Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Circuit Theory

Identification of vocal biomarkers in Parkinson´s disease and related movement disorders via automated acoustic analysis

Disertation thesis

Ing. Vojtěch Illner

Ph.D. programme: Bioengineering
Supervisor: Doc. Ing. Jan Rusz, Ph.D.

Prague, 2024

Thesis Supervisor:
Doc. Ing. Jan Rusz, Ph.D.
Department of Circuit Theory
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2
160 00 Prague 6
Czech Republic

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

In Prague, 2024

.....

Ing. Vojtěch Illner

Contents

Abstract

Early identification of Parkinson's disease (PD) during its prodromal stage is crucial for the advancement of neuroprotective therapies. Unfortunately, accurate biomarkers for prodromal PD are lacking, hindering early detection. Speech dysfunction typically emerges early in PD, suggesting potential for vocal assessments in patients with isolated rapid eye movement sleep behavior disorder (iRBD), a prodromal PD condition, and PD. This approach could serve as a diagnostic and progressive biomarker for PD and related synucleinopathies, with the opportunity of a remote, passive monitoring via smartphones. However, challenges remain, such as developing reliable automated algorithms to assess speech features and ensuring robustness against poor microphone quality or background noise.

The current study encompasses multiple investigations into using smartphones to capture speech as a biomarker for PD. Firstly, reliable automated methods were established to assess various physiological aspects of speech production. These methods demonstrated deficits in pathological utterances, including impairments in phonation, prosody, speech timing, and articulation. Secondly, a smartphone application and data acquisition system were developed to monitor subjects' speech unobtrusively through calls and active tasks. Finally, a cross-sectional study involving iRBD and PD patients was conducted using the developed system, supporting the use of smartphones to detect speech abnormalities. This approach not only aids in diagnosis but also has potential applications in enhancing current treatment strategies for diagnosed PD patients, providing feedback in neuropsychiatry, mitigating speech-related side effects of deep brain stimulation through parameter optimization, population screening, and more.

Keywords: Prodromal synucleinopathy biomarker, Parkinson's disease, speech, voice, dysarthria, smartphone, telehealth, machine learning

Abstrakt

Včasná identifikace Parkinsonovy nemoci (PN) v jejím prodromálním stadiu má zásadní význam pro rozvoj neuroprotektivní léčby. Bohužel, v současnosti nejsou známy žádné přesné biomarkery prodromální PN, což brání jejímu včasnému odhalení. Dysfunkce řeči se obvykle objevuje v brzkém stadiu PN, což naznačuje potenciál pro hodnocení řeči u pacientů s izolovanou poruchou chování ve fázi spánku s rychlými očními pohyby (iRBD), což je prodromální stav PN, a PN. Tento přístup by mohl sloužit jako diagnostický a progresivní biomarker pro PN a příbuzné synukleinopatie s možností pasivního monitorování na dálku prostřednictvím chytrých telefonů. Přetrvávají však výzvy, jako je vývoj spolehlivých automatizovaných algoritmů pro hodnocení fyziologických řečových vzorců a zajištění odolnosti proti špatné kvalitě mikrofonu nebo šumu na pozadí.

Tato studie zahrnuje několik výzkumů vedoucích k využití chytrých telefonů k zachycení řeči jako biomarkeru PN. Nejprve byly ustanoveny spolehlivé automatizované metody pro výpočet různých fyziologických aspektů produkce řeči. Tyto metody prokázaly deficity v patologické řeči, zahrnující poruchy fonace, prozodie, časování řeči a artikulace. Za druhé byla vyvinuta aplikace pro chytré telefony a systém sběru dat, který umožňuje neinvazivně a eticky nahrávat řeč subjektů prostřednictvím hovorů a aktivních úloh. Nakonec byla pomocí vyvinutého systému provedena průřezová studie zahrnující pacienty s iRBD a PN, která podpořila využití chytrých telefonů k detekci řečových abnormalit. Tento přístup může pomoci nejen při brzké diagnostice, ale má také potenciální využití při vývoji současných léčebných metod pro pacienty diagnostikované s PN, poskytování zpětné vazby v neuropsychiatrii, zmírňování vedlejších účinků hluboké mozkové stimulace pomocí optimalizace parametrů, populačním screeningu a dalších.

Klíčová slova: Biomarker prodromální synucleinopatie, Parkinsonova nemoc, řeč, hlas, dysarthrie, smartphone, telehealth, strojové učení

Acknowledgements

I express my gratitude to my supervisor, Doc Ing. Jan Rusz PhD., for his support and insightful research guidance throughout my doctoral studies. Additionally, I am appreciative of the entire Signal Analysis and Modeling team at the Faculty of Electrical Engineering, Czech Technical University in Prague, for their assistance, support, and companionship. Special thanks go to my former classmates, now colleagues, for our daily lunches and coffee breaks, despite the messy office environment. Their valuable insights greatly contributed to my doctoral journey. Furthermore, I extend my thanks to Prof. Ing. Pavel Sovka CSc. for his support in teaching the Adaptive Signal Processing course and for providing comprehensible technical explanations.

I would also like to express my thanks to my girlfriend and family for their support and understanding. Lastly, I am thankful to my friends for being themselves.

Foreword

This dissertation, submitted to fulfill the requirements for the Ph.D. degree in Bioengineering at the Czech Technical University of Prague, Faculty of Electrical Engineering, is the culmination of six studies conducted at the Department of Circuit Theory. Four articles have been published in impacted journals, one is currently in a peer review, and one was presented at a conference of a rank A[1]. Additionally, one study unrelated to the thesis topic was presented at a conference of a rank A.

List of author's publications related to the doctoral thesis

Articles in impacted journals

Smartphone voice calls provide early biomarkers of parkinsonism in REM sleep behaviour disorder (2024)
V. Illner, M. Novotný, T. Kouba, T. Tykalová, M. Šimek, P. Sovka, J. Švihlík, E. Růžička, K. Šonka, P. Dušek, J. Rusz
Currently in a peer review in Movement Disorders (Q1, IF 8.6).

Automated Vowel Articulation Analysis in Connected Speech Among Progressive Neurological Diseases, Dysarthria Types, and Dysarthria Severities (2023)
V. Illner, T. Tykalova, D. Skrabal, J. Klempir, J. Rusz
Journal of Speech, Language, and Hearing Research: JSLHR, 66(8), 2600–2621, 2023
Impact factor (2022): 2.6
Quartile in category[2]: Q1
Number of citations[3]: 1

Toward Automated Articulation Rate Analysis via Connected Speech in Dysarthrias (2022)
V. Illner, T. Tykalová, M. Novotný, J. Klempíř, P. Dušek, J. Rusz
Journal of Speech, Language, and Hearing Research, 65(4), 1386-1401, 2022.
Impact factor (2022): 2.6
Quartile in category: Q1
Number of citations: 0

[1] According to CORE23 ranking.
[2] According to Web of Science
[3] According to Web of Science, without auto citations.

Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies (2022)
T. Kouba[§], V. Illner[§], J. Rusz
BMJ Open, 12(6), 2022
Impact factor (2022): 2.9
Quartile in category: Q2
Number of citations: 5

Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease (2020)
V. Illner, P. Sovka, J. Rusz
Biomedical Signal Processing and Control, 58, 2020.
Impact factor (2022): 5.1
Quartile in category: Q2
Number of citations: 11

Conference proceedings

Which aspects of motor speech disorder are captured by Mel Frequency Cepstral Coefficients? Evidence from the change in STN-DBS conditions in Parkinson's disease (2023)
V. Illner, P. Krýže, J. Švihlík, M. Sousa, P. Krack, E. Tripoliti, R. Jech, J. Rusz
INTERSPEECH 2023, 5027–5031.
CORE23 ranking: A
Number of citations: 0

List of author's publications not related to the doctoral thesis

Conference proceedings

Relationship between LTAS-based spectral moments and acoustic parameters of hypokinetic dysarthria in Parkinson's disease (2023)
J. Svihlik, V. Illner, P. Kryze, M. Sousa, P. Krack, E. Tripoliti, R. Jech, J. Rusz
INTERSPEECH 2023, 1758–1762.
CORE23 ranking: A
Number of citations: 0

[§]The authors contributed equally.

Chapter 1

Introduction

Parkinson's disease (PD) is a neurological condition marked by the loss of dopaminergic neurons in the brain's substantia nigra [1]. It affects approximately 1.8% of individuals over 65 years old, and the figure is expected to rise due to longer life expectancy [2], [3]. Currently, there are no treatments capable of halting or slowing PD's progression; available pharmacological and surgical interventions only alleviate specific symptoms. Diagnosis typically occurs when key motor symptoms manifest, such as severe tremor, by which time up to half of the substantia nigra's neurons may already be irreversibly damaged [4]. Unfortunately, reliable biomarkers for PD are lacking, hindering both diagnosis and the assessment of experimental treatments' effectiveness in slowing disease progression. Similarly, there's no dependable method for identifying individuals at high risk of developing PD. Developing such biomarkers would represent a significant breakthrough, greatly impacting diagnosis and treatment strategies in PD research [5].

Isolated rapid eye movement sleep behavior disorder (iRBD) is a type of parasomnia characterized by dream-enactment behaviors, occurring during REM sleep but without the usual muscle atonia [6]. This disorder serves as an early warning sign or prodromal marker for neurodegenerative synucleinopathies, particularly Parkinson's disease (PD) and dementia with Lewy bodies [7]. Individuals with iRBD have an exceptionally high risk (>80%) of developing a neurodegenerative condition [8], [9]. Given that iRBD symptoms precede the onset of Parkinsonism, research into this disorder is crucial for developing therapies that can protect against synucleinopathy, as no other preclinical marker holds predictive significance comparable to iRBD [10].

The emergence of digital health introduces the possibility of remotely and noninvasively identifying and monitoring early indicators of PD using technologies like smartphones [11]–[14]. However, many current tests, such as finger tapping or walking predetermined distances, require active, instructed participation [15]. An optimal digital biomarker should be measured passively, without any additional effort from the subject or investigator. In this context, speech analysis presents intriguing potential advantages, as a significant portion of the population communicates via smartphones daily. Therefore, analyzing speech patterns from smartphone calls in real world scenarios offers a unique opportunity to establish a passive biomarker. This approach enables continuous assessment of experimental treatment effectiveness in natural settings and opens the door to large-scale screening possibilities.

The complex coordination of speech involves a elaborate interplay of over 100 muscles, making it particularly susceptible to alterations in neural structures governing motor functions [16]. *Hypokinetic dysarthria*, a collection of speech and voice disorders, affects

up to 90% of individuals diagnosed with PD, resulting in a diminished voice quality, hypokinetic articulation, hypophonia, monopitch, monoloudness and deficits in timing and phrasing [17]. Based on the findings of a recent multilanguage, multicentric study using an objective acoustical analysis of 150 patients with iRBD, it is evident that speech disorders are one of the earliest motor signs of PD [18]. Specifically, dysprosody and imprecise vowel articulation have been detected in iRBD subjects with impaired olfactory function but still largely functional nigrostriatal dopaminergic transmission [19], [20]. Studies utilizing a murine model of PD have identified deficits in ultrasonic vocalizations as among the initial signs of motor dysfunction [21]. In humans, longitudinal observations indicate that alterations in voice characteristics may manifest as the earliest motor symptoms, emerging up to a decade before formal diagnosis and preceding typical PD symptoms like rigidity and gait abnormalities [22].

Unfortunately, these findings rely on speech recordings actively conducted with a professional condenser microphone within controlled laboratory environments, significantly constraining the wider usability of speech evaluation [23]. Furthermore, in current practice, evaluation of PD is often subjective, rater-dependent, based on a lengthy and expensive manual labelling [24]. Given that the amount of data acquired in-the-wild is in principle impossible to evaluate manually, assessing the symptoms via robust, automated methods is recognized as the future direction of the research [25]. However, numerous challenges must be addressed, such as an inferior quality of microphone, ambient noise prevalent in everyday surroundings, and the unstable direction and distance of microphone from the mouth caused by diverse holding positions, in order to facilitate remote monitoring of speech [26], [27]. Moreover, the reliability of smartphones in detecting prodromal PD (that is, iRBD) via smartphone calls in realistic scenarios has not yet been investigated.

1.1 Aims of the study

The development of a fully automated vocal evaluation brings many substantial challenges, such as finding sensitive acoustic vocal biomarkers to neurodegeneration, automatizing their analysis process based on digital signal processing and machine learning techniques, estimating their precision via statistical analysis, and testing their robustness against corruptive noise, recording device, and conditions.

Therefore, the study aims to (i) establish suitable automated digital biomarkers that will accurately monitor selected, physiologically interpretable speech patterns, which can be deployed from any recording device and thus allow both in-clinic and remote assessment. Next (ii), to develop a system which can reliably, remotely, unobtrusively, and in accordance to ethical guidelines measure these features. And lastly (iii), use such a system in a remote, cross sectional study to monitor prodromal parkinsonism through speech in a real word setting. Such a tool has the potential for a broadly application in neuroprotective trials, deep brain stimulation optimalization, neuropsychiatry, speech therapy, population screening, and beyond.

Chapter 2

Methods & Results

2.1 Study design & literature

Hypokinetic dysarthria arises from dysfunction within the basal ganglia motor circuit. The impairment results in difficulties regulating the initiation, amplitude, and velocity of movements [16], [17]. Consequently, specific acoustic features associated with the motor aspects of speech, which have a clear link to Parkinson's disease pathophysiology [15], align with perceptual descriptions of hypokinetic dysarthria as outlined by Darley et al [28]. These features are the prime candidates for automated voice analysis and include:

1. Disruptions in phonation caused by dysfunctions in the vocal folds. The impairment can be captured using acoustic measures such as Cepstral Peak Prominence, which correlates with the auditory perception of decreased voice quality/breathiness [27].

2. Dysprosody is reflected by the reduced amplitude of vocal cord movements, correlating with reduced pitch variability, called *monopitch* [18].

3. Timing deficits, represented by a decreased ability to maintain the speech motor sequence or to alternate quickly between responses. The dysfunction can be reflected by the acoustic measures of Net Speech Rate and Duration of Pause Intervals, which reflect the perceived auditory timing of speech and may describe deficits such as a slow articulation rate and a reduced ability to intermit and initiate speech [29].

4. Articulation deficits are perceived as a decrease in intelligibility. Most often, they are described using metrics related to vowel production triangle, such as the Vowel Space Area [30]. Articulation characteristics are also partially described by complex speech parametrizations primarily used in speech recognition. These might include Mel Frequency Cepstal Coefficients (MFCCs) or deep neural network embeddings. Insight into the detailed network behavior would be supportive and could reveal critical physiological details [31].

Several methods for the automatic analysis of the mentioned key dimensions of speech in patients with PD have already been developed [29]. However, many methods have been tailored for brief, specific tasks with predetermined content, such as sustained phonation, syllable repetition, or reading text. Consequently, these algorithms may struggle when confronted with spontaneous speech, as they weren't originally intended for such a purpose [27]. Moreover, techniques adapted for spontaneous speech often falter in noisy environments and when recording quality is inferior [26]. Hence, both existing and newly

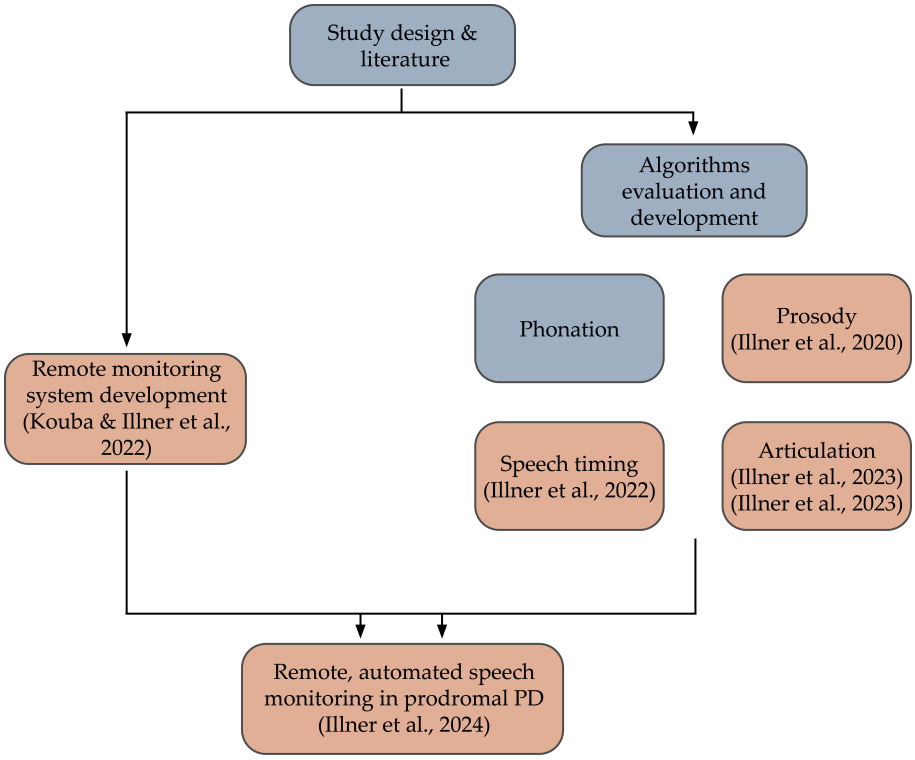


Figure 2.1: Structure of the thesis.

developed methods must undergo experimental and theoretical testing to ascertain their robustness against noise and their overall reliability, thereby confirming their practical utility.

The thesis comprises a collection of published articles aligned with the study's objectives (see Figure 2.1). Each article is presented in the following sections preceded with a brief introduction, contextual background, and its significance. These articles include:

- A study focused on automated assessment of prosody in PD (section 2.3, page 5).
- A study evaluating automated approach of measuring speech rate in PD and prodromal PD (section 2.4.2, page 20).
- A study focused on automated estimation of articulation deficits in a wide range of neurological diseases (section 2.5, page 38).
- A study exploiting the clinical interpretability of changes in Mel Frequency Cepstral Coefficients (section 2.5.1, page 61).
- A protocol of a developed smartphone application for remote monitoring and data acquisition (section 2.6, page 67).
- A cross-sectional study on capturing speech impairment in prodromal PD using a remote, automated approach, currently in a peer review (section 2.7, page 79).

2.2 Phonation dysruptions

Measures like jitter, shimmer, and harmonic-to-noise ratio are frequently employed to screen for voice impairments, serving as established diagnostic biomarkers indicating a decline in vocal fold function [32]. However, these perturbation measures are primarily used during sustained vowel phonation tasks, as they tend to be less accurate and robust in the context of spontaneous speech. In contrast, cepstral patterns emerge as more suitable markers for detecting phonation disruptions across various tasks and recording conditions, given their greater resilience against adverse effects [33].

Cepstral peak prominence (CPP) and its smoothed variant (CPPS) have emerged as key markers in acoustic analysis for evaluating voice quality in relation to dysphonia [34]. Studies have demonstrated a strong correlation between CPP/CPPS values and the severity of dysphonia and breathiness across various languages [35]–[37]. Widely accepted within the speech-language and acoustic communities, CPP was recognized by the American Speech-Language-Hearing Association in 2018 as a general acoustic measure of dysphonia, indicating the overall level of noise in the vocal signal [37]. Previous research has examined the effects of CPP/CPPS measurements both before and after voice treatments, as well as the influence of Parkinson's disease tremor phenotype on the parameter [38].

Previously, evidence of disruptions in CPPs primarily showed in advanced-stage Parkinson's disease (PD) patients undergoing dopaminergic medication [39]. Only one study has reported significant differences in CPPs between early-stage PD patients and controls [40].

Recently, in [27], the sensitivity of the CPPs across a wide range of disease severity, the dependency of the patterns on speech tasks (sustained vowel phonation, reading passages, and monologues), and robustness against additive non-stationary urban noise were investigated. The results showed significant differences in CPPs between controls and early-stage PD for sustained phonation and monologue tasks. Nevertheless, no contrast was demonstrated to capture possible prodromal dysphonia in iRBD and, additionally, the presence of corruptive noise substantially influenced the measures. Hence, the results showed that CPPs patterns might prove vital for early-stage PD assessment but only for a scenario where the recording conditions can be controlled.

2.3 Dysprosody

A comprehensive study from 2020 has thoroughly investigated the state-of-the-art methods for capturing dysprosody in patients with PD, focusing on their in-the-wild conditions reliability and practical utility [26].

Specifically, the focus of the study was on a parameter called *monopitch*, which refers to reduced intonation, indicating lower variability in the fundamental frequency of the voice. Previous research has consistently identified monopitch as a fundamental characteristic of hypokinetic dysarthria, even in the early stages of the disease [41], [42]. Numerous automated speech processing methods have been developed over time to track fundamental frequency (pitch detection algorithms, PDAs), each with its own design. Some studies have concentrated on monopitch in individuals with PD, using automated methods to analyze short prepared utterances [42]–[44]. While it's been confirmed that monopitch measure is robust across different recording devices, it's unclear if there's a universal PDA for measuring it that isn't significantly influenced by thr nature of spontaneous speech, background noise, or severe dysarthria. Therefore, this study aimed to assess

and compare the effectiveness of various PDAs when applied to connected natural speech from PD patients recorded using a smartphone. The study also evaluated the robustness of these trackers against different levels of non-stationary background noise commonly found in urban and household environments by adding varying signal-to-noise levels to the original recordings.

In total, 20 PDAs were identified and studied on a cohort of 60 PD patiens matched with healthy controls. Ten PDAs were subjected to further analysis. From the methods, one outperformed the others in terms of precise accuracy in measuring fundamental frequency, even in detrimental conditions with a low signal-to-noise levels. The method [45] estimates the fundamental frequency contour which maximize a normalized inner product of a warped spectrum of the input signal and a created kernel with given spectral characteristics. A decaying weigh factor is applied to the kernels. The results showed that fundamental frequency estimation from connected speech can be accurate and reliable even when a smartphone is used for recording in an urban environment with the presence of noise.

The findings presented several novel opportunities. Previous research was mostly focused on highly functional vocal paradigms such as sustained phonation. However, tracking pitch changes from connected speech may provide a very natural digital biomarker of disease progression as connected speech reflects the complexity of speech production including a combination of speech motor execution and cognitive-linguistic processing, and has been shown to be superior in capturing subtle PD-related speech changes compared to functional vocal tasks [42]. Monitoring disease progression over time using monopitch as a specific biomarker can thus be done remotely, in the subjects natural environments suing their smartphones. The preprint of the article is attached below.

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease

Vojtech Illner, Pavel Sovka, Jan Rusz*

Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 160 00, Prague 6, Czech Republic

ARTICLE INFO

Article history:
Received 26 June 2019
Received in revised form
29 November 2019
Accepted 17 December 2019
Available online 8 January 2020

Keywords:
Pitch
Fundamental frequency
Speech
Voice
Dysarthria
Smartphones
Parkinson's disease

ABSTRACT

Measuring the fundamental frequency of the vocal folds F_0 is recognized as an important parameter in the assessment of speech impairments in Parkinson's disease (PD). Although a number of F_0 trackers currently exist, their performance in smartphone-based evaluation and robustness against background noise have never been tested. Monologues from 30 newly-diagnosed, untreated PD patients and 30 matched healthy control participants were collected. Additive non-stationary urban and household noise at different SNR levels was added to the recordings, which were subsequently assessed by 10 freely-available and widely-used pitch-tracking algorithms. According to the comparison of all investigated pitch detectors, sawtooth inspired pitch estimator (SWIPE) was the most robust and accurate method in estimating mean F_0 and its standard deviation. However, at a low 6 dB SNR level, a combination of more algorithms may be needed to achieve the desired precision. Monopitch, calculated as F_0 standard deviation and estimated by SWIPE, proved to be robust in distinguishing between the PD and healthy control groups ($p < 0.001$). We anticipate that monopitch may serve as a quick and inexpensive biomarker of disease progression based on longitudinal data collected via smartphone, without any logistical or time constraints for patients and physicians.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease [1]. It is estimated that the incidence of PD is roughly 1.8 % in persons 65 years of age and older [2], and the incidence is expected to grow due to prolonged life expectancy [3]. The disease is characterized by the loss of vulnerable neuronal populations in the brain including *dopaminergic neurons* in the substantia nigra. As a consequence, bradykinesia and other motor disorders such as rigidity, resting tremor and postural instability occur in PD [1]. Although neuroprotective therapies are under development, there is currently no treatment which can fully stop or slow disease progression. Currently available pharmacotherapy methods offer alleviation of PD motor manifestations but do not treat the actual disease process. The cardinal motor manifestations leading to the establishment of diagnosis arise relatively late in the course of neurodegeneration, i.e., at the time when up to 50 % of the neurons in the substantia nigra have already been irretrievably damaged and up to 80 % of striatal dopamine has been depleted [4,5]. The main reason why neuroprotective therapy cannot be developed may be that PD progresses over many years before the appearance of evident motor manifestations and diagnosis, and it is simply too late for intervention. Thus, the recognition of PD *early*, within its prodromal stages, is crucial for the further development of neuroprotective therapy [6,7]. However, no biomarker is available that allows the measurement of experimental treatment efficacy with regard to slowing disease progression.

Speech, the most complex motor skill involving over 100 muscles, is highly sensitive to detrimental changes in neural structures controlling motor abilities [8]. Up to 90 % of people diagnosed with PD develop speech and voice disorders which are collectively called *hypokinetic dysarthria* [9]. A decrease of voice quality, imprecise articulation, monopitch, monoloudness or deficits in timing and phrasing are typical [9]. Therefore, speech disorders may represent one of the earliest motor signs of PD. In the murine model of PD, ultrasonic vocalization deficits are among the first prodromal markers of motor dysfunction [10]. In humans, longitudinal voice changes in subjects at high risk for developing PD were estimated to be the first motor signs, developing up to 10 years before diagnosis,

* Corresponding author at: Department of Circuit Theory, Czech Technical University in Prague, Technická 2, 160 00, Praha 6, Czech Republic.
E-mail address: rusz.mz@gmail.com (J. Rusz).

https://doi.org/10.1016/j.bspc.2019.101831
1746-8094/

well before the appearance of rigidity, gait abnormalities and limb bradykinesia [11]. In addition, pilot cross-sectional studies comparing vocal performance in subjects at high risk of developing PD and healthy controls using quantitative and objective acoustic voice analysis confirmed that speech impairment represents a sensitive prodromal marker of neurodegeneration [12,13]. Thus, the vocal assessment provides intriguing advances as it is inexpensive, non-invasive, and simple to administer. As the recording and processing of human speech is an area that has been extensively investigated, speech changes represent an excellent candidate as a preclinical diagnostic and progressive biomarker of PD.

Voice recordings in PD-related research typically take place in a quiet room with a guiding clinician and are obtained with a professional condenser microphone. This limits a broader and more quantitative vocal assessment in PD. Conversely, speech assessment using a recording obtained by smartphone offers a potentially simple-to-administer and inexpensive solution, scalable to the entire population. Moreover, speech assessment can be performed anywhere, including the patient's home, without the need to visit a clinic. The recordings can then be sent via the wireless network and processed on a remote server, or even directly processed by the smartphone, with fully automated acoustic analysis [13,14]. In both cases, recordings can be processed without signal frequency content degradation. Given that smartphone technology is rapidly advancing, collecting data through mobile devices continues to be a growing focus not only for speech biomarkers in PD [15–17].

Measuring connected speech via smartphone would represent a significant breakthrough as it does not require any additional effort by the investigated subjects. However, several important issues must be resolved before pursuing a further detailed investigation in smartphone-evaluated speech biomarkers in PD. The application brings challenges as the quality of the smartphone microphone is much lower compared to a professional condenser microphone and *differs from device to device*. Moreover, the presence of background noise, such as traffic, voices of other speakers or surrounding sounds, limits the use of PD-related speech assessment in common environments. Therefore, it is crucial to determine the appropriate features that are robust in this setting and determine their accuracy in terms of detecting PD.

One feature of interest is *monopitch* (or reduced intonation), which is represented by lower variability of the fundamental frequency (F_0). In 1969, Darley et al. [18] were the first to identify monopitch as the most prevalent aspect of hypokinetic dysarthria in PD. Subsequently, a number of studies confirmed the findings of Darley et al. [18] and reported monopitch as a core feature of hypokinetic dysarthria, present from the early stages of the disease [19]. Recently, from various investigated features of hypokinetic dysarthria, Rusz et al. [20] showed that monopitch is the most sensitive feature to the presence of speech disorder even in patients at high risk for developing PD. In the same study [20], monopitch assessed through connected speech was also the most resistant feature with respect to low microphone quality. Interestingly, similar findings were reported by Uloza et al. [21] in healthy subjects, but only on short segments of sustained vowel sounds. Maryn et al. [22] also showed that F_0 is robust with respect to both the recording system and environmental noise, but again only for sustained vowel segments. Admittedly, features based on F_0 estimation are frequently used as patterns of expert systems designed for both early PD detection or PD severity evaluation via acoustic voice analysis (for example see [23–27]).

Currently, a number of methods for F_0 estimation, termed "Pitch Determination Algorithms" (PDAs), have already been developed. However, their reliability in the detection of monopitch in PD across various noise conditions has not yet been tested. Only one previous study by Tsanas et al. [28] compared 10 widely-used PDAs and their adaptive fusion via very short sustained vowel segments recorded by a professional microphone. Also, robustness against noise was tested for just one signal-to-noise ratio (SNR) with the inclusion of Gaussian noise, which may be inadequate in representing real environments [28].

Therefore, the present study aimed to examine and compare the performances of several PDAs tested on connected natural speech obtained with a smartphone device. Robustness against additive non-stationary urban and house background noise was evaluated by adding various SNR levels to the original signals. For comparison, we selected only freely-available PDAs, allowing easy and rapid transcription of our findings to clinical practice and voice research. Our results provide information about the robustness of PDAs in detecting the mean and standard deviation (SD) of the F_0 contour, with future potential to provide progressive smartphone-based speech biomarkers in PD to assess experimental treatment efficacy.

2. Methods

2.1. Participants

During 2015–2017, a total of 60 Czech native speakers was recruited. The study was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic and a written, informed acquiescence was provided by all the participants. Thirty attendees (26 men and 4 women) with a mean age of 62 (SD 11, range 41–79) years were diagnosed with de-novo PD based on the Movement Disorders Society clinical diagnostic criteria [29]. As a healthy control (HC) group, 30 participants (26 men and 4 women), with a comparable mean age of 65 (SD 10, range 41–79, t -test: $p = 0.42$) years, were without a record of any neurological of communication disorder. The investigated subjects also participated in a former study focused on a detailed assessment of speech disorder through a smartphone [20]. However, the comparison of the performance of different PDAs and their reliability against noise was not previously investigated.

Disease duration was estimated based on the self-reported occurrence of first motor symptoms. All patients were rated according to the Hoehn and Yahr disability scale (comprised of stages 1 through 5, where 5 is most severe) as well as motor score of the Movement Disorder Society – Unified Parkinson's disease rating scale (MDS–UPDRS III) which ranges from 0 (no motor disorder) to 132 (severe motor disturbance) [30]. The MDS–UPRDS III speech item was used for the clinical description of speech severity, ranging from 0 to 4 where 0 marks normal and 4 unintelligible speech. The evaluation and diagnosis of these scales were performed by a neurologist with experience in movement disorders. All the participants were examined before the start of any symptomatic treatment and none of them had a record of a therapy with antiparkinsonian medication involved. As a result, PD patients manifested disease duration of 1.9 (SD 1.3, range 0.5–6) years, Hoehn and Yahr score of 2.1 (SD 0.4, range 1–3), MDS-UPDRS III of 29.4 (SD 12.8, range 8–63), and MDS-UPDRS III speech item of 0.60 (SD 0.56, range 0–2). See Table 1 for PD subject details.

2.2. Speech assessment

The recordings took place in a closed room with a low ambient noise level (< 50 dB). The data were gathered using *a smartphone* Sony Xperia Z1 Compact (amplitude and the polar response was reported previously [20]). The participants were instructed to hold the device close to their ear as it would be the case during a regular call. The sampling frequency of 48 kHz and the resolution 16-bit were used as these are the highest setting available by the smartphone system. For the recording, a basic application was developed

Table 1
List of PD subjects with clinical characteristics.

PD patient ID	Gender	Age	Disease duration	MDS-UPDRS III	MDS-UPDRS III
		(years)	(years)		speech item
1	M	75	2.0	29	0
2	M	61	1.0	63	1
3	M	60	2.0	32	1
4	M	73	1.0	38	1
5	M	58	0.5	35	1
6	M	60	0.5	49	1
7	M	43	6.0	14	0
8	M	70	1.0	20	0
9	F	75	2.5	28	0
10	F	73	2.0	32	1
11	F	66	1.5	8	0
12	M	52	2.0	24	1
13	M	51	1.5	17	0
14	M	73	2.0	40	1
15	M	61	1.0	25	0
16	M	41	3.0	56	1
17	F	71	5.0	36	1
18	M	76	2.0	42	0
19	M	58	1	26	0
20	M	63	0.5	32	0
21	M	41	1.5	14	0
22	M	61	1	29	1
23	M	63	1	26	1
24	M	43	2.5	16	1
25	M	52	1	42	2
26	M	73	4.0	14	1
27	M	79	1.0	29	0
28	M	66	4.0	22	1
29	M	71	2.0	28	0
30	M	60	1.0	16	1

PD = Parkinson's disease, MDS-UPDRS = Movement Disorders Society - Unified Parkinson's Disease Rating Scale.

running on an Android 5.1 system. No other settings were adjusted during the recording.

Each participant was recorded in a single session accompanied by a speech specialist who guided through the standardized protocol. The participants were instructed to perform a monologue lasting approximately 90 (mean 92.6, SD 12.1) seconds where they narrated a short fictional story.

2.3. Signal-to-noise ratio

To evaluate the robustness of PDAs against the environmental additive noise we added five types of non-stationary noises to each recording on SNR levels 20, 10 and 6 dB, respectively. The boundary of 6 dB was chosen as it represents the worst scenario that is likely possible to occur in a common environment. Lower levels are more unlikely due to processing in the actual device as a quick investigation showed. The first noise was recorded close to a street with heavy traffic - including cars, trams and motorcycles (hereafter, Noise condition 1). The second one is from a busy shopping mall with amplified music, passing-by people talking and kids shouting (hereafter, Noise condition 2). These two noises were chosen as they contain the most typical sources of disruptive non-white noise in an urban environment differing in the rate of frequency changes and the content of reverberations. The third noise consisted of another speaker present in the room, talking about a different topic (hereafter, Noise condition 3). Fourth noise was a recording of a vacuum cleaner (Noise condition 4) and fifth a sound of an ambulance siren (Noise condition 5). These three additional noises were chosen to represent a possible home or hospital environment, where an older adult with PD might typically spend considerable time. This approach resulted in 16 types of signals with different level and type of noise per one speech record including the original recording (clear signal without noise) and 3 noisy signals on 20, 10, and 6 dB SNRs per each of the 5 noise types.

2.4. PDAs search and selection strategy

A systematic literature search of articles written in English before March 2019 was conducted in the Web of Science. In addition, we explored Google Scholar and IEEE Xplore as these typically indexes more studies focused on PDAs. A wide range of keywords was used: pitch, tracking, PDA, fundamental frequency , F_0 estimation, glottal closure instants... By combining these keywords, we identified and studied 20 freely available PDAs [31–47]. We excluded PDA from analysis if it: (i) used very similar computational principle as another available algorithm, (ii) the authors recommended its use for short voiced segments only, (iii) the authors directly stated that the algorithm is not robust against noise, (iv) the method performed *comparably* worse then the others in first conducted trial testing. As a result, we identified 10 widely used and freely available PDAs that were subsequently included in testing. Table 2 shows all the PDAs we studied altogether with links to the software and reasons why the particular algorithm was excluded.

2.5. Tested PDAs

Here we did a review of 10 selected PDAs that were the subjects of our testing. Several algorithms represent longstanding standards used in speech processing, whereas others represent recently introduced methods based on new approaches. The computations mostly took place in MATLAB (MathWorks, Natick, MA) environment although in some cases an interface to other programs, such as PRAAT [33], was utilized.

There have been many attempts of categorizing PDAs mainly for methodological reasons [32]. Popular way is to group them accord-

ing to the domain they are working in including time-domain approaches (for example PRAAT or RAPT) and frequency-domain approaches using spectral or cepstral characteristics (for example BaNa or SWIPE). However, such categorization is limited with some methods falling to either category.

Most of the algorithms have similar general scheme following the stages including (1) – pre-processing, (2) – the actual calculation of F_0 estimates, and (3) – post-processing. The pre-processing usually shapes the signal to match the inner mechanism of the estimator and reduces error, for example by low-pass filtering for decreasing the effects of formants. Post-filtering typically smooths the estimated F_0 contour such as removal of sudden jumps in consecutive estimates which is not physiologically possible, for example by using Viterbi algorithm (PRAAT or BaNa).

Default parameters were used in all cases. The timestep duration was set to 10 ms, where the adjustment was possible, for detailed F_0 contour. The estimates were allowed to attain values between 60–400 Hz, covering the majority of the population. Most of the algorithms were equipped with voiced/unvoiced speech detectors. These detect whether the current timeframe contains voiced or unvoiced segment and therefore decide if F_0 is to be estimated or not. For those algorithms without voiced/unvoiced decision, we used the detector which is a part of the WORLD speech system [48] and was found previously to provide robust results [49].

2.5.1. Harvest

A high-performance fundamental frequency estimator from speech signals (Harvest) Proposed by Morise [31] Harvest is a frequency-domain PDA which obtains the candidates using a band-pass filter bank with different center frequencies. The estimates are then scored and refined using instantaneous frequency. In the second step, a connection algorithm using neighboring F_0 candidates is deployed to smooth the contour and eliminate errors. Harvest is part of the WORLD speech system.

2.5.2. RAPT

A Robust Algorithm for Pitch Tracking (RAPT) is a time-domain PDA developed by Talkin [32]. It utilizes a normalized cross-correlation function of frames of the original signal and its *sub-sampled* version, given by Eq. (1):

$$F_{i,k} = \frac{\sum_{j=m}^{m+n-1} s[j]\, s[j+k]}{\sqrt{e_m e_{m+k}}},$$
$$k = 0, \ldots, K-1; m = iz; i = 0, \ldots, M-1, \quad (1)$$

where s is the sampled speech signal, k is the lag index, i is the frame index, z is the frame length, M is the total number of frames, n is the size of cross-correlation window, K is the length of the cross-correlation, e is a sum of squared samples of the signal, and s in the given window. Then maxima of the cross-correlation with mutual delay close to 1 are searched for, first in the case of $F_{i,k}$ being computed from the sub-sampled signal and then for the original data. After the F_0 candidates are computed, a dynamic programming approach is used to determine the most probable estimates.

2.5.3. PRAAT

PRAAT [33] from Dutch [pra:t] (i.e., "talk") is one of the standardized and most widely used PDAs. Originally proposed by Boersma [33] it divides the signal into frames using appropriate window function and using autocorrelation the F_0 estimates are computed. The autocorrelation is normalized by the division of the autocorrelation of the window function. Boersma later indicated that a Gaussian window produces better results than the originally used Hanning one [50]. In this study, we used PRAAT using a Gaussian window. In the end, a Viterbi algorithm is applied to reduce errors in F_0 contour.

2.5.4. SHS

Sub-harmonic summation (SHS) proposed by Hermes [34] SHS estimates F_0 for each frame as a frequency that maximizes the sum of the spectrum harmonics $H(f)$, given as Eq. (2):

$$H(f) = \sum_{n=1}^{N} h_n P(nf), \quad (2)$$

where P is a smoothed, filtered amplitude spectrum, and n is the number of harmonics. Usually, around 5–11 harmonics are used and a decay factor h is applied as this prevents choosing the subharmonics as a F_0 candidate. We used SHS implementation in PRAAT software.

2.5.5. REAPER

Robust Epoch And Pitch EstimatoR (REAPER) developed by Talkin [35], estimates the position of glottal closure instants (GCI) using autocorrelation and then determines F_0 candidates as an inverse of the time between successive GCI cycles. Dynamic programming is then used in post-processing to reduce errors.

2.5.6. YANGSAF

Yet ANother Glottal Source Analysis Framework (YANGSAF) proposed by Kahawara et al. [36], this method time-warps the speech signal to remove F_0 fluctuations and process it by a filter bank that decomposes the signal by harmonics. For each harmonic instantaneous frequency and aperiodicity features are extracted. F_0 contour is estimated and then smoothed according to candidate variances computed from the aperiodicity features.

2.5.7. SHRP

Subharmonic-to-Harmonic Ratio Pitch algorithm (SHRP), developed by Sun [37], works in the frequency domain where it uses sub-harmonics to harmonics ratio to determine F_0 estimates, given by Eqs. (3–4):

$$SHR(f) = \frac{SH(f)}{SS(f)}, \quad (3)$$

where SH is similar to (2) and SS is defined as

$$SS(f) = \sum_{n=1}^{N} P\left(\left(n - 1/2\right) f\right). \quad (4)$$

To increase the algorithm effectivity, the frequency is logarithmically scaled, and the spectrum-shifting technique is applied to match the human perception of pitch.

2.5.8. SWIPE

Sawtooth inspired pitch estimator (SWIPE) proposed by Camacho and Harris [38] works as well in the frequency domain. It estimates the F_0 candidates as those which maximize a normalized inner product of a warped spectrum of the input signal and a created *kernel* with given spectral characteristics. A decaying weigh factor is applied to the kernels. In this study, we used the algorithm SWIPE', an extension of the original method where only first and prime harmonics are used to estimate the pitch, but refer to it as SWIPE for simplicity.

2.5.9. BaNa

A noise resilient F_0 detection algorithm (BaNa) is a recently proposed method by Yang et al. [39] which especially aims to achieve robust results even in environments with a high noise level. It uses harmonic ratios and Ceptral analysis to gather F_0 candidates and their score. A Viterbi algorithm is then used to choose the most

Table 2
List of the freely-available PDAs.

Abbreviation	Authors [reference]	Link to software	Reason for exclusion
Used for the analyses			
HARVEST	M. Morise [31]	http://www.kisc.meiji.ac.jp/~mmorise/world/english/	
RAPT	D. Talkin [32]	https://www.phon.ucl.ac.uk/resource/sfs/	
PRAAT AC	P. Boersma [33]	http://www.fon.hum.uva.nl/praat/	
PRAAT SHS	D. J. Hermes [34]	http://www.fon.hum.uva.nl/praat/	
REAPER	D. Talkin [35]	https://github.com/google/REAPER	
YANG	H. Kawahara et al. [36]	https://github.com/google/yang_vocoder	
SHRP	X. Sun [37]	https://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm	
SWIPE	A. Camacho, J. G. Harris [38]	https://github.com/kylebgorman/swipe	
BANA	N. Yang et al. [39]	http://www2.ece.rochester.edu/projects/wcng/code.html	
YAAPT	K. Kasi, S. A. Zahorian [40]	http://www.ws.binghamton.edu/zahorian/yaapt.htm	
Excluded			
BPT	L. Shi et al. [41]	https://github.com/LimingShi/Bayesian-Pitch-Tracking-Using-Harmonic-model	Poor results in trial testing.
DIO	M. Morise et al. [42]	http://www.kisc.meiji.ac.jp/~mmorise/world/english/	Not suitable for noisy signals, a similar approach as REAPER, poor results in trial testing.
FXAC	Mark Huckvale	https://www.phon.ucl.ac.uk/resource/sfs/	A similar approach as PRAAT AC, poor results in trial testing.
FXANAL	B. Secrest, G. Doddington [43]	https://www.phon.ucl.ac.uk/resource/sfs/	A very similar approach as PRAAT AC, poor results in trial testing.
FXCEP	L.C.Whitaker et al.	https://www.phon.ucl.ac.uk/resource/sfs/	Similar to a part of BaNa algorithm, poor results in trial testing.
DYPSA	P. A. Naylor et al. [44]	http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html	Similar to REAPER, only for short voiced segments, not robust against noise.
PEFAC	S. Gonzalez, M. Brookes [45]	http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html	Poor results in trial testing.
PRAAT CC	P. Boersma	http://www.fon.hum.uva.nl/praat/	A very similar approach to PRAAT AC, using cross-correlation instead, similar results.
YIN	A. de Cheveigné, H. Kawahara [46]	http://audition-backend.ens.fr/adc/	Not suitable for noisy signals and connected speech, poor results in trial testing.
MBSC	L. N. Tan, A. Alwan [47]	http://www.seas.ucla.edu/spapl/shareware.html	Poor results in trial testing.

likely trajectory between the estimates. The authors even created an app for android platform running this algorithm which suggests motivation exactly matching our case.

2.5.10. YAAPT

Yet another algorithm for pitch tracking (YAAPT) developed by Kasi [40] has a similar approach as Talkin with RAPT using normalized cross-correlation function to determine the F_0 estimates. The differences are that here both the original and nonlinearly warped signals are processed to restore weak F_0 components, application of more sophisticated peak picking methods and incorporation of robust pitch contours obtained from smoothed versions of low-frequency parts of spectrograms. Dynamic programming is then used to find the best F_0 trajectory.

2.6. The reference values

The performance of the algorithms was evaluated with respect to the *mean* and *standard deviation* of the output F_0 contour from the current speech.

The reference values for F_0 trajectory (Gold standard) were estimated by manual analysis obtained using PRAAT software with the standard autocorrelation method (algorithm described in 2.5.3). We used the original smartphone recordings for the calculation of F_0 parameters (description in 2.2.). The maximum and minimum value allowed for the pitch to be manually adjusted per each recording to avoid pitch doubling and halving. Subsequently, a visual and listening verification was done for each recording and individual F_0 segments were manually corrected as necessary to obtain reliable F_0 sequence estimates.

Table 3
Performance of F_0 estimation algorithms on original speech without the presence of noise.

Clear speech											
		HARVEST	RAPT	PRAAT AC	PRAAT SHS	REAPER	YANGsaf	SHRP	SWIPE	BANA	YAAPT
Mean	MAE	0.31	0.49	0.56	0.52	0.90	0.12	1.07	0.24	0.48	0.42
	NRMSE	0.03	0.08	0.06	0.07	0.10	0.01	0.14	0.02	0.05	0.08
	Spearman *r*	0.99	0.92	0.96	0.96	0.93	0.99	0.90	0.99	0.97	0.92
SD	MAE	0.96	0.78	0.84	1.13	1.06	0.20	2.63	0.16	0.65	0.32
	NRMSE	0.29	0.22	0.22	0.29	0.31	0.09	0.67	0.06	0.23	0.13
	Spearman *r*	0.89	0.75	0.50	0.82	0.81	0.93	0.55	0.95	0.73	0.85

MAE = mean absolute error, NRMSE = normalized root mean square error, SD = standard deviation. All correlations reached significance $p < 0.001$.

Table 4
Performance of F_0 estimation algorithms with the presence of additive noise at SNR 20 dB.

SNR 20 dB											
		HARVEST	RAPT	PRAAT AC	PRAAT SHS	REAPER	YANGsaf	SHRP	SWIPE	BANA	YAAPT
Noise condition 1											
Mean	MAE	0.33	0.53	0.56	0.85	1.13	0.13	1.08	0.32	0.46	0.41
	NRMSE	0.03	0.08	0.06	0.09	0.12	0.01	0.15	0.03	0.04	0.08
	Spearman *r*	0.99	0.92	0.96	0.96	0.91	0.99	0.91	0.99	0.97	0.92
SD	MAE	1.15	0.71	0.75	1.21	1.23	0.31	2.54	0.22	0.59	0.31
	NRMSE	0.40	0.23	0.22	0.35	0.35	0.16	0.84	0.09	0.22	0.12
	Spearman *r*	0.88	0.76	0.54	0.77	0.75	0.84	0.69	0.96	0.74	0.85
Noise condition 2											
Mean	MAE	0.33	0.51	0.80	0.60	1.07	0.14	1.08	0.29	0.68	0.41
	NRMSE	0.03	0.08	0.11	0.07	0.12	0.01	0.14	0.03	0.06	0.08
	Spearman *r*	0.99	0.92	0.92	0.95	0.92	0.99	0.90	0.99	0.96	0.93
SD	MAE	1.16	0.78	1.02	1.10	1.17	0.30	2.64	0.18	0.94	0.31
	NRMSE	0.36	0.24	0.25	0.31	0.34	0.15	0.91	0.07	0.36	0.13
	Spearman *r*	0.88	0.74	0.45	0.82	0.79	0.85	0.63	0.96	0.59	0.86
Noise condition 3											
Mean	MAE	0.32	0.50	0.57	0.55	0.91	0.13	1.03	0.23	0.46	0.40
	NRMSE	0.03	0.08	0.06	0.07	0.11	0.01	0.13	0.02	0.04	0.08
	Spearman *r*	0.99	0.92	0.96	0.96	0.93	0.99	0.92	0.99	0.97	0.93
SD	MAE	0.99	0.80	0.84	1.08	1.04	0.20	2.53	0.17	0.56	0.30
	NRMSE	0.30	0.22	0.22	0.30	0.31	0.09	0.69	0.06	0.22	0.13
	Spearman *r*	0.88	0.75	0.49	0.82	0.82	0.93	0.56	0.96	0.73	0.85
Noise condition 4											
Mean	MAE	0.28	0.50	0.62	0.48	0.79	0.12	0.98	0.24	0.53	0.41
	NRMSE	0.02	0.08	0.07	0.06	0.09	0.01	0.12	0.02	0.05	0.08
	Spearman *r*	0.99	0.92	0.95	0.97	0.93	0.99	0.93	0.99	0.97	0.93
SD	MAE	0.92	0.72	0.88	0.98	0.94	0.20	2.50	0.17	0.49	0.30
	NRMSE	0.25	0.22	0.23	0.26	0.28	0.09	0.68	0.07	0.21	0.12
	Spearman *r*	0.85	0.75	0.47	0.84	0.84	0.93	0.55	0.95	0.71	0.87
Noise condition 5											
Mean	MAE	0.31	0.49	0.62	0.50	0.90	0.12	1.08	0.23	0.52	0.40
	NRMSE	0.03	0.07	0.07	0.07	0.10	0.01	0.14	0.02	0.05	0.08
	Spearman *r*	0.99	0.93	0.96	0.96	0.92	0.99	0.91	0.99	0.97	0.92
SD	MAE	0.99	1.24	0.89	1.15	1.06	0.21	2.65	0.16	0.70	0.31
	NRMSE	0.30	0.32	0.23	0.32	0.32	0.09	0.67	0.06	0.23	0.13
	Spearman *r*	0.89	0.62	0.47	0.81	0.80	0.93	0.54	0.95	0.74	0.85

MAE = mean absolute error, NRMSE = normalized root mean square error, SD = standard deviation. All correlations reached significance $p < 0.001$.

To minimize the effects of differences in pitch between individual speakers the received F_0 values were converted into the logarithmic tonal scale (*semitones*). This way, for instance, different pitch ranges such as 100–200 Hz and 200–400 Hz will be represented by equal semitone intervals [51].

2.7. The performance validation and statistical analysis

Three performance measures were used for evaluation: a) mean absolute error (MAE), b) normalized root mean square error (NRMSE), and c) Spearman correlation coefficient *r*. As NRMSE is particularly sensitive to the presence of large errors, we expect the difference between MAE and NRMSE to grow larger as the variability of errors increases. The metrics are defined by (5–6) as follows

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{x}_i - x_i\right|, (\text{semitones}) \tag{5}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{x}_i - x_i\right)^2}}{\max\left(\hat{x}\right) - \min\left(\hat{x}\right)}, (-) \tag{6}$$

where N is the number of speech utterances ($N = 60$) and $\hat{x}_i$, x_i represent a computed statistical value (mean or SD) of the estimated F_0 trajectory in semitones of a given speech and its respective reference (Gold standard).

Mann-Whitney *U* test was applied to evaluate between-group differences (PD vs HC) for the F_0 algorithm with the most robust performance in detecting F_0. The use of the non-parametric statistics was preferred as it produces more reliable results for random variables without normal distribution and is less vulnerable to the possible presence of outliers.

We anticipate that a robust performance of PDA is particularly represented by a very strong correlation and reliability obtained

Table 5
Performance of F_0 estimation algorithms with the presence of additive noise at SNR 10 dB.

SNR 10 dB											
		HARVEST	RAPT	PRAAT AC	PRAAT SHS	REAPER	YANGsaf	SHRP	SWIPE	BANA	YAAPT
Noise condition 1											
Mean	MAE	0.55	0.79	0.70	1.44	2.25	0.41	1.81	0.56	0.61	0.78
	NRMSE	0.06	0.10	0.08	0.18	0.23	0.04	0.29	0.05	0.05	0.12
	Spearman *r*	0.99	0.91	0.94	0.94	0.87	0.98	0.90	0.98	0.97	0.86
SD	MAE	1.53	1.14	0.63	1.67	2.06	1.30	2.82	0.31	0.75	0.64
	NRMSE	0.58	0.33	0.21	0.39	0.53	0.76	0.72	0.12	0.26	0.20
	Spearman *r*	0.77	0.64	0.66	0.65	0.50	0.64	0.60	0.93	0.69	0.70
Noise condition 2											
Mean	MAE	0.44	0.53	1.27	0.81	1.92	0.52	1.46	0.55	1.62	0.47
	NRMSE	0.04	0.09	0.14	0.11	0.20	0.04	0.26	0.05	0.13	0.08
	Spearman *r*	0.99	0.92	0.90	0.95	0.85	0.99	0.87	0.99	0.91	0.92
SD	MAE	1.63	1.28	1.38	1.65	1.89	1.18	3.03	0.29	2.11	0.37
	NRMSE	0.56	0.32	0.30	0.50	0.50	0.45	0.98	0.11	0.55	0.14
	Spearman *r*	0.78	0.64	0.46	0.78	0.58	0.54	0.62	0.83	0.25*	0.80
Noise condition 3											
Mean	MAE	0.41	0.53	0.63	0.76	1.04	0.13	0.99	0.24	0.41	0.40
	NRMSE	0.04	0.08	0.08	0.09	0.12	0.01	0.13	0.02	0.08	0.08
	Spearman *r*	0.99	0.92	0.95	0.96	0.92	0.99	0.94	0.99	0.92	0.93
SD	MAE	1.06	0.77	0.88	1.09	1.06	0.22	2.37	0.17	0.30	0.30
	NRMSE	0.33	0.21	0.23	0.32	0.32	0.10	0.74	0.07	0.13	0.13
	Spearman *r*	0.88	0.76	0.49	0.80	0.80	0.91	0.63	0.96	0.84	0.85
Noise condition 4											
Mean	MAE	0.25	0.48	0.63	0.41	0.67	0.12	0.91	0.28	0.66	0.40
	NRMSE	0.02	0.08	0.06	0.05	0.09	0.01	0.11	0.02	0.06	0.08
	Spearman *r*	0.99	0.92	0.96	0.97	0.93	0.99	0.95	0.99	0.96	0.92
SD	MAE	0.87	0.61	0.71	0.88	0.79	0.20	2.35	0.20	0.48	0.29
	NRMSE	0.23	0.20	0.22	0.23	0.25	0.09	0.70	0.07	0.22	0.12
	Spearman *r*	0.90	0.74	0.52	0.82	0.85	0.94	0.57	0.96	0.65	0.89
Noise condition 5											
Mean	MAE	0.35	1.28	1.08	0.50	1.14	0.13	1.08	0.23	0.82	0.41
	NRMSE	0.03	0.12	0.12	0.07	0.13	0.01	0.14	0.02	0.07	0.08
	Spearman *r*	0.99	0.88	0.90	0.96	0.92	0.99	0.91	0.99	0.95	0.92
SD	MAE	1.17	3.50	1.50	1.15	1.32	0.22	2.65	0.16	1.18	0.31
	NRMSE	0.37	0.48	0.33	0.32	0.35	0.10	0.67	0.06	0.33	0.12
	Spearman *r*	0.84	0.29	0.31	0.81	0.72	0.92	0.54	0.95	0.53	0.85

MAE = mean absolute error, NRMSE = normalized root mean square error, SD = standard deviation. All correlations reached significance $p < 0.001$ except for * which refers to $p < 0.01$.

between Gold standard and a given speech signal (Spearman's $r \geq 0.90$ and NRMSE ≤ 0.10).

3. Results

3.1. Performance of PDAs

Table 3 shows the results for original speech without the presence of any additive noise. A considerable difference in estimating the mean and SD of the F_0 contour can be seen. In the case of the mean F_0, all algorithms achieved very good correlation with reference labels ($r > 0.90$, $p < 0.001$), with very low NRMSE ≤ 0.03 in some cases (Harvest, YANGsaf and SWIPE). However, the performance of the algorithms dropped significantly for the estimation of the F_0 SD. Only YANGsaf and SWIPE reached NRMSE < 0.10, with strong correlations between estimated and manual labels ($r = 0.93$, $p < 0.001$ and $r = 0.95$, $p < 0.001$, respectively).

Table 4 shows results with an SNR level of 20 dB. With respect to mean F_0 estimation, the accuracy was nearly the same as for clear speech for all types of noise, with NRMSE mostly below 0.10 (all $r \geq 0.90$, $p < 0.001$). For F_0 SD, accuracy decreased in some cases only slightly, for example in the performance of RAPT where NRMSE dropped to 0.23 ($r = 0.76$, $p < 0.001$), 0.24 ($r = 0.74$, $p < 0.001$), 0.22 ($r = 0.75$, $p < 0.001$), 0.22 ($r = 0.75$, $p < 0.001$), 0.32 ($r = 0.62$, $p < 0.001$) for Noise conditions 1–5, respectively. The decrease was more rapid for instance in the case of SHRP with NRMSE of 0.84 ($r = 0.69$, $p < 0.001$), 0.91 ($r = 0.63$, $p < 0.001$), 0.69 ($r = 0.56$, $p < 0.001$), 0.68 ($r = 0.55$, $p < 0.001$), 0.67 ($r = 0.54$, $p < 0.001$) for Noise conditions 1–5, respectively. Only SWIPE was able to reach NRMSE < 0.10 ($r = 0.95$-0.96, $p < 0.001$) for all noise conditions. YANGsaf reached similar results as YAAPT with an NRMSE of 0.16 ($r = 0.84$, $p < 0.001$), 0.15 ($r = 0.85$, $p < 0.001$), 0.09 ($r = 0.93$, $p < 0.001$), 0.09 ($r = 0.93$, $p < 0.001$), 0.09 ($r = 0.93$, $p < 0.001$) for Noise conditions 1–5, respectively. No other algorithm achieved NRMSE < 0.20. Except in special cases such as BaNa and RAPT, differences in performance for the all noise types were generally low at the 20 dB SNR level.

