact that poor control of the laryngeal muscles related to *ataxia* may lead to an abnormally high or low quantile difference.

Unvoiced stop mean duration

The mean duration the speaker spent pronouncing the phonemes/p/,/t/,/k/,/c/, which are prototypically pronounced with a full closure of two articulatory organs.

Motivated by the fact that individuals struggling with *spasticity* may hold this closure for a longer time.

CSI of intensity

CSI calculated from the *sound intensity* contour of the recording measured in decibels. May informally be interpreted as the total rate of change of speech loudness.

Motivated by the fact that *fatigue* may result in a flatter intensity contour due to breathing muscle weakness, leading to an unusually small CSI value; alternatively, *spasticity* of the breathing muscles may lead to a greater CSI value than normal.

Standard deviation of the spectral centroid of/s/

Standard deviation of the center of mass of the acoustic spectrum of all pronunciations of the/s/phoneme across the recording. Can informally be understood from a perceptual viewpoint as the *overall variation in sound sharpness* when pronouncing this phoneme.

Motivated by the fact that the sharpness of/s/is determined by the ability to produce and maintain a shallow groove in one's tongue whilst it is pressed against the roof of the mouth as a sufficiently strong flow of air from the lungs is maintained. Since this is a difficult task from a muscle coordination perspective, discrepancies may reflect *ataxia*.

Standard deviations of F1, F2 and F3

Standard deviations of the frequencies of the first three spectral maxima apart from the base frequency (formants) measured using the Burg method.

Motivated by the fact that the variability of the first three formants across time strongly correlates with the range of motion of the jaw and tongue and overall muscle tenseness.

A low SD of F1 could therefore indicate difficulty utilizing the full range of motion of the jaw, a relatively heavy organ, pointing to *fatigue*; conversely, a high SD of the same could point to a high amount of correctional movements, indicating *ataxia*.

A low SD of F2 could point to an unusually immobile tongue and thus *spasticity*; a high SD of F2 could again point to a high amount of correctional movements, indicating *ataxia*.

References

[1] M.M. Goldenberg, Multiple sclerosis review, Pharmacol. Ther. 37 (3) (2012 Mar) 175–184.

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

[3] Z. Pavelek, L. Sobíšek, J. Šarláková, P. Potužník, M. Peterka, I. Štětkárová, et al., Comparison of therapies in MS patients after the first demyelinating event in real clinical practice in the Czech republic: data from the national registry ReMuS, Front. Neurol. 11 (2021) 1833.

[4] M.P. McGinley, C.H. Goldschmidt, A.D. Rae-Grant, Diagnosis and treatment of multiple sclerosis: a review, JAMA 325 (8) (2021 Feb 23) 765–779.

[5] K. Bjornevik, M. Cortese, B.C. Healy, J. Kuhle, M.J. Mina, Y. Leng, et al., Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis, Science 375 (6578) (2022 Jan 21) 296–301.

[6] C. Krienke, L. Kolb, E. Diken, M. Streuber, S. Kirchhoff, T. Bukur, et al., A noninflammatory mRNA vaccine for treatment of experimental autoimmune encephalomyelitis, Science 371 (6525) (2021 Jan 8) 145–153.

[7] A.J. Thompson, B.L. Banwell, F. Barkhof, W.M. Carroll, T. Coetzee, G. Comi, et al., Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, Lancet Neurol. 17 (2) (2018 Feb 1) 162–173.

[8] F.J.F. Gerald, B.E. Murdoch, H.J. Chenery, Multiple sclerosis: associated speech and language disorders, Aust. J. Hum. Commun. Disord. 15 (2) (1987 Dec 1) 15–35.

[9] J. Rusz, Detecting Speech Disorders in Early Parkinson's Disease by Acoustic Analysis, 2018.

[10] J. Rusz, B. Benová, H. Růžičková, M. Novotný, T. Tykalová, J. Hlavnicka, et al., Characteristics of motor speech phenotypes in multiple sclerosis, Mult. Scler. Relat. Disord. 19 (2018 Jan) 62–69.

[11] L. Hartelius, B. Runmarker, O. Andersen, Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: relation to neurological data, Folia Phoniatrica Logop. 52 (4) (2000) 160–177.

[12] G. Noffs, F.M.C. Boonstra, T. Perera, H. Butzkueven, S.C. Kolbe, F. Maldonado, et al., Speech metrics, general disability, brain imaging and quality of life in multiple sclerosis, Eur. J. Neurol. 28 (1) (2021 Jan) 259–268.

[13] A. Waleed Salehi, P. Baglat, G. Gupta, Review on machine and deep learning models for the detection and prediction of Coronavirus, Mater. Today Proc. 33 (2020) 3896–3901.