Table 5 shows results with an SNR level of 10 dB, where deviation from the results in clear speech became more apparent. These differences were not as evident for mean F_0 estimation, where most algorithms still performed well with an NRMSE around 0.10 ($r \geq 0.90$, $p < 0.001$). In F_0 SD, the difference was substantial as most of the algorithms showed serious deficiencies with noise at this power. Only a few algorithms showed NRMSE < 0.30, with only SWIPE and YAAPT reaching NRMSE < 0.20. Differences in performance for the noise types became more striking. YANGsaf was no longer accurate at the 10 dB noise level with an NRMSE of 0.76 ($r = 0.64$, $p < 0.001$) for Noise condition 1 and 0.45 ($r = 0.54$, $p < 0.001$) for Noise condition 2, although maintained satisfactory results for Noise conditions 3–5 with NRMSE 0.10 ($r = 0.91$, $p < 0.001$), 0.09 ($r = 0.94$, $p < 0.001$) and 0.10 ($r = 0.92$, $p < 0.001$), respectively. Similar behavior was manifested by HARVEST, SHS and REAPER. Also, Noise condition 5 led to worse performance, for example RAPT with mean NRMSE 0.27 for Noise condition 1–4 increased to 0.48 for Noise condition 5. Algorithms that reached similar performances for all noises were PRAAT AC, SWIPE and YAAPT.

Table 6 shows results with the lowest SNR level of 6 dB. When estimating mean F_0, most algorithms showed an NRMSE < 0.20 ($r >$

Table 6
Performance of F_0 estimation algorithms with the presence of additive noise at SNR 6 dB.

SNR 6 dB											
		HARVEST	RAPT	PRAAT AC	PRAAT SHS	REAPER	YANGsaf	SHRP	SWIPE	BANA	YAAPT
Noise condition 1											
Mean	MAE	0.75	1.32	0.88	1.87	3.21	0.76	2.38	0.64	0.80	1.82
	NRMSE	0.09	0.17	0.10	0.27	0.37	0.07	0.43	0.05	0.07	0.22
	Spearman *r*	0.98	0.90	0.93	0.93	0.85	0.97	0.90	0.98	0.97	0.82
SD	MAE	1.87	1.72	0.69	1.96	2.47	1.75	3.02	0.29	0.98	1.51
	NRMSE	0.58	0.34	0.21	0.43	0.55	1.08	0.77	0.10	0.31	0.33
	Spearman *r*	0.72	0.52	0.64	0.59	0.42	0.65	0.55	0.92	0.58	0.59
Noise condition 2											
Mean	MAE	0.60	0.72	1.79	0.97	2.98	0.94	1.81	0.75	2.54	0.89
	NRMSE	0.07	0.11	0.21	0.15	0.36	0.07	0.39	0.06	0.20	0.12
	Spearman *r*	0.99	0.92	0.88	0.95	0.82	0.98	0.86	0.98	0.86	0.88
SD	MAE	1.92	1.82	1.70	2.01	2.41	1.50	3.25	0.50	2.98	0.46
	NRMSE	0.58	0.42	0.35	0.57	0.52	0.60	1.04	0.19	0.65	0.12
	Spearman *r*	0.66	0.55	0.33*	0.74	0.43	0.55	0.57	0.68	0.02*	0.83
Noise condition 3											
Mean	MAE	0.48	0.56	0.66	0.91	1.17	0.14	1.01	0.25	0.40	0.42
	NRMSE	0.05	0.08	0.08	0.11	0.14	0.01	0.14	0.02	0.04	0.08
	Spearman *r*	0.99	0.92	0.95	0.95	0.92	0.99	0.95	0.99	0.98	0.92
SD	MAE	1.11	0.76	0.93	1.13	1.10	0.25	2.27	0.19	0.53	0.28
	NRMSE	0.36	0.21	0.24	0.34	0.31	0.12	0.74	0.07	0.20	0.12
	Spearman *r*	0.87	0.76	0.48	0.78	0.77	0.90	0.65	0.95	0.78	0.87
Noise condition 4											
Mean	MAE	0.26	0.49	0.66	0.39	0.65	0.12	0.87	0.32	0.74	0.38
	NRMSE	0.02	0.08	0.06	0.05	0.09	0.01	0.11	0.03	0.06	0.07
	Spearman *r*	0.99	0.92	0.96	0.97	0.93	0.99	0.95	0.99	0.97	0.93
SD	MAE	0.89	0.54	0.56	0.85	0.74	0.21	2.29	0.23	0.55	0.31
	NRMSE	0.24	0.20	0.20	0.23	0.24	0.09	0.68	0.08	0.23	0.11
	Spearman *r*	0.87	0.76	0.60	0.81	0.85	0.93	0.61	0.96	0.55	0.88
Noise condition 5											
Mean	MAE	0.37	2.25	2.01	0.49	1.25	0.13	1.90	0.25	1.09	0.43
	NRMSE	0.04	0.21	0.17	0.07	0.14	0.01	0.24	0.02	0.09	0.07
	Spearman *r*	0.99	0.85	0.85	0.95	0.91	0.99	0.90	0.99	0.93	0.92
SD	MAE	1.26	4.70	2.79	1.48	1.49	0.24	3.39	0.15	1.58	0.29
	NRMSE	0.38	0.56	0.49	0.39	0.39	0.11	0.86	0.06	0.33	0.12
	Spearman *r*	0.82	0.16*	0.10**	0.75	0.69	0.92	0.31*	0.95	0.35	0.87

MAE = mean absolute error, NRMSE = normalized root mean square error, SD = standard deviation. All correlations reached significance $p < 0.001$ except for * and ** which refers to $p < 0.01$, respectively $p < 0.1$.

0.80, $p < 0.001$). Even under these conditions we were able to get very close to the Gold standard with Harvest, YANGsaf and SWIPE. When estimating F_0 SD, the only acceptably accurate results were achieved by SWIPE with an NRMSE of 0.10 ($r = 0.92$, $p < 0.001$) for Noise condition 1. The performance of NRMSE dropped to 0.19 for Noise condition 2 ($r = 0.68$, $p < 0.001$) but remained reliable for Noise conditions 3–5 with NRMSE 0.07 ($r = 0.95$, $p < 0.001$), 0.08 ($r = 0.96$, $p < 0.001$), and 0.06 ($r = 0.95$, $p < 0.001$), respectively. Conversely, YAAPT showed sufficient accuracy for Noise conditions 2–5 with an NRMSE of 0.12 ($r = 0.83$, $p < 0.001$), 0.12 ($r = 0.87$, $p < 0.001$), 0.11 ($r = 0.88$, $p < 0.001$), and 0.12 ($r = 0.87$, $p < 0.001$) but failed for Noise condition 1 with an NRMSE of 0.33 ($r = 0.59$, $p < 0.001$). YANGsaf showed the same accuracy as in the case of 10 dB SNR scenario with satisfactory results for Noise conditions 3–5 with mean NRMSE 0.11 ($r = 0.90$-0.93, $p < 0.001$), but insufficient for Noise conditions 1–2 with mean NRMSE 0.84 ($r = 0.55$-0.65, $p < 0.001$). Other algorithms were beyond these results with NRMSE usually much greater than 0.30 ($r = 0.50$-0.60, $p < 0.05$). Differences in performance for the noise types were substantial and differed between algorithms.

SWIPE was the least affected algorithm by different noise types through all the SNR levels. To make the orientation within a number of comparisons more straightforward, the results listed in the Tables 3–6 were also visualized by Figs. 1 and 2 for mean F_0 and F_0 SD where the individual points represent mean MAE and NRSME across all the noise conditions for different SNR levels.

3.2. SWIPE performance evaluation on the distinction between study groups

According to the results of comparisons (Table 2–6), SWIPE was found to be the most robust algorithm for the detection of F_0 SD in different noise conditions. Fig. 3 demonstrates statistically significant group differences between the PD and HC groups for F_0 SD across all investigated noise conditions ($p < 0.01$). No statistically significant differences were found for the mean F_0 across various noise conditions (Fig. 4).

4. Discussion

Our results show that F_0 estimation from connected speech can be accurate and reliable even when a smartphone is used for recording in an urban environment with the presence of noise. While previous methods were mostly designed for highly functional vocal paradigms such as sustained phonation [52,53], our current findings present several new opportunities. Tracking pitch changes from connected speech may provide a very natural digital biomarker of disease progression based on longitudinal data acquired without any cost or burden to the patient and investigator. Moreover, connected speech reflects the complexity of speech production including a combination of speech motor execution and cognitive-linguistic processing, and therefore has been shown to be superior in capturing subtle PD-related speech changes compared to functional vocal tasks [20]. Observing disease progres-

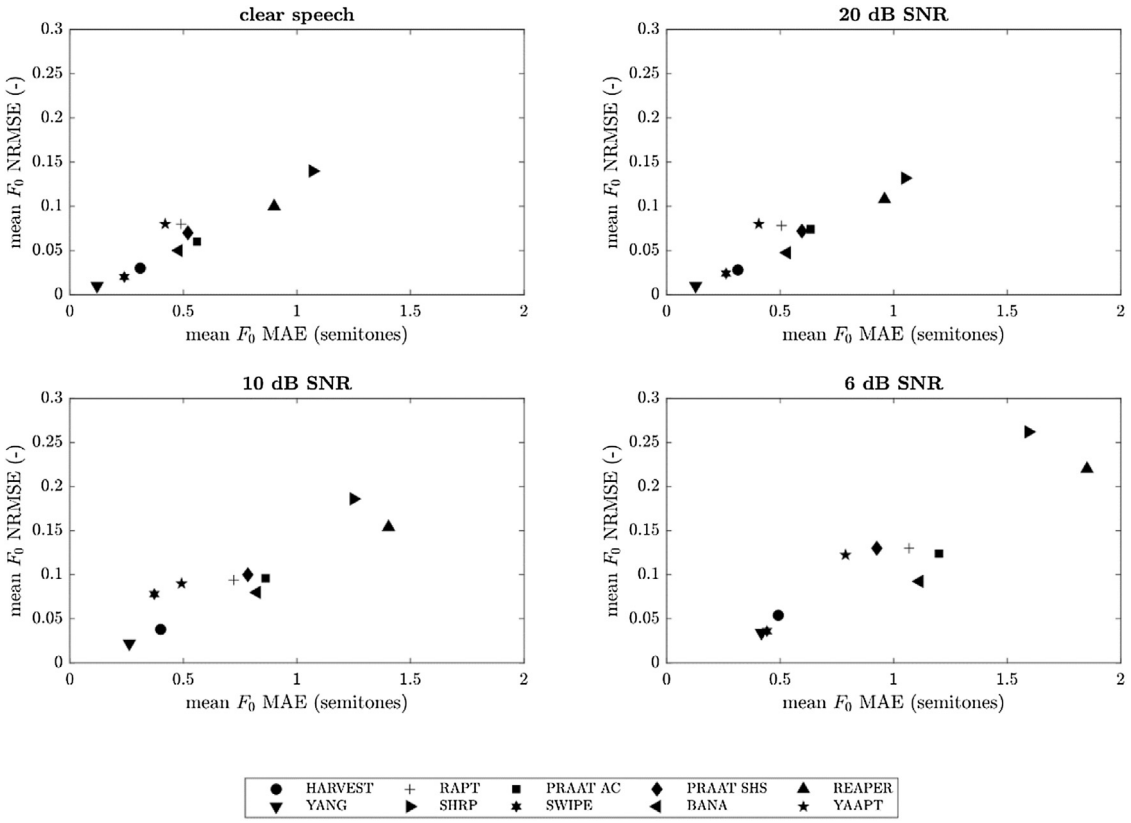


Fig. 1. Mean F_0 MAE and NRMSE results for the algorithms on different SNR levels. For 20, 10, and 6 dB SNR, each point corresponds to mean MAE or NRMSE across all the noise types for a given level.

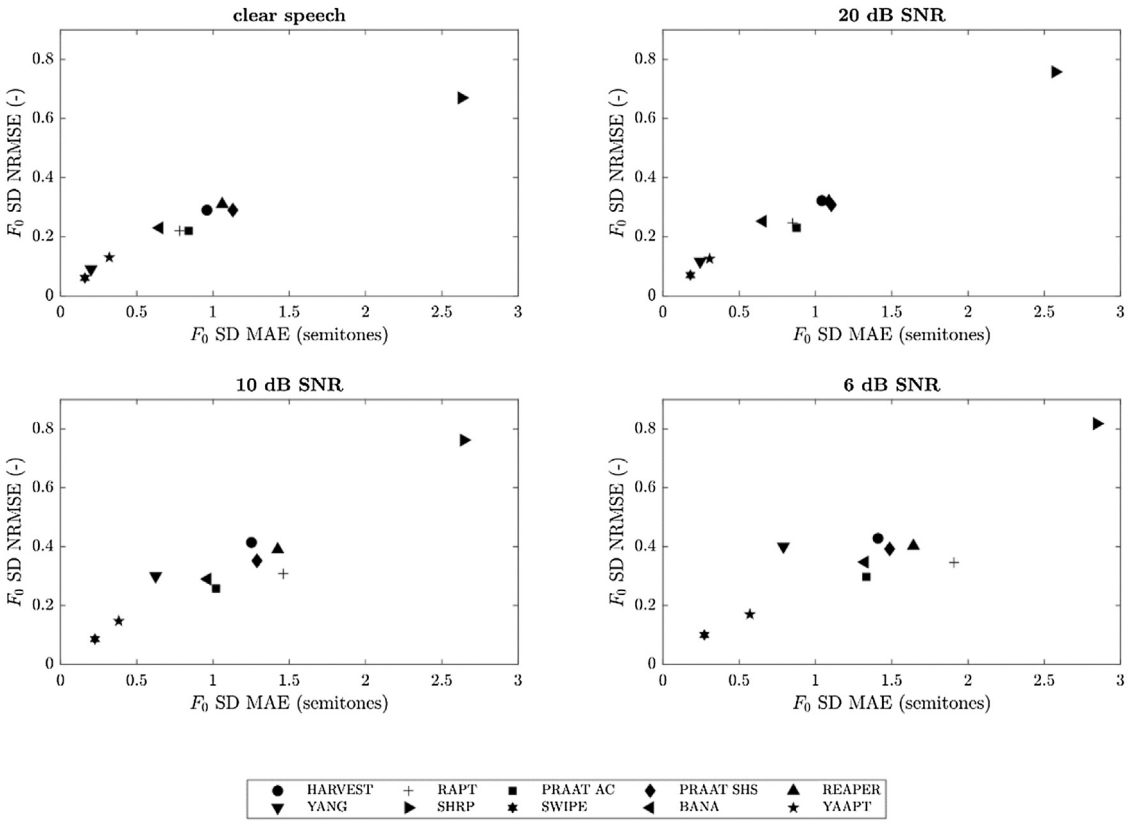


Fig. 2. F_0 SD MAE and NRMSE results for the algorithms on different SNR levels. For 20, 10, and 6 dB SNR, each point corresponds to mean MAE or NRMSE across all the noise types for a given level.

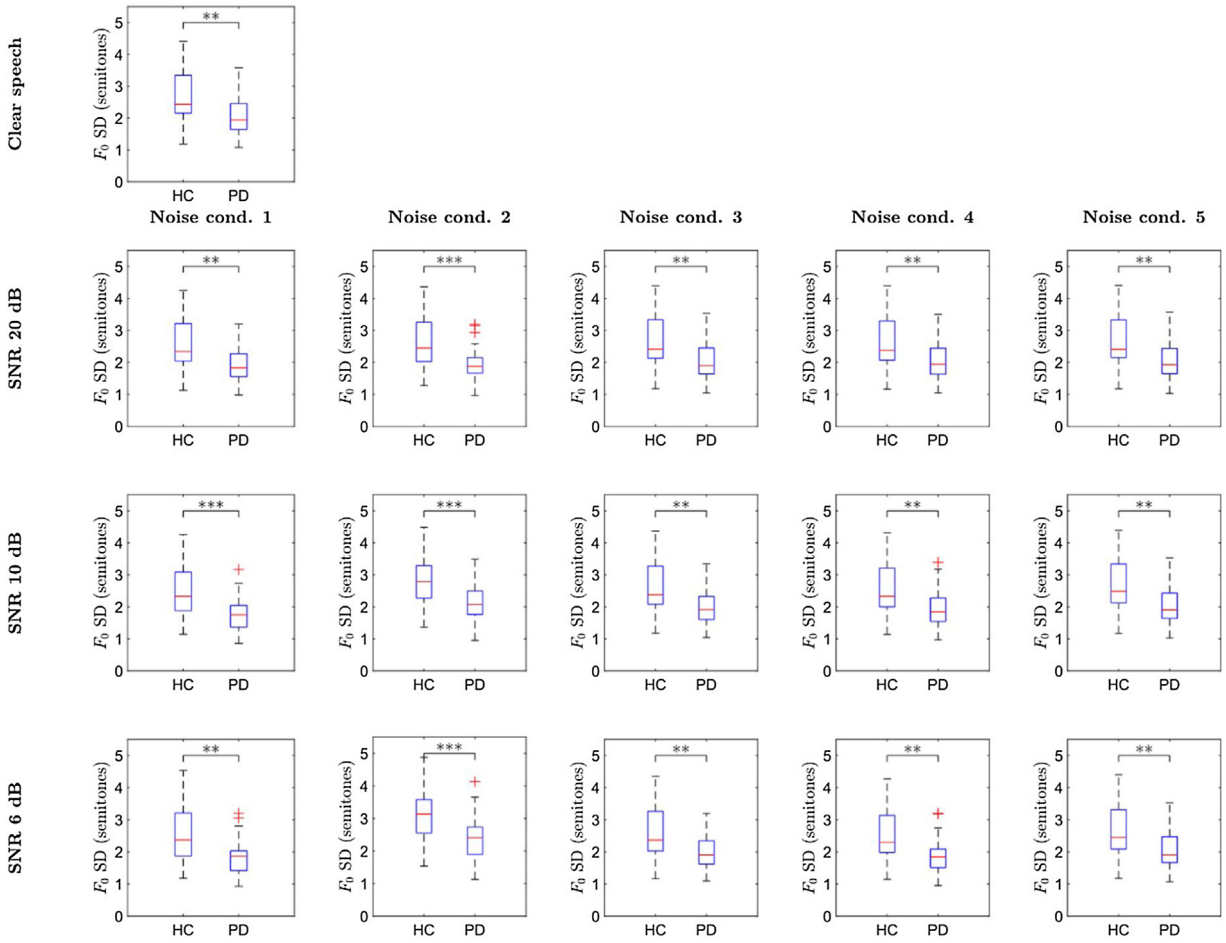


Fig. 3. Group differences in SWIPE F_0 SD estimation between the HC and PD set with ** referring to $p < 0.01$ and *** to $p < 0.001$. Red bars represent the median of F_0 SD, rectangles represent interquartile range. Maximum and minimum values are by error bars. Outliers are marked as red crosses. Captions: PD = Parkinson's disease, HC = healthy controls, SNR = signal-to-noise ratio.

sion over a short period using well-defined and disease-specific biomarkers such as monopitch may thus significantly aid in the recruitment of appropriate cases into large studies of innovative therapies for prodromal PD, and in the future may also bolster early presymptomatic diagnosis and enable rapid access to neuroprotective therapy once available. Quick, inexpensive, and non-invasive vocal assessment by smartphone may also allow for personalized implementation of therapeutic strategies providing (i) rapid feedback after exercise, (ii) monitoring the effects of pharmacological therapies including advanced drug delivery systems and making it possible to modify medication doses according to immediate needs, or (iii) modification of effects and speech-related side-effects of deep brain stimulation by re-programming and optimizing stimulation parameters [54]. In addition, it could be extended to improve diagnosis and monitoring disease progression in other disorders affecting the fundamental frequency such as in Huntington's disease, amyotrophic lateral sclerosis, and even various neuropsychiatric disorders [55–57]. In the future, as computational power has enabled a higher level of automation, PDA can be implemented directly to smartphones and speech data analyzed on the device. Thus, only statistical trends related to pitch performance would be stored and available for interested patients or their neurologists, whereas no personal data need to be shared or transferred.

The estimation of mean F_0 was very accurate across all SNR levels and almost independent of the type of noise for most of the PDAs. This is in agreement with a previous study by Maryn et al. [22], where mean F_0 was accurately estimated across various SNR levels (mean 31.84 dB, SD 10.66 dB), which were however higher compared to those used in the present study. The performances in measuring F_0 SD heavily rely on which PDA is used particularly for lower SNR levels. To recommend the most robust PDA for all situations, the choice would clearly be SWIPE. Harvest, YAAPT and YANG were also sufficiently accurate. These outperformed the other algorithms mainly in terms of high resistance to the effects of noise and reliable SD estimates. In addition, algorithms such as REPAER were sufficiently accurate for clear speech but failed in noisy scenarios.

Considering F_0 SD estimation by SWIPE, the process was highly independent of the presence of non-stationary noise with almost no difficulties even at 10 dB SNR. At the lower 6 dB SNR level (Table 6), the presence of noise substantially decreased the accuracy of SWIPE, especially for Noise condition 2 captured in an acoustically enclosed space. We might assume that such decreased accuracy is caused particularly by two factors including (i) similarity between F_0 contours estimated from monologue and background speech and (ii) reverberations involved in noise due to the acoustically-enclosed space. Improvement of final PDA accuracy might be achieved by a fusion of more algorithms as was done in a previous study by Tsanas et al. [28]. However, the SWIPE-estimated F_0 SD marker was still able to separate between PD and healthy control groups at a high level of statistical significance *for all noise levels*. In agreement with our findings, the previous study by Tsanas et al. [28] recommended SWIPE as an accurate F_0 estimator of pitch contours from sustained vowels in environments without the presence of noise.

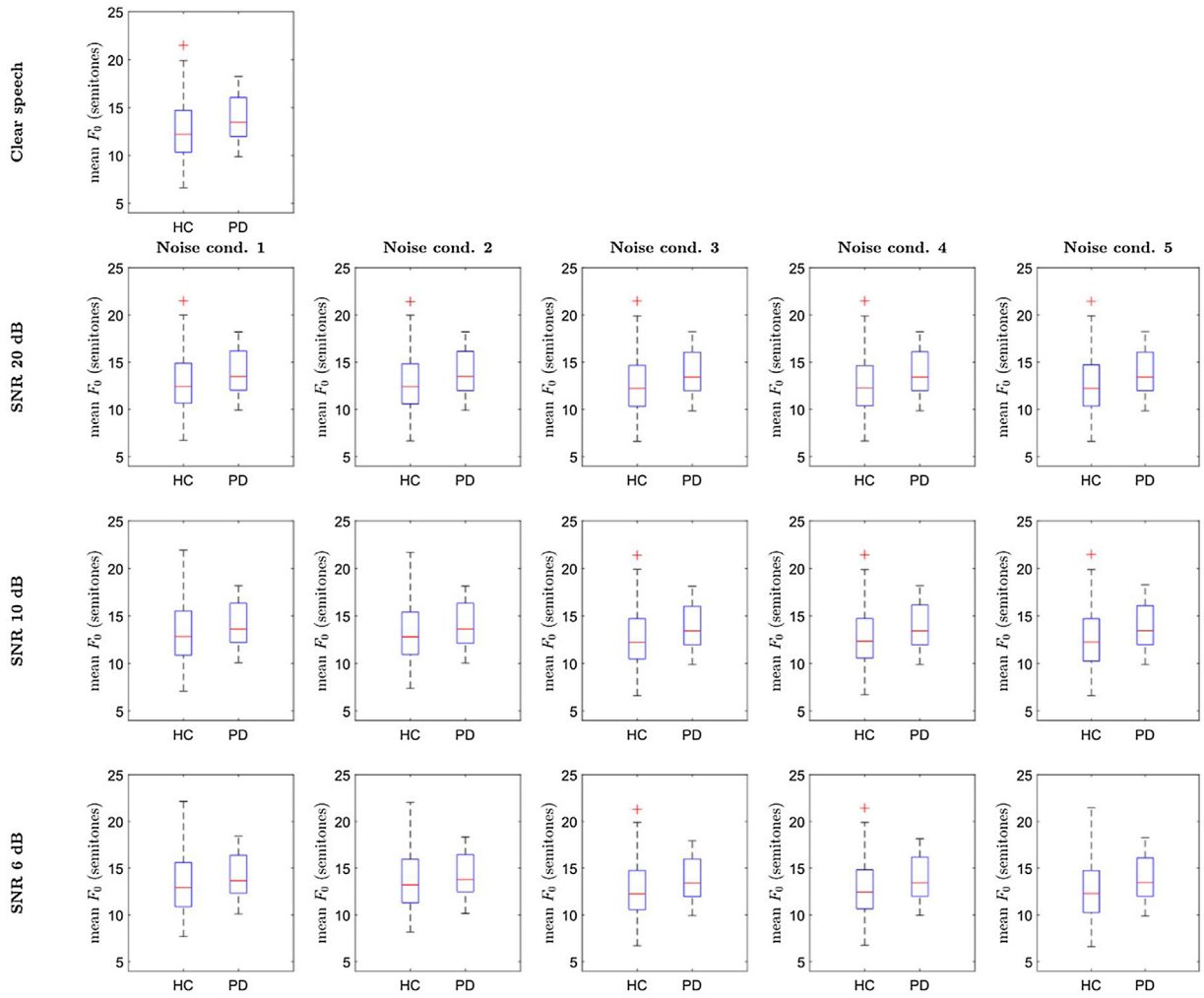


Fig. 4. Group differences in SWIPE mean F_0 estimation between the HC and PD set. No statistically significant differences were found between the groups. Red bars represent the median of F_0 SD, rectangles represent interquartile range. Maximum and minimum values are by error bars. Outliers are marked as red crosses. Captions: PD = Parkinson's disease, HC = healthy controls, SNR = signal-to-noise ratio.

As we found high statistically significant differences between our PD and HC groups for F_0 SD using SWIPE across various noise levels, reduced intonation variability can be considered as a reliable marker of early speech disorder in PD suitable for the smartphone application. Further support to measure intonation variability comes from previous literature, showing the occurrence of monopitch even in patients with rapid eye movement sleep behavior disorder [20], which is considered one of the most important clinical phenotypes for predicting future conversion to PD [11]. In addition, the previous study reported cases with reduced intonation variability detectable several years before the onset of the first PD motor manifestations [19]. From a practical point of view, features such as intonation variability extracted from longer connected speech material better ensure the stability of speech assessment compared to short and functional vocal tasks [58]. Yet, the stability of speech assessment is important to highlight even small speech changes due to neurodegeneration or effects of therapy introduction during longitudinal monitoring. Also, previous research documented that analysis of spontaneous utterances is the best way to assess the impact of PD on speech [59,60]. Considering mean F_0, we did not find any significant differences between PD and HC groups in our cohort. This finding is in agreement with a previous study by Holmes et al. [61], which demonstrated that a higher speaking F_0 was associated with advanced PD only in males. It is noteworthy that our dataset was composed primarily of male participants and thus we cannot exclude the possibility that the observed monopitch and no differences in mean pitch may be influenced by gender-specific aspects.

The present study has certain limitations. Only freely-available F_0 estimation algorithms were tested, allowing easy transcription into clinical practice. Some promising methods previously recommended by Tsanas et al. [28] were not part of this study and their sensitivity to connected speech in a noisy environment should be tested in future studies. Only one type of smartphone was used for the recordings and thus differences between various devices could not be evaluated. Although high reliability between professional condenser and low-quality smartphone microphones for pitch estimation has already been shown [20], future studies are encouraged to confirm our findings and test new solution, preferably using different devices and other languages (for example using freely-available datasets such as Ïtalian Parkinson's Voice and Speech Dataset[62]). In addition, albeit several types of additive noise were used, they do not contain all possible noise sources and noise types in a natural environment. However, based on research conducted using smartphones, Lebacq et al. [63], Jannets et al. [64] and Manfredi et al. [65] showed that F_0 measures are sufficiently robust with respect to the recording device. In addition, Maryn et al. [22] showed that measuring F_0 contour is robust with respect to both the recording system and environmental noise, as well as their combination. It should also be noted that we did not evaluate the influence

of convolutive noise on PDA performance as speech recorded with a smartphone does not need to undergo signal frequency content attenuation via transmission and processing.

5. Conclusion

This work represents a further step in using smartphones to evaluate speech disorders due to the presence of PD. We found that monopitch can be reliably measured by the SWIPE algorithm even when a smartphone device is used for the recording and non-stationary urban noise up to 10 dB SNR is present. At a lower SNR level, a combination of more algorithms may be needed to achieve sufficient robustness. Monopitch estimated by SWIPE was able to significantly distinguish between PD and healthy control groups and may serve as a useful digital biomarker in monitoring the effectiveness of speech therapy or experimental treatments on slowing the progression of PD. Future longitudinal studies should show the sensitivity of tracking pitch changes through smartphones as a potential diagnostic and progressive digital biomarker of PD.

Funding

This study was supported by the Czech Ministry of Health (grant no. NV19-04-00120) and by the OP VVV MEYS project Ŕesearch Center for Informatics(grant no. CZ.02.1.01/0.0/0.0/16_019/0000765).

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Vojtech Illner: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft. **Pavel Sovka:** Conceptualization, Methodology, Writing - review & editing. **Jan Rusz:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing - original draft.

References

[1] W. Poewe, K. Seppi, C.M. Tanner, G.M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, A.E. Lang, Parkinson disease, Nat. Rev. Dis. Primers 23 (3) (2017) 17013.
[2] M.C. de Rijk, L.J. Launer, K. Berger, M.M. Breteler, J.F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martiney-Lage, J. Trenkwalder, A. Hofman, Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts, Neurologic Diseases in the Elderly Research Group, Neurology 54 (11) (2000) S21–S23.
[3] L.J. Findley, The economic impact of Parkinson's disease, Park. Relat. Disord. 13 (2007) S8–S12.
[4] M.C. Rodriguez-Oroz, M. Jahanshahi, P. Krack, I. Litvan, R. Macias, E. Bezard, J.A. Obeso, Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms, Lancet Neurol. 8 (2009) 1128–1139.
[5] H. Bernheimer, W. Birkmayer, O. Hornykiewicz, K. Jellinger, F. Seitelberger, Brain dopamine and the syndromes of Parkinson and Huntington. Clinical, morphological and neurochemical correlations, J. Neurol. Sci. 20 (1973) 415–455.
[6] C.H. Schenck, J.Y. Montplaisir, B. Frauscher, B. Hogl, J.F. Gagnon, R. Postuma, K. Sonka, P. Jennum, M. Partinen, I. Arnulf, V. Cochen de Cock, Y. Dauvilliers, P.H. Luppi, A. Heidbreder, G. Mayer, F. Sixel-Döring, C. Trenkwalder, M. Unger, P. Young, Y.K. Wing, L. Ferini-Strambi, R. Ferri, G. Plazzi, M. Zucconi, Y. Inoue, A. Iranzo, J. Santamaria, C. Bassetti, J.C.M.ö ller, B.F. Boeve, Y.Y. Lai, M. Pavlova, C. Saper, P. Schmidt, J.M. Siegel, C. Singer, E. St Louis, A. Videnovic, W. Oertel, Rapid eye movement sleep behavior disorder: devising controlled active treatment studies for symptomatic and neuroprotective therapy a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group, Sleep Med. 14 (2013) 795–806.
[7] B. Högl, A. Stefani, A. Videnovic, Idiopathic REM sleep behavior disorder and neurodegeneration – an update, Nat. Rev. Neurol. 14 (2018) 40–55.
[8] J.R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis and Management, 3rd ed, Elsevier Health Sciences, 2013.
[9] A.K. Ho, R. Iansek, C. Marigliani, J.L. Bradshaw, S. Gates, Speech impairment in a large sample of patients with Parkinson's disease, Behav. Neurol. 11 (1999) 131–137.
[10] L.M. Grant, F. Richter, J.E. Miller, S.A. White, C.M. Fox, C. Zhu, M.R. Ciucci, Vocalization deficits in mice over-expressing alpha-synuclein, a model of pre-manifest Parkinson's disease, Behav. Neurol. 128 (2) (2014) 110–121.
[11] R.B. Postuma, A.E. Lang, J.F. Gagnon, A. Pelletier, J.Y. Montplaisir, How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder, Brain 135 (6) (2012) 1860–1870.
[12] J. Rusz, J. Hlavnička, T. Tykalová, J. Bušková, O. Ulmanová, E. Růžička, K. Šonka, Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder, Sleep Med. 19 (2016) 141–147.
[13] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, J. Rusz, Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder, Sci. Rep. 7 (1) (2017) 12.
[14] M. Novotný, J. Rusz, R. Čmejla, E. Růžička, Automatic evaluation of articulatory disorders in Parkinsons disease, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2014) 1366–1378.
[15] S. Arora, et al., Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study, Parkinsonism Relat. Disord. 21 (6) (2015) 650–653.
[16] A. Zhan, S. Mohan, C. Tarolli, R.B. Schneider, J.L. Adams, S. Sharma, M.J. Elson, K.L. Spear, A.M. Glidden, M.A. Little, A. Terzis, E.R. Dorsey, S. Saria, Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score, JAMA Neurol. 75 (7) (2018) 876–880.
[17] F. Lipsmeier, K.I. Taylor, T. Kilchenmann, D. Wolf, A. Scotland, J. Schjodt-Eriksen, W.Y. Cheng, I. Fernandez-Garcia, J. Siebourg-Polster, L. Jin, J. Soto, et al., Evaluation of smartphone-based testing to generate exploratory measures in a Phase 1 Parkinson's disease clinical trial, Mov. Disord. 33 (8) (2018) 1287–1297.
[18] F.L. Darley, A.E. Aronson, J.R. Brown, Differential diagnostic patterns of dysarthria, J. Speech Hear. Res. 12 (1969) 246–269.
[19] B.T. Harel, M.S. Cannizzarom, H. Cohen, N. Reilly, P.J. Snyder, Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment, J. Neurolinguistics 17 (2004) 439–453.
[20] J. Rusz, J. Hlavnička, T. Tykalová, M. Novotný, P. Dušek, K. Šonka, E. Růžička, Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (8) (2018) 1495–1507.
[21] V. Uloza, E. Padervinskis, A. Vegiene, R. Pribuisiene, V. Saferis, E. Vaiciukynas, A. Gelzinis, A. Verikas, Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening, Arch. Otorhinolaryngol. 272 (11) (2015) 3391–3399.
[22] Y. Maryn, F. Ysenbaert, A. Zarowski, R. Vanspauwen, Mobile communication devices, ambient noise, and acoustic voice measures, J. Voice 31 (2) (2016) 248.e11–248.e23.
[23] W. Caesarendra, F.T. Putri, M. Ariyanto, J.D. Setiawan, Pattern recognition methods for multi stage classification of Parkinson's disease utilizing voice features, in: 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Busan, Korea: IEEE, 2015, pp. 802–807.
[24] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Detecting Parkinson's disease from sustained phonation and speech signals, PLoS One 12 (2017), e0185613.
[25] B.E. Sakar, M.E. Isenkul, C.O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings, IEEE J. Biomed. Health Inform. 17 (2013) 828–834.
[26] M. Asgari, I. Shafran, Predicting severity of Parkinson's disease from speech, Conf. Proc. IEEE Eng. Med. Biol. Soc. (2010) 5201–5204.
[27] L. Ali, C. Zhu, Z. Zhang, Y. Liu, Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network, IEEE J. Transl. Eng. Health Med. 7 (2019) 2000410.
[28] A. Tsanas, M. Zañartu, M.A. Little, C. Fox, L.O. Ramig, G.D. Clifford, Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering, J. Acoust. Soc. Am. 135 (5) (2014) 2885–2901.
[29] R.B. Postuma, D. Berg, M. Stern, W. Poewe, C.W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A.E. Lang, G. Halliday, C.G. Goetz, T. Gasser, B. Dubois, P. Chan, B.R. Bloem, C.H. Adler, G. Deuschl, MDS clinical diagnostic criteria for Parkinson's disease, Mov. Disord. 30 (12) (2015) 1591–1601.
[30] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A.E. Lang, A. Lees, S. Leurgans, P.A. LeWitt, D. Nyenhuis, C.W. Olanow, O. Rascol, A. Schrag, J.A. Teresi, J.J. van Hilten, N. LaPelle, Movement Disorder Society UPDRS Revision Task Force, "Movement disorder society-sponsored revision of the Unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results, Mov. Disord. 23 (2008) 2129–2170.
[31] M. Morise, Harvest: a High-performance Fundamental Frequency Estimator From Speech Signals, INTERSPEECH, 2017.
[32] D. Talkin, A robust algorithm for pitch tracking (RAPT), in: W.B. Kleijn, K.K. Palatal (Eds.), Speech Coding and Synthesis, Elsevier Science B.V., 1995, pp. 497–518.
[33] P. Boersma, Praat, a system for doing phonetics by computer, Glot. Int. 5 (2002) 341–345.

[34] D.J. Hermes, Measurement of pitch by subharmonic summation, J. Acoust. Soc. Am. 83 (1988) 257–264.
[35] D. Talkin, REAPER: Robust Epoch And Pitch EstimatoR, 2015 https://github.com/google/REAPER.
[36] H. Kawahara, Y. Agiomyrgiannakis, H. Zen, Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis, 9th ISCA Workshop on Speech Synthesis (2016).
[37] X. Sun, Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, IEEE, 2002.
[38] A. Camacho, J.G. Harris, A sawtooth waveform inspired pitch estimator for speech and music, J. Acoust. Soc. Am. 124 (3) (2008) 1638–1652.
[39] N. Yang, H. Ba, W. Cai, I. Demirkol, W. Heinzelman, BaNa: a noise resilient fundamental frequency detection algorithm for speech and music, IEEEACM Trans. Audio Speech Lang. Process. 22 (2014) 1833–1848.
[40] K. Kasi, S.A. Zahorian, Yet another algorithm for pitch tracking, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 2002, pp. I-361-I-364.
[41] L. Shi, J.K. Nielsen, J.R. Jensen, M.A. Little, M.G. Schristensen, Robust bayesian pitch tracking based on the harmonic model, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (11) (2019) 1737–17351.
[42] M. Morise, H. Kawahara, H. Katayose, Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, AES 35th International Conference (2009).
[43] B. Secrest, G. Doddington, An integrated pitch tracking algorithm for speech systems, Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (1983) 1352–1355.
[44] P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, IEEE Trans. Speech Aud. Proc. 15 (2007) 34–43.
[45] S. Gonzalez, M. Brookes, PEFAC - a pitch estimation algorithm robust to high levels of noise, IEEEACM Trans. Audio Speech Lang. Process. 22 (2) (2014) 518–530.
[46] A. Cheveigné, H. Kahawara, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Amer. 114 (4) (2002), 1970-1930.
[47] L. Tan, A. Alwan, Multi-band summary correlogram-based pitch detection for noisy speech, Speech Commun. 55 (2019) 841–856.
[48] M. Morise, Y. Fumiya, O. Kenji, WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE Trans. 99-D (2016) 1877–1884.
[49] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3.
[50] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled signal, IFA Proc. 17 (1993) 97–110.
[51] A.P. Simpson, Phonetic differences between male and female speech, Lang. Linguist. Compass 3/2 (2009) 621–640.
[52] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, Suitability of dysphonia measurement for telemonitoring of Parkinson's disease, IEEE Trans. Biomed. Eng. 56 (2009) 1015–1022.
[53] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Eng. 59 (2012) 1264–1271.
[54] S. Sarkar, J. Raymick, S. Imam, Neuroprotective and therapeutic strategies against Parkinson's disease: recent perspectives, Int. J. Mol. Sci. 17 (6) (2016) 904.
[55] J. Rusz, C. Saft, U. Schlegel, R. Hoffman, S. Skodda, Phonatory dysfunction as a preclinical symptom of huntington disease, PLoS One 9 (11) (2014), e113412.
[56] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E.M. Christensen, O. Winther, J.E. Bardram, L.V. Kessing, Voice analysis as an objective state marker in bipolar disorder, Transl. Psychiatry 6 (856) (2016) 1–8.
[57] A. Bandini, J.R. Green, J. Wang, T.F. Campbell, L. Zinman, Y. Yunusova, Kinematic features of jaw and lips distinguish symptomatic from presymptomatic stages of bulbar decline in amyotrophic lateral sclerosis, J. Speech Lang. Hear. Res. 61 (5) (2018) 1118–1129.
[58] A.P. Vogel, J. Fletcher, P.J. Snyder, A. Fredrickson, P. Maruff, Reliability, stability, and sensitivity to change and impairment in acoustic measures of timing and frequency, J. Voice 25 (2011) 137–149.
[59] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task, J. Acoust. Soc. Amer. 134 (2013) 2171–2181.
[60] D. Kempler, D. Van Lancker, Effect of speech task on intelligibility in dysarthria: a case study of Parkinson's disease, Brain Lang. 80 (2002) 449–464.
[61] R.J. Holmes, J.M. Oates, D.J. Phyland, A.J. Hughes, Voice characteristics in the progression of Parkinson's disease, Int. J. Lang. Commun. Disord. 35 (2000) 407–418.
[62] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of speech intelligibility in parkinson's disease using a speech-to-Text system, IEEE Access 5 (2017) 22199–22208.
[63] J. Lebacq, J. Schoentgen, G. Cantarella, F. Thomas Bruss, C. Manfredi, P. DeJonckere, Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research, J. Voice 31 (5) (2017) 550–556.
[64] S. Jannetts, F. Schaeffler, J. Beck, S. Cowen, Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types, Int. J. Lang. Commun. Disord. 54 (2019) 292–305.
[65] C. Manfredi, J. Lebacq, G. Cantarella, J. Schoentgen, S. Orlandi, A. Bandini, P.H. Dejonckere, Smartphones offer new opportunities in clinical voice research, J. Voice 31 (1) (2017), 111.e1-111.e7.

2.4 Speech timing

2.4.1 Speech pauses

Considering speech pauses, both development of automated methods and the effect of parkinsonism has been thoroughly investigated [29], [46]. From the motor side, prolongation of pauses was associated with dysrhythmia patterns in iRBD and PD cohorts and partially with a decreased ability to stop voicing properly, which may reflect weak abduction of the vocal folds due to bradykinesia and rigidity of laryngeal muscles. Furthermore, prolonged pauses might be a sign of a mild cognitive decline, a risk factor of conversion into a synucleinopathy such as PD or dementia with Lewy bodies [47], [48].

Detecting speech pauses has been a longstanding engineering challenge, leading to the development of various robust techniques over the years. These range from traditional power spectrum processing to more recent advancements like deep neural network approaches [49], [50].

2.4.2 Speech rate

Considering the rate of speech, it represents of key dimensions in dysarthria. In PD, disparate findings have been reported, including no changes in the speech rate, decreased speech rate, or even an accelerated speech rate [51]–[53]. Nonetheless, most patients are expected to develop more severe articulation rate abnormalities as the disease progresses [54]. In people with iRBD, evidence is limited to the trend toward a slower rate [18].

Several methods for measuring speech rate have been developed, ranging from manual annotation to fully automatic techniques. However, the reliability of automated methods in analyzing spontaneous speech has not been thoroughly explored, with most of the research focusing on *read text* utterances. Since speech production of patients with PD was even more affected during extemporaneous speech than it was in nonspontaneous speech tasks, a comprehensive study was conducted to evaluate several automated approaches for estimating spontaneous speech rate in PD, iRBD, and multiple system atrophy, a rare, severe parkinsonism [55].

A variety of approaches for estimating NAR and NSR are available; apart from methods that classify the speech rate according to discrete categories such as *slow* and *fast* [56], most algorithms assess the articulation rate in terms of speech units per time. A widely used metric is the words per minute measurement [57], but this approach has several disadvantages, such as a specific word spanning from a single isolated speech sound to a multisyllabic expression or combination to two or more words to produce a composite form (such as “work” and “man” to produce “workman”). Given the variations in a word’s acoustic and syllabic complexity, the words per minute measurement appears to be suitable for a short, functional vocal task but not for a real-world scenario in which the input may vary substantially in terms of length or content. As a result, counting syllables per minute has become popular and is currently perceived as a standard measurement [58].

The speech rate evaluation can thus be considered as a syllable detection task. However, automating the process has several challenges. Syllable localization techniques are not fully established, and none of the developed algorithms is universally applicable [59]. Usually, the methods are developed for a specific, short context only, for healthy speech without the presence of dysarthria [60]. Therefore, this study aimed to evaluate the reliability of different approaches for estimating the articulation rates in connected speech

of Parkinsonian patients with different stages of neurodegeneration compared to healthy controls.

In total, 21 different approaches were identified and tested in a cohort of 25 PD, 25 multiple system atrophy patients, and 25 healthy controls. The results showed that the speech rate features of connected speech using syllables as units in time could be measured accurately based on data from patients with different synucleinopathies and various degrees of severity of speech disorders. The estimation accuracy heavily depends on the particular method used. It was found that the most precise estimates produce a developed framework of speech recognition tool followed by hyphenation procedure. Moreover, the speech recognition tools, such as Whisper, are extensively robust across various conditions and languages [61]. However, a minor deterioration in accuracy was observed in multiple system atrophy group, probably due to the algorithm's lack of familiarity with highly dysarthric speech. If the study cohort contains subjects with severe dysarthria conditions, or there are ethical or technical issues that prohibit the use of automated speech recognition a method based on intensity characteristic function followed by a balanced peak pruning [62] might be used instead.

Using the validated methods, measures of speech rate can be obtained remotely and automatically from the subjects. The findings suggest that the automatic evaluation and tracking of changes in the speech rate of both reading passages and spontaneous speech may be able to provide a natural biomarker of disease progression. The preprint of the article is provided below.

ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Toward Automated Articulation Rate Analysis via Connected Speech in Dysarthrias

Vojtěch Illner,[a] Tereza Tykalová,[a] Michal Novotný,[a] Jiří Klempíř,[b] Petr Dušek,[b] and Jan Rusz[a,b]

[a] Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic [b] Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic

ARTICLE INFO

Article History:
Received October 12, 2021
Revision received December 7, 2021
Accepted December 30, 2021

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton

https://doi.org/10.1044/2021_JSLHR-21-00549

ABSTRACT

Purpose: This study aimed to evaluate the reliability of different approaches for estimating the articulation rates in connected speech of Parkinsonian patients with different stages of neurodegeneration compared to healthy controls.
Method: Monologues and reading passages were obtained from 25 patients with idiopathic rapid eye movement sleep behavior disorder (iRBD), 25 de novo patients with Parkinson's disease (PD), 20 patients with multiple system atrophy (MSA), and 20 healthy controls. The recordings were subsequently evaluated using eight syllable localization algorithms, and their performances were compared to a manual transcript used as a reference.
Results: The Google & Pyphen method, based on automatic speech recognition followed by hyphenation, outperformed the other approaches (automated vs. hand transcription: $r > .87$ for monologues and $r > .91$ for reading passages, $p < .001$) in precise feature estimates and resilience to dysarthric speech. The Praat script algorithm achieved sufficient robustness (automated vs. hand transcription: $r > .65$ for monologues and $r > .78$ for reading passages, $p < .001$). Compared to the control group, we detected a slow rate in patients with MSA and a tendency toward a slower rate in patients with iRBD, whereas the articulation rate was unchanged in patients with early untreated PD.
Conclusions: The state-of-the-art speech recognition tool provided the most precise articulation rate estimates. If speech recognizer is not accessible, the freely available Praat script based on simple intensity thresholding might still provide robust properties even in severe dysarthria. Automated articulation rate assessment may serve as a natural, inexpensive biomarker for monitoring disease severity and a differential diagnosis of Parkinsonism.

Parkinson's disease (PD) is a neurological disorder characterized by the abnormal accumulation of aggregates of alpha-synuclein protein in the neurons, nerve fibers, or glial cells (McCann et al., 2014; Poewe et al., 2017). At present, the available pharmacotherapy methods only mitigate PD motor symptoms and do not treat the actual process of the disease. Although neuroprotective therapies are currently being developed (Devos et al., 2021), no treatment can stop or slow the progression of the disease at the moment. The diagnosis is typically made with the appearance of cardinal motor manifestations, including bradykinesia, rigidity, and resting tremor (Poewe et al., 2017). The fact that PD progresses over many years before the appearance of obvious motor manifestations may be the main reason that neuroprotective therapy has not been discovered because, when a formal diagnosis is made, it is simply too late for an intervention. Therefore, early identification of PD in its prodromal stages is crucial for the development of future therapies (Högl et al., 2018; Schenck et al., 2013).

Idiopathic rapid eye movement sleep behavior disorder (iRBD) is characterized by the loss of physiologic muscle atonia and abnormal behavior during rapid eye movement sleep, such as significant, uncontrolled body movements and vocalizations (St Louis et al., 2017). The

Correspondence to Jan Rusz: rusz.mz@gmail.com. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

patients have a high risk (> 80%) of developing a neurodegenerative disease in the synucleinopathy group, predominantly PD or dementia with Lewy bodies, but possibly also multiple system atrophy (MSA; Arnaldi et al., 2017). In people with MSA, alpha-synuclein aggregates spread into the glial cells as well as into the neurons, and affect the brainstem, the cerebellum, the basal ganglia, and the cortex (Longo et al., 2015). The initial symptoms may be similar to PD, but the progression of the disease is faster, the symptoms are less responsive to dopaminergic therapy, and the patients develop severe autonomic nervous system dysfunctions and postural disorders at an early stage (Longo et al., 2015).

Speech Rate Impairment in PD, MSA, and iRBD

As the most complex acquired human motor skill entailing over 100 muscles, speech is highly sensitive to degenerative processes in the brain's motor system (Duffy, 2019). Speech and voice disorders, referred to as *hypokinetic dysarthria*, develop in up to 90% of patients with PD (Ho et al., 1999). Typical characteristics of hypokinetic dysarthria are decreased voice quality, imprecise articulation, monopitch, monoloudness, and rhythm deficits in timing and phrasing (Ho et al., 1999). Compared to PD, MSA typically manifests more severe dysarthria with mixed hypokinetic, ataxic, and spastic features. Voice changes in patients with iRBD have been suggested to be the first motor signs of the disease and may appear up to 10 years before the cardinal motor symptoms (Postuma et al., 2012). The findings of a recent multilanguage study based on an objective acoustic analysis of 150 patients with iRBD indicated that subliminal speech dysfunction may be a potential prodromal and progressive biomarker of Parkinsonism (Rusz et al., 2021). As vocal assessments have intriguing advantages because they are inexpensive, noninvasive, and easy to administer, changes in speech are promising candidates as preclinical diagnostic and progressive biomarkers of Parkinsonism.

A hallmark of dysarthric speech is alterations in the rate of speech. Due to more widespread neurodegeneration, patients with MSA usually exhibit substantial slowness of articulation rate (Rusz et al., 2015; Skrabal et al., 2020). In PD, no speech rate alterations were observed in early stages (Rusz et al., 2011, 2021), whereas disparate findings have been reported in the later stages, including no changes in the speech rate, decreased speech rate, or even an accelerated speech rate (Delval et al., 2016; Konstantopoulos et al., 2021; Skodda & Schlegel, 2008). Nonetheless, most patients are expected to develop more severe articulation rate abnormalities as the disease progresses (Skodda & Schlegel, 2008; van Nuffelen et al., 2009). In people with iRBD, evidence is limited to the trend toward a slower rate (Rusz et al., 2021).

Most importantly, knowledge about speech rate impairments in PD and other synucleinopathies is based mainly on *read text* utterances, whereas little is known about patterns of timing deficits in spontaneous speech. It has been found that the speech production of patients with PD was even more affected during extemporaneous speech than it was in nonspontaneous speech tasks (Kempler & Lancker, 2002; Rusz et al., 2013). For example, simple speech tasks such as reading sentences or a diadochokinetic test (e.g., producing the /pa/, /ta/, /ka/ sequence) do not require the patient's full attention because the production is more automatic than are the complex procedures involved in spontaneous speech.

Moreover, most speech pathologists currently evaluate the articulation rate by hand-labeling or using perceptual tests in their medical practices (Sussman & Kris, 2012). Since hand-labeling is time consuming and subjective, it does not allow for the assessment of spontaneous recordings with arbitrary content and length (Kent, 1996). Therefore, the design of a system allowing the automatic assessment of speaking rate is necessary.

Toward Automated Assessment of Speech Rate

Concerning timing deficits, the net articulation rate (NAR) or net speech rate (NSR) are speech features that are consistently affected in dysarthria (Delval et al., 2016; Konstantopoulos et al., 2021; Rusz et al., 2021; Skodda & Schlegel, 2008). While NSR describes the rate of speech computed from the entire utterance, NAR is defined as the total number of speech units (such as words and syllables) divided by the speech duration after the removal of pauses. Both features capture respiration and articulation impairments in dysarthria; these elements significantly affect the form of speaking, such as a gradual slowing of the speech tempo, stuttering, or sudden acceleration (Skodda & Schlegel, 2008; van Nuffelen et al., 2009). These lead to unintelligible parts of speech and unfinished words.

A variety of approaches for estimating NAR and NSR are available; apart from methods that classify the speech rate according to discrete categories such as "slow" and "fast" (Zellner, 1998), most algorithms assess the articulation rate in terms of speech units per time. A widely used metric is the *words per minute* measurement (Maclay & Osgood, 2015), but this approach has several disadvantages. Notably, a specific word can span from a single isolated speech sound (such as the first person singular "I") to a multisyllabic expression. Moreover, most languages allow for two or more words to be combined to produce a composite form (such as "work" and "man" to produce "workman"), and this form is considered to be a single word in the words per minute approach. Given the variations in a word's acoustic and syllabic complexity,

the words per minute measurement appears to be suitable for a short, functional vocal task such as reading a sentence with fixed, predescribed content, but not for a real-world scenario in which the input may vary substantially in terms of length or content. As a result, a method for counting syllables per minute, called *mrate,* was developed (Morgan & Fosler-Lussier, 1998). This approach has become popular and is currently perceived as a standard and putatively language-independent measurement. Therefore, we focused on the techniques that use syllable counts in this study.

The NAR and NSR evaluation tasks can, thus, be considered to be syllable detection tasks. However, automating the process has several challenges. Syllable localization techniques are not fully established, and none of the developed algorithms is universally applicable (Jiao et al., 2015). Typically, a particular algorithm is designed to work reasonably well in a specific scenario (e.g., short utterances with fixed content; Huici et al., 2016), but might not perform adequately in a different context. The presence of dysarthria might be a notable disruptive factor; for example, the algorithm may struggle with dysarthria-specific impairments such as phonatory disruptions, imprecise articulation, hypernasality, or low voice intensity (Kent et al., 1999). However, almost all the existing methods were developed based on healthy speech, with only a few studies (Martens et al., 2015; Huici et al., 2016; Jiao et al., 2015; Carmichael, 2017) focusing on patients with dysarthria. Moreover, only one study (Carmichael, 2017) has analyzed a speech task involving more than 20 words. To the best of our knowledge, no research related to the reliability of automatic speech rate estimations using a monologue and a broader range of dysarthria severity has been conducted.

Aims of This Study

This study aimed to evaluate the reliability of different approaches for estimating the articulation rates in connected speech of Parkinsonian patients with different stages of neurodegeneration compared to healthy controls (HCs). We hypothesized that automated assessment of articulation rate in Parkinsonian with reliable performance as compared to human annotation would be possible. We also hypothesized that altered articulation rate would be detectable from prodromal to advanced stages of neurodegeneration.

Method

Participants

This article represents the observational, cross-sectional research study. Ninety-five native Czech speakers were recruited. The research was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic, and written, informed consent was provided by each participant.

Twenty-five de novo, drug-naive patients with PD (10 women and 15 men) with a mean age of 60.8 (*SD* = 11.5, range: 41–75) years were diagnosed based on the Movement Disorders Society's clinical diagnostic criteria (Postuma et al., 2015). The inclusion criteria for PD were as follows: (a) native Czech language speaker, (b) no history of therapy with antiParkinsonian medication, (c) no history of significant communication or neurological disorders unrelated to PD, and (d) no current or past involvement in any speech therapy. Twenty-five participants with iRBD (nine women and 16 men) with a mean age of 61.8 (*SD* = 8.3, range: 40–73) years were diagnosed according to the diagnostic criteria in the third edition of the International Classification of Sleep Disorders based on video-polysomnography (Sateia, 2014). The inclusion criteria for iRBD were as follows: (a) native Czech language speaker, (b) no history of therapy with antiparkinsonian medication, and (c) no history of significant communication or neurological disorders. Twenty patients with MSA and the Parkinsonian subtype (10 women, 10 men) with a mean age of 61.2 (*SD* = 7.4, range: 45–72) years were diagnosed using the consensus diagnostic criteria for MSA (Gilman et al., 2008). The inclusion criteria for MSA were as follows: (a) native Czech language speaker, (b) no history of therapy with antipsychotic medication, (c) no history of significant communication or neurological disorders unrelated to MSA, (d) no current or past involvement in any speech therapy, and (e) no severe cognitive decline that would interfere with recording procedure. A movement disorder specialist (Jiří Klempíř and Petr Dušek) performed the diagnosis and clinical evaluation. For the iRBD and PD groups, the severity of the disease was rated (Petr Dušek) according to the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS III) motor part (which ranges from 0 to 132, with 0 indicating *no motor manifestations* and 132 representing *severe motor disturbance*; Goetz et al., 2008). The patients with MSA were scored (Jiří Klempíř) using Neuroprotection and Natural History in the Parkinson Plus Syndromes Scale (ranging from 0 indicating *no manifestations* to 309 representing *severe dysfunction*; Payan et al., 2011); interrater reliability for the total score has been shown to be high (intraclass coefficient = 0.94; Payan et al., 2011). Patients with MSA had been treated with levodopa, either on its own or with dopamine agonists and/or amantadine. The clinical descriptions of the severity of the speech disorders of the individuals with iRBD, PD, and MSA were determined perceptually based on speech Item 3.1. in the MDS-UPDRS III, which ranges from 0 to 4 with 0 indicating *normal* and 4 *unintelligible speech*. All the patients with PD were recruited consecutively during their first visit to

the clinic and were examined before symptomatic treatment began. The duration of the disease was estimated based on the self-reported evidence of the first motor symptoms. See Table 1 for the clinical characteristics of the patients.

The HC group consisted of 25 age- and gender-matched (10 women, 15 men) participants with a mean age of 62.0 (*SD* = 8.5, range: 40–74) years. The inclusion criteria for HC were as follows: (a) native Czech language speaker and (b) no history of significant communication or neurological disorders.

Speech Assessment

The recordings were captured in a closed room with a low ambient noise level. A head-mounted condenser microphone (Beyerdynamic Opus 55) was used to record the data. The sampling frequency was set to 48 kHz and the resolution to 16-bit. Each participant was accompanied by a guiding speech specialist (Tereza Tykalová, Michal Novotný, and Jan Rusz), and the recordings were made in a single session. The participants were asked to present a monologue about an arbitrary topic of approximately 90 s in duration and perform a reading passage task of a standardized text of 80 words. The same settings applied to subjects in all groups.

The monologues were not altered significantly during the processing, although it was necessary to make minor modifications. In the event that the examiner's speech was recorded, it was carefully removed from the waveform. In the case of one of the participants with MSA, isolated short segments that contained speech so severely unintelligible as to make a direct transcription impossible had to be removed; the extracted percentage of the removed text did not exceed 5% of the recorded duration. The average final duration of the monologues used for the analyses, given in seconds, was 125.4 (*SD* = 13.3) for the HC, 129.5 (*SD* = 16.4) for the iRBD, 122.1 (*SD* = 17.0) for the PD, and 130.8 (*SD* = 44.5) for the MSA groups. The reading passages were not necessary to modify in any way, and the average duration was 33.4 (3.9) for the HC, 36.9 (*SD* = 5.3) for the RBD, 35.7 (*SD* = 5.1) for the PD, and 42.9 (*SD* = 11.5) for the MSA groups.

Table 1. List of the clinical characteristics of the groups.

iRBD (*n* = 25; 9 women and 16 men)	***M/SD* (range)**
Age (years)	61.8/7.4 (40–73)
Symptom duration (years)	5.4/4.2 (1–20)
MDS-UPDRS III total	7.6/5.8 (0–24)
MDS-UPDRS III speech item	0.0/0.2 (0–1)
PD (*n* = 25; 10 women and 15 men)	
Age (years)	60.8/11.5 (41–75)
Symptom duration (years)	3.0/2.4 (1–11)
MDS-UPDRS III total	31.1/12.0 (10–56)
MDS-UPDRS III speech item	0.8/0.4 (0–1)
MSA (*n* = 20; 10 women and 10 men)	
Age (years)	61.2/7.4 (45–72)
Symptom duration (years)	4.0/1.8 (2–7)
NNIPPS total	76.9/33.9 (35–123)
MDS-UPDRS III speech item	1.5/0.7 (0–3)

Note. iRBD = idiopathic rapid eye movement sleep behavior disorder; MDS-UPDRS III = Movement Disorder Society–Unified Parkinson's Disease Rating Scale; PD = Parkinson's disease; MSA = multisystem atrophy; NNIPPS = Neuroprotection and Natural History in the Parkinson Plus Syndromes Scale.

Speech Rate Algorithm Search and Selection Strategy

A systematic search of English articles written before December 2020 was conducted on the Web of Science, IEEE Xplore, and Google Scholar for more relevant studies of speech rate estimation. Multiple keywords were used for the search, namely, "speech rate," "articulation rate," "speech rhythm," "syllable detection," "dysarthria," "syllable count estimation," and "word count estimation." Two methods by Dekens et al. (2014) and Martens et al. (2015) as well as Huici et al. (2016), aimed specifically at speech rate estimation in dysarthria, were identified. In addition, we identified and analyzed 19 different methods (Aharonson et al., 2017; Carmichael, 2017; de Jong & Wempe, 2009; Heinrich & Schiel, 2011; Jiao et al., 2015, 2016; Mannem et al., 2020; Morgan & Fosler-Lussier, 1998; Nayak et al., 2019; Pfitzinger et al., 1996; Räsänen , Doyle, & Frank, 2018; Schwarz & Černocký, 2008; Seshadri & Räsänen, 2019; Villing, 2004; Wang & Narayanan, 2007; Yarra et al., 2016, 2019; Yuan & Liberman, 2010; Zhang & Glass, 2009). A particular method was excluded from further consideration if

(i) the algorithm was designed for functional tasks only (i.e., it would not be suitable for connected speech),
(ii) its computation principle was very similar to another selected algorithm,
(iii) the authors had subsequently published a more advanced algorithm that outperformed the initial one,
(iv) the authors did not compare the performance of their algorithm to any other approaches,
(v) the published platform did not offer a sufficient description of the method used to allow for its proper implementation,
(vi) the algorithm delivered a *substantially* worse performance than did the others in the trials that were conducted, or
(vii) the method required training or tuning data sets that were not available.

Moreover, we added one extra procedure employing an automatic speech recognition system as a proof of the concept. In total, we identified eight methods that were

subsequently included in the testing. See Table 2 for the list of selected and excluded methods with links to their sources and the reasons for their exclusion.

Tested Methods for Estimating Speech Rate

A review of the eight selected techniques is provided in this section. Some algorithms were a popular choice for comparison and were used frequently in the related studies, such as in the work of Wang and Narayanan (2007), whereas others represent methods with novel approaches that were introduced very recently, such as in the work of Seshadri and Räsänen (2019). The computations were performed in MATLAB (MathWorks) and Python environments, although an interface to other programs, such as Praat (Boersma, & Weenink, 2001), was employed in some cases.

Google and Pyphen

This straightforward strategy employs automatic speech recognition for estimating the speech rate, as suggested previously (Wang & Narayanan, 2007; Yuan & Liberman, 2010). However, its potential has not been explored because the speech rate is commonly used as a parameter to improve the performance of automatic speech recognition systems (Heinrich & Schiel, 2011; Nayak et al., 2019). One particular service, *Google Speech-to-Text* (Google Speech-to-Text, 2013), was chosen because it is in the top-tier of speech recognition software and employs complex neural structures; it also has comparably the largest training database and can be used for most of the world's languages.

In the procedure, the data are sent to the service, where the online processing occurs. The outputs are recognized words with given timestamps. The single words are then divided into syllables using Hunspell hyphenation libraries (Németh, 2002), a widely used hyphenation tool across many platforms and languages. We used a Python implementation module in Hunspell called Pyphen (Berendsen, 2015).

Since the only timestamps available were the start and the end of each word, the syllable locations (nuclei) were determined based on their equal distribution throughout a given word. For example, when a word spanned 600 ms and contained three syllables, its nuclei timestamps were allocated as three 200-ms intervals.

Praat Script

The algorithm (de Jong & Wempe, 2009), named after the environment it was created in, uses intensity as the characteristic function from which the syllable position candidates are derived. The intensity contour is computed as a convolution of a squared input signal and a Gaussian type window of a length L. L was set to $3.2 \cdot f_s/f_{min}$ where f_s was the sampling frequency and f_{min} corresponded to minimal signal periodicity, which was set to 50 Hz.

Peaks in the intensity contour with a more extensive value than the intensity median were marked as syllable candidates. Their height was checked relative to the preceding valley, followed by pitch verification using the Praat autocorrelation method (Boersma, 1993).

Phoneme Recognizer

This approach was inspired by Yuan and Liberman (2010), in which a broad phonetic class recognizer was used to estimate the speech rate. Here, we used a phoneme recognizer based on split temporal context feature extraction classified by a neural network with a Viterbi algorithm applied in the decoding phase (Schwarz & Černocký, 2008). The algorithm was trained on the SpeechDat-E database (SpeechDat-E database, 1999) available for multiple languages. The output phoneme stream was then processed via a phonologic classifier to determine the eventual syllables based on a set of fixed patterns.

SylNet

The SylNet (Seshadri & Räsänen, 2019) algorithm is a syllable count estimator that uses a neural network approach. It employs the WaveNet (Oord et al., 2016) model used for speech synthesis in conjunction with a long short-term memory (LSTM) method to predict the aggregated syllable count based on the log-Mel spectrogram (25-ms window with 10-ms time steps) features of speech. The input features pass through an input layer with 128 channels, each consisting of a gated convolutional unit followed by another 10 layers utilizing similar units. The outputs, together with the residual and skip connections, are then fed into a PostNet layer with a layer-specific affine transform. The PostNet gathers all the information via summation. Finally, the sum is sent via a rectified linear unit to an LSTM layer that has 128 cells and a dense layer at its output, thus accumulating syllable count estimation over time.

SylNet was trained on Estonian (Lippus et al., 2013) and Korean (Yun et al., 2015) corpora of spontaneous speech with manually annotated syllable counts. Three other corpora, FinDialog (Lennes, 2009; in Finnish), C-PROM (Avanzi et al., 2010; in French), and Switchboard (Godfrey et al., 1992; in English), were used for testing and further adaptation. We used SylNet in its original pretrained form without any adaptations.

WN

The method, developed by Wang and Narayanan (Wang & Narayanan, 2007), usually referred to as WN (the authors' initials) in the literature, utilizes a speech subband correlation approach. The speech waveform is

Table 2. A list of the identified Net Articulation Rate and syllable detection algorithms.

Algorithm	Authors (reference)	Software available (link)	Reason for exclusion
Used for the analyses			
Google + Pyphen	V. Illner	Yes (https://github.com/vojtaiii/google_speech_to_syllables)	—
Praat script	N. H. de Jong, T. Wempe (de Jong & Wempe, 2009)	Yes (https://sites.google.com/site/speechrate)	—
Phoneme recognizer	P. Schwarz, J. Černocký (Schwarz & Černocký, 2008)	Yes (https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context)	—
SylNet	S. Seshadri, O. Räsänen (Seshadri & Räsänen, 2019)	Yes (https://github.com/shreyas253/SylNet)	—
WN	D. Wang, S. Narayanan (Wang & Narayanan, 2007)	No (supplied by the authors on request)	—
ThetaSeg	O. Räsänen et al. (Räsänen, Doyle, & Frank, 2018)	Yes (https://github.com/orasanen/thetaOscillator)	—
LFME	T. Dekens et al. (Dekens et al., 2014)	No (was replicated)	—
Landmark-based	H. D. Huici et al. (Huici et al., 2016)	No (was replicated)	—
Excluded			
Mode-shape classificator	C. Yarra et al. (Yarra et al., 2016)	No	(ii), (iii)
A real-time phoneme counting algorithm	V. Aharonson (Aharonson et al., 2017)	No	(i), (iv)
Automatic blind syllable segmentation	R. Villing (Villing, 2004)	No	(i), (ii), (iv)
Convex weighting criteria for speech rate estimator	Y. Jiao et al. (Jiao et al., 2015)	No	(i), (iii), (v)
Enhanced speech rate estimation technique	J. N. Carmichael (Carmichael, 2017)	No (commercial)	(ii)
Rhythmicity parameters speech rate estimator	C. Heinrich, F. Schiel (Heinrich & Schiel, 2011)	No	(iv)
Noise robust speech rate estimator	C. Yarra et al. (Yarra et al., 2019)	No	(ii), (v)
Online speech rate estimation using RNNs	Y. Jiao et al. (Jiao et al., 2016)	No	(v), (vii)
Broad phonetic class recognition for speech rate estimator	J. Yuan, M. Liberman (Yuan & Liberman, 2010)	No	(ii), (v)
Speech rate estimator using RNN representations	R. Mannem et al. (Mannem et al., 2020)	Yes (https://github.com/mannemrenuka/sr-cdnn-conv1D)	(i), (ii)
Speech rhythm guided syllable nuclei detection	Y. Zhang, J. R. Glass (Zhang & Glass, 2009)	No	(i), (iv)
Syllable detector in read and spontaneous speech	H. R. Pfitzinger et al. (Pfitzinger et al., 1996)	No	(iv), (vii)
Zero resource speech rate estimator	S. Nayak et al. (Nayak et al., 2019)	No	(iv), (vii)
Combining multiple estimators of speech rate	N. Morgan, E. Fosler-Lussier (Morgan & Fosler-Lussier, 1998)	No	(ii), (vi)

Note. Dashes indicate not applicable. (i) the algorithm was designed for functional tasks only (i.e., it would not be suitable for connected speech), (ii) its computation principle was very similar to another selected algorithm, (iii) the authors had subsequently published a more advanced algorithm that outperformed the initial one, (iv) the authors did not compare the performance of their algorithm to any other approaches, (v) the published platform did not offer a sufficient description of the method to allow for its proper implementation, (vi) the algorithm delivered a substantially worse performance than did the others in the trials that were conducted, and (vii) the method required training or tuning data sets that were not available. WN = Wang and Narayanan; LFME = low-frequency modulated energy; RNN = recurrent neural network.

transformed into a 1-D envelope through a process consisting of several stages. First, a spectral 19-subband procedure is performed with only the top M bands considered by subband energy. The energy is then weighted in the time domain, and the temporal correlation is computed from the subband energies vector using a length of the frame K.

Subsequently, a spectral subband correlation is applied to the selected bands, and the result is smoothed by Gaussian window filtering to obtain the characteristic function. The syllabic positions are then derived using a threshold-based peak localization aided by pitch verification.

The authors initially used a pitch estimation algorithm based on a normalized cross-correlation and dynamic programming (Talkin, 1983). In this study, we applied the sawtooth waveform-inspired estimator (Camacho & Harris, 2008), which was previously found to provide superior results (Illner et al., 2020). The parameter M was set to 12 and K to 11, as suggested by the authors.

ThetaSeg

The ThetaSeg algorithm (Räsänen, Doyle, & Frank, 2018) utilizes *sonority* as a characteristic function and is based on the sonority sequencing principle (Bertoncini & Mehler, 1981). The sonority trajectory is computed using a Gammatone filter bank and a set of harmonic oscillators, as these closely resemble the human hearing apparatus. The cth band output of the filter bank was down-sampled to f_s = 1000 Hz and fed as an input to the harmonic oscillator system.

The sonority trajectory $S[n]$ was obtained by taking eight oscillator amplitudes with the most considerable energy and computing their logarithmic sum, $S[n]$. The $S[n]$ trajectory was then normalized to the range [0, 1], and a peak-picking mechanism was applied to search for the syllabic boundaries.

Low-Frequency Modulated Energy

The low-frequency modulated energy (LFME) algorithm (Dekens et al., 2014; Martens et al., 2015) targets explicitly dysarthric speech. It assumes that substantial information about syllable distribution lies in the low-frequency bands.

The number of bands was chosen as four, and the energy for each band was computed from the input speech signal using the short-time Fourier transform (STFT). The lowest frequency band spanned the range of 50–800 Hz, and the other bands were logarithmically distributed up to 4 kHz. The LFME trajectory was computed as a product of the squared lowest frequency band energy and the sum of the remaining bands.

As the peaks in the LFME[n] trajectory are not caused solely by syllables but also by other speech events, the algorithm applies a complex thresholding and peak-picking algorithm to select only those candidates that are most likely to belong to a syllable segment.

Landmark-Based Detector

This approach is based on a speech landmark detector (Huici et al., 2016; Liu, 1996). The landmarks represent abrupt signal changes and are classified as glottis, sonorants, and bursts. The detector uses six spectral bands ranging from 0 to 8 kHz, which are analyzed via STFT. The STFT outputs are processed by the six-band rate-of-rise (ROR) trajectory detector, which uses the estimation of the first difference in each band's dB energy. Additional smoothing is then applied, and thresholding is performed to identify candidates for the peaks. The ROR detection is performed in two parallel threads for fine and coarse resolution, differing only in their settings. The final set of candidates is an aggregate of the fine and coarse peak candidate subsets.

The candidates serve as inputs to the glottis detector, where additional thresholding and pitch verification are performed using the average magnitude difference function method. The sonorant landmarks are localized using the identified glottis positions and the energy bands' peak-picking structure. Syllable locations are then estimated as segments containing glottis intervals, and sonorant segments are only located inside the glottis segments.

Speech Rate Features

The speech rate features were computed as

$$\mathrm{NSR} = \frac{N_{\mathrm{syll}}}{t_s}, \mathrm{NAR} = \frac{N_{\mathrm{syll}}}{t_s - t_p}, \quad (\mathrm{syll/s}), \qquad (1)$$

where N_{syll} refers to the total number of syllables, t_s is the time span of the utterance, and t_p is the duration of pauses (nonspeech segments).

The Reference Values

The reference values for NSR and NAR were calculated in the same way, with the exception that N_{syll} and t_s values were obtained using a manually annotated transcript of each recording. The value of t_p was determined using an automatic segmentation tool for connected speech (Hlavnička et al., 2017).

The Performance Validation and Statistical Analysis

The normalized root-mean-square error (NRMSE) and the Spearman correlation coefficient r were used for the validation. The NRMSE enables a description of several variables by describing the error as a fraction of the observed variable range. The NRMSE is defined as

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}}{(\max(\hat{x}_i) - \min(\hat{x}_i))}, \quad (-), \qquad (2)$$

where N is the number of utterances, $\hat{x}_i$ represents an estimated feature, and x_i is its respective reference. The r coefficient was computed as a nonparametric measure of the rank correlation between the estimated and reference values in each group. A one-way analysis of variance with Fisher's least significant difference type post hoc tests were applied as a standard tool to evaluate group differences based on the automated estimation provided by the algorithm with the most robust NSR and NAR assessment performance.

Data normality was verified via the Shapiro–Wilcoxon and Bartlett (equality of variance) tests. If the conditions of normality were not met, the Kruskal–Wallis test was applied as a form of nonparametric analysis of variance.

Results

Evaluation of the Algorithms' Performances

Table 3 shows the performance results of the algorithms for NSR and NAR estimated from monologues in each group. The best overall precision was achieved using the Google & Pyphen methods, which outperformed the others for the HC, iRBD, and PD groups, respectively, with NRMSE 0.11 (r = .99, p < .001), 0.17 (r = .95, p < .001), 0.13 (r = .97, p < .001) for NSR, and 0.15 (r = .98, p < .001),

Table 3. Algorithm performance results for the NSR and the NAR estimated from monologues in each group.

	NSR			NAR		
HC	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.11	.99	< .001	0.15	.98	< .001
Praat script	0.22	.85	< .001	0.30	.82	< .001
Phon. recognizer	0.31	.86	< .001	0.82	.77	< .001
SylNet (not adapted)	0.90	.76	< .01	1.28	.56	< .01
WN	0.51	.69	< .001	0.66	.59	< .01
ThetaSeg	0.54	.75	< .001	0.55	.64	< .01
LFME	0.81	.17	.42	0.98	.11	.61
Landmark-based	1.21	–.11	.60	1.35	–.22	.29
iRBD	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.17	.95	< .001	0.22	.92	< .001
Praat script	0.29	.69	< .001	0.36	.65	< .001
Phon. recognizer	0.38	.76	< .001	0.69	.73	< .001
SylNet (not adapted)	0.56	.82	< .001	0.66	.67	< .001
WN	0.47	.71	< .001	0.64	.72	< .001
ThetaSeg	0.49	.56	< .01	0.52	.57	< .01
LFME	1.24	.14	.50	1.46	.06	.79
Landmark-based	1.23	.17	.43	1.56	.16	.43
PD	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.13	.97	< .001	0.17	.98	< .001
Praat script	0.31	.85	< .001	0.43	.84	< .001
Phon. recognizer	0.34	.87	< .001	0.59	.87	< .001
SylNet (not adapted)	0.84	.77	< .001	1.24	.61	< .01
WN	0.66	.46	< .1	0.94	.46	< .1
ThetaSeg	0.71	.69	< .001	0.85	.48	< .1
LFME	0.90	.24	.25	1.10	.12	.55
Landmark-based	1.04	.03	.90	1.22	.01	.96
MSA	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.35	.87	< .001	0.47	.90	< .001
Praat script	0.28	.91	< .001	0.38	.82	< .001
Phon. recognizer	0.41	.91	< .001	0.69	.82	< .001
SylNet (not adapted)	0.57	.90	< .001	0.70	.70	< .001
WN	0.52	.83	< .001	0.80	.74	< .001
ThetaSeg	0.48	.85	< .001	0.53	.64	< .01
LFME	1.06	.56	< .1	1.40	.16	.49
Landmark-based	1.27	–.29	.22	1.40	–.22	.35

Note. NSR = net speech rate; NAR = net articulation rate; HC = healthy controls; NRMSE = normalized root-mean-square error; WN = Wang and Narayanan; LFME = low-frequency modulated energy; iRBD = idiopathic rapid eye movement sleep behavior disorder; PD = Parkinson's disease; MSA = multisystem atrophy.

0.22 (r = .92, p < .001) and 0.17 (r = .98, p < .001) for NAR. The Google & Pyphen methods were only outperformed by the Praat script for the MSA group, with NRMSE 0.35 (r = .87, p < .001) compared to 0.28 (r = .91, p < .001) for NSR and 0.47 (r = .90, p < .001) compared to 0.38 (r = .82, p < 0.001) for NAR. Praat script was the second-best algorithm overall, with solid results for the HC, PD, and MSA groups, aggravated only for the iRBD group for which the algorithm performance dropped to NRMSE 0.29 (r = .69, p < .001) for NSR and to 0.36 (r = .65, p < .001) for NAR.

With regard to the other methods, the phoneme recognizer also provided acceptable results for NSR with a mean NRMSE of 0.36 (r = .76 – 0.91, p < .001). For NAR, it maintained a high correlation coefficient, but the NRMSE rose notably with a mean value of 0.70 (r = .73–0.87, p < .001). SylNet, WN, and ThetaSeg produced a markedly worse performance with occasional exceptions; for example, ThetaSeg attained NRMSE 0.48 (r = .85, p < .001) for NSR and 0.53 (r = .64, p < .01) for NAR in the MSA group. LFME and landmark-based did not produce satisfactory results in any group.

Table 4 shows the performance results estimated from reading passages. The outcomes are very similar or slightly better compared to the results from monologue task; that is, Google & Pyphen had the highest precision across HC, iRBD, PD, and MSA groups with NRMSE 0.13 (r = .98, p < .001), 0.11 (r = .99, p < .001), 0.10 (r = .99, p < .001) and 0.21 (r = .95, p < .001) for NSR, and 0.15 (r = .98, p < .001), 0.13 (r = .99, p < .001), 0.12 (r = .99, p < .001) and 0.30 (r = .91, p < .001) for NAR. Praat script showed

Table 4. Algorithm performance results for the NSR and the NAR estimated from reading passages in each group.

	NSR			NAR		
HC	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.13	.98	< .001	0.15	.98	< .001
Praat script	0.34	.82	< .001	0.45	.86	< .001
Phon. recognizer	0.80	.84	< .001	1.03	.87	< .001
SylNet (not adapted)	1.87	–.72	< .001	2.35	–.45	< .1
WN	0.61	.55	< .01	0.76	.56	< .01
ThetaSeg	0.65	.63	< .001	0.84	.62	< .01
LFME	0.94	.07	.72	1.09	.15	.47
Landmark-based	1.75	.02	.93	2.08	–.15	.48
iRBD	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.11	.99	< .001	0.13	.99	< .001
Praat script	0.34	.80	< .001	0.41	.78	< .001
Phon. recognizer	0.93	.79	< .001	1.09	.77	< .001
SylNet (not adapted)	1.63	–.48	< .1	1.95	–.46	< .1
WN	0.63	.78	< .001	0.80	.75	< .001
ThetaSeg	0.79	.54	< .01	0.91	.39	< .1
LFME	1.37	.25	.22	1.68	.32	.12
Landmark-based	1.47	–.04	.86	1.75	–.08	.71
PD	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.10	.99	< .001	0.12	.99	< .001
Praat script	0.31	.80	< .001	0.37	.91	< .001
Phon. recognizer	0.85	.74	< .001	0.85	.74	< .001
SylNet (not adapted)	1.78	–.73	< .001	2.18	–.59	< .01
WN	0.52	.77	< .001	0.59	.82	< .001
ThetaSeg	0.70	.69	< .001	0.79	.74	< .001
LFME	0.97	.15	.46	1.16	.18	.40
Landmark-based	1.24	.07	.73	1.44	.11	.60
MSA	**NRMSE**	***r***	***p***	**NRMSE**	***r***	***p***
Google & Pyphen	0.21	.95	< .001	0.30	.91	< .001
Praat script	0.25	.93	< .001	0.35	.90	< .001
Phon. recognizer	0.70	.89	< .001	0.86	.89	< .001
SylNet (not adapted)	1.37	–.08	.75	1.62	–.48	< .1
WN	0.59	.85	< .001	0.78	.68	< .01
ThetaSeg	0.44	.85	< .001	0.59	.75	< .001
LFME	1.25	.49	< .1	1.47	.37	.11
Landmark-based	1.39	–.33	.15	1.63	–.40	< .1

Note. NSR = net speech rate; NAR = net articulation rate; HC = healthy controls; NRMSE = normalized root-mean-square error; WN = Wang and Narayanan; LFME = low-frequency modulated energy; iRBD = idiopathic rapid eye movement sleep behavior disorder; PD = Parkinson's disease; MSA = multisystem atrophy.

second-best performance in all groups with a maximum NRMSE value of 0.45 ($r = 0.86$, $p < .001$) for NAR in the HC group. The other methods did not reach sufficient performance across reading passages.

Group Differences

Based on the results presented in Table 3 and Table 4, Google & Pyphen and Praat script were found to be reliable estimation methods for NSR and NAR both from monologues and reading passages. Figure 1 demonstrates statistically significant group differences among the HC, PD, iRBD, and MSA groups for the NSR and NAR features estimated by Google & Pyphen and Praat script from monologues. Substantial differences were found between the HC and MSA groups ($p < .001$ for NSR, NAR, Google & Pyphen estimated, $p < .001$ for NSR and $p < .01$ for NAR, Praat script estimated). Moreover, the MSA group was differentiated from the iRBD group ($p < .001$ for NSR, $p < .01$ for NAR, Google & Pyphen estimated, $p < .05$ for NSR, Praat script estimated) and from the PD group ($p < .001$ for NSR and NAR, Google & Pyphen estimated and $p < .05$ for NSR,

Figure 1. Group differences for NSR (net speech rate) and NAR (net articulation rate) features estimated from arbitrary monologues by two most accurate approaches including Google & Pyphen and Praat script methods with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$. Middle bars represent median, rectangles represent interquartile range. Maximum and minimum values are by error bars. Outliers are marked as crosses. HC = healthy controls; iRBD = idiopathic rapid eye movement sleep behavior disorder; PD = Parkinson's disease; MSA = multisystem atrophy.

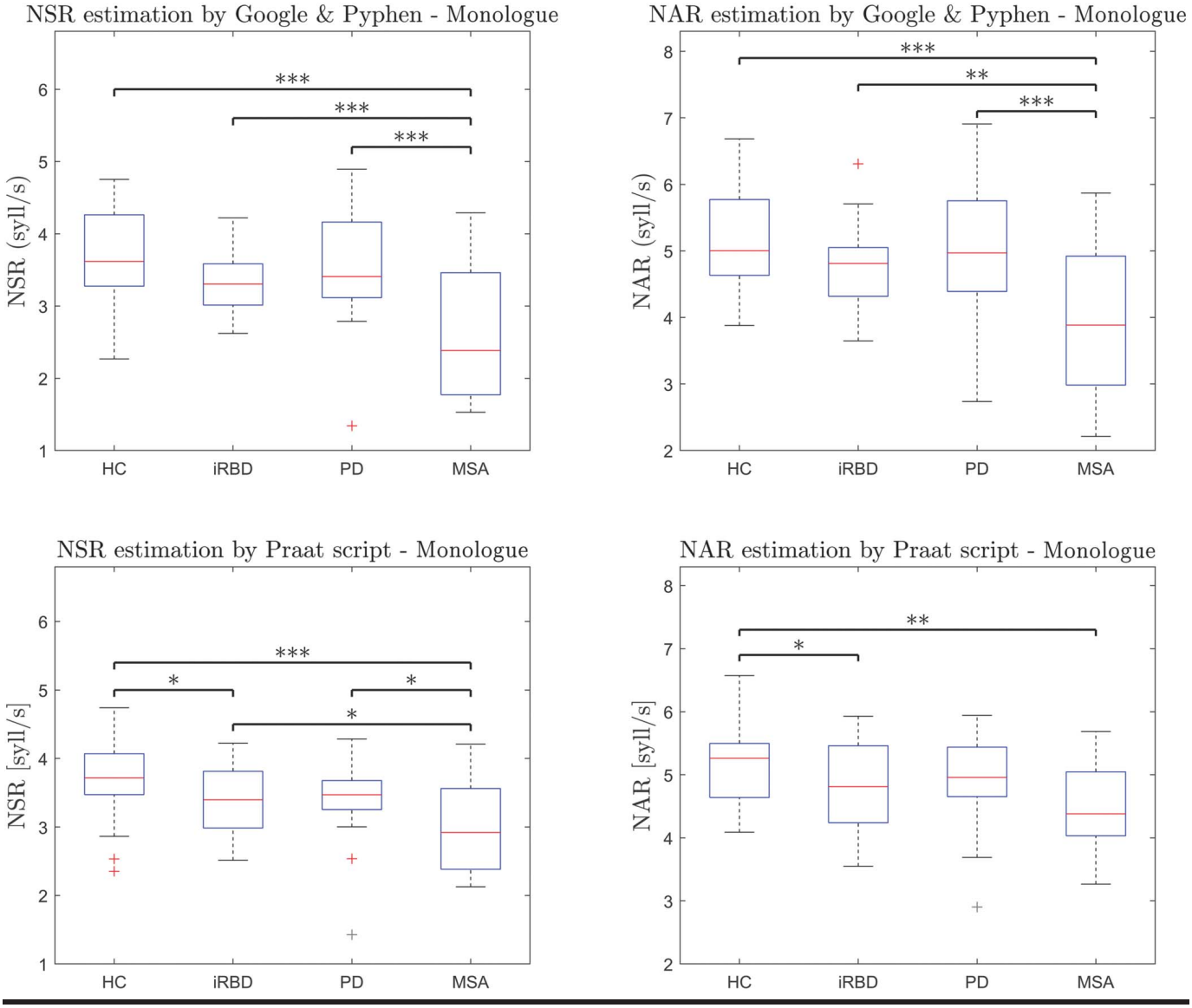
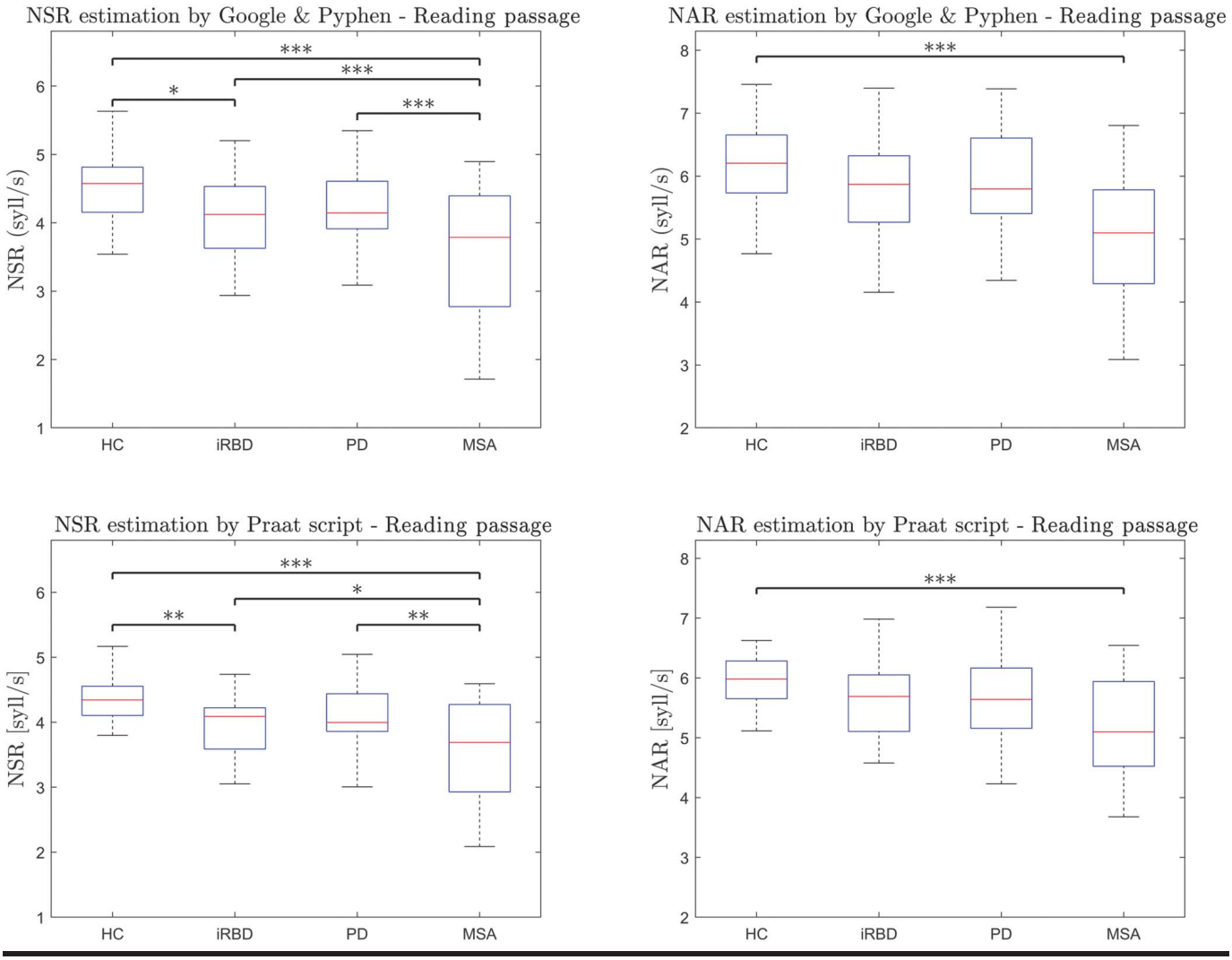


Figure 2. Group differences for NSR (net speech rate) and NAR (net articulation rate) features estimated from reading passages by two most accurate approaches including Google & Pyphen and Praat script methods with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$. Middle bars represent median, rectangles represent interquartile range. Maximum and minimum values are by error bars. Outliers are marked as crosses. HC = healthy controls; iRBD = idiopathic rapid eye movement sleep behavior disorder; PD = Parkinson's disease; MSA = multisystem atrophy.



Praat script estimated). The iRBD group demonstrated a significant difference compared to the HC in the Praat script NSR and NAR estimates ($p < .05$).

Figure 2 shows significant differences for NSR and NAR features estimated by Google & Pyphen and Praat script from reading passages. A notable NSR decrease can be observed in the MSA group compared to HC, iRBD, and PD groups ($p < .001$) based on Google & Pyphen, and to HC ($p < .001$), iRBD ($p < .05$), and PD ($p < .01$) based on Praat script. Moreover, the iRBD group demonstrated a significant NSR decline compared to the HC estimated by both Google & Pyphen ($p < .05$) and Praat script ($p < .01$). For NAR, only a significant difference between HC and MSA was observed ($p < .001$) estimated by both Google & Pyphen and Praat script methods.

Discussion

Our results showed that the speech rate features of connected speech using syllables as units in time could be measured accurately based on a large sample of data from patients with different synucleinopathies and various degrees of severity of speech disorders. In general, the detection performance of algorithms was similar or slightly better in reading passage compared to the monologue task. Using a fully automated approach, we captured a slow speech rate in patients with MSA and a tendency toward a slower speech rate in patients with iRBD from both arbitrary monologues and reading passages. Most of the previous studies have focused on specific vocal tasks such as reading a carefully selected short

utterance lasting a few seconds (Mannem et al., 2020; Mendoza Ramos et al., 2020; Zhang & Glass, 2009); however, our current study proposed a general approach that allowed for more possibilities and advances in the estimation of speech rates. Our findings suggest that the automatic evaluation and tracking of changes in the speech rate of both reading passages and spontaneous speech may be able to provide a natural biomarker of disease progression. Importantly, as spontaneous speech encompasses the complexity of human speech production, including motor execution and cognitive–linguistic processing, it has likely been shown to be superior for capturing subtle PD-related speech disruptions than functional vocal tasks (Rusz et al., 2013, 2018). Moreover, when a wearable device, such as a smartphone or a recorder, is used for the recordings of connected speech, the data may be acquired by the patients themselves at home, without any further cost or time burden on either the patients or the investigators. Using well-defined and disease-specific biomarkers such as the speech rate during a short-time observation of disease progression might help to recruit carefully selected subjects to participate in studies examining prodromal PD. Such biomarkers would facilitate an early presymptomatic diagnosis, as well as rapid access to neuroprotective therapy once it is available.

The estimation accuracy of NSR and NAR relies heavily on the particular method used. Google & Pyphen outperformed the other approaches in terms of reliable feature estimates and resilience to dysarthric speech. To the best of our knowledge, this study is the first to use the Google & Pyphen approach to assess the speech rate of the natural, connected speech of dysarthric patients. For monologues, the results, based on the NSR and NAR estimation by Google & Pyphen for the HC, PD, and iRBD groups, were extremely precise compared to other approaches that followed a similar scenario (Abdelwahab & Busso, 2014; Dekens et al., 2007; Heinrich & Schiel, 2011; Pfitzinger et al., 1996; Räsänen , Sesadri, & Casillas, 2018). However, this method had a minor deterioration in accuracy for syllable detection in the MSA group, also observed in the performance for the reading passage task. In agreement with Dekens et al. (2007), we might assume that the decreased precision was caused by the algorithm's lack of familiarity with highly dysarthric speech; in other words, it had difficulty recognizing words that were on the border of intelligibility, as the data on which it is trained are not as comprehensive as they could be. This factor might indicate the algorithm's level of sensitivity to the input data, such as nonstandard speech and other aggravating conditions in the scenario. However, the results for the automatic speech rate features estimated via Google & Pyphen showed a correlation with manual hand-labeling that was greater than 0.91 for reading passages and 0.87 for monologues across all the groups of interest. This fact is crucial for clinical practice, as it is more important to achieve a correct estimation of the patient's speech performance than it is to obtain the precise position of individual syllables.

The Praat script method was sufficiently accurate, considering its simplicity. This finding is in accordance with other studies scrutinizing its performance (Jiao et al., 2016; Nayak et al., 2019). The method achieved similar results across all the investigated groups, regardless of the dysarthria severity degree, and demonstrated the best precision of all the examined methods for the monologue task in the MSA group. This might be attributed to a considered choice of an advantageous intensity characteristic function and balanced peak pruning that showed robust properties even in highly dysarthric speech. It should be noted that the Google & Pyphen and Praat script methods are likely to be able to be applied to different languages, or even to multilanguage studies, in a straightforward way without exhaustive tuning and data set preparation.

The performance of the phoneme recognizer was also within acceptable boundaries for practical use. However, adjusting the procedure to be incorporated into a comprehensive, multilanguage system might be challenging due to extensive database preparation and model development. The SylNet algorithm demonstrated lesser robustness of a direct neural net approach for connected speech tasks. Apart from the training process, adapting the algorithm to a given language, vocal tasks, and different conditions seems necessary for the precise estimation of the results. The other tested methods failed to provide sufficient accuracy in both monologue and reading passage tasks due to inappropriate design for the given scenario.

Considering the methods' outcomes and their stability across the given conditions, we anticipate that an automatic articulation rate assessment system that is suitable for widespread use must be designed based on automatic speech recognition, such as the Google & Pyphen algorithm, to achieve the best accuracy possible. Nevertheless, if ideal circumstances, such as sufficient computational power and online processing, are not available, or the analysis of severely dysarthric patients is of interest, the Praat script algorithm might be an adequate substitute due to its simplicity and greater robustness in varying conditions.

With regard to the group differences among the iRBD, PD, and MSA groups, similar trends were observed for spontaneous speech and reading passages. The speech of people with MSA was characterized as having the slowest rate, which may have been caused by the more widespread neuronal atrophy and occurrence of spastic dysarthria elements (Rusz et al., 2015; Skrabal et al., 2020). Our findings concerning the unchanged speech rate in monologues and reading passages produced by early PD patients are in agreement with those of previous studies that have investigated only the reading passage

task (Rusz et al., 2011, 2021). It would appear that characteristic changes in the articulation rate appear later in PD progression. Both slower and faster speech rates can be observed in advanced stages of PD (Skodda & Schlegel, 2008); faster speech is likely to reflect a physiological tendency to accelerate speech due to the impaired motor planning (oral festination) that is frequent in early stage PD patients (Delval et al., 2016). The tendency toward a slower speech rate may be theoretically attributed to the degeneration of nondopaminergic pathways (Skodda et al., 2009). Of interest, we also observed a tendency for speech to slow in patients with iRBD, mainly from the NSR feature. A slower rate represents a typical speech change due to mild cognitive impairment (De Looze et al., 2018) and has been reported previously in patients with dementia with Lewy bodies (Ash et al., 2012). Thus, the tendency toward a slower articulation rate in our patients with iRBD may reflect the fact that approximately 40% of iRBD patients later convert to dementia with Lewy bodies.

It must be admitted that this study has limitations. First, the current findings are based solely on the Czech language; thus, the language independence of the applied methods should be verified in future studies. Nonetheless, most of the state-of-the-art automatic speech recognition systems support the majority of the world's languages; Praat script has been evaluated using different language data sets (de Jong & Wempe, 2009; Jiao et al., 2016; Nayak et al., 2019) and has produced comparable performances. Finally, patients in the early stage of PD were compared to patients who had been treated and who had more advanced MSA. Thus, it is not certain whether the differences in the articulation rates would also apply to the differential diagnosis of de novo patients with MSA and PD.

Conclusions

This study represents a further step toward the automatic evaluation of speech disorders in PD and other synucleinopathies. We found that features of the speech and articulation rates could be reliably estimated from reading passages and spontaneous speech using a state-of-the-art speech recognition system and hyphenation rules. For robust results in more severe conditions, such as in the advanced stages of dysarthria, the Praat script algorithm might be a reasonable choice. This study also sheds light on the mechanisms that underlie potential speech rate differences in Parkinsonian speech production. Speech rate metrics may thus provide an applicable digital biomarker for the assessment of the severity of the disease, to monitor the effects of speech therapy, to differentiate among synucleinopathy subtypes, or to assess the efficacy of experimental disease-modifying treatments for PD. Future longitudinal studies should confirm changes related to speech rates as a potential diagnostic and progressive biomarker of PD.

Acknowledgments

This study was supported by the Czech Ministry of Health (Grant NU20-08-00445 awarded to Vojtech Illner, Tereza Tykalova, Michal Novotny, Petr Dusek, and Jan Rusz; and Grant MH CZ–DRO-VFN64165 awarded to Jiri Klempir, Petr Dusek, and Jan Rusz), OP VVV MEYS project "Research Center for Informatics" (Grant CZ.02.1.01/0.0/0.0/16_019/0000765 awarded to Jan Rusz), and Czech Technical University in Prague (Grant SGS20/168/OHK3/3T/13 awarded to Vojtech Illner).

References

Abdelwahab, M., & Busso, C. (2014). Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition. *2014 IEEE Spoken Language Technology Workshop*, pp. 472–477. https://doi.org/10.1109/SLT.2014.7078620

Aharonson, V., Aharonson, E., Raichlin-Levi, K., Sotzianu, A., Amir, O., & Ovadia-Blechman, Z. (2017). A real-time phoneme counting algorithm and application for speech rate monitoring. *Journal of Fluency Disorders, 51,* 60–68. https://doi.org/10.1016/j.jfludis.2017.01.001

Arnaldi, D., Antelmi, E., St. Louis, E., Postuma, R., & Arnulf, I. (2017). Idiopathic REM sleep behavior disorder and neurodegenerative risk: To tell or not to tell to the patient? How to minimize the risk? *Sleep Medicine Reviews, 36,* 82–95. https://doi.org/10.1016/j.smrv.2016.11.002

Ash, S., McMillan, C., Gross, R., Cook, P., Gunawardena, D., Morgan, B., Boller, A., Siderowf, A., & Grossman, M. (2012). Impairments of speech fluency in Lewy body spectrum disorder. *Brain and Language, 120*(3), 290–302. https://doi.org/10.1016/j.bandl.2011.09.004

Avanzi, M., Simon, A., Goldman, J., & Auchlin, A. (2010). C-PROM: An annotated corpus for French prominence study. In *Proceedings of the 5th International Conference on Speech Prosody*.

Berendsen, W. (2015). *Pyphen*. Retrieved 2021-03-12, from https://pyphen.org/

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior & Development, 4,* 247–260. https://doi.org/10.1016/S0163-6383(81)80027-6

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings, 17*(17), 97–110. https://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345. https://www.researchgate.net/publication/208032992_PRAAT_a_system_for_doing_phonetics_by_computer

Camacho, A., & Harris, J. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the*

Acoustical Society of America, 124(3), 1638–1652. https://doi.org/10.1121/1.2951592

Carmichael, J. (2017). Enhancing speech rate estimation techniques to improve dysarthria diagnosis. *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 309–313. https://doi.org/10.1109/IEMCON.2017.8117233

de Jong, N., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

De Looze, C., Kelly, F., Crosby, L., Vourdanou, A., Coen, R., Walsh, C., Lawlor, B., & Reilly, R. (2018). Changes in speech chunking in reading aloud is a marker of mild cognitive impairment and mild-to-moderate Alzheimer's disease. *Current Alzheimer Research, 15*(9), 828–847. https://doi.org/10.2174/1567205015666180404165017

Dekens, T., Demol, M., Verhelst, W., & Verhoeve, P. (2007). A comparative study of speech rate estimation techniques. *Proceedings of Interspeech,* 510–513. https://doi.org/10.21437/Interspeech.2007-237

Dekens, T., Martens, H., Van Nuffelen, G., De Bodt, M., & Verhelst, W. (2014). Speech rate determination by vowel detection on the modulated energy envelope. In *European signal processing conference proceedings*. IEEE.

Delval, A., Rambour, M., Tard, C., Dujardin, K., Devos, D., Bleuse, S., Defebvre, L., & Moreau, C. (2016). Freezing/festination during motor tasks in early-stage Parkinson's disease: A prospective study. *Movement Disorders, 31*(12), 1837–1845. https://doi.org/10.1002/mds.26762

Devos, D., Hirsch, E., & Wyse, R. (2021). Seven solutions for neuroprotection in Parkinson's disease. *Movement Disorders, 36*(2), 306–316. https://doi.org/10.1002/mds.28379

Duffy, J. (2019). *Motor speech disorders: Substrates, differential diagnosis, and management (Fourth)*. Elsevier.

Gilman, S., Wenning, G., Low, P., Brooks, D., Mathias, C., Trojanowski, J., Wood, N., Colosimo, C., Durr, A., Fowler, C., Kaufmann, H., Klockgether, T., Lees, A., Poewe, W., Quinn, N., Revesz, T., Robertson, D., Sandroni, P., Seppi, K., & Vidailhet, M. (2008). Second consensus statement on the diagnosis of multiple system atrophy. *Neurology, 71*(9), 670–676. https://doi.org/10.1212/01.wnl.0000324625.00404.15

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 517–520). IEEE. https://doi.org/10.1109/ICASSP.1992.225858

Goetz, C., Tilley, B., Shaftman, S., Stebbins, G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A., Lees, A., Leurgans, S., LeWitt, P., Nyenhuis, D., . . . LaPelle, N. (2008). Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders, 23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Google Speech-to-Text. (2013). Retrieved 2021-03-12, from https://cloud.google.com/speech-to-text/

Heinrich, C., & Schiel, F. (2011). *Estimating speaking rate by means of rhythmicity parameters*. http://www.isca-speech.org/archive/interspeech_2011/i11_1873.html

Hlavnička, J., Čmejla, R., Tykalová, T., Šonka, K., Růžička, E., & Rusz, J. (2017). Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports, 7*(1), 12. https://doi.org/10.1038/s41598-017-00047-5

Ho, A., Iansek, R., Marigliani, C., Bradshaw, J., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11*(3), 131–137. https://doi.org/10.1155/1999/327643

Högl, B., Stefani, A., & Videnovic, A. (2018). Idiopathic REM sleep behaviour disorder and neurodegeneration—An update. *Nature Reviews Neurology, 14*(1), 40–55. https://doi.org/10.1038/nrneurol.2017.157

Huici, H., Kairuz, H., Martens, H., Van Nuffelen, G., & De Bodt, M. (2016). Speech rate estimation in disordered speech based on spectral landmark detection. *Biomedical Signal Processing and Control, 27,* 1–6. https://doi.org/10.1016/j.bspc.2016.01.005

Illner, V., Sovka, P., & Rusz, J. (2020). Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. *Biomedical Signal Processing and Control,* 58, 101831. https://doi.org/10.1016/j.bspc.2019.101831

Jiao, Y., Berisha, V., Tu, M., & Liss, J. (2015). Convex weighting criteria for speaking rate estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(9), 1421–1430. https://doi.org/10.1109/TASLP.2015.2434213

Jiao, Y., Tu, M., Berisha, V., & Liss, J. (2016). Online speaking rate estimation using recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5245–5249).https://doi.org/10.1109/ICASSP.2016.7472678

Kempler, D., & Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Kent, R. (1996). Hearing and believing. *American Journal of Speech-Language Pathology, 5*(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Kent, R., Weismer, G., Kent, J., Vorperian, H., & Duffy, J. (1999). Acoustic studies of dysarthric speech. *Journal of Communication Disorders, 32*(3), 141–186. https://doi.org/10.1016/S0021-9924(99)00004-0

Konstantopoulos, K., Vogazianos, P., Christou, Y., & Pisinou, M. (2021). Sequential motion rate and oral reading rate: Normative data for Greek and clinical implications. *Logopedics Phoniatrics Vocology,* 1–6. https://doi.org/10.1080/14015439.2021.1901309

Lennes, M. (2009). Segmental features in spontaneous and read-aloud Finnish. In V. de Silva & R. Ullakonoja (Eds.), *Phonetics of Russian and Finnish general introduction: Spontaneous and read-aloud speech* (pp. 145–166). Peter Lang.

Lippus, P., Tuisk, T., Salveste, N., & Teras, P. (2013). *Phonetic corpus of Estonian spontaneous speech*. Institute of Estonian and General Linguistics. https://www.keeletehnoloogia.ee/et/ekt-projektid/eesti-keele-spontaanse-kone-foneetilise-korpuse-arendused

Liu, S. (1996). Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America, 100*(5), 3417–3430. https://doi.org/10.1121/1.416983

Longo, D., Fanciulli, A., & Wenning, G. (2015). Multiple system atrophy. *New England Journal of Medicine, 372*(3), 249–263. https://doi.org/10.1056/NEJMra1311488

Maclay, H., & Osgood, C. (2015). Hesitation phenomena in spontaneous English speech. *WORD, 15*(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

Mannem, R., Jyothi, H., Illa, A., & Ghosh, P. (2020). Speech rate estimation using representations learned from speech with convolutional neural network. In *2020 International Conference on Signal Processing and Communications (SPCOM)* (pp. 1–5). https://doi.org/10.1109/SPCOM50965.2020.9179502

Martens, H., Dekens, T., Van Nuffelen, G., Latacz, L., Verhelst, W., & De Bodt, M. (2015). Automated speech rate measurement in dysarthria. *Journal of Speech, Language, and Hearing Research, 58*(3), 698–712. https://doi.org/10.1044/2015_JSLHR-S-14-0242

McCann, H., Stevens, C., Cartwright, H., & Halliday, G. (2014). α-Synucleinopathy phenotypes. *Parkinsonism & Related Disorders, 20,* S62–S67. https://doi.org/10.1016/S1353-8020(13)70017-8

Mendoza Ramos, V., Kairuz Hernandez-Diaz, H., Hernandez-Diaz Huici, M., Martens, H., Van Nuffelen, G., & De Bodt, M. (2020). Acoustic features to characterize sentence accent production in dysarthric speech. *Biomedical Signal Processing and Control, 57,* 101750. https://doi.org/10.1016/j.bspc.2019.101750

Morgan, N., & Fosler-Lussier, E. (1998). Combining multiple estimators of speaking rate. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No. 98CH36181)* (pp. 729–732). IEEE. https://doi.org/10.1109/ICASSP.1998.675368

Nayak, S., Bhati, S., & Rama Murty, K. (2019). Zero resource speaking rate estimation from change point detection of syllable-like units. In *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6590–6594). IEEE. https://doi.org/10.1109/ICASSP.2019.8683462

Németh, L. (2002). *Hunspell*. Retrieved 2021-03-12, from https://hunspell.github.io/

Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*.

Payan, C., Viallet, F., Landwehrmeyer, B., Bonnet, A., Borg, M., Durif, F., Lacomblez, L., Bloch, F., Verny, M., Fermanian, J., Agid, Y., Ludolph, A., Leigh, P., Bensimon, G., & Guo, M. (2011). Disease severity and progression in progressive supranuclear palsy and multiple system atrophy: Validation of the NNIPPS—PARKINSON PLUS SCALE. *PLOS ONE, 6*(8), Article e22293. https://doi.org/10.1371/journal.pone.0022293

Pfitzinger, H., Burger, S., & Heid, S. (1996). Syllable detection in read and spontaneous speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* (pp. 1261–1264). IEEE. https://doi.org/10.1109/ICSLP.1996.607838

Poewe, W., Seppi, K., Tanner, C., Halliday, G., Brundin, P., Volkmann, J., Schrag, A., & Lang, A. (2017). Parkinson disease. *Nature Reviews Disease Primers, 3*(1), 17013. https://doi.org/10.1038/nrdp.2017.13

Postuma, R., Berg, D., Stern, M., Poewe, W., Olanow, C., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A., Halliday, G., Goetz, C., Gasser, T., Dubois, B., Chan, P., Bloem, B., Adler, C., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders, 30*(12), 1591–1601. https://doi.org/10.1002/mds.26424

Postuma, R., Lang, A., Gagnon, J., Pelletier, A., & Montplaisir, J. (2012). How does Parkinsonism start? Prodromal Parkinsonism motor changes in idiopathic REM sleep behaviour disorder. *Brain, 135*(6), 1860–1870. https://doi.org/10.1093/brain/aws093

Räsänen, O., Doyle, G., & Frank, M. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition, 171,* 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Räsänen, O., Seshadri, S., & Casillas, M. (2018). Comparison of syllabification algorithms and training strategies for robust word count estimation across different languages and recording conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association*. INTERSPEECH. https://doi.org/10.21437/Interspeech.2018-1047

Rusz, J., Bonnet, C., Klempíř, J., Tykalová, T., Baborová, E., Novotný, M., Rulseh, A., & Růžička, E. (2015). Speech disorders reflect differing pathophysiology in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of Neurology, 262*(4), 992–1001. https://doi.org/10.1007/s00415-015-7671-1

Rusz, J., Čmejla, R., Ruzickova, H., & Růžička, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America, 129*(1), 350–367. https://doi.org/10.1121/1.3514381

Rusz, J., Čmejla, R., Tykalová, T., Ruzickova, H., Klempíř, J., Majerova, V., Picmausova, J., Roth, J., & Růžička, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. *The Journal of the Acoustical Society of America, 134*(3), 2171–2181. https://doi.org/10.1121/1.4816541

Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., Gagnon, J., Dušek, P., Galbiati, A., Marelli, S., Timm, P., Teigen, L., Janzen, A., Habibi, M., Stefani, A., Holzknecht, E., Seppi, K., Evangelista, E., Rassu, A., . . . Šonka, K. (2021). Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson Disease. *Annals of Neurology, 90*(1), 62–75. https://doi.org/10.1002/ana.26085

Rusz, J., Hlavnička, J., Tykalová, T., Novotný, M., Dušek, P., Šonka, K., & Růžička, E. (2018). Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(8), 1495–1507. https://doi.org/10.1109/TNSRE.2018.2851787

Sateia, M. (2014). International Classification of Sleep Disorders–Third Edition. *Chest, 146*(5), 1387–1394. https://doi.org/10.1378/chest.14-0970

Schenck, C., Montplaisir, J., Frauscher, B., Högl, B., Gagnon, J., Postuma, R., Šonka, K., Jennum, P., Partinen, M., Arnulf, I., Cochen de Cock, V., Dauvilliers, Y., Luppi, P., Heidbreder, A., Mayer, G., Sixel-Döring, F., Trenkwalder, C., Unger, M., Young, P., . . . Oertel, W. (2013). Rapid eye movement sleep behavior disorder: Devising controlled active treatment studies for symptomatic and neuroprotective therapy—A consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group. *Sleep Medicine, 14*(8), 795–806. https://doi.org/10.1016/j.sleep.2013.02.016

Schwarz, P., & Černocký, J. (2008). *Phoneme recognition based on long temporal context [Doctoral thesis, Brno University of Technology]*.

Seshadri, S., & Räsänen, O. (2019). SylNet: An adaptable end-to-end syllable count estimator for speech. *IEEE Signal Processing Letters, 26*(9), 1359–1363. https://doi.org/10.1109/LSP.2019.2929415

Skodda, S., Rinsche, H., & Schlegel, U. (2009). Progression of dysprosody in Parkinson's disease over time—A longitudinal study. *Movement Disorders, 24*(5), 716–722. https://doi.org/10.1002/mds.22430

Skodda, S., & Schlegel, U. (2008). Speech rate and rhythm in Parkinson's disease. *Movement Disorders, 23*(7), 985–992. https://doi.org/10.1002/mds.21996

Skrabal, D., Tykalová, T., Klempíř, J., Růžička, E., & Rusz, J. (2020). Dysarthria enhancement mechanism under external clear speech instruction in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of Neural Transmission, 127*(6), 905–914. https://doi.org/10.1007/s00702-020-02171-5

SpeechDat-E database. (1999). Retrieved 2021-03-12, from http://www.fee.vutbr.cz/SPEECHDAT-E/

St Louis, E., Boeve, A., & Boeve, B. (2017). REM sleep behavior disorder in Parkinson's disease and other synucleinopathies. *Movement Disorders, 32*(5), 645–658. https://doi.org/10.1002/mds.27018

Sussman, J. E., & Kris, T. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Talkin, D. (1983). A robust algorithm for pitch tracking (RAPT). *Proceedings of ICASSP,* 1352–1355.

van Nuffelen, G., De Bodt, M., Wuyts, F., & Van de Heyning, P. (2009). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatrica et Logopaedica, 61*(2), 69–75. https://doi.org/10.1159/000208805

Villing, R. (2004). Automatic blind syllable segmentation for continuous speech. *Irish Signals and Systems Conference, 2004,* 41–46. https://doi.org/10.1049/cp:20040515

Wang, D., & Narayanan, S. (2007). Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech and Language Processing, 15*(8), 2190–2201. https://doi.org/10.1109/TASL.2007.905178

Yarra, C., Deshmukh, O., & Ghosh, P. (2016). A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection. *Speech Communication, 78,* 62–71. https://doi.org/10.1016/j.specom.2016.01.004

Yarra, C., Nagesh, S., Deshmukh, O., & Kumar Ghosh, P. (2019). Noise robust speech rate estimation using signal-to-noise ratio dependent sub-band selection and peak detection strategy. *The Journal of the Acoustical Society of America, 146*(3), 1615–1628. https://doi.org/10.1121/1.5124473

Yuan, J., & Liberman, M. (2010). Robust speaking rate estimation using broad phonetic class recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4222–4225). IEEE. https://doi.org/10.1109/ICASSP.2010.5495686

Yun, W., Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., Byun, K., Hahn, H., & Kim, J. (2015). The Korean corpus of spontaneous speech. *Phonetics and Speech Sciences, 7*(2), 103–109. https://doi.org/10.13064/KSSS.2015.7.2.103

Zellner, B. (1998). Fast and slow speech rate: A characterisation for French. *Proceedings of the International Conference on Spoken Language Proceedings., Sydney, Australia, 7,* 3159–3163.

Zhang, Y., & Glass, J. (2009). Speech rhythm guided syllable nuclei detection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3797–3800). https://doi.org/10.1109/ICASSP.2009.4960454

2.5 Articulation deficits

The imprecise articulation of vowels is a key factor contributing to reduced speech intelligibility, resembling characteristics often seen in dysarthric speech [63]. This impairment reflects a decrease in both the amplitude and velocity of articulators such as the lips, tongue, and jaw, resulting in what is termed undershooting of articulatory gestures [64]. Vowel articulation abnormalities have been observed across various progressive neurological diseases, particularly in PD [65]. While previous studies have documented these abnormalities, their interpretation and comparability are limited due to small sample sizes and variations in methodologies used for analysis. Nonetheless, these findings suggest that assessing vowel articulation may potentially serve as a surrogate marker for neurodegeneration.