[14] A.W. Salehi, G. Gupta, Sonia, A prospective and comparative study of machine and deep learning techniques for smart healthcare applications, Mob. Health Adv. Res. Appl. (2021) 163–189.
[15] R. Yousef, G. Gupta, C.H. Vanipriya, N. Yousef, A comparative study of different machine learning techniques for brain tumor analysis, Mater. Today Proc. (2021 Apr 1).
[16] C.H. Polman, S.C. Reingold, B. Banwell, M. Clanet, J.A. Cohen, M. Filippi, et al., Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria, Ann. Neurol. 69 (2) (2011) 292–302.
[17] J.F. Kurtzke, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS), Neurology 33 (11) (1983 Nov 1), 1444–1444.
[18] D.M.A. Gronwall, Paced auditory serial-addition task: a measure of recovery from concussion, Percept. Mot. Skills 44 (2) (1977 Apr 1) 367–373.
[19] A. Smith, Symbol Digit Modalities Test (SDMT). Manual, Western Psychological Services, Los Angeles, CA, 1973.
[20] J. Rusz, T. Tykalova, L.O. Ramig, E. Tripoliti, Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders, Mov. Disord. 36 (4) (2021) 803–814.
[21] K. Čapek, Měl Jsem Psa a Kočku [Internet], Městská knihovna v Praze, 1939 [cited 2020 May 16]. Available from: https://search.mlp.cz/cz/titul/mel-jsem-psa-a-kocku/3347549/.
[22] P. Pollák, J. Volín, R. Skarnitzl, HMM-based phonetic segmentation in Praat environment, in: Proceedings of the VII Th International Conference "Speech and Computer–SPECOM, 2007, pp. 537–541.
[23] P. Boersma, D. Weenink, Praat: doing phonetics by computer (Version 5.1.13) [Internet], Available from: http://www.praat.org, 2009.
[24] P. Machač, R. Skarnitzl, Fonetická Segmentace Hlásek. [Internet], Epocha, 2010. Available from: https://search.ebscohost.com/login.aspx?authtype=shib&custid=s1240919&profile=eds.
[25] M. Kuhn, topepo/caret [Internet] [cited 2020 May 10]. Available from: https://github.com/topepo/caret, 2020.
[26] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Software 28 (1) (2008 Nov 10) 1–26.
[27] L. Moro-Velazquez, J.A. Gomez-Garcia, J.D. Arias-Londoño, N. Dehak, J.I. Godino-Llorente, Advances in Parkinson's Disease detection and assessment using voice and speech: a review of the articulatory and phonatory aspects, Biomed. Signal Process Control 66 (2021 Apr 1), 102418.
[28] K. Rozenstoks, M. Novotný, D. Horáková, J. Rusz, Automated assessment of oral diadochokinesis in multiple sclerosis using a neural network approach: effect of different syllable repetition paradigms, IEEE Trans. Neural. Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc. 28 (1) (2020) 32–41.
[29] G.D. Vavougios, T. Doskas, K. Konstantopoulos, An electroglottographical analysis-based discriminant function model differentiating multiple sclerosis patients from healthy controls, Neurol. Sci. 39 (5) (2018 May 1) 847–850.
[30] G. Noffs, F.M.C. Boonstra, T. Perera, S.C. Kolbe, J. Stankovich, H. Butzkueven, et al., Acoustic speech analytics are predictive of cerebellar dysfunction in multiple sclerosis, Cerebellum Lond Engl 19 (5) (2020 Oct) 691–700.
[31] Y. Zhao, T. Wang, R. Bove, B. Cree, R. Henry, H. Lokhande, et al., Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study, Npj Digit Med. 3 (1) (2020 Oct 16) 1–8.
[32] R. Seccia, S. Romano, M. Salvetti, A. Crisanti, L. Palagi, F. Grassi, Machine learning use for prognostic purposes in multiple sclerosis, Life 11 (2) (2021 Feb) 122.
[33] E. Bates, S.M. Wilson, A.P. Saygin, F. Dick, M.I. Sereno, R.T. Knight, et al., Voxel-based lesion-symptom mapping, Nat. Neurosci. 6 (5) (2003 May) 448–450.
[34] S.M. Wilson, Lesion-symptom mapping in the study of spoken language understanding, Lang Cogn Neurosci. 32 (7) (2017) 891–899.
[35] J. Rusz, M. Vaněčková, B. Benová, T. Tykalová, M. Novotný, H. Ruzickova, et al., Brain volumetric correlates of dysarthria in multiple sclerosis, Brain Lang. 194 (2019) 58–64.
[36] H.M. Clark, J.R. Duffy, J.L. Whitwell, J.E. Ahlskog, E.J. Sorenson, K.A. Josephs, Clinical and imaging characterization of progressive spastic dysarthria, Eur. J. Neurol. 21 (3) (2014) 368–376.
[37] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, E. Ruzicka, Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy and multiple system atrophy, Brain Lang. 165 (2017 Feb 1) 1–9.
[38] J. Rusz, J. Hlavnička, M. Novotný, T. Tykalová, A. Pelletier, J. Montplaisir, et al., Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease, Ann. Neurol. 90 (1) (2021) 62–75.
[39] J. Volín, T. Tykalová, T. Boril, in: Stability of Prosodic Characteristics across Age and Gender Groups, 2017, 3902–6.
[40] Time step settings. [Internet]. [cited 2021 Feb 24]. Available from: https://www.fon.hum.uva.nl/praat/manual/Time_step_settings__.html.