However, most current methods for evaluating vowel articulation in dysarthria patients rely on precise yet time-consuming manual labeling of predefined speech utterances [66]. Only two attempts have been made to automate this analysis using acoustic methods [67], [68]. However, these attempts were constrained by their focus on analyzing predetermined reading sentences from predominantly healthy controls and PD patients with mild hypokinetic dysarthria.

Therefore, a novel, fully-automated method had been designed to analyze vowel articulation impairment from spontaneous speech and assessed in a wide range of synucleinopathies with distinct dysarthria subtypes and severities, including PD and iRBD [69].

A novel method was developed, utilizing phoneme recognition and segmentation, together with formant computation algorithm. The output was combined and refined using outliers detection strategy and k-means clustering. The result were formant frequencies of distinguished corner vowels. The approach was validated in a cohort of 459 of patients with various neurodegenerative disorders and 306 healthy controls. The algorithm reached a resulting accuracy of 77%, based on F-score, a promising result given a large number of etiologies and dysarthria severities involved. A novel, complex features based on the formant ratios of corner vowels were proposed. These features demonstrated imprecise vowel articulation in a broad spectrum of progressive neurodegenerative diseases, in all dysarthria subtypes such as hypokinetic, hyperkinetic, ataxic, spastic, and their mixed variants, including the flaccid–spastic subtype, and was influenced by dysarthria severity.

The study established a proper methodology and confirmed that objectively analyzing vowel articulation using developed measures could offer a universally applicable approach to screening articulation impairment in neurological diseases that affect movement abilities. This method can be applied using everyday speech, without restricting analysis to short, guided speech tasks. The preprint of the article is provided below. The supplementary material to the article is displayed in the Appendix A.

ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Automated Vowel Articulation Analysis in Connected Speech Among Progressive Neurological Diseases, Dysarthria Types, and Dysarthria Severities

Vojtech Illner,[a] Tereza Tykalova,[a] Dominik Skrabal,[b] Jiri Klempir,[b] and Jan Rusz[a,b,c]

[a] Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic [b] Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic [c] Department of Neurology and ARTORG Center, Inselspital, Bern University Hospital, University of Bern, Switzerland

ARTICLE INFO

Article History:
Received September 9, 2022
Revision received January 3, 2023
Accepted April 20, 2023

Editor-in-Chief: Cara E. Stepp
Editor: Jun Wang

https://doi.org/10.1044/2023_JSLHR-22-00526

ABSTRACT

Purpose: Although articulatory impairment represents distinct speech characteristics in most neurological diseases affecting movement, methods allowing automated assessments of articulation deficits from the connected speech are scarce. This study aimed to design a fully automated method for analyzing dysarthria-related vowel articulation impairment and estimate its sensitivity in a broad range of neurological diseases and various types and severities of dysarthria.
Method: Unconstrained monologue and reading passages were acquired from 459 speakers, including 306 healthy controls and 153 neurological patients. The algorithm utilized a formant tracker in combination with a phoneme recognizer and subsequent signal processing analysis.
Results: Articulatory undershoot of vowels was presented in a broad spectrum of progressive neurodegenerative diseases, including Parkinson's disease, progressive supranuclear palsy, multiple-system atrophy, Huntington's disease, essential tremor, cerebellar ataxia, multiple sclerosis, and amyotrophic lateral sclerosis, as well as in related dysarthria subtypes including hypokinetic, hyperkinetic, ataxic, spastic, flaccid, and their mixed variants. Formant ratios showed a higher sensitivity to vowel deficits than vowel space area. First formants of corner vowels were significantly lower for multiple-system atrophy than cerebellar ataxia. Second formants of vowels /a/ and /i/ were lower in ataxic compared to spastic dysarthria. Discriminant analysis showed a classification score of up to 41.0% for disease type, 39.3% for dysarthria type, and 49.2% for dysarthria severity. Algorithm accuracy reached an F-score of 0.77.
Conclusions: Distinctive vowel articulation alterations reflect underlying pathophysiology in neurological diseases. Objective acoustic analysis of vowel articulation has the potential to provide a universal method to screen motor speech disorders.
Supplemental Material: https://doi.org/10.23641/asha.23681529

Imprecise vowels represent one of the core articulatory deficits contributing to reduced intelligibility due to dysarthria (H. Kim, Hasegawa-Johnson, & Perlman, 2011). Impairment of vowel articulation reflects reduced amplitude and velocity of articulators, including lips, tongue, and jaw (the so-called undershooting of articulatory gestures; Robertson & Hammerstad, 1996). Previous studies have documented the presence of vowel articulation abnormalities in a number of progressive neurological diseases (Whitfield, 2019), particularly in Parkinson's disease (PD; Lam & Tjaden, 2016; Skodda et al., 2011; Tjaden et al., 2013; Whitfield & Goberman, 2014; Whitfield & Mehta, 2019) and sporadically in progressive supranuclear palsy (PSP), multiple-system atrophy (MSA), Huntington's disease (HD), essential tremor (ET), cerebellar ataxia (CA), multiple sclerosis (MS), and amyotrophic lateral sclerosis (ALS; Rusz et al., 2014, 2015; Tjaden et al., 2005; Tykalova et al., 2016; Yunusova et al., 2013).

Correspondence to Jan Rusz: rusz.mz@gmail.com. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

In addition, distinctive progressive neurological diseases typically comprehend differing subtypes of dysarthria, with the most prevalent hypokinetic, hyperkinetic, spastic, ataxic, or flaccid variant (Duffy, 2019). These dysarthria subtypes reflect the underlying pathophysiology of the disease and may give us clues for differential diagnosis (Duffy, 2019). In some cases, such as PD, where most patients develop pure hypokinetic dysarthria (Ho et al., 1999), there is good correspondence between the type of disease and type of dysarthria. Contrary, the correspondence might be weaker in other cases as multiple dysarthria subtypes may occur for a single disease type due to more than one component of the motor system being affected. For instance, patients with atypical parkinsonism such as MSA or PSP typically manifest various combinations of hypokinetic, spastic, and ataxic dysarthria components (Rusz et al., 2015). However, previous studies have not addressed whether the vowel articulation impairment is differentially valuable by directly comparing several disease etiologies or dysarthria subtypes. Moreover, the previous evidence is limited due to the small sample sizes available and different methodologies used for analysis (Lam & Tjaden, 2016; Rusz et al., 2014, 2015; Skodda et al., 2011; Tjaden et al., 2005; Tykalova et al., 2016; Whitfield & Goberman, 2014; Whitfield & Mehta, 2019; Yunusova et al., 2013).

Additionally, dysarthria severity varies across neurological diseases depending on their stage and rate of progression (Y. Kim, Kent, & Weismer, 2011). In particular, higher dysarthria severity could be expected in disorders with faster disease progression (Rusz et al., 2015). Nevertheless, there is no standard measure of speech severity in dysarthria. Estimates of speech intelligibility are frequently used to estimate the extent to which neurological disease affects the speech mechanism (Y. Kim, Kent, & Weismer, 2011). Since the relationships between the severity of vowel articulation impairment and the perceptual impression of unintelligibility in dysarthric speakers have been widely documented (H. Kim, Hasegawa-Johnson, & Perlman, 2011; H. M. Liu et al., 2005; Weismer et al., 2001), automated vowel articulation analysis may have a potential to provide such a measure of speech severity in dysarthria. However, there is a lack of relevant vowel articulation studies with a sufficiently large number of dysarthric speakers on various levels of severity.

A reliable and automatic method applicable to natural, spontaneous speech without any cost or burden to the patient or investigator is necessary to facilitate the use of vowel articulation assessment in common clinical practice. The intelligibility and quality of each vowel can be determined particularly by the distinct acoustic energy peak of the first (F_1) and second (F_2) formant frequency. The acoustic–articulatory relationship is defined such that the F_1 frequency varies inversely with tongue height and the F_2 frequency varies directly with tongue advancement (Kent et al., 1999). The limited articulatory range of motion due to dysarthria may result in various shifts in formant frequencies; most typically, formants with naturally higher frequencies tend toward lower frequencies, whereas formants with naturally lower frequencies tend toward higher frequencies (Kent & Kim, 2003; Roy et al., 2009; Shimon et al., 2010). However, most current methods for evaluating vowel articulation via formants in dysarthrias rely on precise and time-consuming hand-labeling of predefined speech utterances (Shimon et al., 2010; Skodda et al., 2011). Only two attempts have been made to evaluate vowel articulation employing automated acoustic analysis (Y. Liu et al., 2021; Sandoval et al., 2013); these were limited by analysis of only predefined reading sentences obtained from a sample predominantly composed of healthy controls (HCs) and PD patients with mild severity of hypokinetic dysarthria.

Therefore, we aimed to design a fully automated method for analyzing vowel articulation impairment due to dysarthria via detecting formant frequencies from corner vowels. Based on this approach and a large sample of patients with various progressive neurological diseases, we quantitatively assessed the sensitivity of imprecise vowel articulation to different (a) types of neurological disease, (b) types of dysarthria, and (c) severity of dysarthria.

Method

Subjects

Each participant provided written informed consent. This study was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic, in accordance with the ethical standards established in the 1964 Declaration of Helsinki.

Between 2011 and 2021, a total of 459 successive native Czech speakers with Central Bohemia accent were recruited for this study. Considering progressive neurodegenerative diseases, 20 patients with PD (10 women, 10 men; de-novo PD examined before antiparkinsonian treatment was started), 15 with PSP (five women, 10 men; 11 with Richardson's syndrome, two with PSP-parkinsonism, and two with PSP-pure akinesia with gait freezing), 20 with MSA (12 women, eight men; 17 with parkinsonian and three with cerebellar variant), 20 with HD (10 women, 10 men), 20 with ET (10 women, 10 men), 18 with CA (eight women, 10 men; 11 with sporadic late-onset CA other than MSA, seven with spinocerebellar ataxia [Type 1, 2, 7, or 8]), 20 with MS (11 women, nine men; 10 with relapsing–remitting MS, five with primary progressive

MS, five with secondary progressive MS), and 20 with ALS (14 women, six men) were recruited (see Table 1). All patients were examined by a neurologist with an experience in movement, demyelinating, or neuromuscular disorders. The diagnosis of PD was established by the Movement Disorders Society clinical diagnostic criteria (Postuma et al., 2015); PSP by the Movement Disorder society diagnostic criteria for PSP (Höglinger et al., 2017); MSA by the consensus diagnostic criteria for MSA (Gilman et al., 2008); HD by clinical and genetic testing (Huntington Study Group; 1996); ET by published clinical research criteria (Louis et al., 2007); CA by genetic testing or results of neurological, neuropsychological, and magnetic resonance imaging testing; MS by the revised McDonald Criteria (Thompson et al., 2018); and ALS according to the El Escorial Criteria from the World Federation of Neurology (Brooks et al., 2000). Additionally, 306 healthy subjects (158 women, 148 men) with a mean age of 59.1 (SD = 13.2, range: 31–87) years with no history of neurological or communication disorders participated as HCs to match the wide age and gender range of investigated neurodegenerative diseases.

Clinical Evaluation

The disease severity of PD was assessed according to the motor score of the Movement Disorders Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III (Goetz et al., 2008), PSP and MSA by The Natural History and Neuroprotection in Parkinson Plus Syndromes–Parkinson Plus Scale (NNIPPS-PPS; Payan et al., 2011), HD by the motor score of the Unified Huntington's Disease Rating Scale (UHDRS; Huntington Study Group, 1996), ET by the Tremor Research Group Essential Tremor Rating Assessment Scale (TETRAS; Elble et al., 2012), CA by the Scale for the Assessment and Rating of Ataxia (SARA; Schmitz-Hübsch et al., 2006), MS by the Expanded Disability Status Scale (EDSS; Kurtzke, 1983), and ALS by the ALS Functional Rating Scale–Revised (ALSFRS-R; Cedarbaum et al., 1999). Disease duration was estimated based on the self-reported occurrence of the first motor symptoms.

Speech Examination

Each subject was recorded during a single session accompanied by a speech specialist who guided through the standardized protocol. No time limits were imposed during the recording. All participants were willing to cooperate and could repeat their performance if necessary. The participants were instructed to present a monologue about an arbitrary, emotionally neutral topic for at least 90 s (M = 128.3, SD = 27.5, range: 74–312). In addition, all subjects performed a reading passage task of a standardized text of 80 words (Supplemental Material S1). The same settings were applied to subjects in all groups. Speech recordings were performed in a quiet room with a low ambient noise level using a head-mounted condenser microphone (Beyerdynamic Opus 55) placed approximately 5 cm from the subject's mouth. Speech signals were sampled at 48 kHz with a 16-bit resolution.

Auditory–Perceptual Estimates of Dysarthria Presence, Type, and Severity

The dysarthria presence and type, including severity, were made by the consensus auditory–perceptual judgment of two speech-language pathologists with more than 10 years of experience in movement disorders who were aware of each patient's medical diagnosis. The judgment was based on offline audio recordings following the perceptual criteria outlined by Darley et al. (1969a, 1969b). The dysarthria types identified across eight neurological conditions included hypokinetic, hyperkinetic, ataxic, spastic, flaccid–spastic, spastic–ataxic, hypokinetic–spastic, hypokinetic–ataxic, and hypokinetic–spastic–ataxic (see Table 1). In addition, the severity of dysarthria was rated on a 4-point scale (0 = *none*, 1 = *mild*, 2 = *moderate*, 3 = *severe*). The lower average dysarthria severity with a dominant occurrence of mild dysarthria was observed only for PD and MS groups (see Table 1). Potential participants without the presence of perceptual severity of dysarthria, with the presence of language disorders or apraxia of speech, or with a speech dysfunction not related to the diagnosed neurological disorder were excluded from this study.

Automatic Algorithm for Vowel Articulation Features

The algorithm utilizes a formant tracker in combination with a phoneme recognizer and subsequent signal processing analysis (see Figure 1). It processes the connected speech utterance for reading passages and monologues separately, and estimates F_1 and F_2 formant values for each corner vowel /a/, /i/, and /u/. These corner vowels are essential to form a vowel triangle (i.e., triangular F_1–F_2 vowel space), which reflects extreme placements of the tongue. (H. Kim, Hasegawa-Johnson, & Perlman, 2011; Rusz et al., 2013; Skodda et al., 2011).

Formants and Phonemes Estimation (Step A)

The speech input was processed in parallel by a formant tracker and a phoneme recognizer (see Figure 1A). Burg algorithm (Childers, 1978) implementation in Praat (Boersma, 2001) was used for the first two formants contour estimation resulting in $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ vectors over the utterance. After the trial testing, the window length was

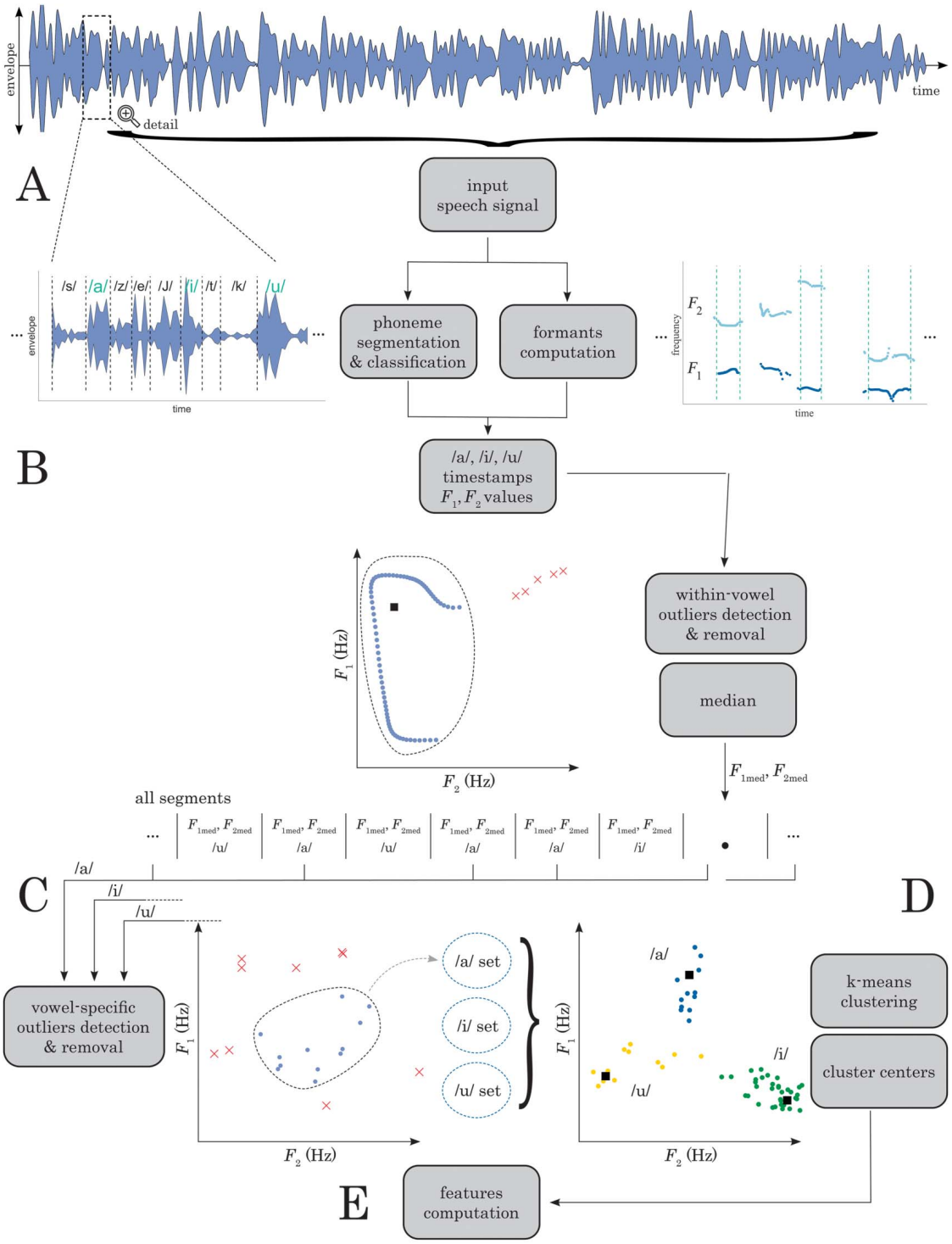
Table 1. Clinical characteristics of the investigated subjects.

Disease	Sex	Motor score (disease severity) *M*/*SD* (range)	Age (years) *M*/*SD* (range)	Symptom duration (years) *M*/*SD* (range)	Dysarthria type (auditory–perceptual)	Dysarthria severity (auditory–perceptual)
PD	F = 10	38.7/14.7[a]	63.5/8.9	1.6/1.3	Hypokinetic (*n* = 20)	Mild (*n* = 13)
	M = 10	(18–70)	(42–79)	(0.3–5.9)		Moderate (*n* = 7) Severe (*n* = 0) Mean severity: 1.35[h]
PSP	F = 5	65.7/28.9[b]	66.0/5.1	4.7/2.7	Hypokinetic (*n* = 3)	Mild (*n* = 3)
	M = 10	(19–132)	(54–71)	(2.0–11.0)	Hypokinetic–spastic (*n* = 4)	Moderate (*n* = 5)
					Hypokinetic–ataxic (*n* = 3)	Severe (*n* = 7) Mean severity: 2.27
					Hypokinetic–spastic–ataxic (*n* = 5)	
MSA	F = 12	79.1/21.1[b]	62.0/7.0	4.4/1.8	Hypokinetic (*n* = 3)	Mild (*n* = 1)
	M = 8	(35–115)	(45–73)	(2.0–7.5)	Spastic–ataxic (*n* = 1)	Moderate (*n* = 12)
					Hypokinetic–spastic (*n* = 8)	Severe (*n* = 7)
					Hypokinetic–ataxic (*n* = 3)	Mean severity: 2.30
					Hypokinetic–spastic–ataxic (*n* = 5)	
HD	F = 10	24.8/9.9[c]	53.1/11.0	5.2/3.6	Hyperkinetic (*n* = 20)	Mild (*n* = 1)
	M = 10	(8–42)	(34–69)	(1.0–16.0)		Moderate (*n* = 13)
						Severe (*n* = 6) Mean severity: 2.25
ET	F = 10	17.5/7.6[d]	64.3/11.1	28.9/17.5	Hyperkinetic (*n* = 18)	Mild (*n* = 5)
	M = 10	(6–35)	(40–82)	(3.0–60.0)	Hypokinetic (*n* = 1)	Moderate (*n* = 9)
					Spastic (*n* = 1)	Severe (*n* = 6) Mean severity: 2.05
CA	F = 8	13.9/4.8[e]	54.7/12.6	11.0/8.5	Ataxic (*n* = 5)	Mild (*n* = 5)
	M = 10	(4–24)	(34–72)	(0.5–28.0)	Spastic (*n* = 1)	Moderate (*n* = 7)
					Spastic–ataxic (*n* = 11)	Severe (*n* = 6) Mean severity: 2.06
					Hypokinetic–ataxic (*n* = 1)	
MS	F = 11	4.6/0.8[f]	52.2/10.1	17.8/8.6	Ataxic (*n* = 7)	Mild (*n* = 16)
	M = 9	(4–7)	(33–74)	(6.0–32.0)	Spastic (*n* = 3)	Moderate (*n* = 3)
					Spastic–ataxic (*n* = 10)	Severe (*n* = 1) Mean severity: 1.25[h]
ALS	F = 14	35.6/6.5[g]	62.1/11.1	1.9/1.2	Spastic (*n* = 4)	Mild (*n* = 5)
	M = 6	(22–45)	(37–85)	(0.5–5.0)	Flaccid–spastic (*n* = 16)	Moderate (*n* = 6)
						Severe (*n* = 9) Mean severity: 2.20
Total	F = 80		59.7/9.6	9.4/5.7	Hypokinetic (*n* = 27)	Mild (*n* = 49)
	M = 73		(33–85)	(0.3–60.0)	Hyperkinetic (*n* = 38)	Moderate (*n* = 62)
					Ataxic (*n* = 12)	Severe (*n* = 42)
					Spastic (*n* = 9)	Mean severity: 2.09
					Flaccid–spastic (*n* = 16)	
					Spastic–ataxic (*n* = 22)	
					Hypokinetic–spastic (*n* = 12)	
					Hypokinetic–ataxic (*n* = 7)	
					Hypokinetic–spastic–ataxic (*n* = 10)	

Note. PD = Parkinson's disease; F = female; M = male; PSP = progressive supranuclear palsy; MSA = multiple system atrophy; HD = Huntington's disease; ET = essential tremor; CA = cerebellar ataxia; MS = multiple sclerosis; ALS = amyotrophic lateral sclerosis; MDS-UPDRS = Movement Disorders Society–Unified Parkinson's Disease Rating Scale; NNIPPS-PPS = Natural History and Neuroprotection in Parkinson Plus Syndromes–Parkinson Plus Scale; UHDRS = Unified Huntington's Disease Rating Scale; TETRAS = Tremor Research Group Essential Tremor Rating Assessment Scale; SARA = Scale for the Assessment and Rating of Ataxia; EDSS = Expanded Disability Status Scale; ALSFRS-R = Amyotrophic Lateral Sclerosis Functional Rating Scale–Revised.

[a]MDS-UPDRS Part III total scale. [b]NNIPPS-PPS total scale. [c]UHDRS total scale. [d]TETRAS score scale. [e]SARA total scale. [f]EDSS total scale. [g]ALSFRS-R total scale. [h]This group was found to have significantly lower disease severity compared to PSP, MSA, HD, ET, CA, and ALS groups with $p < .01$.

Figure 1. Illustrative schema of the automated method for formants estimation. F_1 = first formant frequency; F_2 = second formant frequency; VSA = vowel space area; FRI = formant ratio index; SFRI = second formant ratio index.



$$F_{1a}, F_{2a}, F_{1i}, F_{2i}, F_{1u}, F_{2u}$$

$$\text{VSA} = \tfrac{1}{2}|F_{1i}(F_{2a} - F_{2u}) + F_{1a}(F_{2u} - F_{2i}) + F_{1u}(F_{2i} - F_{2a})|$$

$$\text{FRI} = \frac{F_{1a}+F_{1i}+F_{1u}+F_{2a}+F_{2i}}{F_{2u}} \qquad \text{SFRI} = \frac{F_{2a}+F_{2i}}{F_{2u}}$$

set to 50 ms with a 1-ms overlap and the formant ceiling was set to 5250 Hz, as these values achieved the best precision of the estimates. The maximum number of formants was set to 5 as recommended by the method documentation, even though only the first two formants were subsequently extracted. A phoneme recognizer was employed based on split temporal context feature extraction (Schwarz & Černocký, 2008), pretrained on the Czech version of the SpeechDat-E database (Pollak et al., 2000). The recognizer is available pretrained for several languages with an error rate of 24.2% for both Czech and English (Schwarz et al., 2022). The recordings were subsampled to 8 kHz beforehand to match the training data. The output is represented by recognized phonemes with timestamps marking the corresponding speech segment.

Outlier Detection Across Individual Phonemes Segments (Step B)

The consecutive phoneme segments were further analyzed (see Figure 1B). If the frame was classified as corresponding to either /a/, /i/, or /u/ vowel, the $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ values within were extracted. These might be burdened with formant tracker errors, and thus, an outlier analysis is performed in each segment. Outliers were identified and discarded based on Mahalanobis distance (Mahalanobis, 1936), which calculates the distance of a given point from a chosen distribution. For normally distributed data, the squared distance follows χ^2 distribution. The procedure consists of two phases and is as follows.

First, normalized versions of each formant vector were computed by extracting the mean and dividing by the standard deviation. Then, the Mahalanobis distance was computed between each point on the normalized [$\boldsymbol{F}_1$, $\boldsymbol{F}_2$] grid and χ^2 distribution with two degrees of freedom since we have two formant contours. If the distance was greater then $\chi^2(q)$, where q is a chosen quantile value, it was marked as an outlier. In the second phase, the nonoutlier points formed a new distribution, and Mahalanobis distance was calculated between previously identified outlier points. If the distance was less than $\chi^2(q + 0.1)$, the corresponding point was withdrawn from the outliers set. After the conducted trial testing, the value of q was set to 0.8, making the procedure more benevolent in the outlier decision. It achieved higher effectiveness yet not suffering a decrease in accuracy than choosing harsher settings, that is, lower values of q.

The procedure ensures that the extreme outliers are correctly recognized while preserving most of the information around the formant contour. From each segment, a median value was computed from the first and second formants of the nonnormalized, nonoutlier points resulting in $\boldsymbol{F}_{1\text{med}}$ and $\boldsymbol{F}_{2\text{med}}$ vectors over the whole utterance with information about the particular vowel on each index.

Outlier Detection Across All /A/, /I/, and /U/ Vowels (Step C)

The medians from the segments might still contain false values, for example, when the phoneme recognizer misclassifies a consonant as a vowel. Therefore, the $\boldsymbol{F}_{1\text{med}}$ and $\boldsymbol{F}_{2\text{med}}$ values are grouped to either /a/, /i/, or /u/ set, and another outlier analysis was performed in each group (see Figure 1C).

The procedure is the same as described in the previous section; however, the value of the quantile q is set to 0.5, making the method less benevolent to any deviations, which was found to provide more accurate outcomes while maintaining a reasonable throughput. The nonoutliers for each vowel were then put together for final cluster analysis.

Vowels Clustering (Step D)

The described method is still prone to error when the phoneme recognizer misclassifies the vowel as another, for example, /u/ as /i/. The misclassified vowel might have the formant frequencies close to the original one and thus will not be detected in the outlier analysis.

For this reason, the vowel points were partitioned using the k-means algorithm into three clusters representing the single vowels (see Figure 1D). The distance metric was set to square Euclidean distance, and initial cluster centroid positions were chosen as the maximum value of F_1 and median of F_2 of the vowel /a/ (hence, cluster /a/), the minimum of F_1 and maximum of F_2 of the vowel /i/ (hence, cluster /i/), and the minimum of F_1 and the minimum F_2 of the vowel /u/ (hence, cluster /u/). The resulting clusters /a/, /i/, and /u/ then consisted of [$\boldsymbol{F}_{1\text{a}}$, $\boldsymbol{F}_{2\text{a}}$], [$\boldsymbol{F}_{1i}$, $\boldsymbol{F}_{2i}$], and [$\boldsymbol{F}_{1u}$, $\boldsymbol{F}_{2u}$] points, respectively. The misclassified vowel should be included in its corresponding cluster in this process.

In the final step, one pair of $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ values was calculated from the points of each cluster. For the /a/ cluster, $\boldsymbol{F}_1$ was calculated as an upper (0.75) quantile of the $\boldsymbol{F}_{1\text{a}}$ values and F_2 as the median of the $\boldsymbol{F}_{2\text{a}}$ values. For the /i/ cluster, F_1 was computed as a lower (0.25) quantile of the $\boldsymbol{F}_{1i}$ values and F_2 as an upper quantile of the $\boldsymbol{F}_{2i}$ values. For the /u/ cluster, F_1 and F_2 were selected as lower quantiles of $\boldsymbol{F}_{1u}$ and $\boldsymbol{F}_{2u}$ values, respectively. The choice of the particular quantiles was designed to reflect the corner vowel characteristics (Y. Liu et al., 2021), and the values were tuned in pretesting to achieve maximum estimates precision.

Vowel Articulation Features (Step E)

The outcome of the process is the pair of F_1 and F_2 values for each vowel from which the vowel articulation features were derived (see Figure 1E). Subsequently, the most commonly used features that represent complex

vowel articulation characteristics are vowel space area (VSA) and measures representing various shifts in formant frequencies (Kent & Kim, 2003; Roy et al., 2009; Shimon et al., 2010; Skodda et al., 2011). VSA, expressed in Hz^2, was calculated using the Euclidean distances between the F_1 and F_2 coordinates of the corner vowels /a/, /i/, and /u/ in the triangular [F_1, F_2] vowel space as

$$\mathrm{VSA} = \frac{1}{2}|F_{1i}(F_{2a} - F_{2u}) + F_{1a}(F_{2u} - F_{2i}) + F_{1u}(F_{2i} - F_{2a})|. \tag{1}$$

Formant ratio index (FRI) reflects the shift in formant frequencies based on all corner vowels and can be expressed using the following formula (i.e., expected trend is lowering of F_{1a}, F_{1i}, F_{1u}, F_{2a}, and F_{2i} and rising of F_{2u} due to the presence of dysarthria):

$$\mathrm{FRI} = \frac{F_{1a} + F_{1i} + F_{1u} + F_{2a} + F_{2i}}{F_{2u}}. \tag{2}$$

Finally, the second formant ratio index (SFRI) reflects the shift of the second formants only and was computed using the following formula (i.e., expected trend is lowering of F_{2a} and F_{2i} and rising of F_{2u} due to the presence of dysarthria)

$$\mathrm{SFRI} = \frac{F_{2a} + F_{2i}}{F_{2u}}. \tag{3}$$

All analyses were conducted in MATLAB (MathWorks).

Reference Hand Labels

The hand-labeled reference values of F_1 and F_2 formant frequencies and time event of the vowel occurrence for each corner vowel were obtained from 20 randomly selected recordings of the reading passage (1,760 vowels; 660 vowels of /a/, 720 vowels of /i/, and 380 vowels of /u/) with the representative distribution regarding gender, etiology, and dysarthria severity (11 men and nine women; four HC, two PD, two PSP, two MSA, two HD, two ET, two CA, two MS, and two ALS speakers; four *none*, seven *mild*, five *moderate*, and four *severe dysarthria* severity). All corner vowels of /a/, /i/, and /u/ were selected; the position of the selected vowel for the reading passage is in bold in Supplemental Material S1. Formants were extracted according to widely accepted previously published methodology validated in several languages (Roy et al., 2009; Rusz et al., 2013; Shimon et al., 2010; Skodda et al., 2011); F_1 and F_2 frequencies were determined by employing a 30-ms segment at the temporal midpoint of the stable part of each vowel (in order to avoid the influence of vowels preceding or following). The corresponding timestamps including the start and end times of the segment were recorded. The formant frequencies were not possible to extract in 23 cases of /a/, 43 of /i/, and 77 of /u/ due to (a) coarticulation with other phonemes leading to the indistinct formants in the target band (68%), (b) coarticulation with other phonemes leading to too many formants in the target band (10%), (c) the word with target vowel is not pronounced properly (16%), and (d) the vowel duration is shorter than 30 ms (6%). All analyses were performed in the Praat software (Boersma, 2001) using both the combined wideband spectrographic display and the power spectral density.

Algorithm Performance Metrics

F-score was used as the primary outcome to assess the algorithm accuracy and was defined as

$$F = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}, \tag{4}$$

where

$$\mathrm{precision} = \frac{\mathrm{true\ positives}}{\mathrm{true\ positives} + \mathrm{false\ positives}}, \tag{5}$$

and

$$\mathrm{recall} = \frac{\mathrm{true\ positives}}{\mathrm{true\ positives} + \mathrm{false\ negatives}}. \tag{6}$$

In other cases, normalized root-mean-square error (NRMSE) and the Spearman correlation coefficient r were utilized. The NRMSE enables a description of several variables by describing the error as a fraction of the observed variable range and is defined as

$$\mathrm{NRSME} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_l - x_i)^2}}{(\max(\hat{x}_l) - \min(\hat{x}_l))}, \tag{7}$$

where N is the number of utterances, $\hat{x}_i$ represents an estimated feature, and x_i is its respective reference. The r coefficient was computed as a nonparametric measure of the rank correlation between the estimated and reference values.

Algorithm Validation Steps

The algorithm incorporates several steps in its procedure. In each step, it can make a different type of error. Therefore, to uncover potential error sources for each step separately, a three-step validation that corresponds to the steps in algorithm design is provided: (a) validation of vowel identification via phoneme recognizer and independent parallel validation of formant values estimation via

formant tracker (the result of the algorithm's Step A), (b) validation of combined accuracy via phoneme recognizer and formant tracker based on outlier detection and vowel clustering (the result of the algorithm's mutual Steps B–D), and (c) validation of algorithm total accuracy via resulting formant features (the result of the algorithm's Step E).

First, the performances of the phoneme recognizer and the formant tracker (the result of the algorithm's Step A) were compared to reference hand labels. The validation was performed across the corner vowels of /a/, /i/, and /u/. The vowel from automatic recognition was searched for within the 30-ms segment corresponding to manual timestamps with a 5-ms tolerance for the start and the end. To evaluate the reliability of phoneme recognizer, the accuracy was evaluated in terms of F-score. *True positive* cases were if the vowel was correctly detected (e.g., /a/ was detected as /a/). *False positive* cases were if the vowel was incorrectly detected (e.g., /a/ was detected as /i/). *False negative* case was if the vowel was not detected (e.g., /a/ was missed). To evaluate the reliability of the formant tracker, a median formant frequency was calculated from automatically obtained formant estimates via a 30-ms window corresponding to the start and end of the hand-label timestamps. These medians were compared to reference hand values in terms of NRMSE and Spearman correlation.

Second, the accuracy of the combination of phoneme recognizer and formant tracker (the result of the algorithm's mutual Steps B–D) was validated using F-score; this evaluation corresponds with the mutual outlier detection and vowel class correction mechanism performed by the algorithm. *True positive* cases were if (a) the vowel was correctly detected (e.g., /a/ was detected as /a/), (b) the vowel was detected as another vowel but automatically corrected (e.g., /a/ was detected as /i/ but corrected back to /a/), and (c) formants were found impossible to estimate by both hand-labeling and automated detection (e.g., formants of /a/ were impossible to determine by hand labels and the automated algorithm was not able to estimate them as well). *False positive* cases were if (a) the vowel was incorrectly detected and not corrected (e.g., /a/ was detected as /i/ and not corrected), (b) the vowel was correctly detected but incorrectly reclassified to a different vowel (e.g., /a/ was detected as /a/ but corrected to /i/), and (c) vowel formants were found impossible to estimate by hand-labeling but were still calculated automatically (e.g., formants of /a/ were impossible to determine by hand labels but automated algorithm produced an estimate). *False negative* case was if the vowel was not detected (e.g., /a/ was missed).

Third, the final averaged formant estimates (i.e., one F_{1a}, F_{2a}, F_{1i}, F_{2i}, F_{1u}, and F_{2u} value per subject/speaking task), as well as complex formant features (i.e., one VSA, FRI, and SFRI value per subject/speaking task) by both automated (the result of the algorithm's step E) and manual analysis were compared using NRMSE and Spearman correlation.

Statistical Analysis

Data extracted from reading passages and monologues were analyzed separately; data related to monologues are presented within the article, whereas data for reading passages can be found in Supplemental Material S1. Data normality was verified via the Shapiro–Wilcoxon and Bartlett (equality of variance) tests. One-way analysis of covariance with post hoc Fisher's least significant difference test was applied to evaluate group differences. All analyses were controlled for age and sex (covariates); intergroup differences among diseases and dysarthria types were in addition controlled for dysarthria severity.

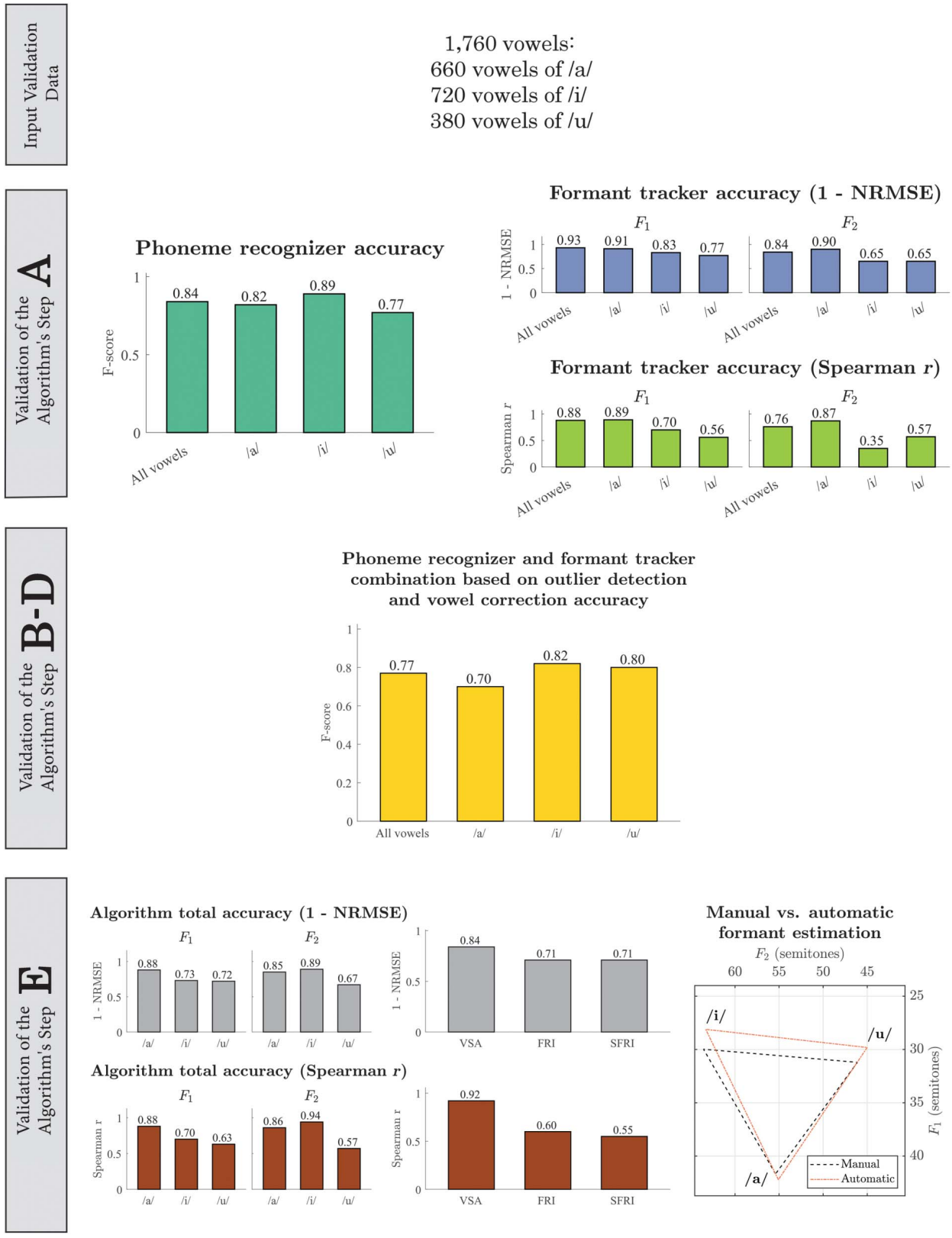
Prompted by primary hypothesis results, we performed a classification experiment based on the discriminant analysis followed by a leave-one-out cross-validation scheme to assess whether the vowel articulation features are best suited to differ between (a) type of neurological disease, (b) type of dysarthria, or (c) severity of dysarthria. In addition, to identify the probability of correct factor identification by chance, we generated a random vector of values ranging from 0 to 100 to substitute vowel articulation features across 459 hypothetical speakers; the average performance was calculated across 100 repetitions.

Results

Algorithm Performance

Compared to manual hand labels (based on 1,760 vowels), the phoneme recognizer attained an F-score of 0.84, whereas the formant tracker achieved 1-NRMSE of 0.93 for F_1 and 0.84 for F_2 across all vowels (see Figure 2, the results of the algorithm's Step A). After combining the error rate of the phoneme recognizer and formant tracker (based on 1,760 vowels), the F-score for all vowels was 0.77 (see Figure 2, the results of the algorithm's Steps B–D). Concerning the final averaged vowel articulation features (based on 20 utterances), the estimation of individual formants achieved 1-NRMSE of 0.88 for F_{1a}, 0.85 for F_{2a}, 0.73 for F_{1i}, 0.89 for F_{2i}, 0.72 for F_{1u}, and 0.67 for F_{2u}, leading to the 1-NRMSE of 0.84 for VSA, 0.71 for FRI, and 0.71 for SFRI (see Figure 2, the results of the algorithm's Step E). In summary, considering the final shape of vowel areas (see Figure 2, VSA plots), the most notable difference between automated and manual labels is due to

Figure 2. Illustrative scheme depicting step-by-step performance results between automated and manual analysis based on 1,760 hand-labeled vowels. NRMSE = normalized root-mean-square error; F_1 = first formant frequency; F_2 = second formant frequency; VSA = vowel space area; FRI = formant ratio index; SFRI = second formant ratio index.



lower estimates of F_1 frequencies of vowel /i/ and /u/ and F_2 of /u/ by the automated approach.

Effect of Neurological Disease Type

Compared to controls, the change in vowel articulation due to neurodegeneration in monologues was primarily demonstrated by trends toward the shift of formants across vowels /i/ and /u/, including an increase in F_{2u} and decrease in F_{1i}, F_{1u}, and F_{2i} frequencies across PD, PSP, MSA, HD, and ALS (see Figure 3 and Table 2). Among diseases, MSA tended to decrease F_1 and CA tended to increase F_1 compared to other neurological diseases, leading to a significantly lower F_1 for MSA than CA across all corner vowels (see Figure 4). Considering complex formant measures, compared to controls, VSA was significantly decreased for MSA, whereas FRI and SFRI were decreased for all neurological diseases except ET and MS (see Figure 5).

Effect of Dysarthria Type

Compared to controls, the trends toward the shift of formants across vowels /i/ and /u/ including increase in F_{2u} and decrease in F_{1i}, $F_{1u,}$ and F_{2i} frequencies in monologues were demonstrated mainly for hypokinetic and hyperkinetic dysarthria, mixed dysarthrias involving hypokinetic components, and flaccid–spastic subtype (see Figure 6 and Table 3). Among dysarthrias, there was a particular difference between ataxic dysarthria manifested by the decrease of F_{1a}, F_{2a}, and F_{2i} compared to spastic dysarthria (and its mixed variants with ataxic and flaccid elements) and in addition by a trend toward increase of F_{1u} to hypokinetic dysarthria (see Figure 7). Additionally, spastic–ataxic dysarthria showed a trend toward increase of F_{1a}, F_{1i}, and F_{1u} compared to hypokinetic dysarthria (and its mixed variants with spastic elements).

Considering complex formant measures, compared to controls, VSA was significantly decreased for ataxic and hypokinetic–spastic dysarthria (see Figure 8). FRI was decreased for hypokinetic, hyperkinetic, ataxic, flaccid–spastic, spastic–ataxic, hypokinetic–spastic, and hypokinetic–spastic–ataxic dysarthria. Finally, SFRI was decreased for hypokinetic, hyperkinetic, ataxic, flaccid–spastic, spastic–ataxic, and hypokinetic–spastic. Among dysarthrias, VSA of ataxic dysarthria was significantly lower than in spastic or spastic–ataxic dysarthria. FRI and SFRI of hypokinetic–spastic dysarthria were lower compared to hyperkinetic, flaccid–spastic, spastic–ataxic, and hypokinetic–ataxic dysarthria.

Effect of Dysarthria Severity

Compared to controls, the shift of formants across vowels /i/ and /u/ in dependence on auditory–perceptual dysarthria severity in monologues was observed, including an increase in F_{2u} and a decrease in F_{1i}, F_{1u}, and F_{2i} frequencies (see Figure 9 and Table 4). Considering complex formant measures, both measures of FRI and SFRI were reduced across all dysarthria severities (see Figure 10).

Classification Analysis

The classification analysis among vowel articulation features in monologues manifested accuracy of up to 39.7% for disease type, 37.3% for dysarthria type, and 49.2% for dysarthria severity (see Table 5); the probability of correct factor identification by chance using a random vector showed 5.3% accuracy for disease type, 4.2% for dysarthria type, and 19.8% for dysarthria severity. Acoustic metrics reflecting the shift in formant frequencies of FRI and SFRI were more sensitive to capturing the change of vowel articulation than VSA.

Effect of Speaking Task Type

The trends toward decrements in complex measures of VSA, FRI, and SFRI in reading passages were demonstrated similarly to those observed in monologues, except for the PD group where imprecise vowels articulation was not affected in reading passages (Supplemental Material S1); the classification experiment showed similar accuracy of up to 41.0% for disease type, 39.3% for dysarthria type, and 47.4% for dysarthria severity.

Discussion

This study is the first to demonstrate a fully automated objective approach to assessing the quality of vowel articulation in a large cohort of 459 speakers, including controls and patients with various neurological diseases and different types and severity of dysarthria, using the natural, unconstrained speech recordings. Based on complex formant measures, we showed that imprecise vowel articulation was presented in a broad spectrum of progressive neurodegenerative diseases, including PD, PSP, MSA, HD, ET, CA, MS, and ALS. Similarly, vowel articulation impairment was presented in all dysarthria subtypes such as hypokinetic, hyperkinetic, ataxic, spastic, and their mixed variants, including the flaccid–spastic subtype. In addition, the extent of vowel articulation impairment was influenced by dysarthria severity. However, we still observed divergent patterns of vowel articulation abnormalities across certain etiologies and dysarthria types independent of dysarthria severity. F_1 of all corner vowels were significantly lower for MSA than CA. In addition, F_2 of vowel /a/ and /i/ was lower in ataxic compared to

Figure 3. Corner vowel production triangles estimated from monologues for individual neurological disease types compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < .001$, $p < .01$, and $p < .05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency; F_2 = second formant frequency; PD = Parkinson's disease; PSP = progressive supranuclear palsy; MSA = multiple system atrophy; HD = Huntington's disease; ET = essential tremor; CA = cerebellar ataxia; MS = multiple sclerosis; ALS = amyotrophic lateral sclerosis.

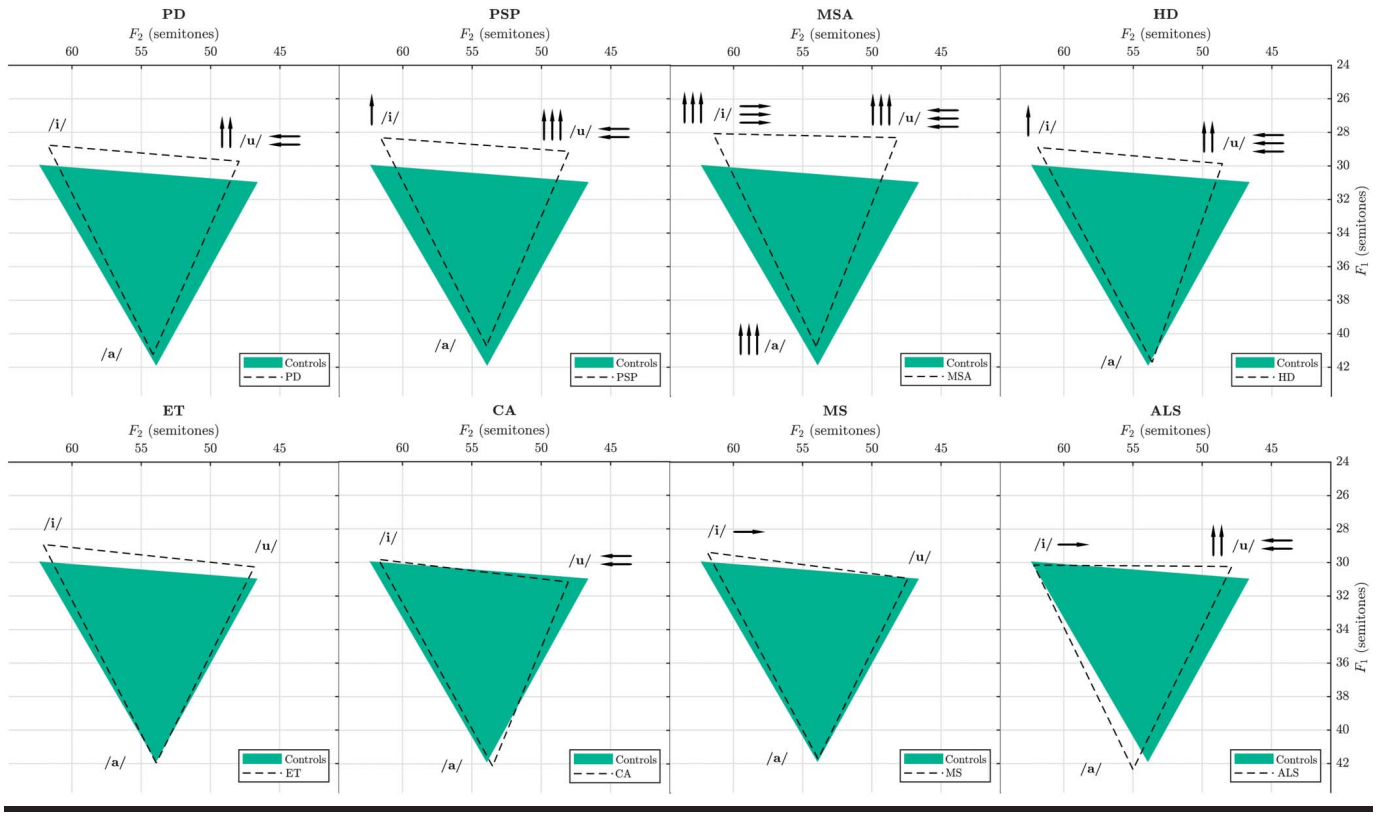
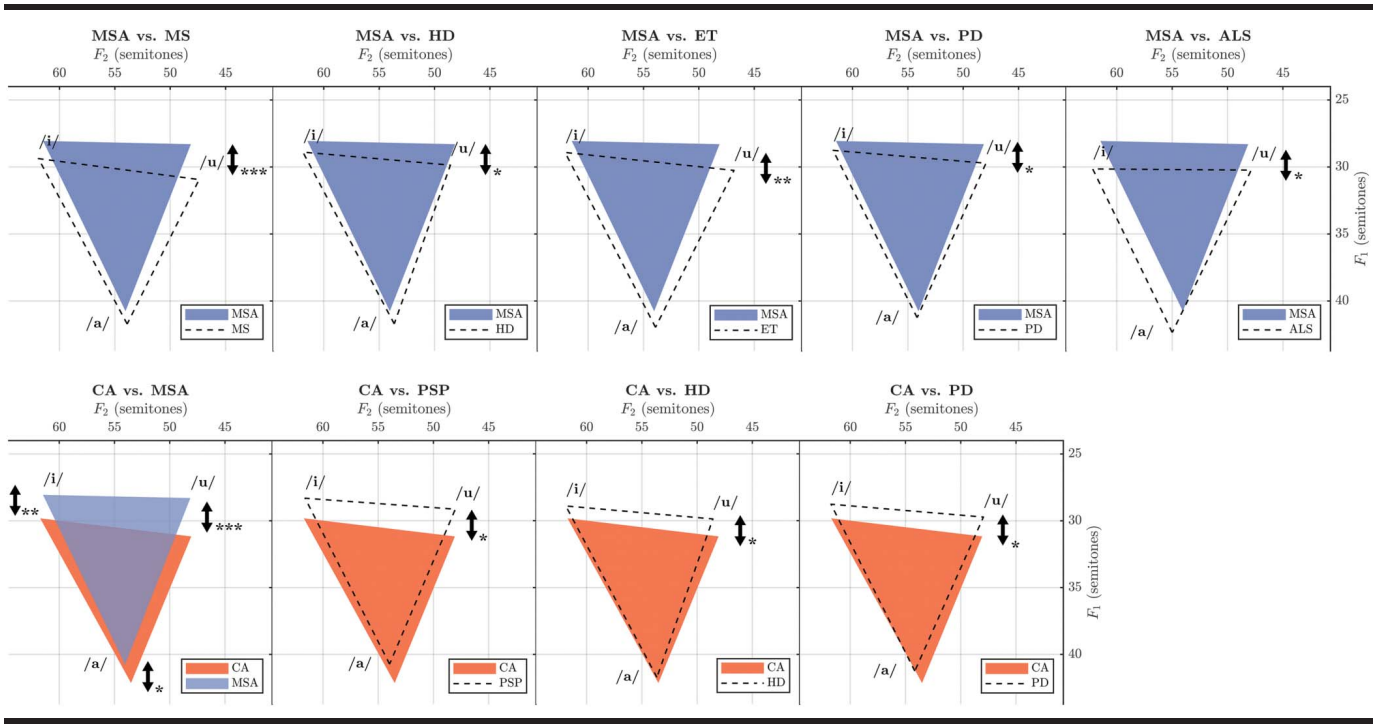


Table 2. Formant frequencies of corner vowels estimated from monologues for individual neurological disease types compared to healthy controls.

Neurological disease type	/a/ *M* (*SD*) Semitones		/i/ *M* (*SD*) Semitones		/u/ *M* (*SD*) Semitones	
	F_1	F_2	F_1	F_2	F_1	F_2
Controls	41.91 (2.6)	53.92 (2.1)	29.93 (2.3)	62.37 (1.6)	30.96 (1.6)	46.60 (1.8)
PD	41.25 (2.9)	54.16 (2.9)	28.76 (2.9)	61.75 (1.6)	29.72 (2.5)	47.95 (1.7)
PSP	40.72 (4.0)	53.97 (1.5)	28.32 (2.6)	61.61 (1.8)	29.13 (2.3)	48.02 (2.3)
MSA	40.76 (3.1)	54.03 (2.0)	28.06 (2.2)	61.48 (1.7)	28.30 (2.5)	48.13 (2.8)
HD	41.71 (3.2)	53.64 (2.0)	28.89 (2.6)	61.88 (1.7)	29.86 (2.3)	48.57 (2.6)
ET	41.97 (3.5)	53.91 (1.7)	28.92 (2.5)	62.08 (1.7)	30.26 (1.7)	46.83 (1.9)
CA	42.14 (3.1)	53.50 (1.7)	29.82 (1.6)	61.72 (1.5)	31.17 (1.6)	48.05 (1.9)
MS	41.72 (2.9)	53.88 (1.8)	29.38 (2.5)	61.89 (1.5)	30.94 (1.7)	47.39 (2.0)
ALS	42.35 (3.8)	54.99 (1.2)	30.15 (2.2)	62.17 (1.4)	30.24 (2.0)	47.86 (2.2)

Note. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). Hertz to semitone formula: f(semitone) = 12*((log*f(Hz)/60)/log(2)). F_1 = first formant frequency; F_2 = second formant frequency; PD = Parkinson's disease; PSP = progressive supranuclear palsy; MSA = multiple system atrophy; HD = Huntington's disease; ET = essential tremor; CA = cerebellar ataxia; MS = multiple sclerosis; ALS = amyotrophic lateral sclerosis.

Figure 4. Corner vowel production triangles estimated from monologues across two pairs of neurological disease types. The double-headed arrows indicate significant differences across diseases adjusted by age, sex, and dysarthria severity with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency; F_2 = second formant frequency; MSA = multiple system atrophy; MS = multiple sclerosis; HD = Huntington's disease; ET = essential tremor; PD = Parkinson's disease; ALS = amyotrophic lateral sclerosis; CA = cerebellar ataxia; PSP = progressive supranuclear palsy.

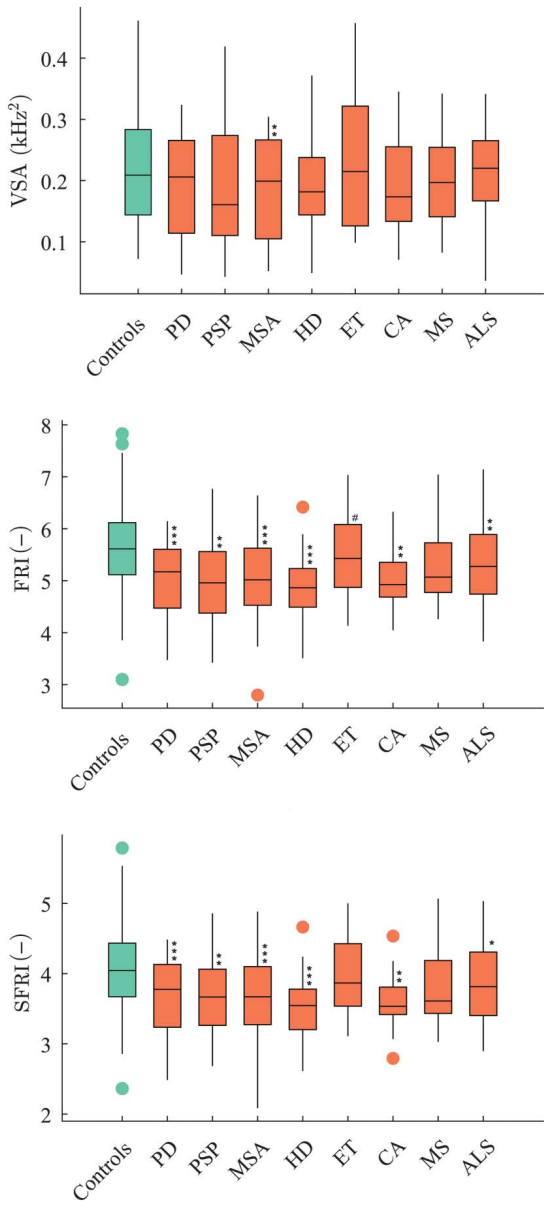


spastic dysarthria. Therefore, objective analysis of vowel articulation has the potential to provide a universally applicable method to screen neurological diseases affecting movement abilities that can be obtained from everyday speech without any cost or burden to the patient and investigator. In the future, vowel articulation deficits could be analyzed via smartphone calls (Kouba et al., 2022), thus significantly aiding in improving innovative neuroprotective therapies' stratification and monitoring effect.

Effect of Neurological Disease Type

Our results confirmed that vowel articulation impairment is exhibited in multiple types of neurological diseases. This finding follows previous acoustic studies (Rusz et al., 2014, 2015; Tjaden et al., 2005; Tykalova et al., 2016; Yunusova et al., 2013), although these have described vowel articulation pertaining to a specific disease rather than comparing these characteristics across diseases. In fact, the presence of vowel articulation deficits across various neurological diseases is not surprising because articulatory impairments represent the most common and distinct characteristics of most dysarthrias (Darley et al., 1969b). However, in this study, certain disease-specific patterns of imprecise vowel articulation have been observed. In general, vowel articulation impairment appeared to be more pronounced in parkinsonian disorders and HD. We might thus assume that the greater extent of vowel articulatory deficits due to tongue movement restriction, reflected mainly by the decrease of F_{2i} and the increase of F_{2u}, is associated with bradykinesia, which represents a common motor sign not only in parkinsonism but also in HD (Reilmann, 2019). Indeed, the previous study on a rat PD model has shown that even unilateral deficits to the nigrostriatal dopamine system leading to bradykinesia substantially contribute to tongue movement restriction responsible for imprecise vowel articulation (Ciucci et al., 2011). Interestingly, the parallel decrease in F_1 of all corner vowels was able to statistically separate MSA from CA even after adjustment for dysarthria severity, presumably as a consequence of damage to basal ganglia structures in addition to cerebellar dysfunction that is typical in both diseases. This finding might have important clinical implications as the differentiation of the cerebellar variant of MSA from idiopathic late-onset CA early in the disease course remains a major diagnostic challenge (Lin et al., 2016). However, although the extent of articulatory disorder appeared to be similar for both MSA subtypes (Rusz et al., 2019), the

Figure 5. Statistically significant group differences for estimated articulation features in monologues among the different types of neurological disease types compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$, respectively. # indicates significant differences to MSA ($p < .05$) after adjusting for age, sex, and dysarthria severity. Middle bars represent median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. VSA = vowel space area; PD = Parkinson's disease; PSP = progressive supranuclear palsy; MSA = multiple system atrophy; HD = Huntington's disease; ET = essential tremor; CA = cerebellar ataxia; MS = multiple sclerosis; ALS = amyotrophic lateral sclerosis; FRI = formant ratio index; SFRI = second formant ratio index.



utility of vowel articulation analysis as such a potential diagnostic marker has to be verified in future studies as the current sample was composed dominantly of the parkinsonian variant of MSA.

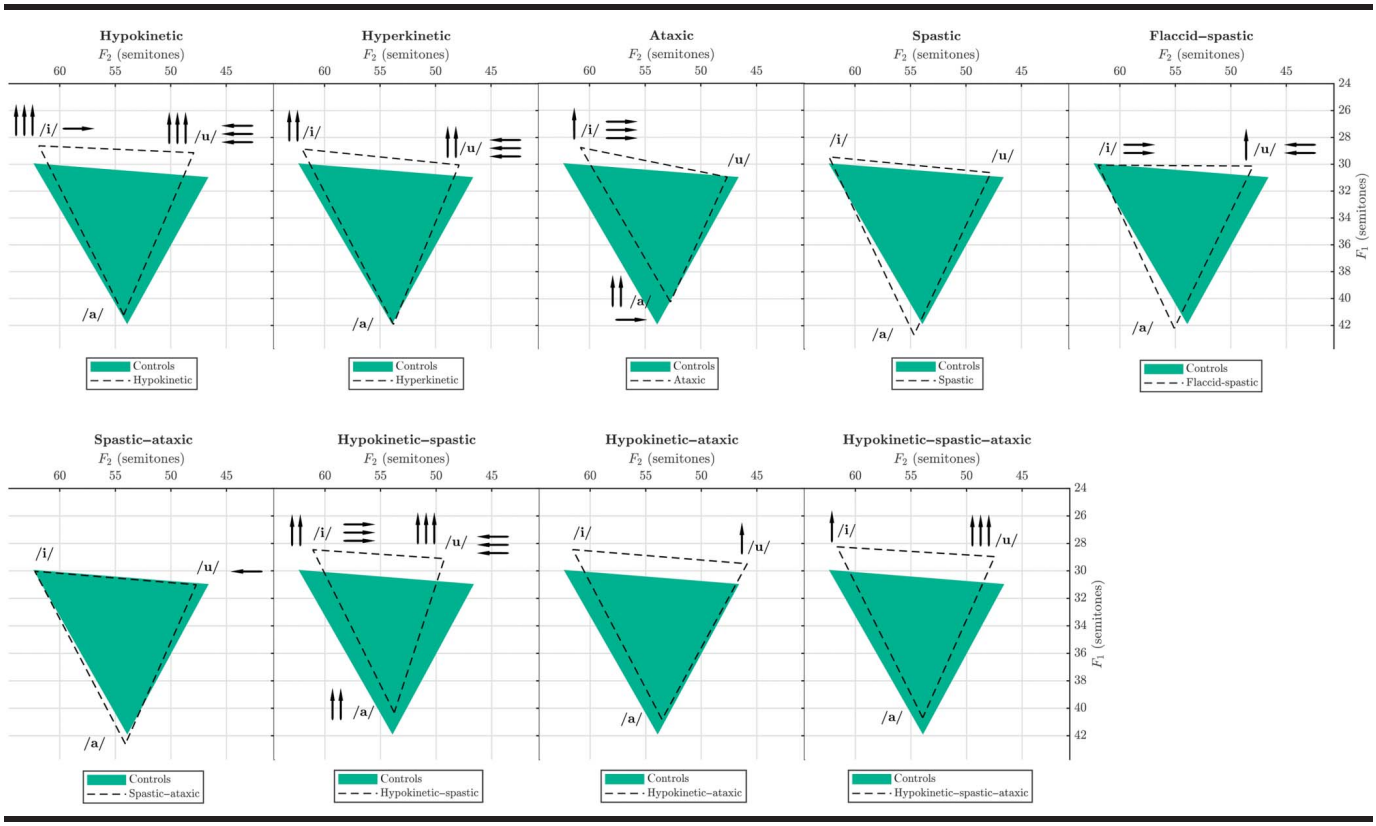
Effect of Dysarthria Type

In line with findings across multiple types of neurological diseases, vowel articulation impairment was observed across various dysarthria types. This result is not surprising as dysarthria type is frequently linked with disease type, and vowel articulation impairment was found in all investigated etiologies. The shift in formants for vowels /i/ and /u/ showed strong similarities for all investigated dysarthrias. This finding follows previous research demonstrating that complex formant-based measures are not sensitive to distinguishing between dysarthria subtypes (Lansford & Liss, 2014). However, one potential phenomenon that might be helpful in the differential diagnosis of dysarthrias is a shift in vowel frequencies for vowel /a/. While both formants of vowel /a/ remained relatively unchanged in patients with hypokinetic or hyperkinetic dysarthria, they tend to be decreased in ataxic dysarthria and increased in spastic dysarthria (as well as in mixed dysarthrias involving spastic elements). However, this finding should be interpreted with caution due to the relatively low number of samples for pure spastic and ataxic speakers in this study. Although the studies on vowel articulation in spastic and ataxic dysarthrias are rare, the shift toward higher vowel /a/ formants in spastic dysarthria seems to align with previous research on patients with poststroke spastic dysarthria (Ge et al., 2021; Mou et al., 2018). This shift might be hypothesized as a consequence of spasticity or weakness of tongue muscles, leading to lower tongue advancement. In addition, a decrease of F_2 for vowel /a/ has been previously reported in patients with spinocerebellar ataxia (Skodda et al., 2014), which might be hypothesized to be a result of inconsistency over the range of tongue movement (Saigusa et al., 2006).

Effect of Dysarthria Severity

Our findings showed that auditory–perceptual dysarthria severity was another factor contributing to the extent of vowel articulation impairment. The result agrees with previous research demonstrating a strong relationship between vowel formant measures and perceptual ratings of dysarthria severity (Fletcher et al., 2017). Further support comes from a recent study that showed a progressive pattern of vowel articulation impairment from the prodromal stages of parkinsonism (Skrabal et al., 2022). Compared to VSA, formant indexes were more effective in capturing dysarthria severity, which follows the previous study showing that vowel articulation index based on changes in individual formants was more stable and reliable over repeated assessments compared to VSA (Caverlé & Vogel, 2020). The effectiveness of formant indexes in contrast to the low sensitivity of VSA suggests that

Figure 6. Corner vowel production triangles estimated from monologues for different dysarthria types compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < .001$, $p < .01$, and $p < .05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency; F_2 = second formant frequency.



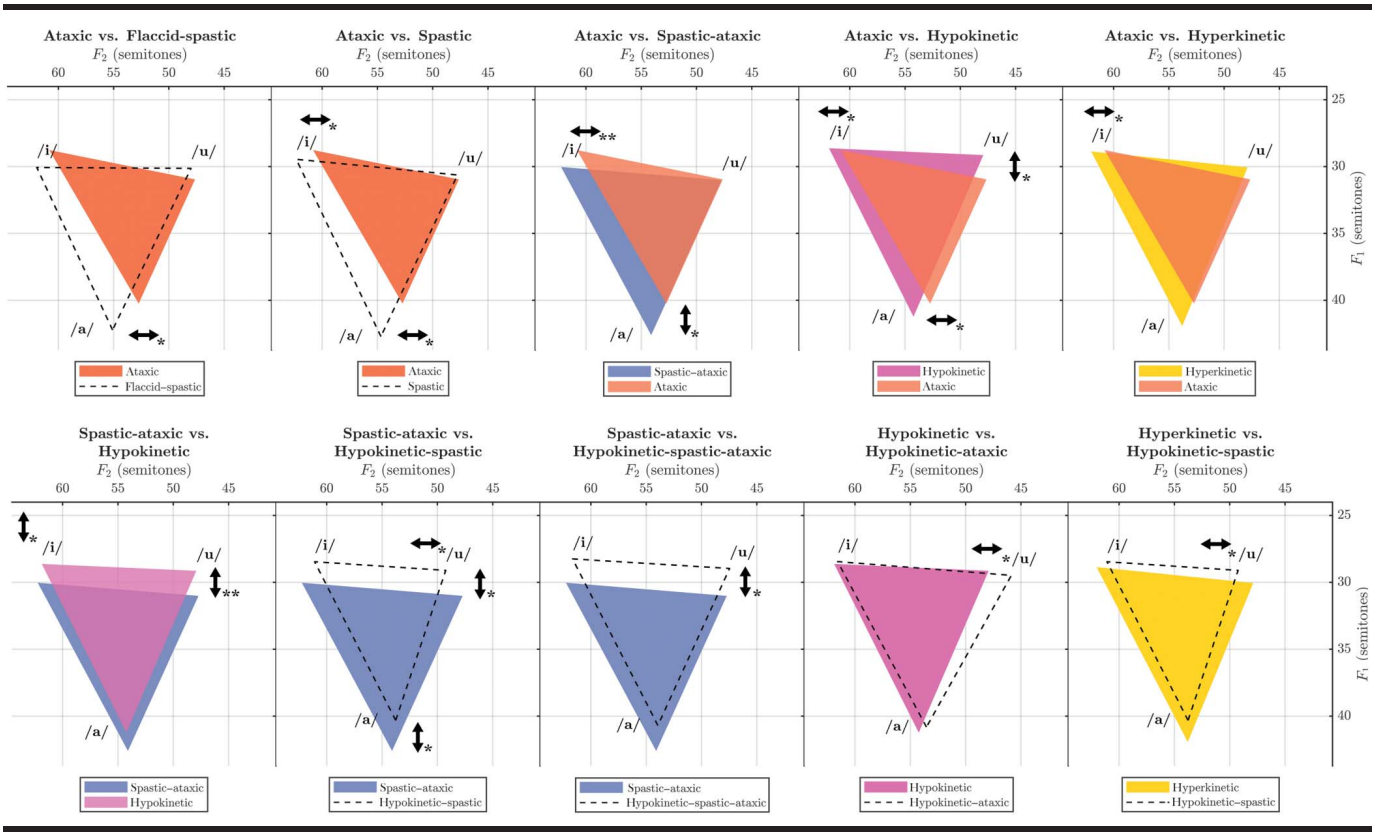
articulatory deficits are due mainly to alterations of the vowel /u/, followed by the vowel /i/, with the vowel /a/ remaining most resistant to change due to dysarthria. This behavior might be a result of different tongue positions and lip posture during individual corner vowels production, where the tongue is positioned low for the vowel /a/, high and forward for the vowel /i/, and high and backward for the vowel /u/, whereas lip posture is spread for

Table 3. Formant frequencies of corner vowels estimated from monologues for different dysarthria types compared to healthy controls.

	/a/ *M* (*SD*) Semitones		/i/ *M* (*SD*) Semitones		/u/ *M* (*SD*) Semitones	
Dysarthria type	F_1	F_2	F_1	F_2	F_1	F_2
Controls	41.91 (2.6)	53.92 (2.1)	29.93 (2.3)	62.37 (1.6)	30.96 (1.6)	46.60 (1.8)
Hypokinetic	41.23 (2.9)	54.24 (2.7)	28.63 (2.8)	61.87 (1.5)	29.14 (2.7)	47.92 (2.3)
Hyperkinetic	41.93 (3.3)	53.79 (1.9)	28.87 (2.6)	62.02 (1.7)	30.05 (2.0)	47.88 (2.3)
Ataxic	40.24 (2.9)	52.73 (2.0)	28.76 (2.2)	60.81 (1.3)	30.96 (1.8)	47.64 (1.6)
Spastic	42.71 (3.5)	54.68 (1.1)	29.45 (2.3)	62.34 (1.7)	30.63 (1.4)	47.75 (2.8)
Flaccid–spastic	42.22 (4.0)	55.09 (1.3)	30.06 (2.3)	61.95 (1.2)	30.13 (2.1)	48.00 (2.3)
Spastic–ataxic	42.60 (2.4)	54.10 (1.4)	30.03 (2.3)	62.24 (1.4)	31.01 (1.9)	47.7 (1.9)
Hypokinetic–spastic	40.32 (3.8)	53.77 (2.0)	28.46 (2.0)	61.09 (1.7)	29.11 (2.8)	49.25 (2.8)
Hypokinetic–ataxic	40.82 (3.1)	53.54 (2.2)	28.45 (1.7)	61.57 (1.7)	29.48 (1.5)	45.87 (1.6)
Hypokinetic–spastic–ataxic	40.70 (3.9)	53.94 (1.5)	28.25 (2.6)	61.65 (1.9)	28.96 (2.1)	47.46 (2.1)

Note. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). Hertz to semitone formula: f(semitone) = 12*((log*f(Hz)/60)/log(2)). F_1 = first formant frequency; F_2 = second formant frequency.

Figure 7. Corner vowel production triangles estimated from monologues across two pairs of dysarthria types. The double-headed arrows indicate significant differences across dysarthria types adjusted by age, sex, and dysarthria severity with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency; F_2 = second formant frequency.



both the /a/ and /i/ vowels and rounded for the vowel /u/ (Hasegawa-Johnson et al., 2003). Therefore, we might assume that the production of the vowel /a/ is less demanding than the production of the vowels /i/ and /u/. Moreover, in comparison to the vowel /i/, the articulation of the vowel /u/ requires more challenging involvement of the orofacial muscles to produce and maintain a tightly rounded lip posture (Hasegawa-Johnson et al., 2003), and its restrictions might also be linked to swallowing deficits in dysarthria (Sapir et al., 2008; Tjaden, 2008).

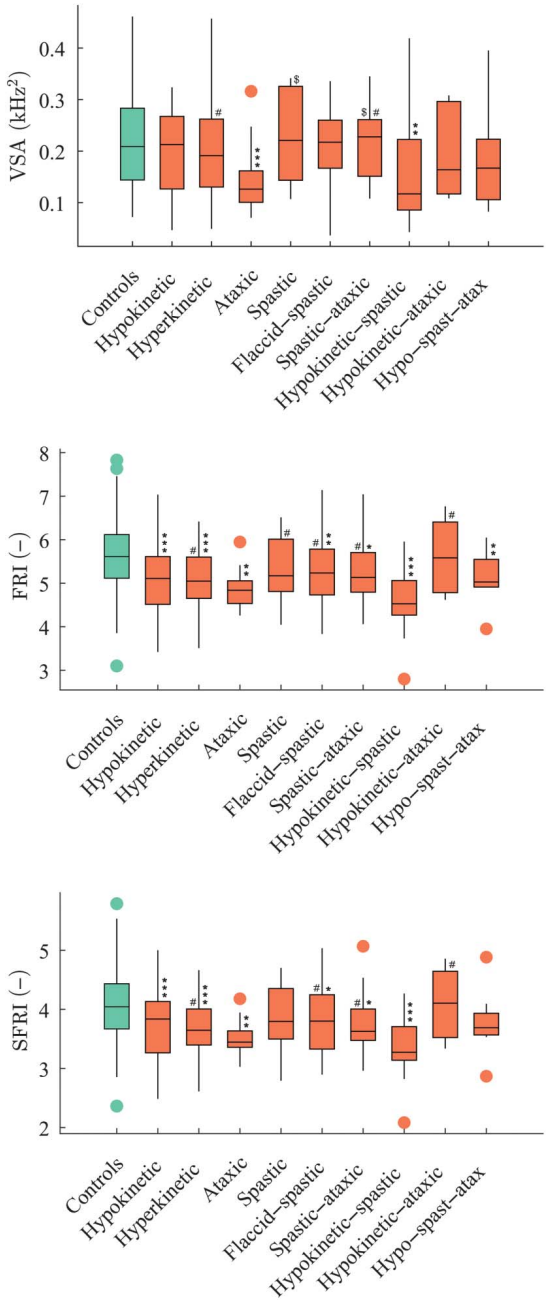
Effect of Speaking Task Type

The results showed that both monologue and reading speech are appropriate for assessing articulation deficits in neurological diseases with similar sensitivity. One notable difference was that only reading passages showed a significant difference between dysarthria severities, although the classification accuracy for dysarthria severity across both tasks was similar. We may thus hypothesize that standardized reading passages might be a better speaking material if capturing speech progression via vowel articulation is the primary endpoint. Considering individual diseases, the only evidence available is from PD, where previous studies have reported the occurrence of a more notable alteration of vowel articulation performance in spontaneous speech compared to nonspontaneous speech (Kempler & Lancker, 2002; Rusz et al., 2013; Weismer, 1984). Indeed, PD was the only group in this study that showed considerably better performance in vowel articulation in reading than monologues. In fact, persons with PD are often highly intelligible in prepared utterances but significantly less intelligible in spontaneous speech, whereas persons with other types of neuromotor disease might be equally intelligible in both forms of utterance (Y. Kim, Kent, & Weismer, 2011). Therefore, this finding might have important implications for future clinical trials in which PD participants should be assessed via spontaneous speech if vowel articulation represents an outcome measure.

Which of the Factors Most Contributes to the Vowel Articulation Impairment?

One of this study's goals was to answer whether vowel articulation impairment is most sensitive to disease

Figure 8. Statistically significant group differences for estimated articulation features in monologues among the different dysarthria types compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$, respectively. # indicates significant differences to hypokinetic–spastic dysarthria ($p < .05$), whereas $ indicates significant differences to ataxic dysarthria ($p < .05$) after adjusting for age, sex, and dysarthria severity. Middle bars represent median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. VSA = vowel space area; FRI = formant ratio index; SFRI = second formant ratio index; Hypo-spast-atax = Hypokinetic–spastic–ataxic dysarthria.

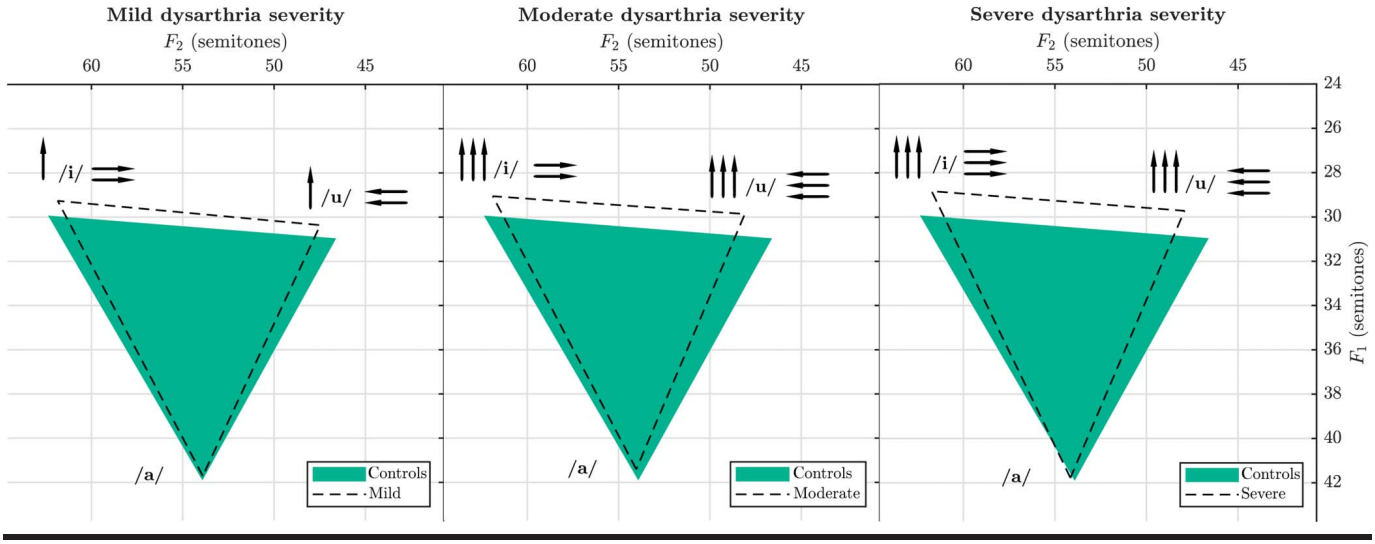


type, dysarthria type, or dysarthria severity. The discriminant analysis classification showed a score of up to 41.0% for the type of neurological disease, 39.3% for dysarthria type, and 49.2% for dysarthria severity. One might thus assume that vowel impairment appears to be more distinctive to dysarthria severity compared to a specific diagnosis of disease or dysarthria subtype. However, these results need to be put in context with the probability of correct factor identification by chance, which showed 5.3% accuracy for disease type, 4.2% for dysarthria type, and 19.8% for dysarthria severity (i.e., approximately equal to the number of groups across each investigated factor). Bearing this in mind, the best ratio between correct classification and identification by chance could be obtained for dysarthria type, although none of the three factors gained superior classification performance. Despite some differences observed in this study that might contribute to the differential diagnosis of dysarthria or disease etiology, we may assume that imprecise vowels represent a universal sign of articulatory disorder showing severity-related variations within several different types of dysarthria. This finding is perhaps not surprising as acoustic similarity across etiologies and types of dysarthria has already been assumed not only for vowel space but also for other acoustic measures such as speaking rate or voice onset time (Weismer, 2006). Indeed, accumulating evidence supports the view that various neuropathologies might similarly affect neuromotor control of speech production, leading to similar manifestations for certain speech aspects at the acoustic surface (Y. Kim, Kent, & Weismer, 2011). On the other hand, the combination of vowel articulation characteristics with other distinct cues that are pathognomic for a specific type of dysarthria, such as strained-strangled voice, slow rate, and reduced loudness variability in spastic dysarthria or normal rate and excessive loudness variability in ataxic dysarthria, might considerably increase correct classification to dysarthria type or disease etiology.

Algorithm Performance

Although articulatory deficits represent the main speech impairment characteristic of most dysarthrias, automated methods for assessing articulatory deficits from connected speech are scarce. In this study, we provided a fully automated approach to assessing the "undershoot of vowels" applicable across various neurological diseases, different dysarthrias, and a wide range of severity, from healthy speech to severe dysarthria. In particular, there are two main sources of errors including incorrect phoneme recognition (16% error based on 1–F-score) and incorrect formant tracking (7% error for F_1 and 16% error for F_2 based on NRMSE). However, the combination of both these error sources leads to an even lower accuracy of the algorithm. Therefore, to provide reliable vowel

Figure 9. Corner vowel production triangles estimated from monologues for different dysarthria severities compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < .001$, $p < .01$, and $p < .05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency; F_2 = second formant frequency.



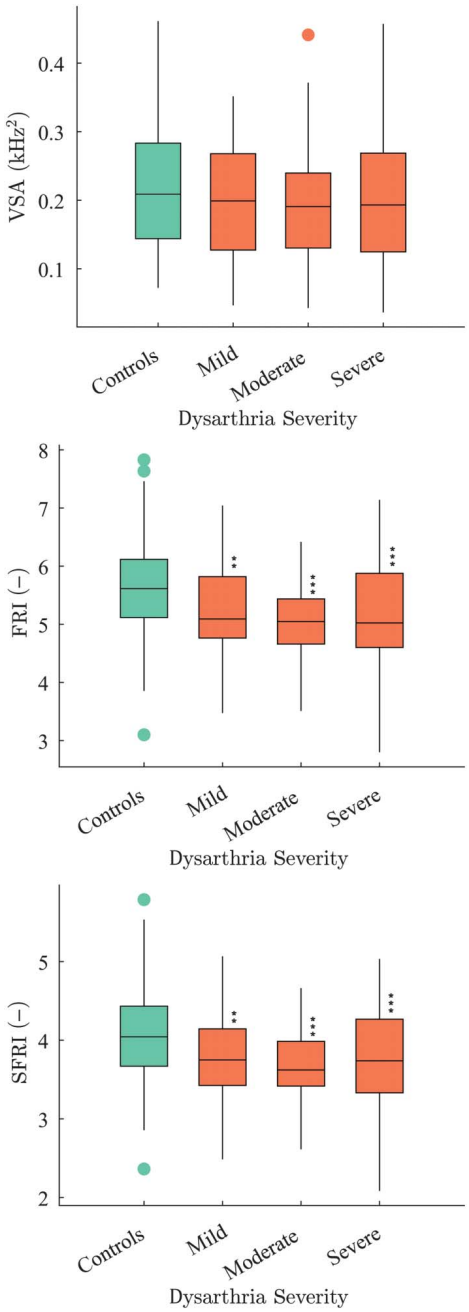
articulation metrics, our algorithm involved multiple levels of error correction such as outlier exclusion and correction of vowel identification by clustering. As a result, we reached the resulting accuracy of 77% (i.e., 23% error based on 1–F-score), which we believe is a very promising accuracy given a large number of etiologies and dysarthria severities involved. In addition, there are limitations in the accuracy of available technologies for phoneme recognition and formant tracking, even for healthy speech. For instance, it might be assumed that the solution toward better accuracy would be to change the formant tracker, yet all available open-source formant trackers were found to have similar detection performance (Schiel & Zitzelsberger, 2018). Considering the shape of the resulting vowel areas, the most considerable discrepancy between automated and manual labels was for F_1 estimation across vowels /i/ and /u/. Whereas the automated method tended to capture lower F_1 of vowels /i/ and /u/ with increasing dysarthria severity, the hand-labeled method did not find any change in F_1 or even increased F_1 due to dysarthria (Roy et al., 2009; Rusz et al., 2013; Skodda et al., 2011). Therefore, in comparison to the vowel articulation index that is most widely used in the literature (Roy et al., 2009; Rusz et al., 2013; Skodda et al., 2011), we proposed an alternative FRI that reflects the dysarthria-related lowering of F_1 in vowels /i/ and /u/ as captured by automated method. The inconsistency between manual and automated labels might be caused by an incidental formant tracker confusion of F_1 as the fundamental frequency and its harmonics close to F_1 of vowel /i/ and /u/. The SFRI, based only on F_2 values showed similar classification accuracy to detect neurological disease type or dysarthria type and even slightly better

Table 4. Formant frequencies of corner vowels estimated from monologues for different dysarthria severities compared to healthy controls.

Dysarthria severity	/a/ *M* (*SD*) Semitones		/i/ *M* (*SD*) Semitones		/u/ *M* (*SD*) Semitones	
	F_1	F_2	F_1	F_2	F_1	F_2
Controls	41.91 (2.6)	53.92 (2.1)	29.93 (2.3)	62.37 (1.6)	30.96 (1.6)	46.60 (1.8)
Mild	41.71 (2.9)	53.90 (1.6)	29.26 (2.0)	61.84 (1.6)	30.36 (1.6)	47.47 (2.2)
Moderate	41.41 (3.2)	54.03 (2.3)	29.05 (2.6)	61.89 (1.7)	29.86 (2.5)	48.09 (1.9)
Severe	41.78 (3.9)	54.15 (1.8)	28.84 (2.8)	61.73 (1.5)	29.72 (2.4)	47.93 (2.6)

Note. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). Hertz to semitone formula: f(semitone) = 12*((log*f(Hz)/60)/log(2)). F_1 = first formant frequency; F_2 = second formant frequency.

Figure 10. Statistically significant group differences for estimated articulation features in monologues among the different dysarthria severities compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < .001$, $p < .01$, and $p < .05$, respectively. Middle bars represent median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. VSA = vowel space area; FRI = formant ratio index; SFRI = second formant ratio index.



accuracy to detect dysarthria severity compared to FRI based on both F_1 and F_2, suggesting SFRI as a suitable alternative to measure vowel articulation deficits. Thus, impairment of vowel articulation in neurodegenerative diseases can be tracked solely by changes in F_2 frequencies that are related to particular deficits in frontward/backward tongue movements. However, automated method achieved an increased inaccuracy in F_2 estimation of vowel /u/, with tendency to capture lower values compared to hand-labeling. Since the observed effects of etiology, dysarthria subtype, and severity were largely reflected by shifts in both formants of vowel /u/, we cannot exclude that these changes could be partially attributed to artifacts related to inaccurate formant tracking rather than actual disease effects. On the other hand, we believe that the automated method's error bias is not specific for etiology or dysarthria subtype and is generally the same for dysarthric and healthy speech, therefore not significantly accounting for the group differences.

Limitations of This Study

This study has certain limitations. Eight groups of patients were selected to cover a wide range of the common movement disorders associated with different pathophysiology responsible for the occurrence of vowel articulation impairment. Several of these etiologies showed different types of mixed dysarthria, leading to a smaller sample size for specific dysarthria subtypes. Additional better-sampled investigations with participants having different disease types, and possibly different dysarthria types, are required to confirm and further extend our findings. The study is based solely on the Czech language; thus, the language independence of the applied methods should be verified in future studies. Nonetheless, a recent multilanguage trial in PD revealed broadly similar profiles of dysarthria across multiple languages (Rusz et al., 2021). In addition, the formant tracker utilizing Burg's algorithm is considered language independent. The phoneme recognizer used in this study can be easily substituted for a universal recognizer that supports most of the world's languages (Li et al., 2020; Y. Liu et al., 2021). Subsequently, it is noteworthy to point out that shifts in formant frequencies and reductions in vowel space might occur due to other conditions than dysarthria such as differing dialect (Williams & Escudero, 2014), behavioral accent (Kamiloğlu et al., 2020), or stuttering-like behavior (Blomgren et al., 1998). We strived to minimize these effects by investigating subjects of the same dialect via an emotionally neutral context of monologue. From the etiologies investigated, the stuttering-like behavior is common only in PSP and very rare in de-novo PD (Rusz et al., 2015; Tykalová et al., 2015). However, the severity of vowel articulation impairment in PSP was not principally different from MSA, which is also atypical parkinsonism without the occurrence of dysfluency but with a similar dysarthria type and severity (Rusz et al., 2015), suggesting that affected vowels are mainly a consequence of dysarthria itself. Finally, our

Table 5. Classification analysis for the formant features for monologues.

%	VSA	FCI	SFCI	Random vector
Neurological disease type	5.0	**39.7**	38.8	5.3
Dysarthria type	5.0	37.0	**37.3**	4.2
Dysarthria severity	39.9	46.8	**49.2**	19.8

Note. The numbers indicate the percentage of subjects correctly identified by the discriminant analysis as original groups. Bold numbers indicate the best accuracy across neurological disease type, dysarthria type, and dysarthria severity. Random vector refers to the experimental results regarding probability of correct factor identification by chance. VSA = vowel space area; FRI = formant ratio index; SFRI = second formant ratio index.

algorithm was tested only with data acquired via a professional microphone without any disruptive noise. Therefore, future studies should evaluate the vowel articulation algorithm performance via a low-quality microphone, such as within smartphones in natural environments (Rusz et al., 2018).

Conclusions

This study represents an insight into the imprecise vowel articulation as a consequence of impairment of fine voluntary movements in a wide range of progressive neurological diseases with various etiologies and stages. We found that an automatized approach could reliably estimate vowel articulation features from natural connected speech regardless of the disease localization in the nervous system (pyramidal tract, basal ganglia, cerebellum, and cranial nerves), etiology (neurodegeneration and autoimmune disorder), and different degrees of disability. However, the specific tongue movement reflected by formant measures differed across some etiologies and dysarthria types independently on dysarthria severity. Therefore, acoustic analysis of vowel articulation may provide a practical tool not only for monitoring the efficacy of future experimental disease-modifying treatments and speech therapy but also for delivering clues for differential diagnosis. Future longitudinal studies should corroborate the sensitivity of vowel articulation deficits to disease progression among progressive disorders.

Authors Contributions

Vojtech Illner: Conceptualization (Equal), Data curation (Lead), Formal analysis (Lead), Methodology (Equal), Software (Lead), Validation (Lead), Visualization (Lead), Writing – original draft (Equal). **Tereza Tykalova:** Investigation (Supporting), Formal analysis (Supporting), Project administration (Equal), Validation (Supporting), Visualization (Supporting), Writing – review & editing (Lead). **Dominik Skrabal:** Formal analysis (Supporting), Investigation (Supporting), Writing – review & editing (Supporting). **Jiri Klempir:** Investigation (Lead), Writing – review & editing (Supporting). **Jan Rusz:** Conceptualization (Equal), Investigation (Supporting), Methodology (Equal), Project administration (Equal), Validation (Supporting), Visualization (Supporting), Funding acquisition (Lead), Writing – original draft (Equal).

Data Availability Statement

Individual participant data that underlie the findings of this study are available upon reasonable request from the corresponding author. The speech data are not publicly available because they contain information that could compromise the privacy of study participants.

Acknowledgments

This study was supported by the Czech Ministry of Health (Grants MH CZ-DRO-VFN64165 to Jiri Klempir and Jan Rusz and NU-20-08-00445 to Vojtech Illner, Tereza Tykalova, Dominik Skrabal, and Jan Rusz), National Institute for Neurological Research (Programme EXCELES; ID Project No. LX22NPO5107), funded by the European Union–Next Generation EU to Tereza Tykalova and Jan Rusz, Czech Technical University in Prague (Grant SGS23/170/OHK3/3 T/13 to Vojtech Illner), and by the Cooperation Program, research area Neuroscience to Jiri Klempir and Jan Rusz. We are obliged to speech-language pathologists Hana Ruzickova and Tereza Listvanova for auditory–perceptual evaluation of dysarthrias.

References

Blomgren, M., Robb, M., & Chen, Y. (1998). A note on vowel centralization in stuttering and nonstuttering individuals. *Journal of Speech, Language, and Hearing Research, 41*(5), 1042–1051. https://doi.org/10.1044/jslhr.4105.1042

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Brooks, B., Miller, R. G., Swash, M., & Munsat, T. (2000). El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and*

Other Motor Neuron Disorders, 1(5), 293–299. https://doi.org/10.1080/146608200300079536

Caverlé, M. W. J., & Vogel, A. P. (2020). Stability, reliability, and sensitivity of acoustic measures of vowel space: A comparison of vowel space area, formant centralization ratio, and vowel articulation index. *The Journal of the Acoustical Society of America, 148*(3), 1436–1444. https://doi.org/10.1121/10.0001931

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences, 169*(1–2), 13–21. https://doi.org/10.1016/S0022-510X(99)00210-5

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Ciucci, M. R., Russell, J. A., Schaser, A. J., Doll, E., Vinney, L. M., & Connor, N. (2011). Tongue force and timing deficits in a rat model of Parkinson disease. *Behavioural Brain Research, 222*(2), 315–320. https://doi.org/10.1016/j.bbr.2011.03.057

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12*(3), 462–496. https://doi.org/10.1044/jshr.1203.462

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*(2), 246–269. https://doi.org/10.1044/jshr.1202.246

Duffy, J. R. (2019). *Motor speech disorders: Substrates, differential diagnosis, and management* (4th ed.). Elsevier.

Elble, R., Comella, C., Fahn, S., Hallett, M., Jankovic, J., Juncos, J. L., LeWitt, P., Lyons, K., Ondo, W., Pahwa, R., Sethi, K., Stover, N., Tarsy, D., Testa, C., Tintner, R., Watts, R., & Zesiewicz, T. (2012). Reliability of a new scale for essential tremor. *Movement Disorders, 27*(12), 1567–1569. https://doi.org/10.1002/mds.25162

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2017). Assessing vowel centralization in dysarthria: A comparison of methods. *Journal of Speech, Language, and Hearing Research, 60*(2), 341–354. https://doi.org/10.1044/2016_JSLHR-S-15-0355

Ge, S., Wan, Q., Yin, M., Wang, Y., & Huang, Z. (2021). Quantitative acoustic metrics of vowel production in Mandarin-speakers with post-stroke spastic dysarthria. *Clinical Linguistics & Phonetics, 35*(8), 779–792. https://doi.org/10.1080/02699206.2020.1827295

Gilman, S., Wenning, G. K., Low, P. A., Brooks, D. J., Mathias, C. J., Trojanowski, J. Q., Wood, N. W., Colosimo, C., Durr, A., Fowler, C. J., Kaufmann, H., Klockgether, T., Lees, A., Poewe, W., Quinn, N., Revesz, T., Robertson, D., Sandroni, P., Seppi, K., & Vidailhet, M. (2008). Second consensus statement on the diagnosis of multiple system atrophy. *Neurology, 71*(9), 670–676. https://doi.org/10.1212/01.wnl.0000324625.00404.15

Goetz, C., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A., Lees, A., Leurgans, S., LeWitt, P., Nyenhuis, D., . . . LaPelle, N. (2008). Movement Disorder Society-Sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders, 23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Hasegawa-Johnson, M., Pizza, S., Alwan, A., Cha, J. S., & Haker, K. (2003). Vowel category dependence of the relationship between palate height, tongue height, and oral area. *Journal of Speech, Language, and Hearing Research, 46*(3), 738–753. https://doi.org/10.1044/1092-4388(2003/059)

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11,* Article 327643. https://doi.org/10.1155/1999/327643

Höglinger, G., Respondek, G., Stamelou, M., Kurz, C., Josephs, K., Lang, A., Mollenhauer, B., Müller, U., Nilsson, C., Whitwell, J., Arzberger, T., Englund, E., Gelpi, E., Giese, A., Irwin, D., Meissner, W., Pantelyat, A., Rajput, A., van Swieten, J., . . . Movement Disorder Society-endorsed PSP Study Group. (2017). Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Movement Disorders, 32*(6), 853–864. https://doi.org/10.1002/mds.26987

Huntington Study Group. (1996). Unified Huntington's Disease Rating Scale: Reliability and consistency. *Movement Disorders, 11*(2), 136–142. https://doi.org/10.1002/mds.870110204

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review, 27*(2), 237–265. https://doi.org/10.3758/s13423-019-01701-x

Kempler, D., & Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Kent, R. D., & Kim, Y. J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics, 17*(6), 427–445. https://doi.org/10.1080/0269920031000086248

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of Communication Disorders, 32*(3), 141–186. https://doi.org/10.1016/S0021-9924(99)00004-0

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 63*(4), 187–194. https://doi.org/10.1159/000318881

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

Kouba, T., Illner, V., & Rusz, J. (2022). Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies: SMART-SPEECH. *BMJ Open, 12*(6), Article e059871. https://doi.org/10.1136/bmjopen-2021-059871

Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology, 33*(11), 1444–1452. https://doi.org/10.1212/WNL.33.11.1444

Lam, J., & Tjaden, K. (2016). Clear speech variants: An acoustic study in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 59*(4), 631–646. https://doi.org/10.1044/2015_JSLHR-S-15-0216

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research, 57*(1), 57–67. https://doi.org/10.1044/1092-4388(2013/12-0262)

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W., & Metze, F. (2020). Universal phone recognition with a multilingual allophone system. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8249–8253). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054362

Lin, D. J., Hermann, K. L., & Schmahmann, J. (2016). The diagnosis and natural history of multiple system atrophy, cerebellar type. *The Cerebellum, 15*(6), 663–679. https://doi.org/10.1007/s12311-015-0728-y

Liu, H. M., Tsao, F. M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America, 117*(6), 3879–3889. https://doi.org/10.1121/1.1898623

Liu, Y., Penttila, N., Ihalainen, T., Lintula, J., Convey, R., & Räsänen, O. (2021). Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 2228–2243. https://doi.org/10.1109/TASLP.2021.3090973

Louis, E. D., Faust, P. L., Vonsattel, J.-P. G., Honig, L. S., Rajput, A., Robinson, C. A., Rajput, A., Pahwa, R., Lyons, K. E., Ross, G. W., Borden, S., Moskowitz, C. B., Lawton, A., & Hernandez, N. (2007). Neuropathological changes in essential tremor: 33 cases compared with 21 controls. *Brain, 130*(12), 3297–3307. https://doi.org/10.1093/brain/awm266

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta), 2,* 49–55.

Mou, Z., Chen, Z., Yang, J., & Xu, L. (2018). Acoustic properties of vowel production in Mandarin-speaking patients with post-stroke dysarthria. *Scientific Reports, 8*(1), Article 14188. https://doi.org/10.1038/s41598-018-32429-8

Payan, C. A. M., Viallet, F., Landwehrmeyer, B. G., Bonnet, A. M., Borg, M., Durif, F., Lacomblez, L., Bloch, F., Verny, M., Fermanian, J., Agid, Y., Ludolph, A. C., Leigh, P. N., Bensimon, G., & NNIPPS Study Group. (2011). Disease severity and progression in progressive supranuclear palsy and multiple system atrophy: Validation of the NNIPPS–PARKINSON PLUS SCALE. *PLOS ONE, 6*(8). https://doi.org/10.1371/journal.pone.0022293

Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Kochanina, J., Majewski, W., Ostroukhov, V., Rusko, M., Sadowski, J., Siemund, R., Staroniewitz, P., Trnka, M., Tropf, H., Vicsi, K., Heuvel, H., & Virag, A. (2000). SpeechDat(E) - Eastern European Telephone Speech Databases. *Proceedings LREC'2000 Satellite Workshop XLDB - Very Large Telephone Speech Databases*, 20–25.

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders, 30*(12), 1591–1601. https://doi.org/10.1002/mds.26424

Reilmann, R. (2019). Parkinsonism in Huntington's disease. *International Review of Neurobiology, 149,* 299–306. https://doi.org/10.1016/bs.irn.2019.10.006

Robertson, L. T., & Hammerstad, J. P. (1996). Jaw movement dysfunction related to Parkinson's disease and partially modified by levodopa. *Journal of Neurology, Neurosurgery, & Psychiatry, 60*(1), 41–50. https://doi.org/10.1136/jnnp.60.1.41

Roy, N., Nissen, S. L., Dromey, C., & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy. *Journal of Communication Disorders, 42*(2), 124–135. https://doi.org/10.1016/j.jcomdis.2008.10.001

Rusz, J., Bonnet, C., Klempíř, J., Tykalová, T., Baborová, E., Novotný, M., Rulseh, A., & Růžička, E. (2015). Speech disorders reflect differing pathophysiology in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of Neurology, 262*(4), 992–1001. https://doi.org/10.1007/s00415-015-7671-1

Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J., & Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. *The Journal of the Acoustical Society of America, 134*(3), 2171–2181. https://doi.org/10.1121/1.4816541

Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., Gagnon, J., Dušek, P., Galbiati, A., Marelli, S., Timm, P., Teigen, L., Janzen, A., Habibi, M., Stefani, A., Holzknecht, E., Seppi, K., Evangelista, E., Rassu, A., … Sonka, K. (2021). Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. *Annals of Neurology, 90*(1), 62–75. https://doi.org/10.1002/ana.26085

Rusz, J., Hlavnicka, J., Tykalova, T., Novotny, M., Dusek, P., Sonka, K., & Ruzicka, E. (2018). Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(8), 1495–1507. https://doi.org/10.1109/TNSRE.2018.2851787

Rusz, J., Klempíř, J., Tykalová, T., Baborová, E., Čmejla, R., Růžička, E., & Roth, J. (2014). Characteristics and occurrence of speech impairment in Huntington's disease: Possible influence of antipsychotic medication. *Journal of Neural Transmission, 121*(12), 1529–1539. https://doi.org/10.1007/s00702-014-1229-8

Rusz, J., Tykalová, T., Salerno, G., Bancone, S., Scarpelli, J., & Pellecchia, M. (2019). Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. *Journal of Neurology, 266*(6), 1394–1404. https://doi.org/10.1007/s00415-019-09271-7

Saigusa, H., Saigusa, M., Aino, I., Iwasaki, C., Li, L., & Niimi, S. (2006). M-Mode color Doppler ultrasonic imaging of vertical tongue movement during articulatory movement. *Journal of Voice, 20*(1), 38–45. https://doi.org/10.1016/j.jvoice.2005.01.003

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America, 134*(5), EL477–EL483. https://doi.org/10.1121/1.4826150

Sapir, S., Ramig, L., & Fox, C. (2008). Speech and swallowing disorders in Parkinson disease. *Current Opinion in Otolaryngology & Head and Neck Surgery, 16*(3), 205–210. https://doi.org/10.1097/MOO.0b013e3282febd3a

Schiel, F., & Zitzelsberger, T. (2018, May). Evaluation of automatic formant trackers. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga, (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 2843–2848). European Language Resources Association.

Schmitz-Hübsch, T., du Montcel, S., Baliko, L., Berciano, J., Boesch, S., Depondt, C., Giunti, P., Globas, C., Infante, J., Kang, J., Kremer, B., Mariotti, C., Melegh, B., Pandolfo, M., Rakowicz, M., Ribai, P., Rola, R., Schols, L., Szymanski, S., … Fancellu, R. (2006). Scale for the assessment and rating of ataxia: Development of a new clinical scale. *Neurology, 66*(11), 1717–1720. https://doi.org/10.1212/01.wnl.0000219042.60538.92

Schwarz, P., & Černocký, J. (2008). *Phoneme recognition based on long temporal context* [Doctoral dissertation, Brno University of Technology].

Schwarz, P., Matejka, P., Burget, L., & Glembek, O. (2022). *Phoneme recognizer based on long temporal context*. Brno

University of Technology. https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

Shimon, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research, 53*(1), 114–125. https://doi.org/10.1044/1092-4388(2009/08-0184)

Skodda, S., Schlegel, U., Klockgether, T., & Schmitz-Hübsch, T. (2014). Vowel articulation in patients with spinocerebellar ataxia. *International Journal of Speech and Language Pathology and Audiology, 1,* 63–71. https://doi.org/10.12970/2311-1917.2013.01.02.3

Skodda, S., Visser, W., & Schlegel, U. (2011). Vowel articulation in Parkinson's disease. *Journal of Voice, 25*(4), 467–472. https://doi.org/10.1016/j.jvoice.2010.01.009

Skrabal, D., Rusz, J., Novotny, M., Sonka, K., Ruzicka, E., Dusek, P., & Tykalova, T. (2022). Articulatory undershoot of vowels in isolated REM sleep behavior disorder and early Parkinson's disease. *NPJ Parkinson's Disease, 8*(1), Article 137. https://doi.org/10.1038/s41531-022-00407-7

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A., Miller, D. H., Montalban, X., . . . Cohen, J.A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology, 17*(2), 162–173. https://doi.org/10.1016/S1474-4422(17)30470-2

Tjaden, K. (2008). Speech and swallowing in Parkinson's disease. *Topics in Geriatric Rehabilitation, 24*(2), 115–126. https://doi.org/10.1097/01.TGR.0000318899.87690.44

Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. *Journal of Speech, Language, and Hearing Research, 56*(5), 1485–502. https://doi.org/10.1044/1092-4388(2013/12-0259

Tjaden, K., Rivera, D., Wilding, G., & Turner, G. (2005). Characteristics of the lax vowel space in dysarthria. *Journal of Speech, Language, and Hearing Research, 48*(3), 554–566. https://doi.org/10.1044/1092-4388(2005/038)

Tykalova, T., Pospisilova, M., Cmejla, R., Jerabek, J., Mares, P., & Rusz, J. (2016). Speech changes after coordinative training in patients with cerebellar ataxia: A pilot study. *Neurological Sciences, 37*(2), 293–296. https://doi.org/10.1007/s10072-015-2379-7

Tykalová, T., Rusz, J., Čmejla, R., Klempíř, J., Růžičková, H., Roth, J., & Růžička, E. (2015). Effect of dopaminergic medication on speech dysfluency in Parkinson's disease: A longitudinal study. *Journal of Neural Transmission, 122*(8), 1135–1142. https://doi.org/10.1007/s00702-015-1363-y

Weismer, G. (1984). Articulatory characteristics of parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal–supraglottal coordination. In M. R. McNeil, J. C. Rosenbek, & A. E. Aronson (Eds.), *The dysarthrias: Physiology, acoustics, perception, management* (pp. 101–130). College-Hill.

Weismer, G. (2006). Philosophy of research in motor speech disorders. *Clinical Linguistics & Phonetics, 20*(5), 315–349. https://doi.org/10.1080/02699200400024806

Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*(1), 1–18. https://doi.org/10.1159/000052649

Whitfield, J. A. (2019). Exploration of metrics for quantifying formant space: Implications for clinical assessment of Parkinson disease. *Perspectives of the ASHA Special Interest Groups, 4*(2), 402–410. https://doi.org/10.1044/2019_PERS-SIG19-2018-0004

Whitfield, J. A., & Goberman, A. (2014). Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease. *Journal of Communication Disorders, 51,* 19–28. https://doi.org/10.1016/j.jcomdis.2014.06.005

Whitfield, J. A., & Mehta, D. D. (2019). Examination of clear speech in Parkinson disease using measures of working vowel space. *Journal of Speech, Language, and Hearing Research, 62*(7), 2082–2098. https://doi.org/10.1044/2019_JSLHR-S-MSC18-18-0189

Williams, D., & Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *The Journal of the Acoustical Society of America, 136*(5), 2751–2761. https://doi.org/10.1121/1.4896471

Yunusova, Y., Green, J., Lindstrom, M. J., Pattee, G. L., & Zinman, L. (2013). Speech in ALS: Longitudinal changes in lips and jaw movements and vowel acoustics. *Journal of Medical Speech-Language Pathology, 21*(1), 1–13.

2.5.1 Exploitation of speech parametrizations

Speech can be parametrized by sets with low interpretability, but high performance, such as MFCCs and their derivatives [70], Relative Spectral Transform - Perceptual Linear Prediction parameters [71], and deep neural networks embeddings [72]. These sets are often used in complex frameworks such as speech recognition where the undisclosed explanation poses no problem. However, it limits the use in clinically related studies and, most notably, in clinical trials [73].

Nonetheless, in recent years, MFCCs specifically have gained interest in studies focused on speech impairments in neurological diseases due to their capacity to capture considerable information from the speech waveform [74]. However, the complete relationship between the MFCCs and particular speech dysfunctions remains unclear. Therefore, a study has been conducted to explore potential links between MFCCs and particular speech impairments [75].

A cohort of 23 individuals with PD who were treated with bilateral Deep Brain Stimulation of the Subthalamic Nucleus (STN-DBS) were recruited for the study, together with 23 healthy controls. The examination in the PD group was held in two conditions, including STN-DBS switched OFF and STN-DBS switched ON, and selected, physiologically interpretable features were calculated from recordings in each condition, together with MFCCs. The stimulation alters several aspects of speech production [76]. The study linked the differences in the features to the differences in MFCCs induced by ON and OFF conditions, hinting which aspects are related to the change.

It was found that changes in lower (2nd to 4th) cepstral coefficients significantly reflect changes in CPP, representing voice quality measure. Higher MFCCs (4th to 9th) highly correlated with measures describing a dynamical ability of articulatory movement. A global parameter, calculated by averaging the individual MFCCs, incorporated the captured speech characteristics from corresponding individual parameters and demonstrated higher sensitivity to distinguish PD and STN-DBS conditions than the standard physiological features.

The findings may shed light on interpreting outcomes from speech assessment for future clinical trials. The preprint of the article is provided below.

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Which aspects of motor speech disorder are captured by Mel Frequency Cepstral Coefficients? Evidence from the change in STN-DBS conditions in Parkinson's disease

Vojtěch Illner[1], *Petr Krýže*[1], *Jan Švihlík*[1,2], *Mario Sousa*[3], *Paul Krack*[3], *Elina Tripoliti*[4], *Robert Jech*[5], *Jan Rusz*[1,3,5]

[1]Faculty of Electrical Engineering, Czech Technical University in Prague, Czechia
[2]Faculty of Chemical Engineering, University of Chemistry and Technology, Czechia
[3]Movement Disorders Center, Department of Neurology, University Hospital of Bern, Switzerland
[4]UCL Queen Square Institute of Neurology, University College London, United Kingdom
[5]Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University, Czechia

Abstract

One of the most popular speech parametrizations for dysarthria has been Mel Frequency Cepstral Coefficients (MFCCs). Although the MFCCs ability to capture vocal tract characteristics is known, the reflected dysarthria aspects are primarily undisclosed. Thus, we investigated the relationship between key acoustic variables in Parkinson's disease (PD) and the MFCCs. 23 PD patients were recruited with ON and OFF conditions of Deep Brain Stimulation of the Subthalamic Nucleus (STN-DBS) and examined via a reading passage. The changes in dysarthria aspects were compared to changes in a global MFCC measure and individual MFCCs. A similarity was found in 2nd to 3rd MFCCs changes and voice quality. Changes in 4th to 9th MFCCs reflected articulation clarity. The global MFCC parameter outperformed individual MFCCs and acoustical measures in capturing STN-DBS conditions changes. The findings may assist in interpreting outcomes from clinical trials and improve the monitoring of disease progression.

Index Terms: Mel Frequency Cepstral Coefficients, Parkinson's disease, speech disorder, dysarthria, acoustic analysis

1. Introduction

Speech represents the most complex quantitative marker of motor function, vastly sensitive to damage to the brain's neural structures [1]. Speech dysfunctions presence has been documented in a number of progressive neurological diseases, such as Parkinson's disease (PD) [2]. In recent years, due to technological and computational advances, there has been an increasing interest in the use of speech for monitoring disease progression, symptoms severity, and a potential diagnostic aid [1, 3]. Improved ease in obtaining voice recordings using either smartphones [4, 5, 6], or telemonitoring homecare systems [7] offers intriguing advances as speech evaluation is inexpensive, non-invasive, simple to administer, and scalable to a large population.

Analysis of the acquired speech data and potential pathology is primarily interpreted using physiological speech patterns describing vocal tract abilities, such as articulation, pitch variability, loudness, rhythm, and phonation [8]. However, speech can also be parametrized by sets with low interpretability, but high performance, such as Mel Frequency Cepstral Coefficients (MFCCs) and their derivatives [9, 10, 11], Relative Spectral Transform - Perceptual Linear Prediction parameters [12], and deep neural networks embeddings [13]. The undisclosed explanation poses no problem for complex frameworks such as speech recognition but limits the use in clinically related studies and, most notably, in clinical trials [14].

Nevertheless, as one of the most popular speech parametrizations, MFCCs remain highly relevant due to their capacity to capture considerable information from the speech waveform. While being a long-standing essential part of frameworks for speech recognition [15], speaker detection [16], speech synthesis [17], and many others, in the last decade, they also gained interest in studies focused on speech impairments in neurological diseases [5, 9, 10, 11, 18]. However, the complete relationship between the MFCCs and particular speech dysfunctions remains clouded. In [9], authors comment that the coefficients detect subtle changes in the motion of the articulators (tongue, lips). Nonetheless, such an assumption has never been validated, while MFCCs can be easily influenced by other factors such as age, gender, speaking style, or recording procedure/microphone quality [19]. Most recently, in Roche's PD Mobile application designed for clinical trial measures in PD [5, 20], the speech performance of the patients was analyzed on a sustained phonation task using only the second coefficient, MFCC2, representing a low-to-high frequency energy ratio [8].

Although the MFCCs are emerging as one of the principal features in assessing speech impairments in neurological diseases, their interpretability remains limited. Therefore, we tested the sensitivity of MFCCs in a scenario covering Parkinsonian patients with ON and OFF conditions of Deep Brain Stimulation of the Subthalamic Nucleus (STN-DBS). Since the STN-DBS might substantially alter the patient's speech abilities [21], we expect to discover changes in MFCCs that might correspond to changed acoustical patterns of hypokinetic dysarthria.

2. Methods

2.1. Participants

A total of 23 individuals with PD (four females), with a mean age of 61.7 years (SD = 5.0, range: 53–72), who were treated with bilateral STN-DBS in combination with dopaminergic medication, were recruited for the study. The examination in the PD group was held in two conditions, including STN-DBS switched OFF (hereafter, the DBS OFF condition) and STN-DBS switched ON (hereafter, the DBS ON condition). Detailed clinical characteristics (clinical scores and DBS settings across individuals with PD) and experimental procedure description can be found in previous study [22]. As a healthy control

10.21437/Interspeech.2023-1744

(HC) group, 23 age- and sex-matched (four females) volunteers, a mean age of 61.5 years (SD = 5.6, range: 52–72), with no history of neurological or communication disorders, were recruited. All participants were native Czech speakers. The study complied with the Helsinki Declaration and was approved by the Ethics Committee of the General University Hospital, Prague, Czech Republic. Each participant provided written informed consent.

2.2. Speech examination

The patients were recorded after the individual therapeutical setting and were asked to perform phonetically balanced reading passage task of a standardized text of 313 words with a familiar, up-to-date vocabulary and grammatical structure. The audio recordings were conducted in a quiet room with a low ambient noise level using a head-mounted condenser microphone (Beyerdynamic Opus 55, Heilbronn, Germany) placed approximately $5\,\mathrm{cm}$ from the subject's mouth. Speech signals were sampled at $48\,\mathrm{kHz}$ with 16-bit resolution.

2.3. MFCCs computation

After the trial testing, the following procedure was established to calculate the first 16 MFCCs. The number 16 was set, similarly to [11, 23], as a tradeoff between studies using fewer coefficients, such as 12 or 13 [9, 24], and longer coefficients vectors, such as 20 [10]. The computations were conducted in MATLAB, Natick, USA.

Similarly, as in [10, 25], the audio input was first downsampled to $16\,\mathrm{kHz}$ with lowpass pre-filtering to guard against aliasing. Next, a pre-emphasis filter was applied to the samples with $\alpha = -0.95$. MFCCs were computed using MATLAB Auditory Toolbox functions. The entire signal is processed in frames using a Hamming window of a length $25\,\mathrm{ms}$ with 5% overlap. The frame's FFT magnitude is converted into Mel filterbank outputs using 13 linearly spaced filters followed by 27 log-spaced filters ranging from approximately 133 Hz to 6864 Hz. Next, a cosine transform of the $\log_{10}$ of the output is computed. The result is a vector of c_0-c_{16} MFCCs standard deviations across frames. The 0th coefficient, c_0, representing signal energy, is discarded.

Necessarily, a voice activity detector has to be applied, so the coefficients are used only in the segments of speech. In this study, dynamical thresholding of the spectral distance of the computed coefficients is utilized to mark segments without speech presence [26]. Coefficients in such segments are discarded.

Apart from analyzing individual c_1-c_{16}, a global MFCC measure is established, inspired by [24], as a mean of the standard deviation (std) of c_1-c_{16}. It is designed to represent the overall dynamic movement ability of individual vocal tract elements, as the individual MFCCs overlap the partitions of the frequency domain.

2.4. Physiological acoustic markers

To link the MFCCs to key dysarthria elements of PD, five acoustic variables with well-known pathophysiological interpretation were extracted from the speech waveform using the framework developed in [8].

Speech impairments in PD can be, for the most part, characterized by decreased voice quality, imprecise articulation, monoloudness, monopitch, deficits in speech timing, and inappropriate pauses [27]. A decrease in voice quality can be reflected by a lower Cepstral Peak Prominence (CPP) measure [28]. Aspects of imprecise articulation can be represented by a decrease in resonant frequency attenuation (RFA) measure, defined as the ratio between local second formant region maxima and local valley region minima. RFA is mainly sensitive to articulatory decay but may also be partly influenced by abnormal nasal resonance [8]. Monoloudness corresponds to a lower std of the speech energy (stdPWR) and monopitch to a lower std of estimated pitch contour (stdF0). Deficits in speech timing and rhythm, such as slowing or accelerating tempo, are reflected by the net speech rate (NSR) measure. Since the information about the length and occurrence of pauses is uncapturable by MFCCs, the measure representative for the description of pauses was omitted.

2.5. Statistical analysis

The following two experiments were conducted to assess the physiological nature of MFCCs.

First, the differences in the variables between DBS ON and OFF, called $\Delta^{\mathrm{ON}}_{\mathrm{OFF}}$, were calculated, representing the change in speech characteristics:

$$\Delta^{\mathrm{ON}}_{\mathrm{OFF}} v_i = v_i^{\mathrm{ON}} - v_i^{\mathrm{OFF}}, \tag{1}$$

where v_i^{ON}, v_i^{OFF} are the variables (MFCCs, global MFCC parameter, acoustical features) from DBS ON and DBS OFF groups, respectively. Then, Spearman correlation was computed between $\Delta^{\mathrm{ON}}_{\mathrm{OFF}}$ MFCC variables and $\Delta^{\mathrm{ON}}_{\mathrm{OFF}}$ acoustical variables.

Subsequently, the individual variables were compared in the three groups (HC, DBS ON, DBS OFF) using repeated measures analysis of variance (RM-ANOVA) followed by Bonferroni post-hoc correction, where the HC group is age- and sex-aligned with the DBS subjects and treated as associated measurement.

3. Results

The results from the first experiment are shown in Figure 1. Change in CPP was correlated with changes in lower MFCCs (c_2-c_3), $\rho > 0.48$, $p < 0.05$. Change in RFA correlated with c_4-c_9 coefficients changes, $\rho > 0.45$, $p < 0.05$. Changes in the global MFCC parameter achieved significant correlations with changes in CPP and RFA, $\rho = 0.46$, $p < 0.05$, partly also reflecting changes in stdF0 and NSR, $\rho = 0.41$, $p = 0.06$, resp. $\rho = -0.40$, $p = 0.06$.

The results from RM-ANOVA for c_1-c_{16} and global MFCC are shown in Figure 2. According to $F(1, 22)$ statistics, the global MFCC parameter achieved the highest overall significance in between-group differences ($F(1, 22) = 53.1$, $p < 0.001$ for HC vs. DBS OFF, $F(1, 22) = 19.1$, $p < 0.001$ for HC vs. DBS ON, $F(1, 22) = 8.4$, $p < 0.05$ for DBS ON vs. DBS OFF). Lower coefficients (c_1-c_5) demonstrate significant differences between HC and DBS ON or DBS OFF groups ($F(1, 22) > 8.0$, $p < 0.05$). However, significant contrast between DBS ON and OFF is present in higher coefficients, c_5-c_8 ($F(1, 22) > 5.8$, $p < 0.05$).

Figure 3 shows boxplots for the global MFCC parameter and acoustical features. Only the global MFCC parameter achieved a significant difference between DBS ON and OFF ($F(1, 22) = 8.4$, $p < 0.05$). RFA and stdF0 demonstrated significant contrast between HC and both DBS ON and OFF ($F(1, 22) > 11.7, p < 0.01$). NSR and stdPWR showed significant differences only between HC and DBS OFF ($F(1, 22) > 6.8, p < 0.05$).

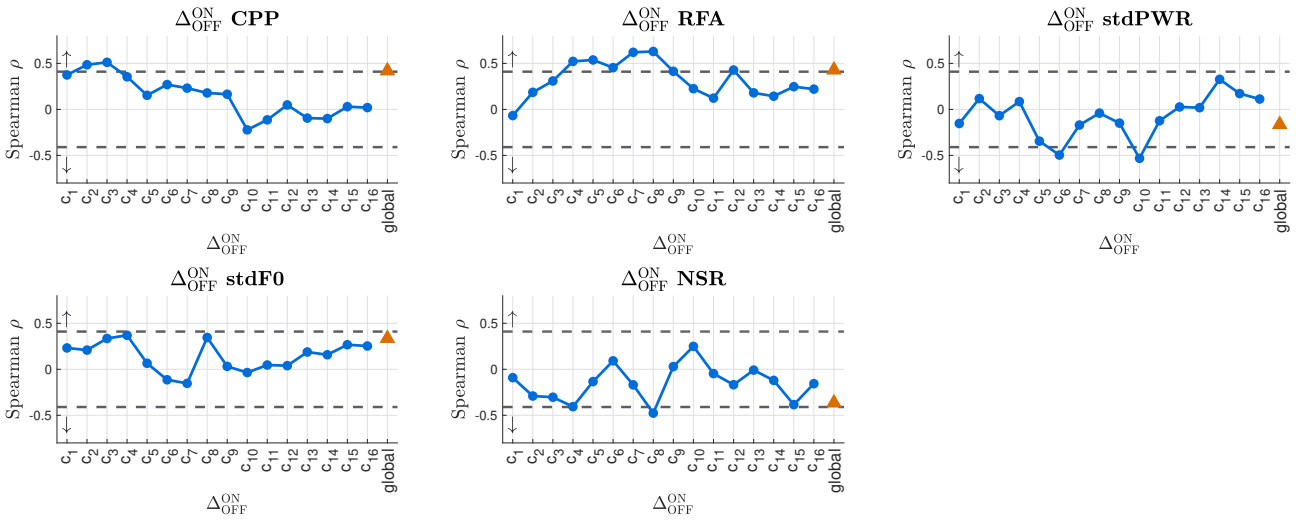


Figure 1: *Spearman correlation of Δ^{ON}_{OFF}, differences in variables between DBS ON and DBS OFF, for individual c_1-c_{16} Mel Frequency Cepstral Coefficients (MFCC) and global MFCC parameter (mean standard deviation of c_1-c_{16}) to acoustical variables. Dashed lines represent the boundary of significant correlation with $p < 0.05$. Captions: CPP=Cepstral Peak Prominence, RFA=Resonant Frequency Attenuation, stdPWR=signal energy standard deviation, stdF0=pitch contour standard deviation, NSR=Net Speech Rate.*

4. Discussion

The present study explored the relationship between individual c_1-c_{16} MFCCs and physiologically interpretable acoustic features, including a global MFCC parameter composed of all the coefficients. Based on changes in speech characteristics between DBS ON and DBS OFF, we could relate the alterations of individual coefficients to specific acoustical markers. Moreover, the global parameter as well as some individual MFCCs were able to capture significant differences in speech production between the HC group and PD groups, and even between DBS ON and OFF, outperforming the traditional acoustical measures.

4.1. Relationship between MFCCs and acoustical measures

Changes in lower cepstral coefficients ($\Delta^{ON}_{OFF}c_2$, $\Delta^{ON}_{OFF}c_3$) significantly reflect changes in CPP, representing voice quality measure. CPP has been shown to strongly correlate with the increase in the severity of dysphonia and breathiness in various languages [28]. CPP integrates multiple acoustical measures related to lower speech frequencies, such as first rahmonic, pitch, waveform deviations, and noise perturbations [28]. Since both c_2 and c_3 are related to the signal energy, covering the corresponding range (approximately 200 - 500 Hz), the relationship with CPP and the ability to capture such characteristics become apparent.

Higher MFCCs, starting from $\Delta^{ON}_{OFF}c_4$ to $\Delta^{ON}_{OFF}c_9$, significantly correlate with the changes in RFA measure. RFA represents the second formant to anti-formant based system [8], i.e., special case of MFCC limited to frequency regions around the second formant. Therefore, RFA obviously provides comparable results to the MFCC system, although MFCCs cover more wide frequency range and are not dependent on the correct estimation of the position of the second formant. Both these MFCCs and RFA metrics (at least considering the frequency range between 4th to 9th cepstral coefficients) might thus provide a measure of the dynamical ability of articulatory movement. Such a measure might supplement the traditional formant-based approaches reflecting the range of movement of articulators, which particularly vary with tongue placement position.

The changes in the global MFCC parameter, designed to represent the overall dynamic movement ability of individual vocal tract elements, are significantly correlated to changes in CPP and RFA as well, thus incorporating the captured speech characteristics from corresponding individual MFCCs. The pitch variability is mildly related and would likely significantly contribute to observed results with increasing sample size. Interestingly, NSR is negatively correlated with the global measure meaning that with an increased articulation rate, the articulation ability and the quality of voice decrease; however, the trend is not significant.

4.2. Capacity of individual MFCCs to capture within-group differences

Individual MFCCs expressed within-group speech characteristics differences with a high significance. Especially lower coefficients (c_1-c_5) achieved an excellent score in distinguishing HC and PD cohorts (including DBS ON and OFF groups). The results might be explained by the close connection between the coefficients and measures of CPP and stdF0, explored in the previous section. On the other hand, higher c_5-c_9 MFCCs, related to the RFA measure, outperformed the lower ones in terms of capturing changes between DBS ON and DBS OFF. The evidence is that distinctive and eminently recognizable speech changes between speech in HC and PD are represented by lower MFCCs, whereas higher (approximately c_5-c_9) reflect subtle changes in articulation ability and formant structure, present between DBS ON and DBS OFF conditions. High coefficients, c_{10}-c_{16}, do not appear to have a significant effect on the group differences between PD and HC groups. However, the statistics are much more powerful between DBS ON and OFF than their comparison to HC. Sporadic significant differences between DBS ON and DBS OFF in coefficients c_{10}, c_{14}, and c_{15} might be due to correspondence with particular high formant structures but also due to noise which is more present in higher frequencies.

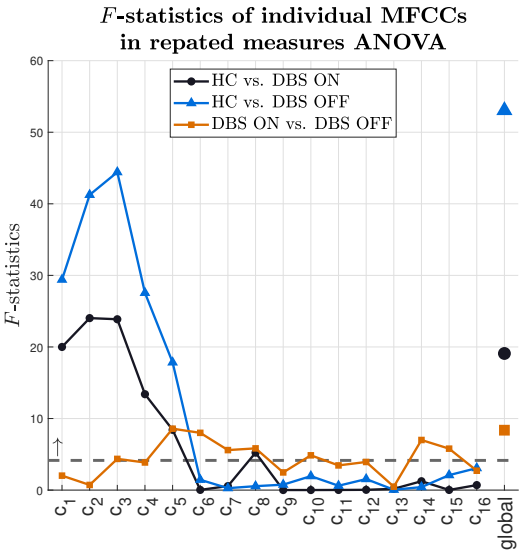


Figure 2: *Computed $F(1, 22)$ statistics for individual c_1-c_{16} Mel Frequency Cepstral Coefficients (MFCC) and global MFCC parameter (mean standard deviation of c_1-c_{16}) according to repeated measures ANOVA. Dashed line represent boundary of significant difference of $p < 0.05$ based on Bonferroni post-hoc correction. Captions: ANOVA=analysis of variance, HC=healthy controls, DBS=deep brain stimulation.*

4.3. The use of the global MFCC parameter

Interestingly, the global MFCC measure demonstrated the most significant overall within-group difference. It achieved the highest score in separating HC and DBS OFF and a comparable score between HC and DBS ON and DBS ON and DBS OFF with the best-achieving coefficients. The fact that the global parameter comprehends the properties of the individual coefficients while maintaining high robustness might prove beneficial for its use in practice. For example, the c_2 coefficient demonstrated a significant, comparable score to the global measure between HC and both DBS states. However, it achieved poor results distinguishing between DBS ON and DBS OFF. The same can be analogously applied to, for example, c_6.

Additionally, compared to acoustical variables used in this study, the global MFCC demonstrates the highest overall significance between the DBS ON and DBS OFF conditions. The evidence might be due to the ability to reflect CPP and RFA, and partly stdF0 and NSR, altogether with capturing additional information about the individual vocal tract elements.

4.4. Limitations of the study

Only Czech-speaking subjects in a small cohort were part of the study. Further investigations should include other languages and larger sample sizes to confirm the findings. Additionally, it has been found that microphone quality and position highly influence amplitude-based features such as RFA [4]. Since we showed that MFCCs work on the same principle, the sensitivity of MFCCs to different experimental recording settings should be recognized and considered for large-scale use [29].

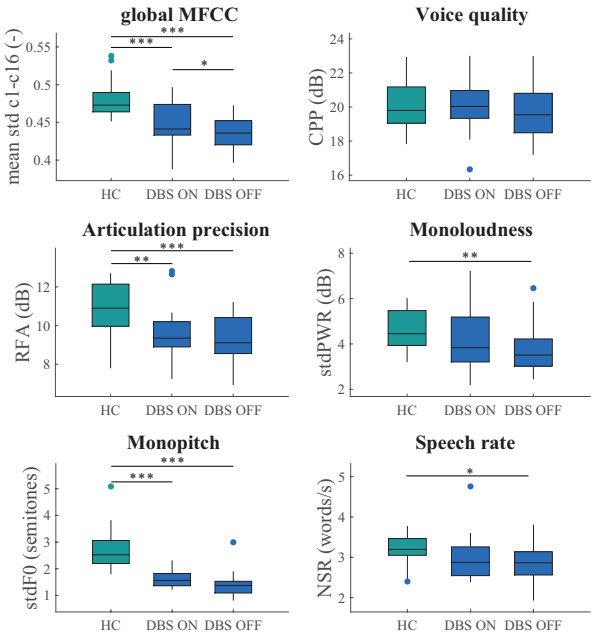


Figure 3: *Statistically significant group differences for the global Mel Frequency Cepstral Coefficients parameter (mean standard deviation of c_1-c_{16}) and acoustical variables with with ***, **, * referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$ according to repeated measures analysis of variance with Bonferroni post-hoc correction. Middle bars represent median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. Captions: DBS=deeb brain stimulation, CPP=Cepstral Peak Prominence, RFA=Resonant Frequency Attenuation, stdPWR=signal energy standard deviation, stdF0=pitch contour standard deviation, NSR=Net Speech Rate.*

5. Conclusions

The present study investigated the relationship between c_1-c_{16} MFCCs and five physiologically interpretable acoustical variables of hypokinetic dysarthria. In addition, a global MFCC parameter was established as mean std of c_1-c_{16}. A high correlation was shown between changes in low c_1-c_3 coefficients and changes in voice quality and signal envelope. Changes in coefficients from approximately c_4-c_9 reflect subtle changes in articulation ability and lower formants structures. The global MFCC measure comprehended the properties of the individual coefficients while maintaining high robustness and achieving significant between-group differences, outperforming all the single coefficients and acoustical measures. The findings may shed light on interpreting outcomes from speech assessment for future clinical trials.

6. Acknowledgments

This study was supported by the Czech Science Foundation (grant no. 21-14216L), SNF (grant no. 32003BL_197709), Czech Ministry of Education: National Institute for Neurological Research Programme EXCELES (ID Project no. LX22NPO5107) - Funded by the European Union – Next Generation EU, and Czech Technical University in Prague (grant no. SGS23/170/OHK3/3T/13).

7. References

[1] J. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*, 4th ed. Elsevier, 2019.

[2] J. Rusz, C. Bonnet, J. Klempíř, T. Tykalová, E. Baborová, M. Novotný, A. Rulseh, and E. Růžička, "Speech disorders reflect differing pathophysiology in parkinson's disease, progressive supranuclear palsy and multiple system atrophy," *Journal of Neurology*, vol. 262, pp. 992–1001, 2015.

[3] J. Rusz, J. Hlavnička, M. Novotný, T. Tykalová, A. Pelletier, J. Montplaisir, J. Gagnon, P. Dušek, A. Galbiati, S. Marelli, P. C. Timm, L. N. Teigen, A. Janzen, M. Habibi, A. Stefani, E. Holzknecht, K. Seppi, E. Evangelista, A. L. Rassu, Y. Dauvilliers, B. Högl, W. Oertel, E. K. S. Louis, L. Ferini-Strambi, E. Růžička, R. B. Postuma, and K. Šonka, "Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease," *Annals of Neurology*, vol. 90, pp. 62–75, 2021.

[4] J. Rusz, J. Hlavnicka, T. Tykalova, M. Novotny, P. Dusek, K. Sonka, and E. Ruzicka, "Smartphone allows capture of speech abnormalities associated with high risk of developing parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 1495–1507, 2018.

[5] F. Lipsmeier, K. I. Taylor, T. Kilchenmann, D. Wolf, A. Scotland, J. Schjodt-Eriksen, W. Y. Cheng, I. Fernandez-Garcia, J. Siebourg-Polster, L. Jin, J. Soto, L. Verselis, F. Boess, M. Koller, M. Grundman, A. U. Monsch, R. B. Postuma, A. Ghosh, T. Kremer, C. Czech, C. Gossens, and M. Lindemann, "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 parkinson's disease clinical trial," *Movement Disorders*, vol. 33, pp. 1287–1297, 2018.

[6] T. Kouba, V. Illner, and J. Rusz, "Study protocol for using a smartphone application to investigate speech biomarkers of parkinson's disease and other synucleinopathies," *BMJ Open*, vol. 12, 2022.

[7] C. L. Payten, D. D. Nguyen, D. Novakovic, J. O'Neill, A. M. Chacon, K. A. Weir, and C. J. Madill, "Telehealth voice assessment by speech language pathologists during a global pandemic using principles of a primary contact model: an observational cohort study protocol," *BMJ Open*, vol. 12, 2022.

[8] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific Reports*, vol. 7, 2017.

[9] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 2012.

[10] A. Benba, A. Jilbab, and A. Hammouch, "Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, pp. 297–307, 2015.

[11] T. Kapoor and R. K. Sharma, "Parkinson's disease diagnosis using mel-frequency cepstral coefficients and vector quantization," *International Journal of Computer Applications*, vol. 14, pp. 975–8887, 2011.

[12] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with parkinson's and neurological diseases using cepstral analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 1100–1108, 2016.

[13] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the reliability of acoustic speech embeddings," *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 4621–4625, 2020.

[14] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, "Quantifying articulatory impairments in neurodegenerative motor diseases: A scoping review and meta-analysis of interpretable acoustic features," *International Journal of Speech-Language Pathology*, 2022.

[15] S. Lokesh and M. R. Devi, "Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method," *Cluster Computing*, vol. 22, pp. 11 669–11 679, 2019.

[16] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using mel frequency cepstral coefficients (mfcc) and vector quantization (vq) techniques," *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, pp. 248–251, 2012.

[17] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 3, pp. 1299–1302, 2000.

[18] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 134, pp. 2171–2181, 2013.

[19] J. Rusz, M. Novotny, J. Hlavnicka, T. Tykalova, and E. Ruzicka, "High-accuracy voice-based classification between patients with parkinson's disease and other neurological diseases may be an easy task with inappropriate experimental design," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, pp. 1319–1321, 2017.

[20] F. Lipsmeier, K. I. Taylor, R. B. Postuma, E. Volkova-Volkmar, T. Kilchenmann, B. Mollenhauer, A. Bamdadian, W. L. Popp, W. Y. Cheng, Y. P. Zhang, D. Wolf, J. Schjodt-Eriksen, A. Boulay, H. Svoboda, W. Zago, G. Pagano, and M. Lindemann, "Reliability and validity of the roche pd mobile application for remote monitoring of early parkinson's disease," *Scientific Reports*, vol. 12, 2022.

[21] E. Tripoliti, M. Zrinzo, I. Martinez-Torres, E. Frost, S. Pinto, T. Foltynie, E. Holl, E. Petersen, M. Roughton, M. I. Hariz, and P. Limousin, "Effects of subthalamic stimulation on speech of consecutive patients with parkinson disease," *Neurology*, vol. 76, pp. 80–86, 2011.

[22] J. Svihlik, M. Novotny, T. Tykalova, K. Polakova, H. Brozova, P. Kryze, M. Sousa, P. Krack, E. Tripoliti, E. Ruzicka, R. Jech, and J. Rusz, "Long-term averaged spectrum descriptors of dysarthria in patients with parkinson's disease treated with subthalamic nucleus deep brain stimulation," *Journal of Speech, Language, and Hearing Research*, vol. 65, pp. 4690–4699, 2022.

[23] R. Fraile, N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex," *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 146–152, 2009.

[24] J. Rusz, J. Klempíř, E. Baborová, T. Tykalová, V. Majerová, R. Čmejla, E. Růžička, and J. Roth, "Objective acoustic quantification of phonatory dysfunction in huntington's disease," *PLoS ONE*, vol. 8, 2013.

[25] A. Slis, N. Lévêque, C. Fougeron, M. Pernon, F. Assal, and L. Lancia, "Analysing spectral changes over time to identify articulatory impairments in dysarthria," *The Journal of the Acoustical Society of America*, vol. 149, pp. 758–769, 2021.

[26] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, vol. 3, 1993, pp. 321–324.

[27] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, pp. 246–269, 1969.

[28] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014.

[29] J. Rusz, T. Tykalova, L. O. Ramig, and E. Tripoliti, "Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders," *Movement Disorders*, vol. 36, pp. 803–814, 2021.

2.6 Smartphone application for remote monitoring

The preceding sections outlined the methods for effectively extracting significant insights from spontaneous, everyday speech to assess the severity and progression of potential neurodegenerative diseases through interpretable physiological patterns. The following section details the design of a developed system capable of unobtrusively, ethically, and securely recording calls and other data using modern smartphones [77].

The application consists of two parts. First part works in the background and records the subject voice calls. After either an incoming or outgoing call was concluded, the user was presented with a screen containing an option to delete or sent the call to research analysis. If the second option was selected, the recording was immediately filtered using a distant-speaker filtering algorithm, and saved to the private application folder on the device. It was possible to replay or delete each recording within the following 24 hours. After 24 hours, the recording was sent to a secured server. The original call was then deleted from the device. Due to ethical reasons and the General Data Protection Regulation law, the subjects' speech partners were needed to be removed entirely from the recordings, which was accomplished by employing a real-time adaptive filtering of the input audio. Specifically, the computational complexity needed to be optimised because intensive background processes tend to be eventually suppressed by the phone's operating system. The algorithm utilized a cross-channel thresholding using smoothed energy estimates from a specific spectral bands, and was based on the Neumann-Pearson Criterion. Once the data arrived to the server a gateway validation was executed. A speaker recognition framework confirmed that the recording contained the right participant and rejected it otherwise, for cases where the participant handed the phone to someone else, the recording was significantly corrupted, or did not contain speech content at all.

The second part of the application consisted of a module that invited participants to record active motor and functional vocal tasks every specified period. The tasks were selected as commonly used in the PD research [78] and included sustained phonation, syllables repetitions, reading a passage, tapping, alternated tapping, writing a sentence, resting hand tremor task, and gait with a turnaround.

The developed framework may serve as a reliable, unobtrusive tool to remotely collect participant‘s spontaneous everyday speech data, together with functional vocal and motor tasks. The tool can be used for cross sectional or longitudinal studies evaluating the severity and progression of parkinsonism via distinguished speech patterns. The preprint of the protocol is provided below.

BMJ Open

Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies: SMARTSPEECH

Tomáš Kouba, Vojtěch Illner, Jan Rusz

To cite: Kouba T, Illner V, Rusz J. Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies: SMARTSPEECH. *BMJ Open* 2022;**12**:e059871. doi:10.1136/bmjopen-2021-059871

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-059871).

TK and VI contributed equally.

Received 03 December 2021
Accepted 14 June 2022

Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic

Correspondence to
Dr Jan Rusz;
rusz.mz@gmail.com

ABSTRACT

Introduction Early identification of Parkinson's disease (PD) in its prodromal stage has fundamental implications for the future development of neuroprotective therapies. However, no sufficiently accurate biomarkers of prodromal PD are currently available to facilitate early identification. The vocal assessment of patients with isolated rapid eye movement sleep behaviour disorder (iRBD) and PD appears to have intriguing potential as a diagnostic and progressive biomarker of PD and related synucleinopathies.

Methods and analysis Speech patterns in the spontaneous speech of iRBD, early PD and control participants' voice calls will be collected from data acquired via a developed smartphone application over a period of 2 years. A significant increase in several aspects of PD-related speech disorders is expected, and is anticipated to reflect the underlying neurodegeneration processes.

Ethics and dissemination The study has been approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic and all the participants will provide written, informed consent prior to their inclusion in the research. The application satisfies the General Data Protection Regulation law requirements of the European Union. The study findings will be published in peer-reviewed journals and presented at international scientific conferences.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study aims to employ the remote longitudinal speech monitoring of prodromal and early-stage patients with Parkinson's disease via smartphone calls.
⇒ An additional set of active speech and motor tasks will be captured.
⇒ The results of this study will be based solely on the Czech language.
⇒ This study will be implemented in one type of smartphone, and the impact of different devices on speech outcomes is yet to be determined.

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder characterised by the loss of dopaminergic neurons in the substantia nigra.[1] The incidence of PD is approximately 1.8% in persons older than 65 years.[2] There is no treatment that halts or slows the progression of PD, and the pharmacotherapy and neurosurgical interventions that are available only mitigate specific symptoms. A PD diagnosis is typically made when cardinal motor manifestations appear; by this point, up to 50% of the neurons in the substantia nigra may already be irrecoverably damaged.[3] Unfortunately, no sufficiently accurate biomarkers of PD are currently available, although the existence of such biomarkers would allow for the measurement of the effectiveness of experimental treatments in slowing the progression of the disease. Furthermore, no reliable method for identifying people who are at high risk of developing PD exists at present. Thus, establishing a suitable biomarker would be a crucial breakthrough that would impact on diagnoses and future PD treatments, and is one of the most important topics in PD-related research.[4] In particular, motor-related dysfunctions are strong predictors of conversion to synucleinopathy, and have one of the most significant hazard ratios of 3.16 among the available predictive markers to date.[5]

Isolated rapid eye movement sleep behaviour disorder (iRBD) is a parasomnia characterised by dream-enactment behaviours associated with REM sleep without muscle atonia,[6] and represents a prodromal marker of neurodegenerative synucleinopathies, mainly PD and dementia with Lewy bodies.[7] The risk of developing a neurodegenerative disease is exceptionally high (>80%) in subjects with iRBD.[8 9] Since symptoms of iRBD precede Parkinsonism, research

focused on iRBD is essential for the development of a neuroprotective therapy to counter synucleinopathy,[10] as no other preclinical marker has a predictive value that is comparable to iRBD.[11]

As the most complex acquired human motor skill, and involving over 100 muscles, speech is a sensitive marker of damage to neural structures engaged in the brain's motor system control.[12] Up to 90% of PD patients develop distinctive speech and voice abnormalities, collectively termed hypokinetic dysarthria, which are mainly characterised by a decrease in the quality of the voice, hypokinetic articulation, hypophonia, monopitch, monoloudness and deficits in timing and phrasing.[13] Based on the recent findings of a multilanguage, multicentric study using an objective acoustical analysis of 150 patients with iRBD, it is clear that speech disorders are one of the earliest motor signs of PD.[14] Thus, vocal assessment appears to have intriguing potential as a preclinical diagnostic and progressive biomarker of PD and related neurodegenerations; it is also inexpensive, non-invasive, easy to administer and has the potential to being conducted remotely from the patient's home (eg, by using a smartphone[15]).

In current practice, measures of PD are primarily subjective, rater-dependent and require in-clinic assessments.[16] The trials using these measures are lengthy and expensive and, as they are usually based on a once-off assessment, might produce false results. Given the technical and accessibility advancements in smartphone technologies, evaluating motor symptoms of PD via mobile devices continues to be an increasing focus among the research community.[17–19] The potential of smartphones has already been demonstrated in quantifying the severity of PD based on the remote assessment of five tasks (voice, finger tapping, gait, balance and reaction time) using an application that was developed for this purpose.[20–22] However, such an assessment requires the patients (or research subjects) to repeatedly perform a set of predefined activities. It is thus reasonable to expect that most patients will not be willing to perform such artificially constructed activities on a daily basis for several years.

By contrast, the extraction of speech patterns from smartphone calls would provide a natural, passive biomarker that does not require additional effort on the part of the subjects. Moreover, such a speech-based application could easily be scaled to a larger population, thus allowing for high-throughput screening, followed by a more detailed analysis if the screen is abnormal. However, the reliability of smartphones in detecting prodromal PD (ie, iRBD) via smartphone calls in realistic scenarios with ambient noise level environments has not yet been investigated.

Therefore, the aim of this study is to develop a fully automated and noise-resistant smartphone-based system that is able to monitor the distinctive speech patterns of neurodegeneration on a daily basis using acoustic data obtained in various environments (SMARTSPEECH). Such a system would have tremendous potential to revolutionise the diagnostic process of PD and could provide a robust biomarker of the progression of the disease. To fully exploit the possibilities of smartphones, a set of active speech and motor tasks is included for sensitivity analysis and for comparison.

METHODS

Objectives

The main objective of the study is to demonstrate that, using the SMARTSPEECH system, speech performance elicited during regular phone calls through a smartphone can provide principal biomarkers for diagnosis and monitoring the progression of prodromal PD.

The specific objectives are to:

1. Develop a smartphone application that will be able to capture the subjects' data,
2. Collect up to 2-year longitudinal speech data from subjects with iRBD, patients with early-stage PD and healthy control subjects of comparable age.
3. Collate existing approaches and develop novel methods allowing assessment speech markers of neurodegeneration in PD, including tests of their robustness against noise and recording conditions and selecting the most appropriate parameters for smartphone-based monitoring.
4. Build the concept of a complex system (SMARTSPEECH) for detection of speech abnormalities in PD and other synucleinopathies, including its statistical power evaluation in differentiation between PD, iRBD and control groups and identify its relationship to essential clinical markers reflecting disease progression such as the Movement Disorder Society-Unified Parkinson's Disease Rating Scale.
5. Analyse the sensitivity of a set of active speech and motor tasks and compare the outcomes to data acquired through SMARTSPEECH passive calls monitoring.

Collection of speech data

To avoid potential conflicts due to different microphone characteristics across various manufacturers, the speech data will be collected using the same smartphone device, HONOR 9X Lite (Shenzhen Zhixin New Information Technology), operating on the Android V.9 system. The application will record the subject's telephone speech with a high degree of quality using a sampling frequency of 44.1 kHz and 16-bit quantification.

The distant-speaker filtering algorithm

Due to ethical reasons and the General Data Protection Regulation law, the subjects' speech partners will need to be removed entirely from the recordings, which will be accomplished by employing real-time adaptive filtering of the input audio. The audio will be collected via two different microphones, representing two channels in the stereo mix (see figure 1). The primary microphone (MIC 1) is the closest to the speaker's mouth and thus captures the subject's speech with absolute power dominance. However, the distant speaker's talk might still be

Figure 1 Schematics of the HONOR 9X smartphone audio inputs and outputs.

present, such as in segments in which the subject is silent or speaks quietly, or when the distant speaker is speaking very loudly. The secondary microphone (MIC 2) mainly captures the speech of the subject, but also captures the distant speaker (coming from a call speaker nearby, hence at a greater power than via MIC1) and surrounding noise. However, these settings are valid only if the loudspeaker is not activated. In this case, the power of both speakers is more or less equal in the channels. Since the algorithm is not guaranteed of functioning properly in such scenario, the recording process will be cancelled immediately when the user activates the loudspeaker.

In order to remove the distant speaker from the final mix completely, a real-time adaptive algorithm was designed based on the hardware and software settings of the smartphone. Specifically, the computational complexity needed to be optimised because intensive background processes tend to eventually be suppressed by the phone's operating system. The principle, which is cross-channel thresholding using smoothed energy estimates from a specific spectral band, is based on the Neumann-Pearson Criterion. The algorithm's illustrative schema is displayed in figure 2, and the procedure is as follows:

The input is a stereo call recording $s[n]$ consisting of channel 1 (from MIC 1), $s_1[n]$, and channel 2 (from MIC 2), $s_2[n]$. Both are processed in subsequent windows of a given length of L seconds. In the current frame, the signals are decimated by two, thus reducing the computational burden of the entire procedure while maintaining adequate resolution. The signals are then processed through a Butterworth band-pass filter with a range of 300–8000 Hz, as this is the range in which the fundamental speech information is expected to be found. A power estimate is computed from the filtered channels using the L-1 norm, which is simply the sum of the absolute values of the samples in a given segment with a length of w. The estimated power trajectory is then smoothed using a normalised integrator with a forgetting factor λ, producing $S_1[n]$ and $S_2[n]$. The segments containing the subject's speech are then detected when the following condition is met:

$$S_1[n] > \mathrm{k} \cdot \mathrm{mean}\left(S_2[n]\right), \quad (1)$$

where k represents the extent of the channels' power difference as a ratio of the $s_1[n]$ and $s_2[n]$ SD, calculated as $\mathrm{k} = \sqrt{\mathrm{std}\left(\mathrm{s}_1[\mathrm{n}]\right)/\mathrm{std}\left(s_2[n]\right)}$. The SD and the mean in equation (1) are both computed per window. At the end of the window processing section, the timestamps are shifted to compensate for the delay introduced by the normalised integrator. The segments that are not considered to consist of the subject's speech are masked by zeros, and only the output from the first channel is subject to further processing.

A final check is made to ensure that the end of the window does not conflict with the subject's speech segment, which might produce disruptive artefacts. If the result is positive, the output from the current window is discarded, and the process is rerun with an adjusted, shorter L to avoid the conflict. The final output is a mono audio recording containing only the subject's speech.

After the trial testing was conducted, the default value for L was set to 30 s as the optimal value for reduced computational power, more precise results and lower possibilities of conflicts at the end of the window. The segment length for the L-1 norm was set to $w = 128$ samples with an overlap of 50%, which resulted in fast processing with sufficient precision. The forgetting factor was set to $\lambda = 0.99$, producing the optimal smoothing effect for the given scenario.

SMARTSPEECH application for monitoring passive speech via regular smartphone calls

After either an incoming or outgoing call is concluded, the user will be presented with a screen containing an option to delete or save the call (figure 3A). If the second option is selected, the recording will be filtered immediately using a distant-speaker filtering algorithm, and will be saved to the private application folder on the device. It will be possible to replay or delete each recording from the recent calls list (figure 3B) in the application within the following 24 hours. After 24 hours, the recording will be compressed and sent to a secured server using SSH and REST API. The original call will then be deleted from the device. The aim is to record at least four 5 min recordings per month from each participant. A setting section on which the user can configure the application to suit their needs will be included (figure 3C).

Active tasks

The application has a module that invites participants to record three active functional vocal tasks once every 14 days; the participants are instructed to capture the data

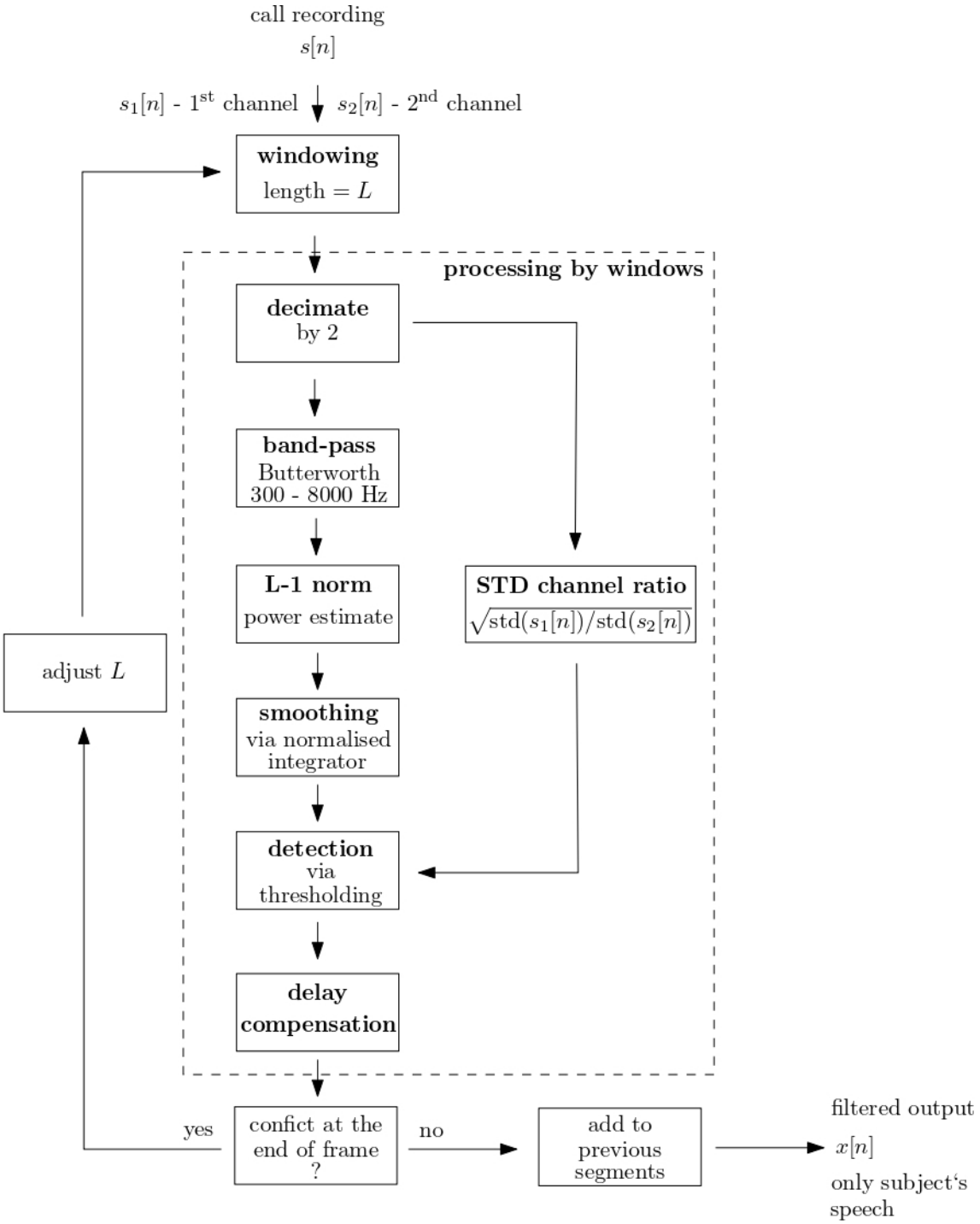


Figure 2 An illustrative schema of the distant-speaker filtering algorithm.

in everyday environments with a low ambient noise level. In addition, the participants are asked to perform five motor tasks that are commonly used in PD research.[20–23] These tasks will be used to compare the sensitivity of the active and the passive data. On launching the active tasks protocol the users are presented with the guidelines and elapsed time since last run (figure 3D). Before each task, the user will be presented with short, written instructions, including an option to replay an audio or video example of the given task. In the voice section, the recordings can be replayed after being recorded. If the user is dissatisfied with the result, an option to repeat the task is available.

The entire set of active tasks consists of prolonged phonation (figure 4A), /pa/-/ta/-/ka/ syllable repetition (figure 4B), reading a passage (figure 4C,D), a tapping game (figure 5A,B), alternated tapping (figure 5C,D), a writing task (figure 5E,F), resting hand tremor (figure 6A,B) and a gait with turnaround (figure 6C,D). Each task will be executed twice, and the motor tasks are

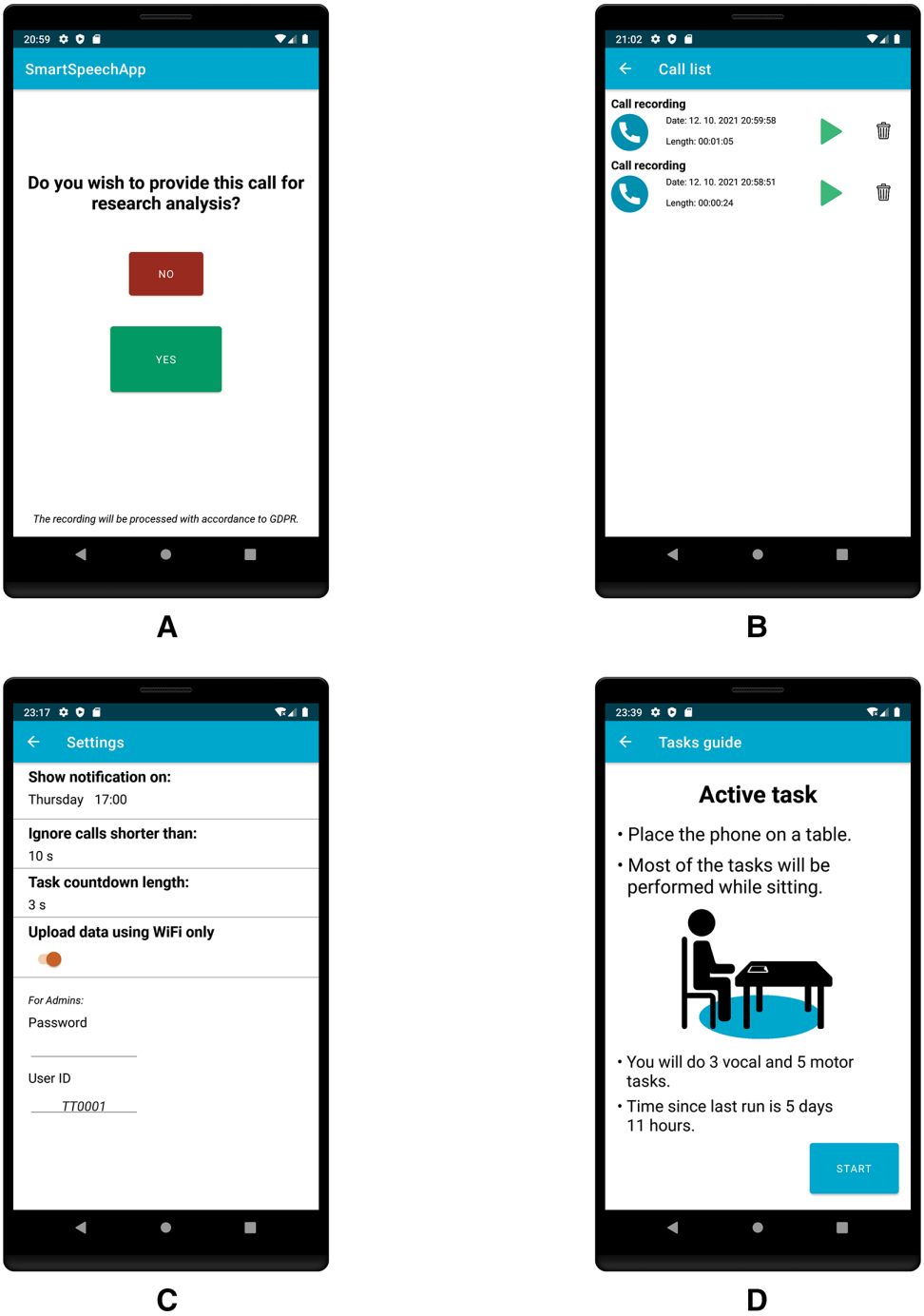


Figure 3 (A) Option screen to delete or save the recently completed call for speech analysis. (B) List of calls from the previous 24 hours with details and the options of replaying and deletion. (C) The setting screen of the app. (D) the active task introduction screen, which shows a guide to performing the active tasks and the time elapsed since the last active task run.

to be executed first using the right hand and then the left hand.

Data validation and monitoring

Once the call and task data arrive on our server, they are saved on a database and the basic information becomes visible on the web interface to allow the data parameters to be reviewed for the quality, frequency and recording duration. In the event that the participant does not make any calls or perform passive tasks within the required period, the participant will be notified via a push notification sent to their email address.

A gateway validation of the incoming data needs to be performed before the data are stored securely on the database to minimise the possibility of corrupted, irrelevant data entering the system, which would produce misleading results.

First, recordings that are shorter than two seconds will be discarded because they do not contain much

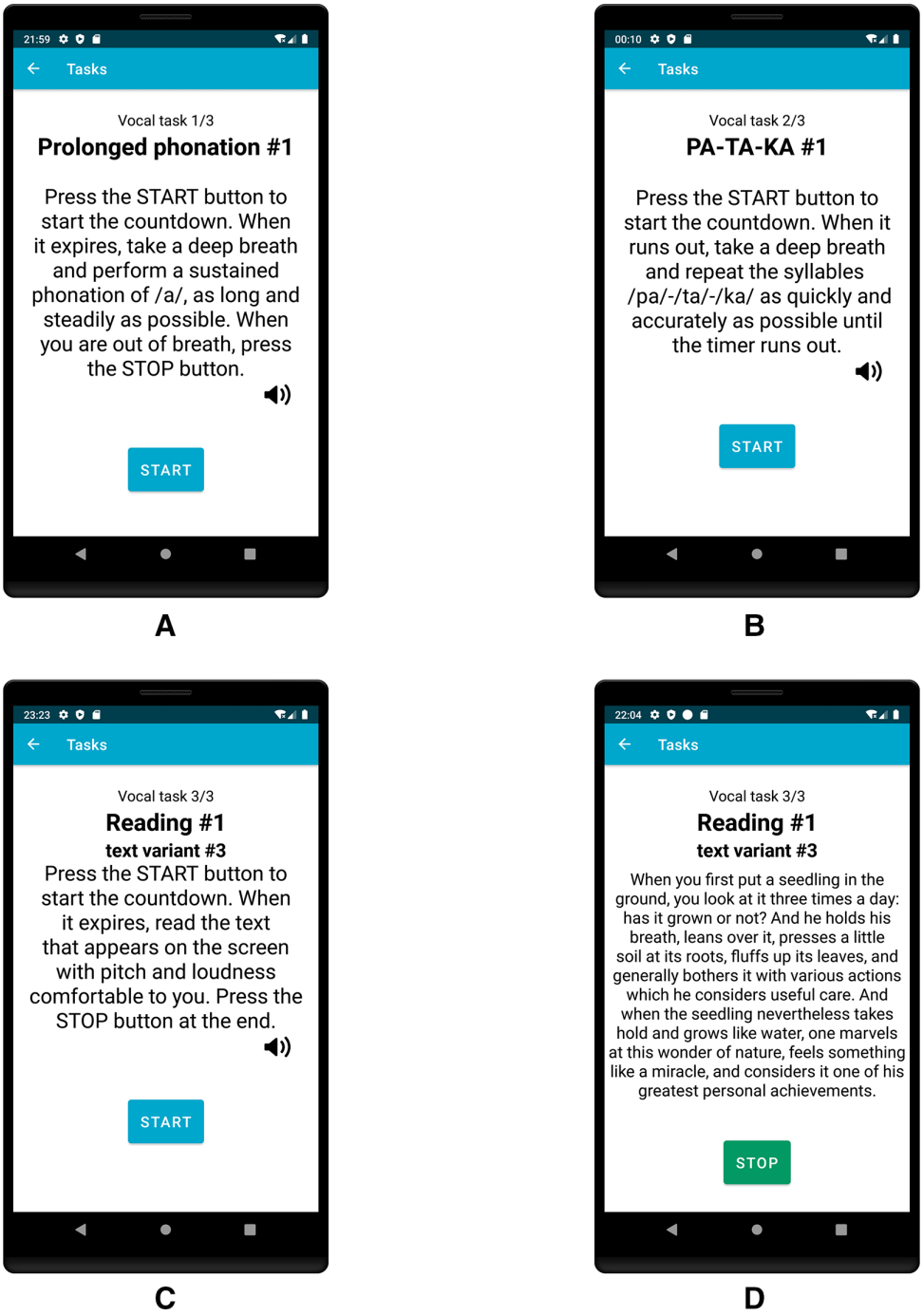


Figure 4 (A) Prolonged phonation task instruction screen (task duration ~15 s). (B) Sequential motion rates instruction screen (task duration 7 s). (C) Reading passage instruction screen. (D) Reading passage sample text that is chosen randomly from six samples and contains approximately 80 words (task duration ~40 s).

information that is relevant and often contain errors. Subsequently, the speaker-detection algorithm will be employed to prevent data contamination when the user hands the smartphone to someone else for the call. The Kaldi toolkit[24] was used to construct the procedure. We employed a high-order Gaussian mixture model trained on the Czech section of the GLOBALPHONE database[25] using a set of mel-frequency cepstral coefficients. New subjects in the study will be enrolled in the model once a sufficient amount of data has been collected. I-vectors are extracted on detection, and are used to compute the probabilistic linear discriminant analysis score. Using this score, a classificator decides whether the recording belongs to the correct subject. The data that will be assessed as not originating from the given subject will be moved to a specific folder and are not included in the data analysis.

Note that the purpose of the algorithm is also a form of quality control, as it excludes recordings that are of poor quality, such as utterances that are significantly

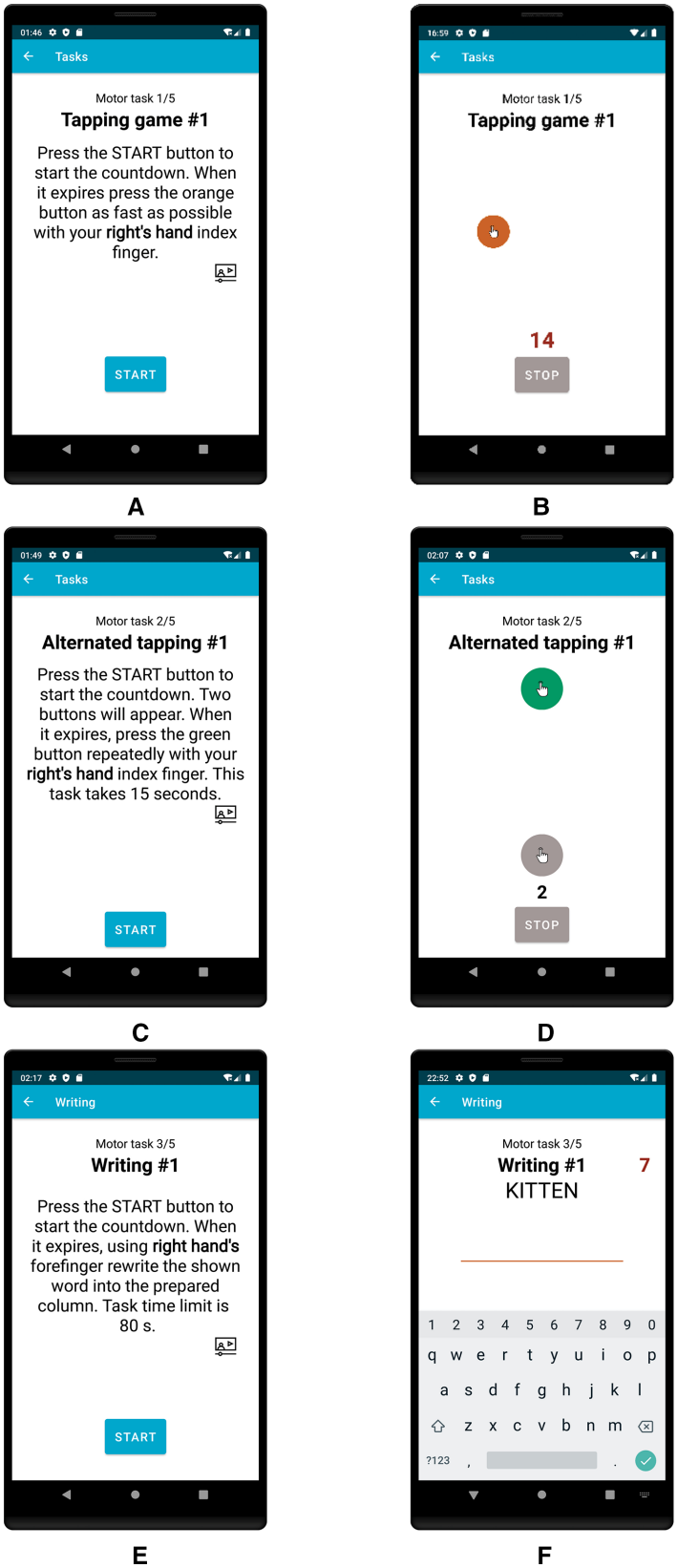


Figure 5 (A) A tapping game introduction screen. (B) A tapping game task. The delay between spawning the circle and a successful tap is measured. The action takes place 20 times (task duration ~20 s). (C) An alternated tapping introduction screen. (D) An alternated tapping task. The timestamps of the individual taps are measured (task duration 15 s). (E) Writing task instruction screen. (F) A writing task. The time spent rewriting each word is measured. The patient is shown five words in each run. Eight different datasets that are assembled in the same manner are chosen randomly in each run. Examples include the Czech words 'SESTRICKA', 'NEJLEPSI', 'LECENI', 'ZABAVA' and 'PREDSTAVA'. The task can be skipped after 80 s have elapsed (task duration ~40 s).

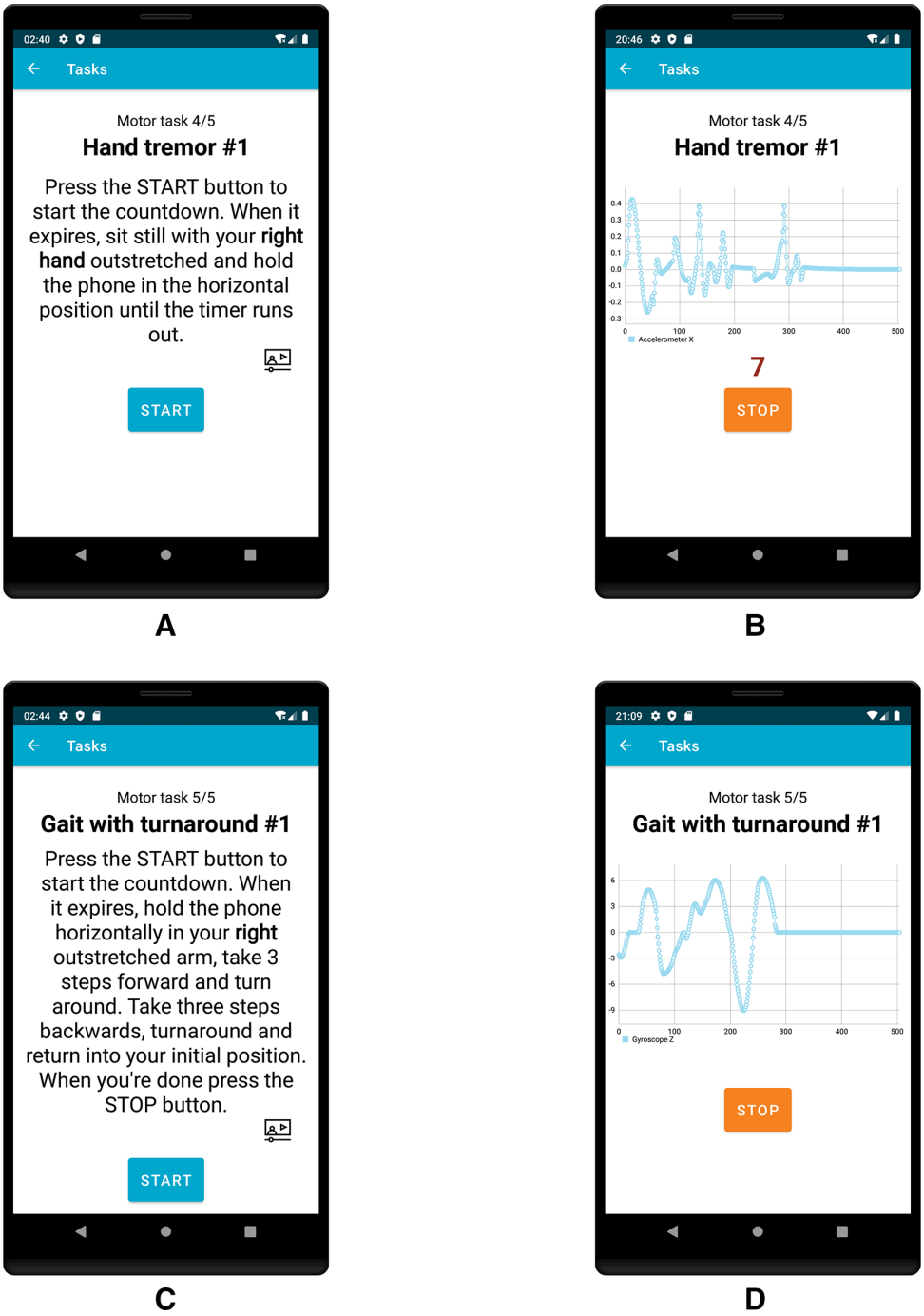


Figure 6 (A) Rest tremor instruction screen. (B) A rest tremor task (task duration 15 s). The patient's hand tremor is measured using an accelerometer in all three axes. The z-axis is live plotted for user feedback. (C) A gait task instruction screen. (D) A gait task (task duration ~30 s). The z-axis rotation is measured and plotted on the screen for user feedback via an internal gyroscope.

affected by disruptive noise at a low signal-to-noise ratio (SNR). In these cases, the subject's speech is so corrupted that the computed features have little similarity to the subject's model. In addition, an SNR estimator based on minimum statistics[26] is employed in the last validation phase. Segments with an instantaneous SNR lower than the preselected threshold will be excluded from the analysis because the methods for extracting speech features cease to be reliable at certain very low SNR levels (such as pitch detection algorithms at SNR less than 6 dB[27]).

A voice activity detector is applied to the samples that have passed the validation procedure. The output indicates long calls during which the subject only speaks for a tiny fraction of the overall duration. This allows to have a better grasp of the amount of speech data gathered via the system and to notify the subject if the amount of data is insufficient.

Extraction of acoustic speech and motor features

Hypokinetic dysarthria is due to dysfunction in the basal ganglia motor circuit, leading to impairments in the regulation of the initiation, the amplitude and the velocity of movements.[12 28] Therefore, this project aims to collect, develop and extract different acoustic features describing motor aspects of speech that have well-defined PD pathophysiology[17] and correspond to the perceptual description of hypokinetic dysarthria defined by Darley *et al*.[29]

The following features represent the most anticipated candidates for the final analysis:

1. Disruptions in phonation caused by dysfunctions in the vocal folds can be captured using the acoustic measure of Cepstral Peak Prominence, which correlates with the auditory perception of decreased voice quality/breathiness.[30]
2. Articulation deficits, which are perceived as a decrease in intelligibility, can be described using metrics such as the vowel space area.[31]
3. Dysprosody is captured by the reduced amplitude of vocal cord movements, which correlates with the impression of monopitch, and can be assessed using the acoustic measure of pitch variability.[14]
4. Decreased ability to maintain the speech motor sequence or to alternate quickly between responses can be reflected by the acoustic measures of the net speech rate and duration of pause intervals, which reflect the perceived auditory timing of speech and may describe deficits such as a slow articulation rate and a reduced ability to intermit and initiate speech.[32]
5. Linguistic deficits, including limited vocabulary and a decrease in its range, which might indicate potential underlying mild cognitive impairment, can be assessed using a set of lexical features such as content density.[33]
6. To assess the overall degree of speech impairment, powerful deep neural networks could be employed, including spectrogram analysis to evaluate specific features when traditional methods might be insufficient. Insight into the detailed method behaviour would be supportive, and would reveal critical physiological details.[34]

Several methods for the automatic analysis of key dimensions of speech in patients with PD have already been developed.[32 35 36] However, all the methods need to undergo experimental and theoretical testing for noise robustness and reliability to validate their usefulness.[27] A set of features commonly used in the existing literature will be extracted from the active motor tasks[20–22]; the set includes tapping velocity, intratap duration variability, reaction time, tremor acceleration skewness and velocity.

Endpoints

The primary endpoint will be represented by composite dysarthria index, reflecting the severity of speech impairment, which will be based on a combination of several distinct acoustic speech features associated with hypokinetic dysarthria in PD. Secondary endpoints will be represented by individual acoustic speech features associated with hypokinetic dysarthria in PD and linguistic features associated with potential cognitive decline.

Study design and participants

During this project, each participant will be given a full explanation of the project's purpose and aims, will be informed about the procedure, and will be given the opportunity to ask questions before deciding whether to sign the informed consent form. In total, we plan to recruit 25 iRBD subjects, 25 early-stage patients with PD and 25 healthy controls as part of this longitudinal study. As we expect a drop-out rate of about 20% per year,[17] up to 50 patients with iRBD and the same number of early-stage patients with PD will be recruited at the baseline. These data might be used for a better-sampled cross-sectional study. All the iRBD subjects will fulfil the criteria listed in the International Classification of Sleep Disorders, third edition diagnostic criteria,[37] including confirmation of REM sleep without atonia via polysomnography. The inclusion criteria for iRBD will be

1. Onset of iRBD after 50 years of age.
2. No history of major neurological disease (such as epilepsy or strokes) or other significant diseases that could affect study participation or voice analysis (eg, active cancer, drug abuse or diseased vocal cords).
3. No significant cognitive decline or severe depression.
4. No history of therapy with antiparkinsonian medication.

All the patients with PD will meet the Movement Disorders Society's clinical diagnostic criteria for PD,[38] and will be investigated during the on-medication state. The inclusion criteria for PD will be:

1. Onset of PD after 50 years of age.
2. Hoehn and Yahr stage 1–2 in the on-medication state.
3. Disease duration more than 5 years after diagnosis.
4. No history of a major neurological disease other than PD (such as epilepsy or strokes) or other significant diseases that could affect study participation or voice analysis (eg, active cancer, drug abuse, or diseased vocal cords).
5. No significant cognitive decline or severe depression.
6. No involvement in any speech therapy during the duration of the project.

The inclusion criterion for controls will be that the participants have no history of neurological or communication disorders, and no history of parasomnias or other sleep disorders. PD and healthy control subjects will be age-matched and gender-matched to the iRBD group.

Each participant will be required to perform passive speech recordings and active tasks using the provided smartphone for 2 years. In addition, each of the subjects will undergo three examinations at the clinic: at the baseline, after 1 year and after 2 years (at the end of the project). The clinical examinations will consist of taking a structured personal history, quantitative testing of motor and non-motor symptoms of PD based on the Movement Disorder Society-Unified Parkinson's Disease Rating Scale,[16] cognitive testing using the Montreal Cognitive

Assessment,[39] autonomic testing using the Scales for Outcomes in Parkinson's Disease-Autonomic Dysfunction,[40] the evaluation of depressive symptoms using the Beck Depression Inventory II,[41] and speech examinations according to the dysarthria research guidelines for acoustic analyses.[23] Recruitment of the participants will begin in October 2021.

Patient and public involvement statement

Most of the patients have been involved in previous studies and are familiar with the team, the research topic and the methods. The SMARTSPEECH protocol was designed with the aid of a short questionnaire that was completed by a selected representative group of 33 patients with iRBD or PD, which provided insight into their requirements (eg, a dual-SIM phone and monetary compensation for mobile phone tariffs), doubts and motivations (such as understanding how to operate the phone and data security) and the expected frequency of phone usage and calls.

Sample size estimation

For the primary endpoint of the project, an ad hoc power analysis for a given large effect size (Cohen's d of 0.8), with the type I error probability (α) set at 0.05 and power of 80%, based on a three-group analysis of variance with one covariate (group), determined a minimum sample size of 66 subjects with at least 22 subjects in each subgroup (ie, 22 patients with iRBD, 22 patients with PD and 22 controls).

Statistical analysis

A one-way analysis of variance with a group (PD vs iRBD vs controls) as a between-subject factor will be used to calculate the differences for each parameter of interest. The Pearson correlation coefficient will be applied to search for correlations among the variables. The minimum level of significance will be set at $p<0.05$ with an appropriate Bonferroni adjustment. In addition, a binary logistic regression followed by a leave-one-subject-out cross-validation will be used to assess the sensitivity/specificity of the proposed features to differentiate the iRBD subjects from controls, patients with PD from controls and patients with PD from iRBD subjects.

ETHICS AND DISSEMINATION

The study has been approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic (no. 30/19 Grant AZV VES 2020 VFN), and will be performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All the participants will provide written, informed consent prior to their inclusion.

The following steps will be taken to mitigate potential ethical concerns: The researchers performing the analyses will have no known relationship with the research subjects, and the recordings will be deidentified to the extent possible. The analyses of audiorecordings will be automated to avoid the recordings being listened to by a human; furthermore, the audiorecordings will be encrypted, will only be available to the authorised researchers, and will be deleted at the request of any participant without the need for justification. All the steps will be conducted according to the directive on personal data protection legislation in the Czech Republic and the approval of the Ethics Committee. Any amendments will be agreed on by the research steering committee and submitted for ethics committee approval prior to implementation.

This project may provide a natural digital biomarker of disease progression based on longitudinal data acquired without any cost or time burden on the patients and investigators. Observing disease progression over a short period using well-defined and disease-specific speech biomarkers may significantly aid in recruiting appropriate cases into large clinical trials for disease-modifying drugs and allows monitoring possible disease-modifying effects of treatment in prodromal PD.[42 43] In the future, speech biomarkers may also bolster early presymptomatic diagnosis and enable rapid access to neuroprotective therapy once available. Results will be presented at national and international conferences, published in peer-reviewed journals, and disseminated to the researchers and Parkinson's community.

Contributors JR conceived of the study and the study design. TK and VI drafted and revised the manuscript. TK and VI designed the application. TK, VI and JR approved the final manuscript.

Funding This study was supported by the Czech Ministry of Health (grant no. NU20-08-00445).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Jan Rusz http://orcid.org/0000-0002-1036-3054

REFERENCES

1 Poewe W, Seppi K, Tanner CM, *et al*. Parkinson disease. *Nat Rev Dis Primers* 2017;3:17013.
2 de Rijk M, Launer L, Berger K. Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurol* 2000;54:S21–3 http://intl.neurology.org/cgi/content/abstract/54/11_suppl_5/S21
3 Rodriguez-Oroz MC, Jahanshahi M, Krack P, *et al*. Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms. *Lancet Neurol* 2009;8:1128–39.
4 Postuma RB, Berg D. Advances in markers of prodromal Parkinson disease. *Nat Rev Neurol* 2016;12:622–34.

5 Postuma RB, Iranzo A, Hu M, *et al*. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain* 2019;142:744–59.
6 Boeve BF. REM sleep behavior disorder. *Ann N Y Acad Sci* 2010;1184:15–54.
7 Miglis MG, Adler CH, Antelmi E, *et al*. Biomarkers of conversion to α-synucleinopathy in isolated rapid-eye-movement sleep behaviour disorder. *Lancet Neurol* 2021;20:671–84.
8 Iranzo A, Tolosa E, Gelpi E, *et al*. Neurodegenerative disease status and post-mortem pathology in idiopathic rapid-eye-movement sleep behaviour disorder: an observational cohort study. *Lancet Neurol* 2013;12:443–53.
9 Schenck CH, Boeve BF, Mahowald MW. Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series. *Sleep Med* 2013;14:744–8.
10 Schenck CH, Montplaisir JY, Frauscher B, *et al*. Rapid eye movement sleep behavior disorder: devising controlled active treatment studies for symptomatic and neuroprotective therapy—a consensus statement from the International Rapid Eye Movement Sleep Behavior Disorder Study Group. *Sleep Med* 2013;14:795–806.
11 Postuma RB, Aarsland D, Barone P, *et al*. Identifying prodromal Parkinson's disease: pre-motor disorders in Parkinson's disease. *Mov Disord* 2012;27:617–26.
12 Duffy J. *Motor speech disorders: substrates, differential diagnosis, and management*. Fourth ed. Maryland Heights: Elsevier, 2019.
13 Ho AK, Iansek R, Marigliani C, *et al*. Speech impairment in a large sample of patients with Parkinson's disease. *Behav Neurol* 1999;11:131–7.
14 Rusz J, Hlavnička J, Novotný M, *et al*. Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. *Ann Neurol* 2021;90:62–75.
15 Rusz J, Hlavnicka J, Tykalova T, *et al*. Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng* 2018;26:1495–507.
16 Goetz CG, Tilley BC, Shaftman SR, *et al*. Movement disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and Clinimetric testing results. *Mov Disord* 2008;23:2129–70.
17 Arora S, Venkataraman V, Zhan A, *et al*. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. *Parkinsonism Relat Disord* 2015;21:650–3.
18 Espay AJ, Bonato P, Nahab FB, *et al*. Technology in Parkinson's disease: challenges and opportunities. *Mov Disord* 2016;31:1272–82.
19 Artusi CA, Mishra M, Latimer P, *et al*. Integration of technology-based outcome measures in clinical trials of Parkinson and other neurodegenerative diseases. *Parkinsonism Relat Disord* 2018;46 Suppl 1:S53–6.
20 Zhan A, Mohan S, Tarolli C, *et al*. Using Smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. *JAMA Neurol* 2018;75:876–80.
21 Lipsmeier F, Taylor KI, Kilchenmann T, *et al*. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov Disord* 2018;33:1287–97.
22 Arora S, Baig F, Lo C, *et al*. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. *Neurology* 2018;91:e1528–38.
23 Rusz J, Tykalova T, Ramig LO, *et al*. Guidelines for speech recording and acoustic analyses in Dysarthrias of movement disorders. *Mov Disord* 2021;36:803–14.
24 Povey D, Ghoshal A, Boulianne G. *The Kaldi speech recognition toolkit*. Proc IEEE Workshop Autom Speech Recognit Underst, 2011.
25 Schultz T, Vu N, Schlippe T. GlobalPhone: A multilingual text & speech database in 20 languages. In: *2013 Proc IEEE Int Conf Acoust Speech Signal Process*, 2013: 8126–30.
26 Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* 2001;9:504–12.
27 Illner V, Sovka P, Rusz J. Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. *Biomed Signal Process Control* 2020;58:101831.
28 Ho AK, Iansek R, Marigliani C, *et al*. Speech impairment in a large sample of patients with Parkinson's disease. *Behav Neurol* 1999;11:131–7.
29 Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Hear Res* 1969;12:246–69.
30 Jannetts S, Lowit A. Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures. *J Voice* 2014;28:673–80.
31 Rusz J, Cmejla R, Tykalova T, *et al*. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. *J Acoust Soc Am* 2013;134:2171–81.
32 Hlavnička J, Čmejla R, Tykalová T, *et al*. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci Rep* 2017;7:12.
33 Beltrami D, Gagliardi G, Rossini Favretti R, *et al*. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci* 2018;10:369.
34 Laganas C, Iakovakis D, Hadjidimitriou S. Parkinson's Disease Detection Based on Running Speech Data From Phone Calls. *IEEE Trans Biomed Eng* 2021
35 Rusz J, Cmejla R, Ruzickova H, *et al*. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J Acoust Soc Am* 2011;129:350–67.
36 Novotny M, Rusz J, Cmejla R. Automatic Evaluation of Articulatory Disorders in Parkinson's Disease. *IEEE/ACM Trans Audio Speech Lang Process* 2014;22:1366–78.
37 American Academy of Sleep Medicine. International Classification of Sleep Disorders. In: *Diagnostic and coding manual*. Third ed. Chicago, Illinois: American Academy of Sleep Medicine, 2014.
38 Postuma RB, Berg D, Stern M, *et al*. MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord* 2015;30:1591–601.
39 Nasreddine ZS, Phillips NA, Bédirian V, *et al*. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005;53:695–9.
40 Visser M, Marinus J, Stiggelbout AM, *et al*. Assessment of autonomic dysfunction in Parkinson's disease: The SCOPA-AUT. *Mov Disord.* 2004;19:1306–12.
41 Beck A, Steer R, Brown G. *Manual for the Beck depression Inventory-II*. San Antonio, TX: Psychological Corporation, 1996.
42 Paolini Paoletti F, Gaetani L, Parnetti L. The Challenge of Disease-Modifying Therapies in Parkinson's Disease: Role of CSF Biomarkers. *Biomolecules* 2020;10:335.
43 Espay AJ, Schwarzschild MA, Tanner CM, *et al*. Biomarker-driven phenotyping in Parkinson's disease: a translational missing link in disease-modifying clinical trials. *Mov Disord* 2017;32:319–24.

2.7 Capturing speech impairments in prodromal PD using a remote, automated approach

The following section describes a conducted cross sectional study in 22 iRBD, 25 PD and 25 control participants using the developed data acquisition application described in section 2.6 and methods developed or evaluated in sections 2.2 through 2.5. The study is currently in a peer review (Movement Disorders, Q1).

Over a period of up to three months following the in-person clinic visit and assessment, data from voice calls were collected from participants, along with periodic recordings of reading passages. The study sought to assess the reliability of passively collected acoustic speech features via everyday smartphone calls for detecting prodromal parkinsonism in individuals with iRBD. It aimed to compare the sensitivity of passive voice monitoring with that of active speech tasks performed using smartphones at home and with a professional microphone in clinical laboratory settings. Additionally, the study aimed to determine the necessary sample duration to achieve optimal sensitivity for detecting prodromal parkinsonism through smartphone-captured speech in real-world conditions.

During the three months of data collection, a total of 3525 calls (mean 49.0, SD 61.1 per participant) were recorded and analyzed. From these, 5990 minutes (mean 83.2, SD 119.7 per participant) of preprocessed speech were extracted for the analysis. On average, one call contained 2.26 minutes (SD 1.96) of preprocessed speech useful for analysis. Considering active assessment, 950 (mean 13.2, SD 7.0 per participant) reading tasks were acquired. 18 minutes of speech (corresponding to approximately 9 calls) was found sufficient to capture prodromal voice changes in-the-wild using smartphones. Interestingly, the results suggested that the higher the severity of dysarthria, the less data is needed. Among the most prominent features of iRBD were monopitch in reading passage and imprecise vowel articulation in phone calls. The combination of passive and active smartphone data captured distinct yet complementary voice information, reaching a high area under curve of 0.85 between iRBD and controls and 0.86 between PD and controls.

The study is the first to evaluate speech characteristics collected in-the-wild in individuals with iRBD and early PD. It revealed that voice calls provide prodromal biomarkers of parkinsonism in iRBD with sensitivity levels comparable to or even exceeding those of laboratory examination using high-quality equipment. Enhancing sensitivity through a combination with active speech tasks amplifies its potential. The findings endorse the feasibility of employing a fully automated and noise-resistant smartphone-based system for passive speech monitoring in real-world scenarios. In the future, the tool might be broadly applied in neuroprotective trials, neurodegeneration screening deep brain stimulation optimalization, neuropsychiatry, speech therapy, population screening, and beyond. The original manuscript is provided below.

Smartphone voice calls provide early biomarkers of parkinsonism in REM sleep behaviour disorder

Vojtěch Illner,[1] Michal Novotný,[1] Tomáš Kouba,[1] Tereza Tykalová,[1] Michal Šimek,[1] Pavel Sovka,[1] Jan Švihlík,[1,2] Evžen Růžička,[3] Karel Šonka,[3] Petr Dušek,[3] Jan Rusz[1,3*]

[1]Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[2]Faculty of Chemical Engineering, University of Chemistry and Technology, Prague, Czech Republic
[3]Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University, Prague, Czech Republic

***Corresponding author address:**
Jan Rusz, MSc, Ph.D., Department of Circuit Theory, Czech Technical University in Prague, Technická 2, 160 00, Praha 6, Czech Republic.
E-mail: rusz.mz@gmail.com
Phone: (+420) 224-352-287
Fax: (+420) 224-311-081

Word count: 3594
Abstract count: 250

Running title: Smartphone calls screening in iRBD.

Keywords: Prodromal synucleinopathy biomarker; Parkinson's disease; speech; wearables; machine learning.

Financial Disclosure/Conflict of Interest: All authors report no conflict of interest concerning the research related to the manuscript.

Funding sources for study: This study was supported by the Czech Ministry of Health (grants no. NU20-08-00445 and MH CZ–DRO-VFN64165), National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107) - Funded by the European Union – Next Generation EU, and by the Czech Technical University in Prague (grant no. SGS23/170/OHK3/3T/13).

Abstract

Background: Speech dysfunction represents one of the initial motor manifestations to develop in Parkinson's disease (PD) and is measurable through smartphone.

Objective: To develop a fully automated and noise resistant smartphone-based system that can unobtrusively screen for prodromal parkinsonian speech disorder in subjects with isolated rapid eye movement sleep behaviour disorder (iRBD) in a real-world scenario.

Methods: This cross-sectional study assessed regular, everyday voice calls data from individuals with iRBD compared to early PD and healthy controls via a developed smartphone application. The participants also performed an active, regular reading of a short passage on their smartphone. Smartphone data was continuously collected for up to three months after the standard in-person assessments at the clinic.

Results: A total of 3525 calls that led to 5990 minutes of preprocessed speech were extracted from 73 participants, including 22 iRBD, 25 PD and 25 controls. With a high area under curve of 0.85 between iRBD and controls, the combination of passive and active smartphone data provided a comparable or even more sensitive evaluation than laboratory examination using a high-quality microphone. The most sensitive features to induce prodromal neurodegeneration in iRBD were monopitch in reading ($p=0.05$) and imprecise vowel articulation in phone calls ($p=0.03$). Eighteen minutes of speech corresponding to approximately nine calls were optimal to obtain the best sensitivity for the screening.

Conclusion: We consider the developed tool widely applicable to dense longitudinal digital phenotyping data with future applications in neuroprotective trials, deep brain stimulation optimization, neuropsychiatry, speech therapy, population screening, and beyond.

INTRODUCTION

Early recognition of Parkinson's disease (PD) has crucial implications for the future development of neuroprotective therapy, as prodromal stages of the disease offer the best opportunity to intervene.[1–3] Therefore, establishing a suitable biomarker effective in prodromal stages would be a game-changing milestone that would impact the diagnosis and future treatments of PD.[4] Isolated rapid eye movement sleep behaviour disorder (iRBD) is now considered an essential prodromal stage of synucleinopathies; patients develop into an overt neurodegenerative disease, particularly PD or dementia with Lewy bodies, after a decade or more.[5–8] Such a long prodromal window provides a unique opportunity to study disease development and design suitable biomarkers.

With the emergence of digital health, there is the potential to remotely and non-invasively detect and track early signs of PD using tools such as smartphones.[9–13] However, many available tests, such as finger tapping or walking a predefined distance, rely on active, instructed involvement,[14] while an ideal digital biomarker should be passively measured without additional effort on the side of the investigated subject or investigator. In this regard, speech analysis has intriguing potential advantages as a large part of the population speaks through smartphones daily. Thus, extracting speech patterns from smartphone calls in a real-world setting has a unique opportunity to provide a passive biomarker, allowing us to continuously measure the effectiveness of experimental treatments in a natural environment, as well as the possibility of large-scale screening.

Since speech represents the most complex quantitative marker of motor function that is highly sensitive to damage to neural structures,[15] it is unsurprising that speech dysfunction has been found to be one of the first motor signs to develop in PD.[16] Specifically, dysprosody and imprecise vowel articulation have been detected in iRBD subjects with impaired olfactory function but still largely functional nigrostriatal dopaminergic transmission,[17,18] that is, in Braak stage 2 before the substantia nigra is affected by synucleinopathy.[19] Unfortunately, these findings are based on actively performed speech recordings obtained using a professional condenser microphone in laboratory settings, which considerably limits the broader applicability of speech assessment.[20] Several challenges must be overcome, including typically lower quality of smartphone microphone, background noise in everyday environments, and the unstable direction and distance of the smartphone from the lips due to various holding positions, to allow passive smartphone speech monitoring.[21,22]

We developed a fully automated and noise-resistant smartphone-based system that can unobtrusively monitor speech in a real-world scenario. We aimed to (i) test the reliability of passively obtained acoustic speech features via everyday smartphone calls to detect prodromal PD in subjects with iRBD, (ii) compare the sensitivity of passive voice monitoring with active speech tasks performed using smartphones at home and professional microphone in the laboratory settings, and (iii) estimate the necessary sample length to reach the optimal sensitivity for the detection of prodromal parkinsonism through speech in a real world setting.

SUBJECTS AND METHODS

Study Design and Participants

From 2021 to 2023, we enroled native Czech iRBD, de novo PD, and healthy control subjects. Patients with iRBD were diagnosed according to the diagnostic criteria of the third edition of the International Classification of Sleep Disorders, including video polysomnography.[23] The exclusion criteria for iRBD were: (i) iRBD onset before 50 years of age; (ii) history of therapy with antiparkinsonian medication, and (iii) iRBD onset within 12 months of introduction of antidepressant treatment. The PD patients were diagnosed based on the Movement Disorder Society clinical diagnostic criteria for PD.[24] The exclusion criteria for PD were: (i) disease duration from diagnosis ≥5 years, (ii) current involvement in any speech therapy and (iii) not on a stable dose of medication over the previous 4 weeks prior start of the study. Exclusion criteria for healthy controls were history of parasomnias or other sleep disorders in adulthood, neurological disorders, or the diagnosis of iRBD on video polysomnography. The exclusion criteria for all groups included: (i) history of communication disorders unrelated to parkinsonism (i.e., problems in speech comprehension or expression) or other neurological disorders, and (ii) unwillingness to achieve at least 10 minutes of phone calls in a month.

The clinical evaluation of each subject included the following: (i) medical history, history of drug and substance intake, and current drug usage; (ii) quantitative testing of motor and nonmotor symptoms of PD with the Movement Disorders Society-Unified Parkinson's Disease Rating Scale, Parts III (MDS-UPDRS);[25] (iii) cognitive testing with the Montreal Cognitive Assessment (MoCA);[26] and (iv) autonomic testing with the Scales for Outcomes in Parkinson's Disease–Autonomic Dysfunction scale.[27] Perceptual speech severity was estimated using the speech item score from the MDS-UPDRS, Part III. Symptom duration was estimated based on the self-reported first occurrence of iRBD/PD symptoms.

Each participant provided written informed consent. The study was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic, in accordance with the ethical standards established in the 1964 Declaration of Helsinki.

Smartphone speech examination

Each subject received a smartphone with preinstalled application,[28] which worked in the background and recorded subject's voice during calls, removing the content from the distant speaker by adaptive filtering (**Figure 1A**). After each incoming or outgoing call, the user was prompted with a screen containing an option to delete the call or send it to confidential analysis. Upon agreement, the audio recording was kept on the device for 24 hours to allow participants to replay and eventually delete it. After 24 hours, the recording was sent to a secure server and validated by a speaker recognition framework. Comprehensive technical details of the application were previously described in the protocol.[28] In addition, the application contained an active part. Subjects were prompted to read a passage twice, selected randomly from six samples of approximately 80 words, displayed on the application screen every 14 days (mean duration 35.1, SD 5.5 seconds). All data was collected from the smartphone during a period of up to three months after the clinic visit. Acquisition and secure data transfer have been carried out in accordance with the directive on the legislation on personal data protection of the European Union.

All subjects possessed HONOR 9X Lite (Shenzhen Zhixin New Information Technology) phone, operating on the Android V.9 system. The phone was chosen as a mainstream product (i.e., commercially available and a relatively inexpensive mid-range smartphone) among the smartphones available in Czechia in 2021. The recordings were sampled at 44.1 kHz with 16-bit quantification.

Laboratory speech examination

Speech recordings were performed in a quiet room with a low ambient noise level using a high-quality head-mounted condenser microphone (Beyerdynamic Opus 55, Heilbronn, Germany) placed approximately 5 cm from the subject's mouth. Speech signals were sampled at 48 kHz with 16-bit quantification. Each subject was recorded during a single session accompanied by a speech specialist who guided the standardized protocol. Participants were instructed to present a monologue about an arbitrary topic of at least 90 seconds (mean duration 123.6, SD 19.3 seconds) and perform a reading passage task twice of a standardized text of 80 words (mean duration 35.7, SD 5.1 seconds).

Smartphone calls preprocessing

The incoming calls contained non-speech periods with no relevant information due to the dialogue nature of a conversation on the phone. Hence, the recordings were stripped of any non-speech segments longer than 0.7 seconds. The threshold was set to preserve natural pauses as a significant aspect of speech production. Subsequently, to normalize the calls in terms of duration, the recordings were partitioned into time frames of equal length, each treated as an individual recording. The frame length was chosen as 20, 30, 45, and 60 seconds to evaluate the impact of different durations. If there was a remainder, it was considered only if longer than 50% of the corresponding segment length (e.g., if a 20 second window was selected for a 32 second call, both 20 and 12 second segments were analysed).

Acoustic speech features

We selected 7 representative acoustic speech features (**Figure 1B**), following three main criteria: (i) representing a unique aspect of speech (the features were found to be only weakly correlated [Pearson: $|r| < 0.48$]), aligning with the perceptual description of the primary patterns of hypokinetic dysarthria.[29], (ii) enabling automated analysis of connected speech, (iii) proven sensitivity in iRBD or early PD in previous studies.[18,30] We limited the number of acoustic parameters included in the experiment to reduce the probability of a Type I error and to reduce potential overfitting for the regression analysis.

Monopitch was assessed by a standard deviation of pitch contour (F0sd),[21] imprecise vowel articulation by a formant ratio index (FRI),[31] voice quality was captured by cepstral peak prominence (CPP),[22] articulatory decay by a standard deviation of mel-frequency cepstral coefficients (global MFCC) ,[32] monoloudness by a standard deviation of intensity contour after removal of pauses (INTsd),[33] prolonged pauses by median duration of pause intervals (DPI),[34] and articulation rate through net speech rate (NSR) acquired via automatic speech recognizer

followed by hyphenation.[35,36] All analyses were performed in MATLAB (MathWorks, Natick, MA) and Python.

Speech sample length estimation

A unified, sufficient speech sample from each participant as well as the most optimal call frame duration (e.g., whether to analyse 10 or 30 minutes of cumulative calls per participant, in 20 or 30 second frames) was determined. A binary logistic regression followed by a leave-one-out cross-validation using a combination of all acoustic features was utilized to determine the classification accuracy. The speech sample was then chosen based on group classification accuracy, at the point when the average accuracy across call frames reached 95% of its maximum value in the cumulative analysed interval. The largest sample length across the three classifications (controls vs. iRBD, controls vs. PD, iRBD vs. PD) was chosen for the statistical analysis. The effect of call frames length was assessed based on accuracy of the selected speech sample. In active speech assessment, the number of reading tasks required for analysis was determined analogously to sample length determination via calls.

Statistical analysis

An ad hoc power analysis for a given large effect size (Cohen's d of 0.8), with the Type I error probability (α) set at 0.05 and power of 80%, based on a three-group analysis of variance with one covariate (group), determined a minimum sample size of 66 subjects (i.e., 22 per group). A one-way analysis of variance with Bonferroni post-hoc test was applied to analyse group differences. The relationships between features were evaluated using Spearman correlation coefficient. To assess the sensitivity between groups, a binary logistic regression model followed by leave-one-out cross-validation was utilized. The features used were determined based on an exhaustive search, providing the best outcome across spontaneous speech (calls and laboratory monologue) and reading tasks (smartphone and laboratory) and their combination, and we compared the receiver operating curve along with its area under the curve (AUC).

RESULTS

Collected data

In this single-centre study, of 52 available iRBD subjects, 22 (42%) met the inclusion criteria and were willing to participate. The main reason for rejection to participate was that subjects (i) made only exceptional phone calls, (ii) were unwilling to use smartphone, and (iii) did not like the purpose of the project and/or the need to share personal voice calls. Additionally, we recruited 25 healthy controls and 25 early PD patients (**Table 1**).
During the three months of data collection, a total of 3525 calls (mean 49.0, SD 61.1 per participant) were recorded and analyzed. From these, 5990 minutes (mean 83.2, SD 119.7 per participant) of preprocessed speech were extracted for the analysis. On average, one call

contained 2.26 minutes (SD 1.96) of pre-processed speech useful for analysis. Considering active assessment, 950 (mean 13.2, SD 7.0 per participant) reading tasks were acquired.

Table 1. Clinical data of the participants.

	controls (n=25)	**iRBD** (n=22)	**PD** (n=25)	***p* value**
Men	24 (96%)	21 (95%)	24 (96%)	0.99
Age (yr)	67.1/7.3 (55-84)	68.3/8.6 (53-86)	58.5/8.6 (45-76)	<0.001[b,c]
Symptom duration (yr)	-	10.3/6.7 (2-29)	5.5/2.1 (2-11)	-
MDS-UPDRS III total	6.5/2.7 (2-11)	9.8/2.5 (5-15)	25.9/9.8 (10-51)	<0.001[b,c]
MDS-UPDRS III speech item	0.3/0.5 (0-1)	0.4/0.5 (0-1)	1.0/0.3 (0-2)	<0.001[b,c]
MoCA	26.2/2.6 (22-30)	25.9/2.2 (21-30)	26.8/2.8 (18-30)	0.46
SCOPA-AUT	7.3/5.1 (1-24)	13.0/8.7 (3-39)	10.0/6.0 (1-24)	<0.05[d]
Antidepressant therapy	1 (4%)	2 (9%)	4 (16%)	0.40
Levodopa equivalent (mg/day)	0	0	621.3/329.5 (0-1440)	<0.001[b,c]
Clonazepam therapy (mg/day)	0	0.2/0.2 (0-0.5)	0	<0.001[c,d]
RBD presence[a]	0 (0%)	22 (100%)	10 (40%)	<0.001[b,c,d]

Captions: Data are the mean/SD (range) or the number (%).
[a]Presence of RBD was diagnosed by videopolysomnography.
[b]Significant difference between PD and controls.
[c]Significant difference between iRBD and PD.
[d]Significant difference between iRBD and controls.
iRBD = idiopathic rapid eye movement sleep behavior; PD = Parkinson's disease; MDS-UPDRS = Movement Disorder Society Unified Parkinson's Disease Rating Scale; MoCA = Montreal Cognitive Assessment; SCOPA-AUT = Scales for Outcomes in Parkinson's Disease-Autonomic Dysfunction.

Speech sample length estimation

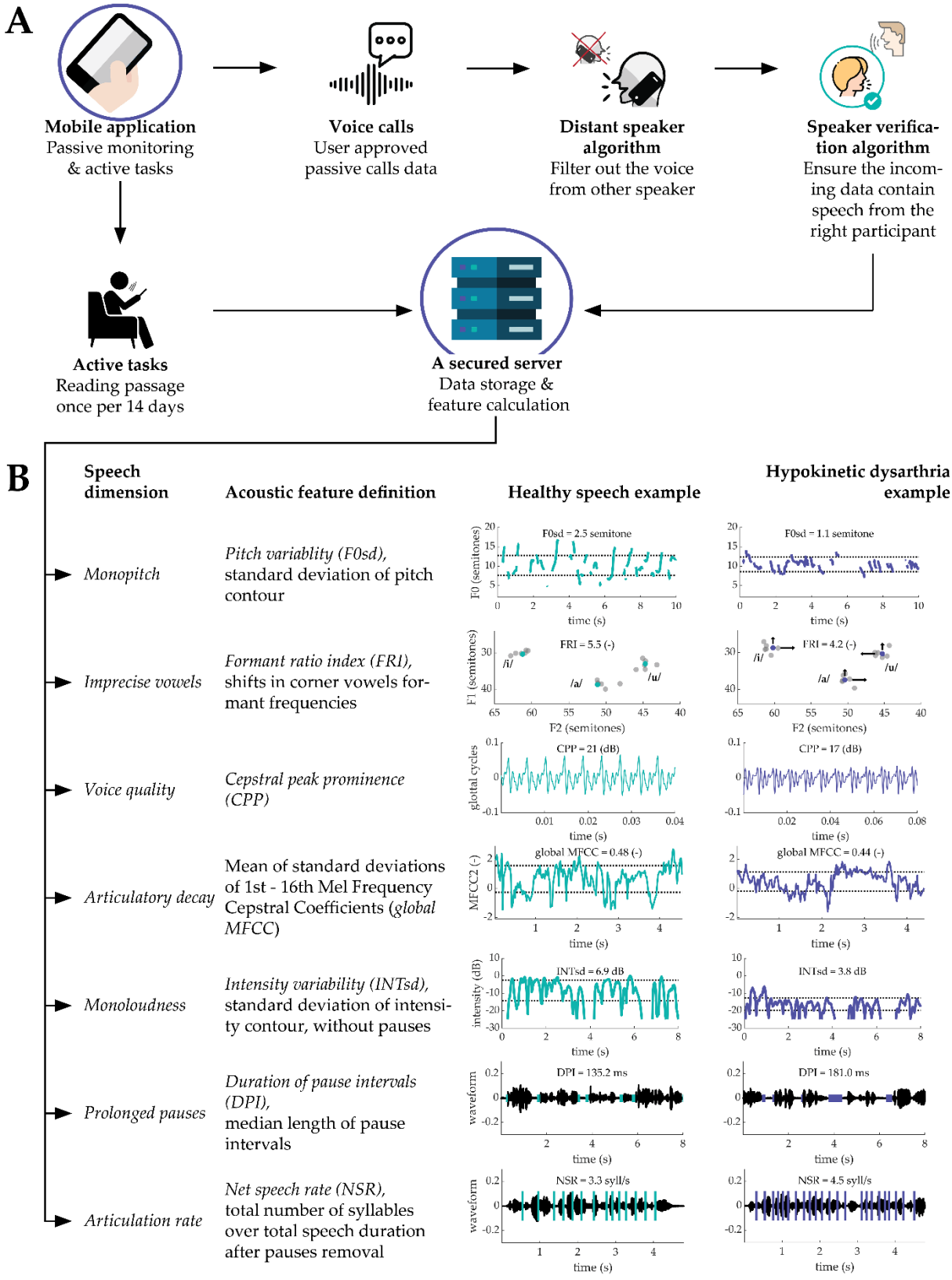
For smartphone calls, the best accuracy reached the threshold at 15 minutes in differentiating between PD and controls, 18 minutes between iRBD and controls, and 3 minutes between iRBD and PD (**Figure 2A**). However, from 8 minutes of sample duration, the performance between iRBD and controls remained stable and continuously increased. Regarding call frame durations, no specific option exhibited notable advantages in accuracy. Since the 20-second time frame provides enhanced flexibility given its brief duration, the analysis was carried out using 18 minutes of calls, pre-processed in 20-second frames.

For active reading tasks, the threshold was reached in 2 tasks (one trial) in differentiating between PD vs. controls, 3 tasks between iRBD and controls, and 1 task between iRBD and PD (**Figure 2B**). As a result, an average of 3 reading tasks were considered for the analysis.

Speech differences

The features demonstrating statistically significant impairment in iRBD and controls were monopitch in laboratory reading task (p=0.049) and imprecise vowels in calls (p=0.03) (**Figure 3**). Compared to controls, PD impairment was captured in monopitch in laboratory reading task (p<0.001), imprecise vowels in calls (p=0.01), articulatory decay in laboratory monologue

Figure 1. The principal speech analysis scheme.



A) Illustrative diagram of the smartphone data acquisition system. B) Illustrative table of speech dimensions described in the study, their definition, and example of healthy and dysarthric speaker.

($p<0.001$), prolonged pauses in laboratory monologue ($p<0.001$), and increased articulation rate in calls ($p=0.01$) and both reading tasks ($p<0.001$). Voice quality and monoloudness did not reach significance between the groups. No significant relationships were observed between individual acoustic features and MDS-UPDRS part III and MoCA in PD, iRBD, or controls.

Correlations among data from different sources

Between calls and laboratory monologue, a high correlation coefficient was achieved only in imprecise vowels ($r=0.67$, $p<0.001$) (**Figure 3**). Between reading tasks, the correlations were generally stronger, with a high correlation coefficient demonstrated in monopitch ($r=0.70$, $p<0.001$), voice quality ($r=0.66$, $p<0.001$), and articulation rate ($r=0.70$, $p<0.001$).

Sensitivity analysis

Based on the exhaustive search, the optimal combination of features for spontaneous speech was monopitch, imprecise vowels, articulatory decay, prolonged pauses, and articulation rate, while for reading tasks monopitch, articulatory decay, monoloudness, and articulation rate. The best AUC between iRBD and controls was 0.79 via calls compared to an AUC of 0.66 via laboratory monologue (**Figure 4**). Between PD and controls, similar AUCs of up to 0.87 were found for both the smartphone and the laboratory setting. In reading, a better AUC of up to 0.83 was obtained in laboratory settings compared to smartphone in separation between controls and both PD and iRBD. In general, the accuracy of prodromal speech disorder detection via smartphone improved to an AUC of up to 0.85 when both passive calls and active reading were combined.

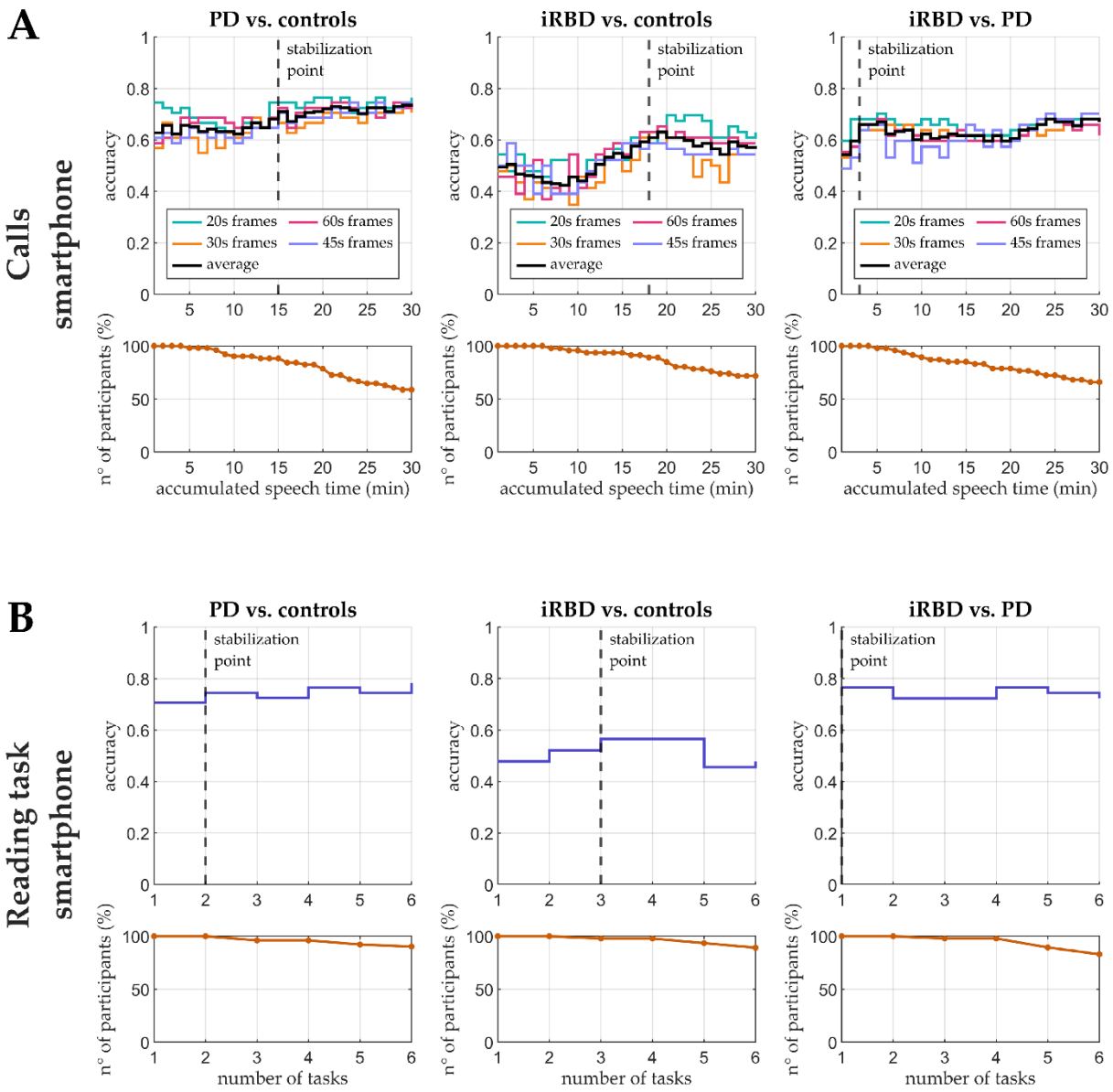
DISCUSSION

The present study is the first to evaluate speech characteristics collected in-the-wild in iRBD and early PD. It revealed that voice calls provide prodromal biomarkers of parkinsonism in iRBD at a level comparable to or even more sensitive than laboratory examination with high-quality equipment. The combination of passive and active smartphone data captured distinct yet complementary voice information, reaching a high AUC of 0.85 between iRBD and controls. Among the most prominent features of iRBD were monopitch in reading and imprecise vowel articulation in phone calls. Our findings support the possibility of employing a fully automated and noise-resistant smartphone-based system that can passively monitor speech in real-world scenarios for future clinical trials.

Speech sample length

The effect of speech sample length on biomarker performance has rarely been investigated. Although not systematically researched, a previous study suggested that 50 smartphone call sessions lasting between 15 and 75 seconds (corresponding to approximately 13 to 65 minutes) are sufficient to detect PD-related speech impairment.[37] The amount is greater than in the current study, where we found 18 minutes of speech (corresponding to approximately 9 calls)

Figure 2. Sample length estimation.



Accuracy of a binary logistic regression followed by a leave-one-out cross-validation using a combination of all the acoustic features for increasing sample size of (A) speech from calls for different frame durations and their average value and (B) smartphone reading tasks. The dashed line corresponds to the point when the average accuracy across call frames reached 95% of its maximum value in the cumulative analysed interval. Below each plot is a percentage of participants able to reach such a sample size. Captions: iRBD - isolated rapid eye movement sleep behaviour disorder, PD – Parkinson's disease.

optimal to capture prodromal voice changes in-the-wild using smartphones. Interestingly, it takes a lesser amount to capture voice impairment in PD, suggesting that the higher the severity of dysarthria, the less data is needed. However, stable but lower accuracy for separating between iRBD and controls was achieved already for 8 minutes of calls.

Considering active smartphone data collection, three reading tasks are sufficient to fully capture prodromal voice characteristics in iRBD, demonstrating that guided tasks require a smaller sample size for effective analysis.[38] This is principally in agreement with previous study showing that at least 120 words are necessary to obtain stable results during reading in

controlled settings.[39] In general, the need for a high-quality of microphone can be replaced by a longer sample size.

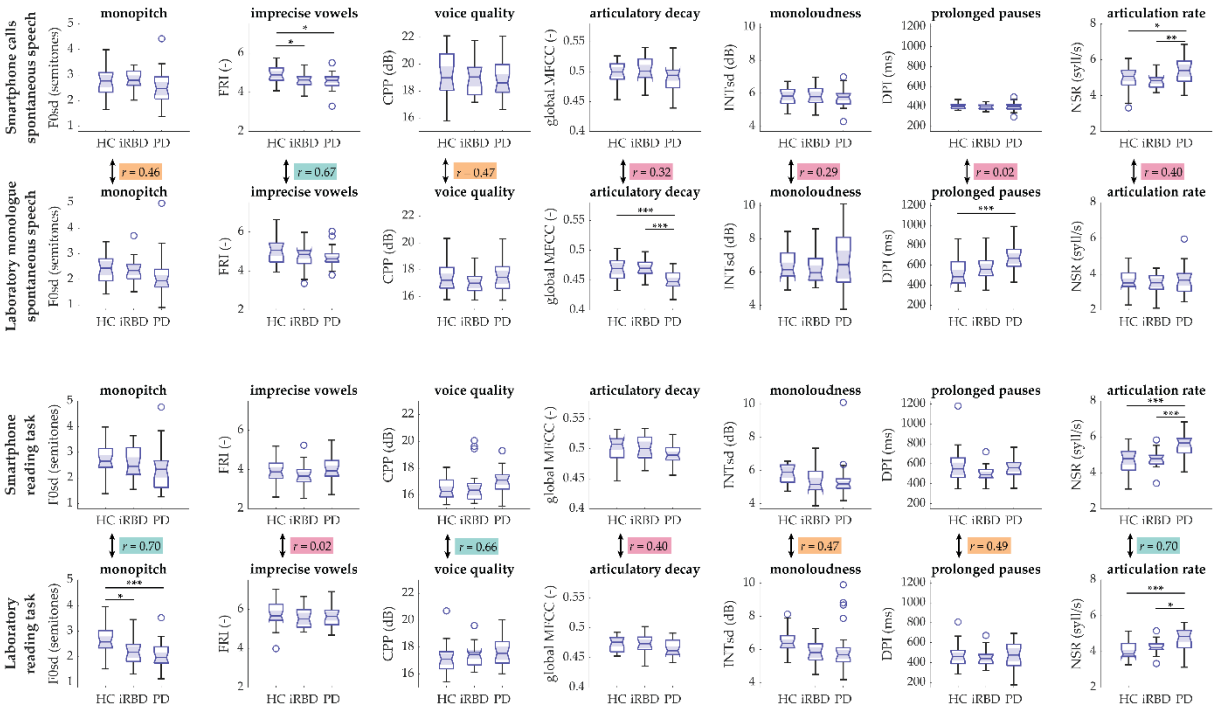
Speech biomarkers

In agreement with previous studies,[17,18,30] speech disorder in iRBD was mainly characterized by monopitch and imprecise vowel articulation. The novel observation is that vowel articulation was particularly affected during spontaneous speech, while intonation was reduced only during reading. This behaviour can likely be explained by different compensatory mechanisms involved.[40] The low intonation pattern that is admissible during a reading of a prepared neutral passage is likely compensated in dialogue to make the speech more compelling for second-side listeners. On the contrary, deficits in internal cueing specific for PD might lead to higher intelligibility in prepared utterances compared to spontaneous speech.[41] The quality of vowel articulation is highly related to intelligibility.[42] Since vowel articulation is a demanding process of articulatory coordination and intelligibility is preserved in the early stages of the disease, it may be difficult or unnecessary for patients to compensate during spontaneous speech. Considering our early PD cohort, the observed trend for worse voice quality, articulatory decay, and prolonged pauses is consistent with previous literature.[30,43] Interestingly, spontaneous speech assessment during phone calls led to the finest sensitivity in increased speech rate in PD, which is presumably a precursor of oral festination.[44] Contributing to palilalia, this is one of the most debilitating and challenging symptoms to assess with no available therapies,[45] leading to social isolation and degradation of interpersonal interactions. Since laboratory speech material is typically short and not representative of everyday situations, it might not be sufficient for advanced analyses. Therefore, spontaneous speech evaluation through calls in the natural environment may provide a novel way to identify markers to predict which patients develop events such as oral festination, potentially leading to better personalized therapies.

Effect of smartphone assessment on individual speech biomarkers

The characteristics of the microphone, environmental noise, position of the microphone, and hardware filtering can all influence the robustness of speech assessment.[46] Many relationships were still surprisingly strong considering that smartphones and laboratory microphone recordings were not done in parallel but at different times and situations. In accordance with previous research,[21] the acoustic measurement of fundamental frequency variability reflecting monopitch demonstrated high resistance against device effect. This is likely due to the nature of the fundamental frequency, which represents a major trend in the frequency domain of a speech signal, and thus can be detected accurately despite the influence of detrimental factors. Imprecise vowels, reflecting the position of resonant frequencies (so-called formants),[31] represent another frequency measure that had good robustness to analysis via smartphone. The voice quality measure was unsurprisingly robust only in a controlled environment without substantial noise.[22] Articulatory decay calculated from MFCCs represents, in principle, an amplitude measure. It demonstrated little resistance against the device effect, as the coefficients tend to be impacted by microphone position and type,[47] and, therefore, is unsuitable for phone screening. Another amplitude measure, monoloudness, was robust only in reading text,

Figure 3. Group differences between speech features in individual tasks and correlations of features between tasks.



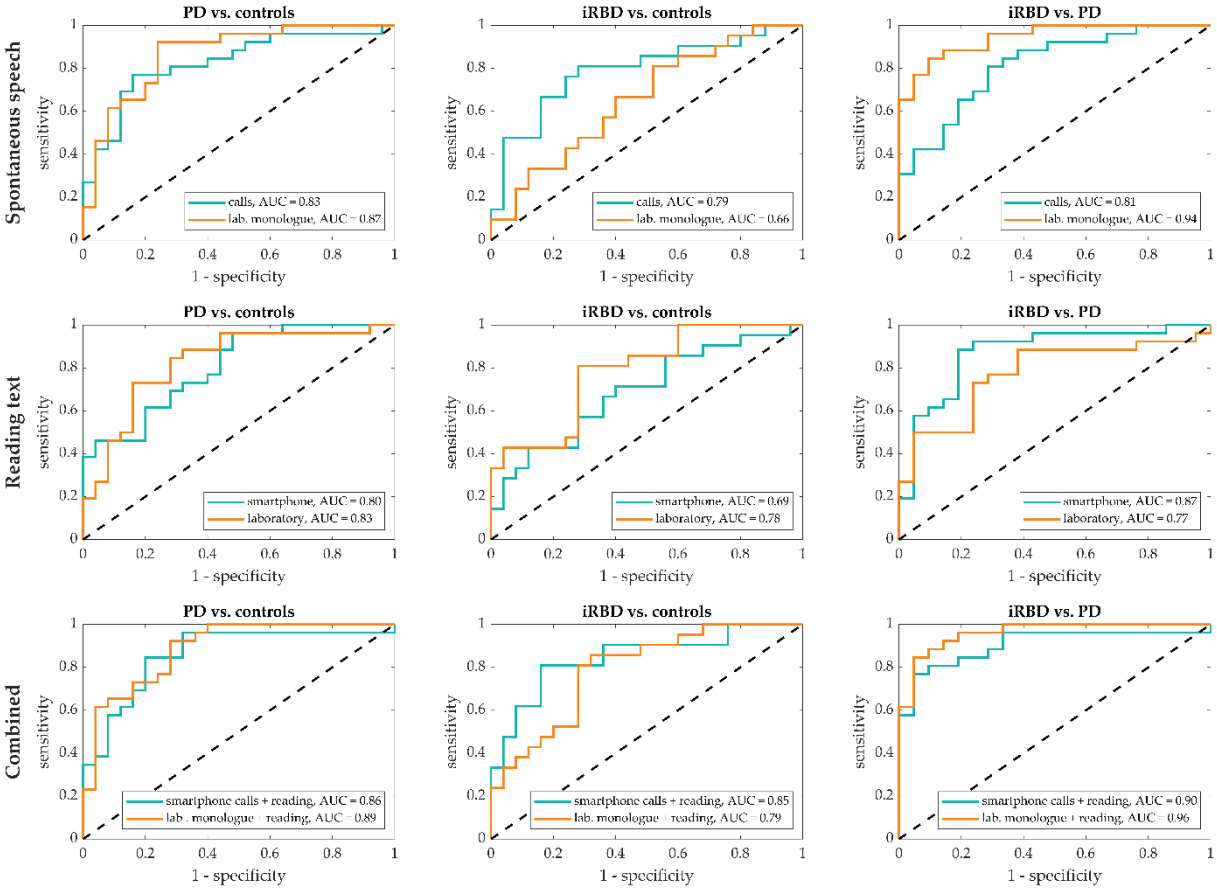
The calls measure is taken from 18 minutes of speech. Horizontal line represents median, the box lower and upper quartiles, the bars minimum and maximum values that are not outliers, the circles outliers. ***, **, and * represent significant differences with $p < 0.001$, 0.1, and 0.5, respectively, after Bonferroni adjustment. Captions: HC – healthy controls, iRBD - isolated rapid eye movement sleep behaviour disorder, PD – Parkinson's disease, r – Spearman correlation coefficient.

probably because of varying conditions in calls. Due to the dialogue nature of calls, pauses cannot be directly compared to those from uninterrupted monologue. In reading, pauses were moderately correlated between the smartphone and the high-quality microphone, which could be due to the insufficient accuracy of speech-pause detection.[34] Articulation rate, calculated as the number of syllables per time, reached high reliability between both devices, indicating the high robustness of the automatic speech-to-text transcription independent of the microphone quality.[35]

Strengths and limitations

Some participants struggled to reach enough speech data from calls, likely due to not sharing all the speech calls or older age leading to potential problems operating the smartphone. However, most of the participants achieved at least the estimated 18 minutes of call speech for optimal sensitivity to detect prodromal voice impairment. In the future scenario, smartphone skills are likely to be widespread among the older population. Additionally, the software can be implemented directly on the smartphone and immediately process a given recording after completion, computing selected features as anonymised numbers. Thus, no audio transfer will be necessary, with only speech features' values stored, thus maintaining the maximum level of privacy.

Figure 4. Sensitivity analysis.



Receiver operating characteristics curves of a binary logistic classification of optimal features based on an exhaustive search. For spontaneous speech, the optimal features were monopitch, imprecise vowels, articulatory decay, prolonged pauses, and articulation rate, whereas for reading task monopitch, articulatory decay, monoloudness, and articulation rate. The calls measure is taken from 18 minutes of speech. Captions: iRBD - isolated rapid eye movement sleep behaviour disorder, PD – Parkinson's disease, AUC – area under curve.

The detection accuracy between the iRBD and PD groups was higher than between the controls and iRBD. Furthermore, our iRBD participants were similar in cognitive performance compared to controls, while the presence of RBD in PD is typically associated with a more impaired cognitive profile.[48] We can thus hypothesise that iRBD participants willing to participate in this study were rather those further away from the disease phenoconversion. This would mean that our diagnostic accuracy of AUC 0.85 between iRBD and controls has the potential to improve. In particular, a very similar diagnostic accuracy of AUC 0.86 was observed between PD and controls. This could be associated with the fact that all PD patients were on stable dopaminergic therapy, which has been shown to ameliorate several speech manifestations in the early stages of the disease.[49] Furthermore, we were unable to recruit enough older PD volunteers with less than 5 years of disease duration, resulting in a 10-year younger PD group on average than the iRBD and controls groups. The inclusion of an older control group likely also negatively affected the reported accuracy of the PD diagnostics.

CONCLUSION

This study has revealed that phone calls provide a novel passive biomarker of prodromal and early parkinsonism and has established a pipeline for the capture of speech biomarkers in real-world settings. Enhancing sensitivity through a combination with active speech tasks amplifies its potential. In the future, our tool might be broadly applied in neuroprotective trials, deep brain stimulation optimalization, neuropsychiatry, speech therapy, population screening, and beyond. Future longitudinal studies should aim to validate the efficacy of phone calls analysis in tracking disease progression.

Authors' Roles

1) Research project: A. Conception, B. Organization, C. Execution; 2) Statistical analysis: A. Design, B. Execution, C. Review and Critique; 3) Manuscript: A. Writing of the first draft, B. Review and Critique.

Vojtěch Illner: 1A, 1B, 1C, 2A, 2B, 3A; ; Michal Novotný: 1B, 1C, 2C, 3B; ; Tomáš Kouba: 1B, 1C, 2C, 3B; Tereza Tykalova: 1B, 1C, 1B, 2C, 3B; Michal Šimek: 1C, 2C, 3B; Jan Švihlík: 1C, 2C, 3B; Evzen Ruzicka: 1C, 2C, 3B; Karel Šonka: 1C, 2C, 3B; Petr Dušek: 1B, 1C, 2C, 3B; Jan Rusz: 1A, 1B, 1C, 2A, 2C, 3A.

Financial Disclosures of all authors

V.I. received funding from the Czech Ministry of Health and Czech Technical University in Prague. M.N. received funding from the Czech Ministry of Education, Czech Ministry of Health and Czech Technical University in Prague. T.K received funding from the Czech Ministry of Education and Czech Technical University in Prague. T.T. received funding from the Czech Ministry of Education, Czech Ministry of Health, and Czech Technical University in Prague. M.Š. received funding from the Czech Ministry of Education, Czech Ministry of Health and Czech Technical University in Prague. P.S. received funding from the Czech Ministry of Health and Czech Technical University in Prague. J.Š received funding from the Czech Technical University in Prague and University of Chemistry and Technology. E.R received funding from the Czech Ministry of Education, Czech Ministry of Health and Charles University and General University Hospital. K.Š received funding from the Czech Ministry of Education, Czech Ministry of Health and Charles University and General University Hospital. P.D received funding from the Czech Ministry of Education, Czech Ministry of Health, and Charles University and General University Hospital. J.R. received funding from the Czech Ministry of Education, Czech Ministry of Health, Czech Technical University in Prague and Charles University and General University Hospital.

REFERENCES

1. Schenck CH, Boeve BF, Mahowald MW. Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder. *Sleep Med*. 2013;14(8):744-748. https://linkinghub.elsevier.com/retrieve/pii/S1389945712003814

2. Postuma RB, Gagnon JF, Bertrand JA, Génier Marchand D, Montplaisir JY. Parkinson risk in idiopathic REM sleep behavior disorder: preparing for neuroprotective trials. *Neurology*. 2015;84(11):1104-1113. http://www.neurology.org/lookup/doi/10.1212/WNL.0000000000001364

3. Videnovic A, Ju YES, Arnulf I, et al. Clinical trials in REM sleep behavioural disorder: Challenges and opportunities. *J Neurol Neurosurg Psychiatry*. 2020;91(7):740-749. doi:10.1136/jnnp-2020-322875

4. Postuma RB, Berg D. Advances in markers of prodromal Parkinson disease. *Nat Rev Neurol*. 2016;12(11):622-634. http://www.nature.com/articles/nrneurol.2016.152

5. Miglis MG, Adler CH, Antelmi E, et al. Biomarkers of conversion to α-synucleinopathy in isolated rapid-eye-movement sleep behaviour disorder. *Lancet Neurol*. 2021;20(issue 8):671-684. https://linkinghub.elsevier.com/retrieve/pii/S1474442221001769

6. Dauvilliers Y, Schenck CH, Postuma RB, et al. REM sleep behaviour disorder. *Nat Rev Dis Primers*. 2018;4(1). doi:10.1038/s41572-018-0016-5

7. Högl B, Stefani A, Videnovic A. Idiopathic REM sleep behaviour disorder and neurodegeneration — an update. *Nat Rev Neurol*. 2018;14(1):40-55. http://www.nature.com/articles/nrneurol.2017.157

8. Fereshtehnejad SM, Yao C, Pelletier A, Montplaisir JY, Gagnon JF, Postuma RB. Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: a prospective study. *Brain*. 2019;142(7):2051-2067. doi:10.1093/brain/awz155

9. Adams JL, Kangarloo T, Tracey B, et al. Using a smartwatch and smartphone to assess early Parkinson's disease in the WATCH-PD study. *NPJ Parkinsons Dis*. 2023;9(1). doi:10.1038/s41531-023-00497-x

10. Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data*. 2016;3. doi:10.1038/sdata.2016.11

11. Lakovakis D, Hadjidimitriou S, Charisis V, et al. Motor impairment estimates via touchscreen typing dynamics toward Parkinson's Disease detection from data harvested in-the-wild. *Frontiers in ICT*. 2018;5. doi:10.3389/fict.2018.00028

12. Papadopoulos A, Kyritsis K, Klingelhoefer L, Bostanjopoulou S, Chaudhuri KR, Delopoulos A. Detecting Parkinsonian Tremor from IMU Data Collected In-The-Wild using Deep Multiple-Instance Learning. *IEEE J Biomed Health Inform*. Published online 2019. doi:10.1109/JBHI.2019.2961748

13. Burton A. Smartphones versus Parkinson's disease: i-PROGNOSIS. *Lancet Neurol*. 2020;19(5):385-386.

14. Arora S, Baig F, Lo C, et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. *Neurology*. 2018;91(16):1528-1538. http://journals.lww.com/00006114-900000000-94055

15. Duffy J. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Vol 4th ed. Fourth. Elsevier; 2019.

16. Postuma RB, Lang AE, Gagnon JF, Pelletier A, Montplaisir JY. How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. *Brain*. 2012;135(6):1860-1870. https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/aws093

17. Rusz J, Janzen A, Tykalová T, et al. Dysprosody in Isolated REM Sleep Behavior Disorder with Impaired Olfaction but Intact Nigrostriatal Pathway. *Movement Disorders*. 2022;37(3):619-623. doi:https://doi.org/10.1002/mds.28873

18. Skrabal D, Rusz J, Novotny M, et al. Articulatory undershoot of vowels in isolated REM sleep behavior disorder and early Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8(1). doi:10.1038/s41531-022-00407-7

19. Braak H, Tredici K Del, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol Aging*. 2003;24(2):197-211. doi:https://doi.org/10.1016/S0197-4580(02)00065-9

20. Brabenec L, Mekyska J, Galaz Z, Rektorova I. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. *J Neural Transm*. 2017;124(3):303-334. http://link.springer.com/10.1007/s00702-017-1676-0

21. Illner V, Sovka P, Rusz J. Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. *Biomed Signal Process Control*. 2020;58. https://linkinghub.elsevier.com/retrieve/pii/S1746809419304124

22. Šimek M, Rusz J. Validation of cepstral peak prominence in assessing early voice changes of Parkinson's disease. *J Acoust Soc Am*. 2021;150(6):4522-4533. https://asa.scitation.org/doi/10.1121/10.0009063

23. American Academy of Sleep Medicine. *International Classification of Sleep Disorders, Third Edition: Diagnostic and Coding Manual*. American Academy of Sleep Medicine; 2014.

24. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*. 2015;30(12):1591-1601. http://doi.wiley.com/10.1002/mds.26424

25. Goetz CG, Fahn S, Martinez-Martin P, et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan. *Movement Disorders*. 2007;22(1):41-47. doi:10.1002/mds.21198

26. Kopecek M, Stepankova H, Lukavsky J, Ripova D, Nikolai T, Bezdicek O. Montreal cognitive assessment (MoCA): Normative data for old and very old Czech adults. *Appl Neuropsychol Adult*. 2017;24(1):23-29. doi:10.1080/23279095.2015.1065261

27. Visser M, Marinus J, Stiggelbout AM, Van Hilten JJ. Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT. *Movement Disorders*. 2004;19(11):1306-1312. https://onlinelibrary.wiley.com/doi/10.1002/mds.20153

28. Kouba T, Illner V, Rusz J. Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies. *BMJ Open*. 2022;12(6). https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2021-059871

29. Darley FL, Aronson AE, Brown JR. Differential Diagnostic Patterns of Dysarthria. *J Speech Hear Res*. 1969;12(2):246-269. http://pubs.asha.org/doi/10.1044/jshr.1202.246

30. Rusz J, Hlavnička J, Novotný M, et al. Speech Biomarkers in Rapid Eye Movement Sleep Behavior Disorder and Parkinson Disease. *Ann Neurol*. 2021;90(1):62-75. https://onlinelibrary.wiley.com/doi/10.1002/ana.26085

31. Illner V, Tykalova T, Skrabal D, Klempir J, Rusz J. Automated Vowel Articulation Analysis in Connected Speech Among Progressive Neurological Diseases, Dysarthria Types, and Dysarthria Severities. *J Speech Lang Hear Res*. 2023;66(8):2600-2621. doi:10.1044/2023_JSLHR-22-00526

32. Illner V, Krýže P, Švihlík J, et al. Which aspects of motor speech disorder are captured by Mel Frequency Cepstral

Coefficients? Evidence from the change in STN-DBS conditions in Parkinson's disease. In: *INTERSPEECH 2023*. ISCA; 2023:5027-5031. doi:10.21437/Interspeech.2023-1744

33. Rusz J, Cmejla R, Ruzickova H, Ruzicka E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J Acoust Soc Am*. 2011;129(1):350-367. http://asa.scitation.org/doi/10.1121/1.3514381

34. Hlavnička J, Čmejla R, Tykalová T, Šonka K, Růžička E, Rusz J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci Rep*. 2017;7(1). http://www.nature.com/articles/s41598-017-00047-5

35. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*. Published online December 6, 2023:1182. http://arxiv.org/abs/2212.04356

36. Illner V, Tykalová T, Novotný M, Klempíř J, Dušek P, Rusz J. Toward Automated Articulation Rate Analysis via Connected Speech in Dysarthrias. *Journal of Speech, Language, and Hearing Research*. 2022;65(4). doi:10.1044/2021_JSLHR-21-00549

37. Laganas C, Iakovakis D, Hadjidimitriou SK, et al. Parkinson's Disease Detection Based on Running Speech Data From Phone Calls. *IEEE Trans Biomed Eng*. 2022;65(9):1573-1584. https://ieeexplore.ieee.org/document/9556632/

38. Rusz J, Tykalova T, Ramig LO, Tripoliti E. Guidelines for Speech Recording and Acoustic Analyses in Dysarthrias of Movement Disorders. *Movement Disorders*. 2021;36(4):803-814. https://onlinelibrary.wiley.com/doi/10.1002/mds.28465

39. Krýže P, Tykalová T, Růžička E, Rusz J. Effect of reading passage length on quantitative acoustic speech assessment in Czech-speaking individuals with Parkinson's disease treated with subthalamic nucleus deep brain stimulation. *J Acoust Soc Am*. 2021;149(5):3366-3374. doi:10.1121/10.0005050

40. Thies T, Mücke D, Geerts N, et al. Compensatory articulatory mechanisms preserve intelligibility in prodromal Parkinson's disease. *Parkinsonism Relat Disord*. 2023;112. doi:10.1016/j.parkreldis.2023.105487

41. Sapir S. Multiple factors are involved in the dysarthria associated with Parkinson's disease: A review with implications for clinical practice and research. *Journal of Speech, Language, and Hearing Research*. 2014;57(4):1330-1343. doi:10.1044/2014_JSLHR-S-13-0039

42. Weismer G, Jeng JY, Laures JS, Kent RD, Kent JF. Acoustic and Intelligibility Characteristics of Sentence Production in Neurogenic Speech Disorders. *Folia Phoniatrica et Logopaedica*. 2001;53(1):1-18. https://www.karger.com/Article/FullText/52649

43. Jeancolas L, Mangone G, Petrovska-Delacrétaz D, et al. Voice characteristics from isolated rapid eye movement sleep behavior disorder to early Parkinson's disease. *Parkinsonism Relat Disord*. 2022;95:86-91. doi:10.1016/j.parkreldis.2022.01.003

44. Moreau C, Ozsancak C, Blatt JL, Derambure P, Destee A, Defebvre L. Oral festination in Parkinson's disease: Biomechanical analysis and correlation with festination and freezing of gait. *Movement Disorders*. 2007;22(10):1503-1506. doi:10.1002/mds.21549

45. Helm Nancy. Management of Palilalia With a Pacing Board. *Journal of Speech and Hearing Disorders*. 1979;44(3):350-353. doi:10.1044/jshd.4403.350

46. Kardous CA, Shaw PB. Evaluation of smartphone sound measurement applications. *J Acoust Soc Am*. 2014;135(4):186-192. doi:10.1121/1.4865269

47. Rusz J, Novotny M, Hlavnicka J, Tykalova T, Ruzicka E. High-Accuracy Voice-Based Classification between Patients with Parkinson's Disease and Other Neurological Diseases May Be an Easy Task with Inappropriate Experimental Design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2017;25(8):1319-1321. doi:10.1109/TNSRE.2016.2621885

48. Jozwiak N, Postuma RB, Montplaisir J, et al. REM sleep behavior disorder and cognitive impairment in parkinson's disease. *Sleep*. 2017;40(8). doi:10.1093/sleep/zsx101

49. Rusz J, Tykalova T, Novotny M, et al. Defining Speech Subtypes in De Novo Parkinson Disease. *Neurology*. 2021;97(21):e2124-e2135. doi:10.1212/WNL.0000000000012878

Chapter 3

Conclusions

The study represents a further step towards sensitive, remote, and unobtrusive monitoring of specific motor and cognitive impairments through everyday phone calls in patients with prodromal PD (i.e. iRBD), PD, or other synucleinopathies. Numerous challenges needed addressing and thorough investigation, particularly regarding the appropriate methodology for automatically measuring physiological speech patterns from recordings of potentially low quality and unguided content. Areas without a thorough research, such as automated measuring of dysprosody in noise, vowel articulation, rate of speech, or physiological explanation of MFCCs were investigated in studies published in impacted journals or a conference proceeding. Once reliable, robust, and precise automated approaches were developed or determined for each feature, a data acquisition software was designed to run on the subject's phone. The application, operating discreetly in the background, was equipped with necessary components, such as algorithms for adaptive filtering of the other call side and speaker verification. Additionally, it included a set of active motor tasks, such as sustained phonation, syllables repetition, reading a text, finger tapping, or walk with a turnaround. Technical details of the application were described in a protocol published in a impacted journal and the design could be directly utilized in clinical practice for further trials.

Finally, a cross-sectional study involving subjects with iRBD and PD was conducted, utilizing data gathered through mobile monitoring over a three-month period. The study, currently in a peer review, evaluated the sensitivity and specificity of proposed features in distinguishing between iRBD subjects and controls, PD patients and controls, and PD patients and iRBD subjects. Results showed that specific speech patterns observed during calls could serve as a biomarker for early parkinsonism, with articulation features notably affected during the calls. Furthermore, combining these passive call observations with active tasks measures, particularly dysprosodic ones, could enhance the sensitivity in detecting subtle voice alterations associated with neurodegeneration. It was revealed that monitoring through phone is at a level comparable to or even more sensitive than laboratory examination with a high-quality equipment.

The validated approach could be applied not only for a diagnosis but also for already diagnosed PD patients to enhance the current treatment strategies. Quick, inexpensive, and non-invasive vocal assessments using smartphones may allow for a personalized implementation of therapeutic strategies by providing rapid feedback after exercise, monitoring the effects of pharmacological therapies (including advanced drug delivery systems and modifying medication doses according to immediate needs), feedback in neuropsychiatry, modification of speech-related side-effects of deep brain stimulation via the

re-programming and optimization of stimulation parameters, population screening and beyond.

Understandably, this approach is not confined to Parkinsonism and could similarly be applied to conditions such as multiple sclerosis, Alzheimer's disease, and other neurodegenerative disorders.

3.1 Future aims

Given the significant benefits of phone call analysis, future studies should validate the findings and monitor disease progression through longitudinal studies in PD and prodromal PD, and extending to other disorders such as multiple sclerosis. Additionally, there should be an focus on exploring the emotional aspects of speech and developing reliable methods for their capture as remote approach in neropsychiatry might offer further advances. Emotional behavior is profoundly influenced by neurodegenerative disorders and could offer valuable insights into disease progression and mechanisms.

Once the connections between physiological, cognitive, and emotional patterns and the mechanism of the disorder are thoroughly understood and sufficient screening data has been collected, a robust disease model can be developed. This model might employ modern deep neural approaches, machine learning, and artificial intelligence to provide optimized personalized treatment for patients and serve as a valuable tool for diagnosis.

Bibliography

[1] W. Poewe, K. Seppi, C. M. Tanner, *et al.*, "Parkinson disease", *Nature Reviews Disease Primers*, vol. 3, 1 2017.

[2] M. C. de Rijk, L. J. Launer, K Berger, *et al.*, "Prevalence of parkinson's disease in europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group", *Neurology*, vol. 54, pp. 21–23, 11 Suppl 5 2000.

[3] L. J. Findley, "The economic impact of parkinson's disease", *Parkinsonism & Related Disorders*, vol. 13, pp. 8–12, 2007.

[4] M. C. Rodriguez-Oroz, M. Jahanshahi, P. Krack, *et al.*, "Initial clinical manifestations of parkinson's disease", *The Lancet Neurology*, vol. 8, pp. 1128–1139, 12 2009.

[5] R. B. Postuma, A. Iranzo, M. Hu, *et al.*, "Risk and predictors of dementia and parkinsonism in idiopathic rem sleep behaviour disorder", *Brain*, vol. 142, pp. 744–759, 3 2019.

[6] Y. Dauvilliers, C. H. Schenck, R. B. Postuma, *et al.*, "Rem sleep behaviour disorder", *Nature Reviews Disease Primers*, vol. 4, 1 2018.

[7] M. G. Miglis, C. H. Adler, E. Antelmi, *et al.*, "Biomarkers of conversion to α-synucleinopathy in isolated rapid-eye-movement sleep behaviour disorder", *The Lancet Neurology*, vol. 20, pp. 671–684, 8 2021.

[8] B. Högl, A. Stefani, and A. Videnovic, "Idiopathic rem sleep behaviour disorder and neurodegeneration — an update", *Nature Reviews Neurology*, vol. 14, pp. 40–55, 1 2018.

[9] S.-M. Fereshtehnejad, C. Yao, A. Pelletier, J. Y. Montplaisir, J.-F. Gagnon, and R. B. Postuma, "Evolution of prodromal parkinson's disease and dementia with lewy bodies: A prospective study", *Brain*, vol. 142, pp. 2051–2067, 7 2019.

[10] R. B. Postuma, J.-A. Bertrand, J. Montplaisir, *et al.*, "Rapid eye movement sleep behavior disorder and risk of dementia in parkinson's disease", *Movement Disorders*, vol. 27, pp. 720–726, 6 2012.

[11] J. L. Adams, T. Kangarloo, B. Tracey, *et al.*, "Using a smartwatch and smartphone to assess early parkinson's disease in the watch-pd study", *npj Parkinson's Disease*, vol. 9, 1 Dec. 2023.

[12] D. Lakovakis, S. Hadjidimitriou, V. Charisis, *et al.*, "Motor impairment estimates via touchscreen typing dynamics toward parkinson's disease detection from data harvested in-the-wild", *Frontiers in ICT*, vol. 5, 2018.

[13] B. M. Bot, C. Suver, E. C. Neto, *et al.*, "The mpower study, parkinson disease mobile data collected using researchkit", *Scientific Data*, vol. 3, Mar. 2016.

[14] A. Papadopoulos, K. Kyritsis, L. Klingelhoefer, S. Bostanjopoulou, K. R. Chaudhuri, and A. Delopoulos, "Detecting parkinsonian tremor from imu data collected in-the-wild using deep multiple-instance learning", *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2559–2569, 9 2020.

[15] S. Arora, F. Baig, C. Lo, *et al.*, "Smartphone motor testing to distinguish idiopathic rem sleep behavior disorder, controls, and pd", *Neurology*, vol. 91, pp. 1528–1538, 16 2018.

[16] J. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*, Fourth. Elsevier, 2019, vol. 4th ed. ISBN: 9780323530545.

[17] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease", *Behavioural Neurology*, vol. 11, pp. 131–137, 3 1999.

[18] J. Rusz, J. Hlavnička, M. Novotný, *et al.*, "Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease", *Annals of Neurology*, vol. 90, pp. 62–75, 1 2021.

[19] J. Rusz, A. Janzen, T. Tykalová, *et al.*, "Dysprosody in isolated rem sleep behavior disorder with impaired olfaction but intact nigrostriatal pathway", *Movement Disorders*, vol. 37, pp. 619–623, 3 2022.

[20] D. Skrabal, J. Rusz, M. Novotny, *et al.*, "Articulatory undershoot of vowels in isolated rem sleep behavior disorder and early parkinson's disease", *npj Parkinson's Disease*, vol. 8, 1 2022.

[21] L. M. Grant, F. Richter, J. E. Miller, *et al.*, "Vocalization deficits in mice overexpressing alpha-synuclein, a model of pre-manifest parkinson's disease", *Behavioral Neuroscience*, vol. 128, pp. 110–121, 2 2014.

[22] R. B. Postuma, A. E. Lang, J. F. Gagnon, A Pelletier, and J. Y. Montplaisir, "How does parkinsonism start? prodromal parkinsonism motor changes in idiopathic rem sleep behaviour disorder", *Brain*, vol. 135, pp. 1860–1870, 6 2012.

[23] L Brabenec, J Mekyska, Z Galaz, and I. Rektorova, "Speech disorders in parkinson's disease: Early diagnostics and effects of medication and brain stimulation", *Journal of Neural Transmission*, vol. 124, pp. 303–334, 3 2017.

[24] C. G. Goetz, B. C. Tilley, S. R. Shaftman, *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results", *Movement Disorders*, vol. 23, pp. 2129–2170, 15 Nov. 2008.

[25] C. A. Artusi, M. Mishra, P. Latimer, *et al.*, "Integration of technology-based outcome measures in clinical trials of parkinson and other neurodegenerative diseases", *Parkinsonism & Related Disorders*, vol. 46, pp. 53–56, 2018.

[26] V. Illner, P. Sovka, and J. Rusz, "Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in parkinson's disease", *Biomedical Signal Processing and Control*, vol. 58, 2020.

[27] M. Šimek and J. Rusz, "Validation of cepstral peak prominence in assessing early voice changes of parkinson's disease", *The Journal of the Acoustical Society of America*, vol. 150, pp. 4522–4533, 6 2021.

[28] F. L. Darley, A. E. Aronson, and J. R. Brown, "Clusters of deviant speech dimensions in the dysarthrias", *Journal of Speech and Hearing Research*, vol. 12, pp. 462–496, 3 1969.

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

[30] J. Rusz, R. Cmejla, T. Tykalova, *et al.*, "Imprecise vowel articulation as a potential early marker of parkinson's disease", *The Journal of the Acoustical Society of America*, vol. 134, pp. 2171–2181, 3 2013.

[31] C. Laganas, D. Iakovakis, S. K. Hadjidimitriou, *et al.*, "Parkinson's disease detection based on running speech data from phone calls", *IEEE Transactions on Biomedical Engineering*, vol. 65, pp. 1573–1584, 9 2022.

[32] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *IFA Proceedings*, vol. 17, pp. 97–110, 17 1993.

[33] B. R. Kumar, J. S. Bhat, and N. Prasad, "Cepstral analysis of voice in persons with vocal nodules", *Journal of Voice*, vol. 24, pp. 651–653, 6 2010.

[34] Y. Maryn, N. Roy, M. D. Bodt, P. V. Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality", *The Journal of the Acoustical Society of America*, vol. 126, pp. 2619–2634, 5 2009.

[35] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis", *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 1 2014.

[36] L. F. Brinca, A. P. F. Batista, A. I. Tavares, I. C. Gonçalves, and M. L. Moreno, "Use of cepstral analyses for differentiating normal from dysphonic voices", *Journal of Voice*, vol. 28, pp. 282–286, 3 2014.

[37] R. R. Patel, S. N. Awan, J. Barkmeier-Kraemer, *et al.*, "Recommended protocols for instrumental assessment of voice", *American Journal of Speech-Language Pathology*, vol. 27, pp. 887–905, 3 2018.

[38] G. G. Alharbi, M. P. Cannito, E. H. Buder, and S. N. Awan, "Spectral/cepstral analyses of phonation in parkinson's disease before and after voice treatment", *Folia Phoniatrica et Logopaedica*, vol. 71, pp. 275–285, 5-6 2019, ISSN: 1021-7762.

[39] S. Jannetts and A. Lowit, "Cepstral analysis of hypokinetic and ataxic voices", *Journal of Voice*, vol. 28, pp. 673–680, 6 2014.

[40] M. Novotný, P. Dušek, I. Daly, E. Růžička, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with parkinson's disease", *Biomedical Signal Processing and Control*, vol. 57, 2020.

[41] V. Uloza, E. Padervinskis, A. Vegiene, *et al.*, "Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening", *European Archives of Oto-Rhino-Laryngology*, vol. 272, pp. 3391–3399, 11 2015.

[42] J. Rusz, J. Hlavnicka, T. Tykalova, *et al.*, "Smartphone allows capture of speech abnormalities associated with high risk of developing parkinson's disease", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 1495–1507, 8 2018, ISSN: 1534-4320. [Online]. Available: `https://ieeexplore.ieee.org/document/8400578/`.

[43] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinsons disease", *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 5 2012.

[44] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels", *The Journal of the Acoustical Society of America*, vol. 135, pp. 2885–2901, 5 2014.

[45] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music", *The Journal of the Acoustical Society of America*, vol. 124, pp. 1638–1652, 3 2008.

[46] V. L. Hammen and K. M. Yorkston, "Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria", *Journal of Communication Disorders*, vol. 29, pp. 429–445, 6 1996.

[47] M. Šubert, M. Novotný, T. Tykalová, *et al.*, "Spoken language alterations can predict phenoconversion in isolated rapid eye movement sleep behavior disorder: A multicentric study", *Annals of Neurology*, vol. 95, pp. 530–543, Mar. 2023.

[48] A. Pistono, J. Pariente, C. Bézy, B. Lemesle, J. L. Men, and M. Jucla, "What happens when nothing happens? an investigation of pauses as a compensatory mechanism in early alzheimer's disease", *Neuropsychologia*, vol. 124, pp. 133–143, Feb. 2019.

[49] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1 1999.

[50] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization", *INTERSPEECH 2023*, pp. 3222–3226, 2023.

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

[52] A. Delval, M. Rambour, C. Tard, *et al.*, "Freezing/festination during motor tasks in early-stage parkinson's disease", *Movement Disorders*, vol. 31, pp. 1837–1845, 12 2016.

[53] K Konstantopoulos, P Vogazianos, Y Christou, and M Pisinou, "Sequential motion rate and oral reading rate", *Logopedics Phoniatrics Vocology*, pp. 1–6, 2021.

[54] S. Skodda and U. Schlegel, "Speech rate and rhythm in parkinson's disease", *Movement Disorders*, vol. 23, pp. 985–992, 7 2008.

[55] V. Illner, T. Tykalová, M. Novotný, J. Klempíř, P. Dušek, and J. Rusz, "Toward automated articulation rate analysis via connected speech in dysarthrias", *Journal of Speech, Language, and Hearing Research*, vol. 65, 4 2022.

[56] B. Zellner, "Fast and slow speech rate: A characterisation for french", *Proc. Int. Conf. Spoken Lang. Proc., Sydney, Australia*, 3159–3163, 7 1998.

[57] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous english speech", *iWORD/i*, vol. 15, pp. 19–44, 1 2015.

[58] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate", *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 729–732, 1998.

[59] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1421–1430, 9 2015, ISSN: 2329-9290.

[60] H.-D. Huici, H. A. Kairuz, H. Martens, G. V. Nuffelen, and M. D. Bodt, "Speech rate estimation in disordered speech based on spectral landmark detection", *Biomedical Signal Processing and Control*, vol. 27, pp. 1–6, 2016.

[61] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision", *Proceedings of the 40th International Conference on Machine Learning*, p. 1182, 2023.

[62] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior Research Methods*, vol. 41, pp. 385–390, 2 2009, ISSN: 1554-351X.

[63] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria", *Folia Phoniatrica et Logopaedica*, vol. 63, pp. 187–194, 4 2011, ISSN: 1421-9972.

[64] L. T. Robertson and J. P. Hammerstad, "Jaw movement dysfunction related to parkinson's disease and partially modified by levodopa", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 60, pp. 41–50, 1 1996.

[65] J. A. Whitfield, "Exploration of metrics for quantifying formant space", *Perspectives of the ASHA Special Interest Groups*, vol. 4, pp. 402–410, 2 2019.

[66] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in parkinson's disease", *Journal of Voice*, vol. 25, pp. 467–472, 4 2011.

[67] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area", *The Journal of the Acoustical Society of America*, vol. 134, pp. 477–483, 5 2013.

[68] Y. Liu, N. Penttila, T. Ihalainen, J. Lintula, R. Convey, and O. Rasanen, "Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2228–2243, 2021.

[69] V. Illner, T. Tykalova, D. Skrabal, J. Klempir, and J. Rusz, "Automated vowel articulation analysis in connected speech among progressive neurological diseases, dysarthria types, and dysarthria severities", *Journal of speech, language, and hearing research : JSLHR*, vol. 66, pp. 2600–2621, 8 2023.

[70] A. Benba, A. Jilbab, and A. Hammouch, "Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines", *International Journal on Electrical Engineering and Informatics*, vol. 7, pp. 297–307, 2 2015.

[71] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with parkinson's and neurological diseases using cepstral analysis", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 1100–1108, 10 2016.

[72] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the reliability of acoustic speech embeddings", *Interspeech 2020*, pp. 4621–4625, 2020.

[73] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, "Quantifying articulatory impairments in neurodegenerative motor diseases: A scoping review and meta-analysis of interpretable acoustic features", *International Journal of Speech-Language Pathology*, vol. 4, pp. 486–499, 25 2022.

[74] F. Lipsmeier, K. I. Taylor, T. Kilchenmann, *et al.*, "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 parkinson's disease clinical trial", *Movement Disorders*, vol. 33, pp. 1287–1297, 8 2018.

[75] V. Illner, P. Krýže, J. Švihlík, *et al.*, "Which aspects of motor speech disorder are captured by mel frequency cepstral coefficients? evidence from the change in stn-dbs conditions in parkinson's disease", *INTERSPEECH 2023*, pp. 5027–5031, 2023.

[76] E Tripoliti, M. L. Zrinzo, I Martinez-Torres, *et al.*, "Effects of subthalamic stimulation on speech of consecutive patients with parkinson disease", *Neurology*, vol. 76, pp. 80–86, 1 2011.

[77] T. Kouba, V. Illner, and J. Rusz, "Study protocol for using a smartphone application to investigate speech biomarkers of parkinson's disease and other synucleinopathies", *BMJ Open*, vol. 12, 6 2022.

[78] A. Zhan, S. Mohan, C. Tarolli, *et al.*, "Using smartphones and machine learning to quantify parkinson disease severity", *JAMA Neurology*, vol. 75, pp. 76–80, 7 2018.

Appendix A: Supplementary material on articulation deficits

The supplementary material for the article "Automated Vowel Articulation Analysis in Connected Speech Among Progressive Neurological Diseases, Dysarthria Types, and Dysarthria Severities".

A detailed description of the results based on reading passage

S1. Reading passage recording

All participants were instructed to present a reading passage of 80 words with a mean duration of 33.9 seconds (SD 4.6, range 23-69). Each reading passage task was performed twice. The performance across both repetitions was averaged for subsequent statistical analyses. The reading passage with the noted position of the selected vowel for the manual analysis in bold is shown in **Figure** S1.

Figure S1. The text of the reading passage task with highlighted /a/, /i/, and /u/ corner vowels used for manual analysis in bold.

Kd**y**[1]ž člověk poprvé vs**a**[1]d**í**[2] do země s**a**[2]zen**i**[3]čk**u**[1], chod**í**[4] se n**a**[3] n**i**[5] d**í**[6]v**a**[4]t tř**i**[7]kr**á**[5]t denně: t**a**[6]k co, pov**y**[8]rostl**a**[7] **u**[2]ž nebo ne? **I**[9] t**a**[8]j**í**[10] dech, n**a**[9]kl**á**[10]n**í**[11] se n**a**[11]d n**í**[12], př**i**[13]tl**a**[12]č**í**[14] troch**u**[3] p**ů**[4]d**u**[5] **u**[6] jej**í**[15]ch koř**í**[16]nk**ů**[7], n**a**[13]čechr**á**[14]v**á**[15] j**í**[17] l**í**[18]stk**y**[19] **a**[16] v**ů**[8]bec j**i**[20] obtěž**u**[9]je r**ů**[10]zn**ý**[21]m kon**á**[17]n**í**[22]m, které pov**a**[18]ž**u**[11]je z**a**[19] **u**[12]ž**i**[23]tečno**u**[13] péč**i**[24]. **A**[20] kd**y**[25]ž se s**a**[21]zen**i**[26]čk**a**[22] přesto **u**[14]jme **a**[23] roste j**a**[24]ko z vod**y**[27], t**u**[15] člověk ž**a**[25]sne n**a**[26]d t**í**[28]mto d**i**[29]vem př**í**[30]rod**y**[31], m**á**[27] poc**i**[32]t čehos**i**[33] j**a**[28]ko z**á**[29]zr**a**[30]k**u**[16] **a**[31] pov**a**[32]ž**u**[17]je to z**a**[33] jeden ze sv**ý**[34]ch největš**í**[35]ch osobn**í**[36]ch **ú**[18]spěch**ů**[19].

S2. Effect of neurological disease type

Compared to controls, the change in vowel articulation due to neurodegeneration in reading passages was primarily demonstrated by trends toward shift of formants across vowels /i/ and /u/, including an increase in F_{2u} and decrease in F_{1i}, F_{1u} and F_{2i} frequencies across PSP, MSA, HD, and ALS (**Figure S2, Table S1**). Among diseases, there was a particular difference between atypical parkinsonism of MSA or PSP compared to other neurological conditions that was mainly demonstrated by trends towards decrease of F_{1a}, F_{1i} and F_{1u} in MSA and decrease of F_{1u} and increase of F_{2a} in PSP (**Figure S3**).

Considering complex formant measures, compared to controls, VSA was significantly decreased for PSP, MSA, and MS (**Figure S3**). Both FRI and SFRI were decreased for all neurological diseases except for PD.

Figure S2. Corner vowel production triangles estimated from reading passages for individual neurological disease types compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency, F_2 = second formant frequency, PD = Parkinson's disease, PSP = progressive supranuclear palsy, MSA = multiple system atrophy, HD = Huntington's disease , ET = essential tremor, CA = cerebellar ataxia, MS = multiple sclerosis, ALS = amyotrophic lateral sclerosis.

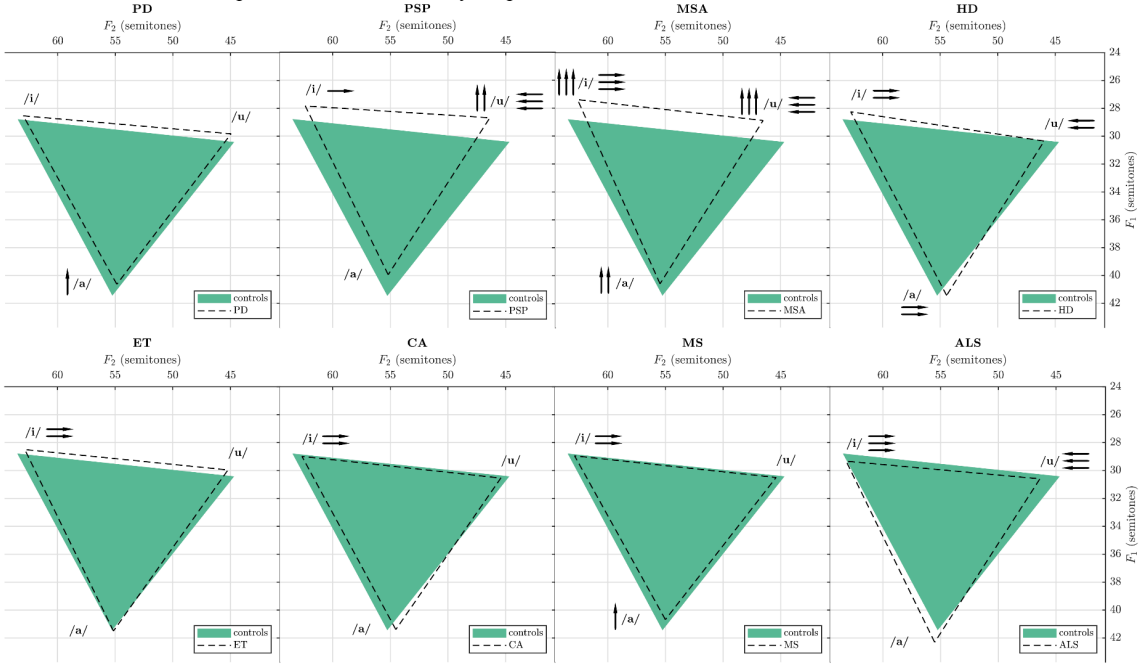


Table S1. Formant frequencies of corner vowels estimated from reading passages for individual neurological disease types compared to healthy controls. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones).

	/a/ mean (SD) semitones		**/i/** Mean (SD) semitones		**/u/** mean (SD) semitones	
Neurological disease type	F_1	F_2	F_1	F_2	F_1	F_2
controls	41.44 (2.4)	55.26 (1.9)	28.79 (2.2)	63.48 (1.5)	30.42 (1.6)	44.68 (1.6)
PD	40.63 (2.5)	54.87 (1.7)	28.54 (2.7)	63.00 (1.5)	29.83 (2.2)	44.95 (1.2)
PSP	39.93 (2.8)	55.18 (1.8)	27.85 (2.5)	62.37 (1.7)	28.69 (2.6)	46.37 (2.9)
MSA	40.59 (2.6)	55.47 (1.8)	27.38 (2.4)	62.63 (1.1)	28.87 (2.2)	46.52 (2.1)
HD	41.23 (3.0)	54.44 (2.3)	28.25 (2.7)	62.77 (2.0)	30.35 (2.5)	45.97 (2.4)
ET	41.50 (3.2)	55.14 (2.0)	28.50 (1.7)	62.83 (1.6)	29.96 (1.8)	45.21 (1.7)
CA	41.39 (2.9)	54.52 (1.4)	28.98 (2.0)	62.66 (1.3)	30.54 (1.6)	45.42 (1.4)
MS	40.68 (2.9)	55.01 (1.8)	28.95 (2.3)	62.88 (1.7)	30.50 (1.7)	45.41 (2.1)
ALS	42.31 (3.7)	55.52 (1.1)	29.33 (2.2)	63.20 (1.8)	30.59 (1.4)	46.40 (2.4)

Captions: SD = standard deviation, F_1 = first formant frequency, F_2 = second formant frequency, PD = Parkinson's disease, PSP = progressive supranuclear palsy, MSA = multiple system atrophy, HD = Huntington's disease , ET = essential tremor, CA = cerebellar ataxia, MS = multiple sclerosis, ALS = amyotrophic lateral sclerosis.

Figure S3. Corner vowel production triangles estimated from monologues across two pairs of neurological disease types. The double-headed arrows indicate significant differences across diseases adjusted by age, sex, and dysarthria severity with *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency, F_2 = second formant frequency, PD = Parkinson's disease, PSP = progressive supranuclear palsy, MSA = multiple system atrophy, HD = Huntington's disease , ET = essential tremor, CA = cerebellar ataxia, MS = multiple sclerosis, ALS = amyotrophic lateral sclerosis.

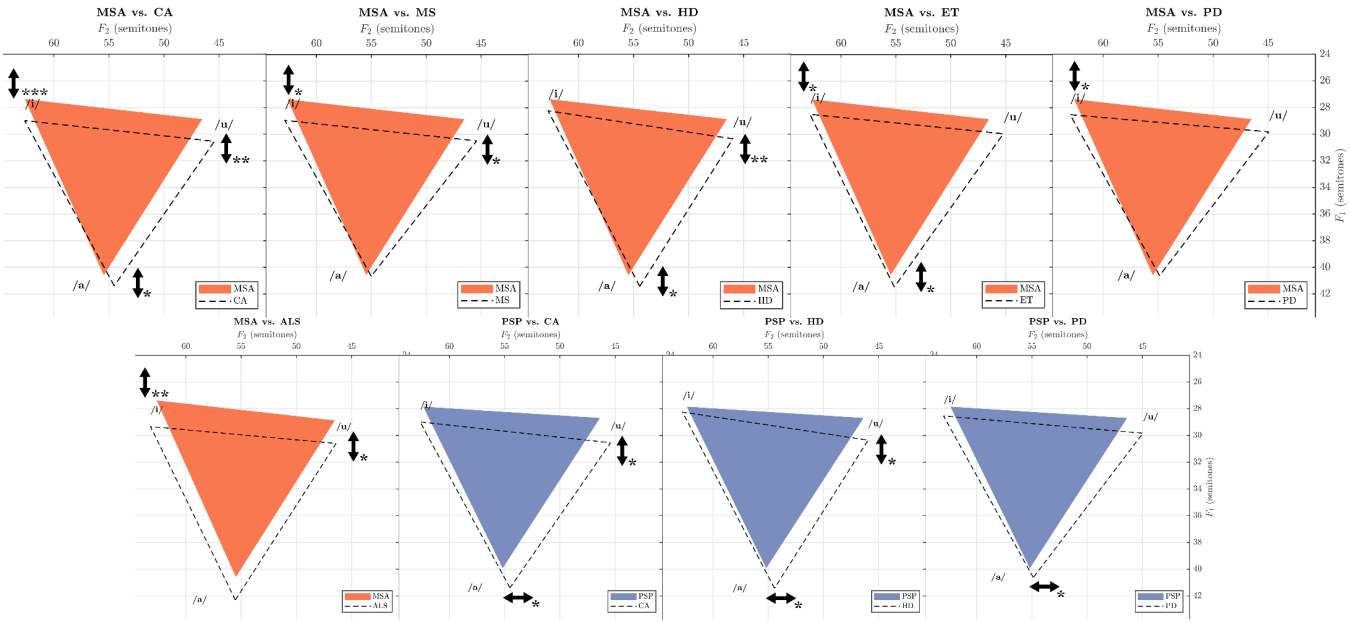
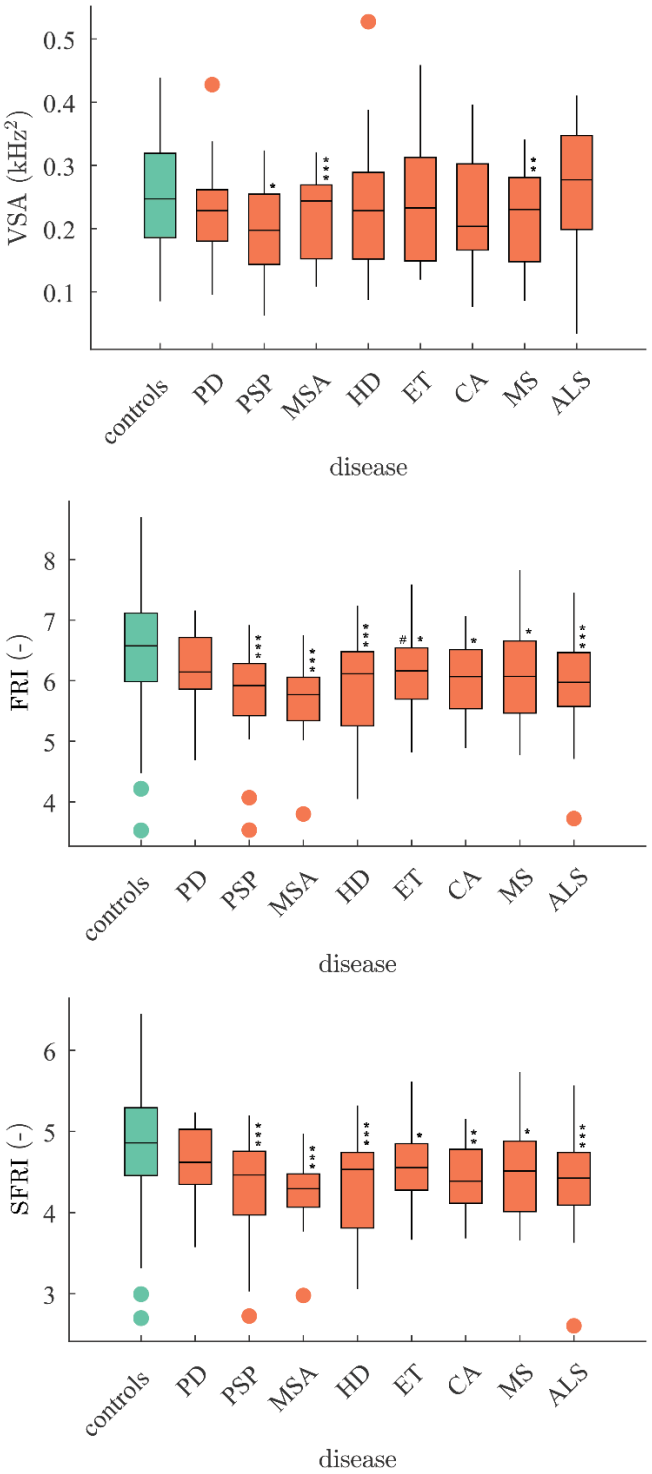


Figure S4. Statistically significant group differences for estimated articulation features among the different types of neurological disease types compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$. # indicates significant differences to MSA ($p < 0.05$) after adjusting for age, sex and dysarthria severity. Middle bars represent the median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. PD = Parkinson's disease, PSP = progressive supranuclear palsy, MSA = multiple system atrophy, HD = Huntington's disease, ET = essential tremor, CA = cerebellar ataxia, MS = multiple sclerosis, ALS = amyotrophic lateral sclerosis, VSA = vowel space area, FRI = formant ratio index, SFRI = second formant ratio index.



S3. Effect of dysarthria type

Compared to controls, the trends toward formants shift across vowels /i/ and /u/ including increase in F_{2u} and decrease in F_{1i}, F_{1u} and F_{2i} frequencies in reading passages were demonstrated mainly for hypokinetic and hyperkinetic dysarthria and all mixed dysarthrias (**Figure S5, Table S2**). Among dysarthrias, there was a particular difference between ataxic dysarthria manifested mainly by increase of F_{1u} compared to hypokinetic dysarthria (and its mixed variants with ataxic and spastic elements) (**Figure S6**). Additionally, the hypokinetic-spastic dysarthria showed particularly increase of F_{2u} compared to hypokinetic, hyperkinetic and spastic ataxic dysarthrias.

Considering complex formant measures, compared to controls, VSA was significantly decreased for hypokinetic, ataxic, flaccid-spastic, spastic-ataxic, hypokinetic-spastic, and hypokinetic-spastic-ataxic in reading passage (**Figure S7**). FRI was decreased for hypokinetic, hyperkinetic, flaccid-spastic, spastic-ataxic, hypokinetic-spastic, and hypokinetic-spastic-ataxic dysarthria. Finally, SFRI was decreased for hypokinetic, hyperkinetic, flaccid-spastic, spastic-ataxic, hypokinetic-spastic, and hypokinetic-spastic-ataxic dysarthria.

Figure S5. Corner vowel production triangles estimated from reading passages for different dysarthria types compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency, F_2 = second formant frequency.

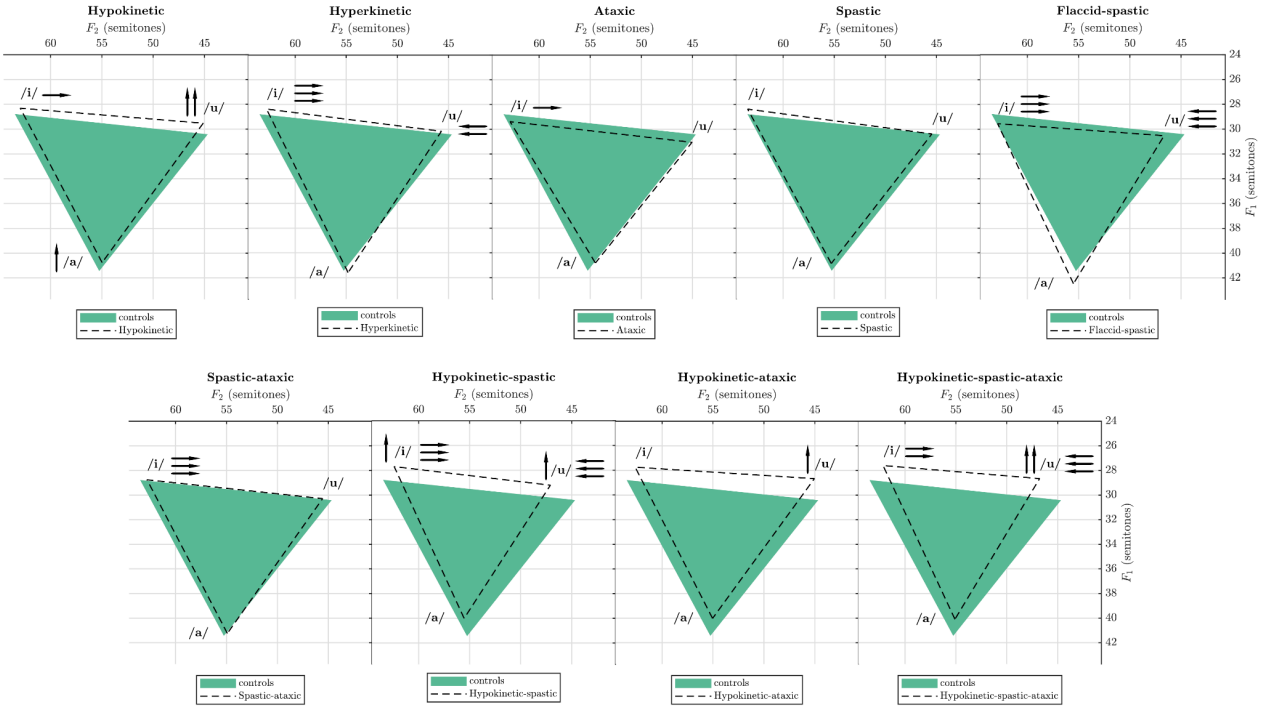


Table S2. Formant frequencies of corner vowels estimated from monologues for different dysarthria types compared to healthy controls. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones).

	/a/ mean (SD) semitones		**/i/** mean (SD) semitones		**/u/** mean (SD) semitones	
Dysarthria type	F_1	F_2	F_1	F_2	F_1	F_2
controls	41.44 (2.4)	55.26 (1.9)	28.79 (2.2)	63.48 (1.5)	30.42 (1.6)	44.68 (1.6)
Hypokinetic	40.76 (2.4)	54.97 (1.6)	28.31 (2.7)	62.96 (1.3)	29.53 (2.2)	45.12 (1.6)
Hyperkinetic	41.59 (3.1)	54.82 (2.1)	28.36 (2.2)	62.81 (1.8)	30.16 (2.2)	45.70 (2.1)
Ataxic	40.82 (3.3)	54.50 (1.6)	29.38 (2.0)	62.84 (1.3)	31.06 (1.0)	45.05 (1.4)
Spastic	40.86 (3.5)	55.30 (1.8)	28.37 (1.8)	63.47 (1.8)	30.38 (1.6)	45.46 (1.4)
Flaccid-spastic	42.46 (3.9)	55.50 (1.0)	29.56 (2.3)	62.90 (1.7)	30.53 (1.5)	46.66 (2.4)
Spastic-ataxic	41.24 (2.6)	54.92 (1.6)	28.76 (2.3)	62.81 (1.6)	30.29 (1.8)	45.56 (2.1)
Hypokinetic-spastic	40.00 (3.6)	55.54 (1.9)	27.71 (2.6)	62.40 (1.4)	29.22 (2.3)	47.10 (2.6)
Hypokinetic-ataxic	40.03 (2.3)	55.04 (2.4)	27.74 (2.2)	62.65 (1.2)	28.67 (2.4)	45.04 (1.9)
Hypokinetic-spastic-ataxic	40.09 (1.8)	55.09 (1.8)	27.60 (2.8)	62.16 (1.7)	28.66 (2.6)	46.77 (2.7)

Captions: SD = standard deviation, F_1 = first formant frequency, F_2 = second formant frequency.

Figure S6. Corner vowel production triangles estimated from monologues across two pairs of neurological disease types. The double-headed arrows indicate significant differences across diseases adjusted by age, sex, and dysarthria severity with *** $p < 0.001$, ** $p < 0.01$, and * $p < 0.05$. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency, F_2 = second formant frequency.

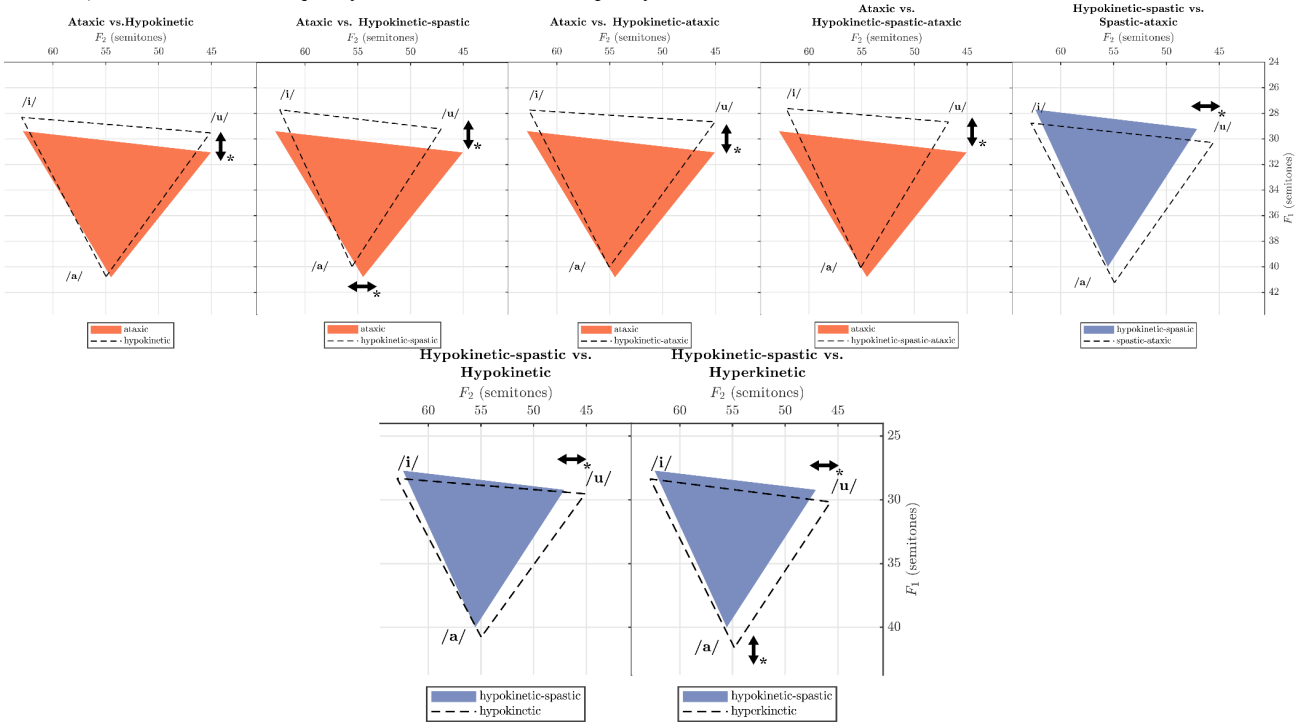
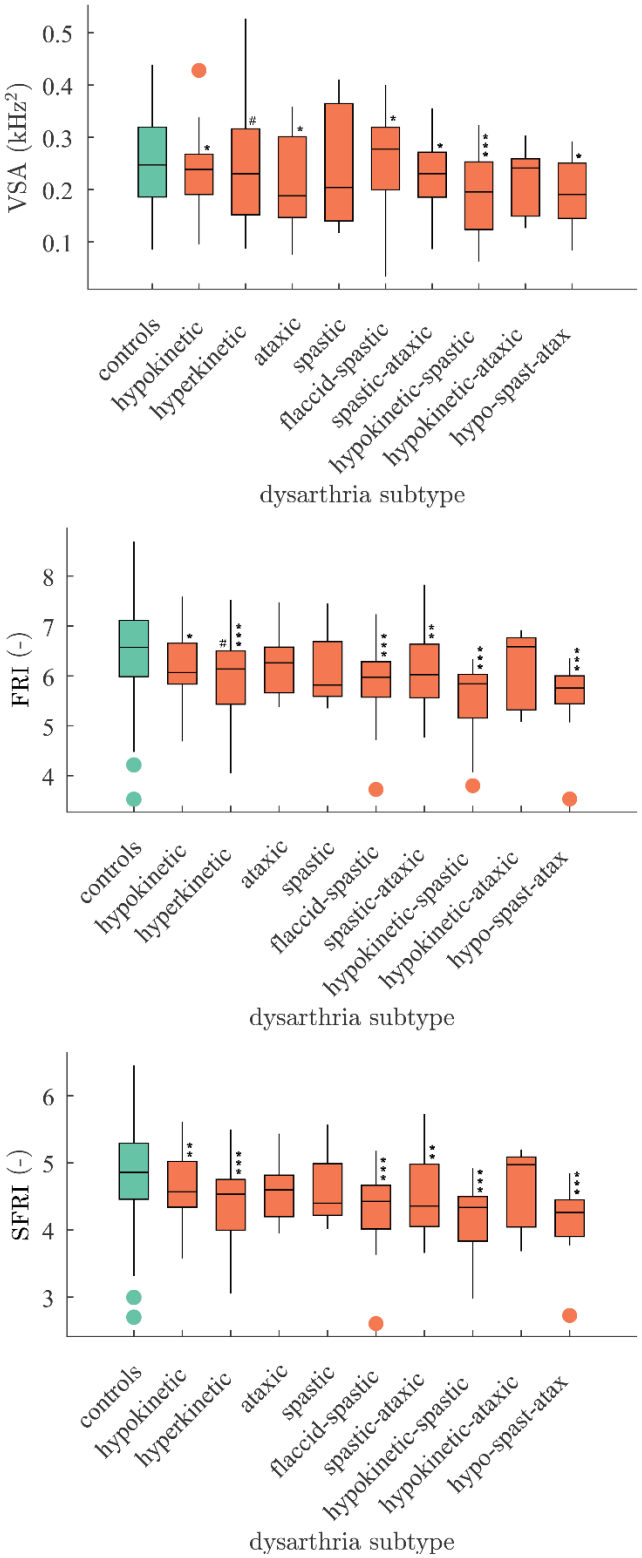


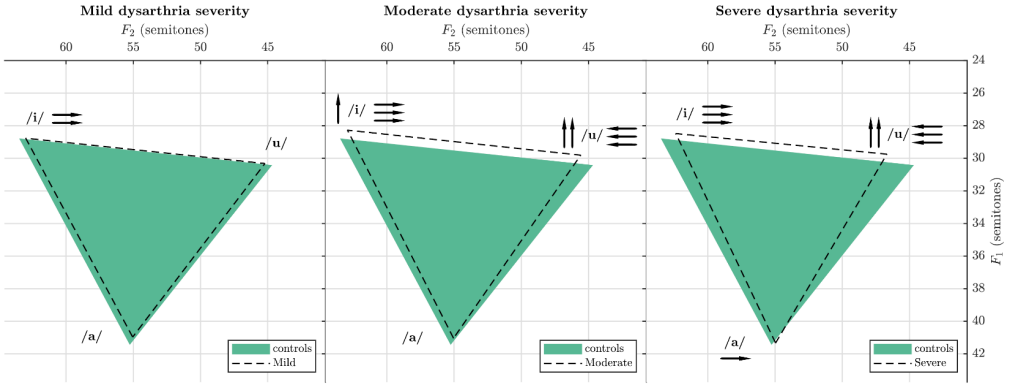
Figure S7. Statistically significant group differences for estimated articulation features among the different dysarthria types compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$. # indicates significant differences to hypokinetic-spastic dysarthria ($p < 0.05$) after adjusting for age, sex and dysarthria severity. Middle bars represent the median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. VSA = vowel space area, FRI = formant ratio index, SFRI = second formant ratio index, Hypo-spast-atax = Hypokinetic-spastic-ataxic dysarthria.



S4. Effect of dysarthria severity

Compared to controls, the formants shift across vowels /i/ and /u/ in dependence on dysarthria severity in reading passages was observed, including an increase in F_{2u} and a decrease in F_{1i}, F_{1u}, and F_{2i} frequencies (**Figure S8, Table S3**). Considering complex formant measures, mainly measures of FRI and SFRI were progressively reduced with increasing dysarthria severity (**Figure S9**).

Figure S8. Corner vowel production triangles estimated from reading passages for different dysarthria severities compared to healthy controls. The arrows indicate significant differences in the values to healthy controls adjusted by age and sex, with three, two, and one arrows referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones). F_1 = first formant frequency, F_2 = second formant frequency.

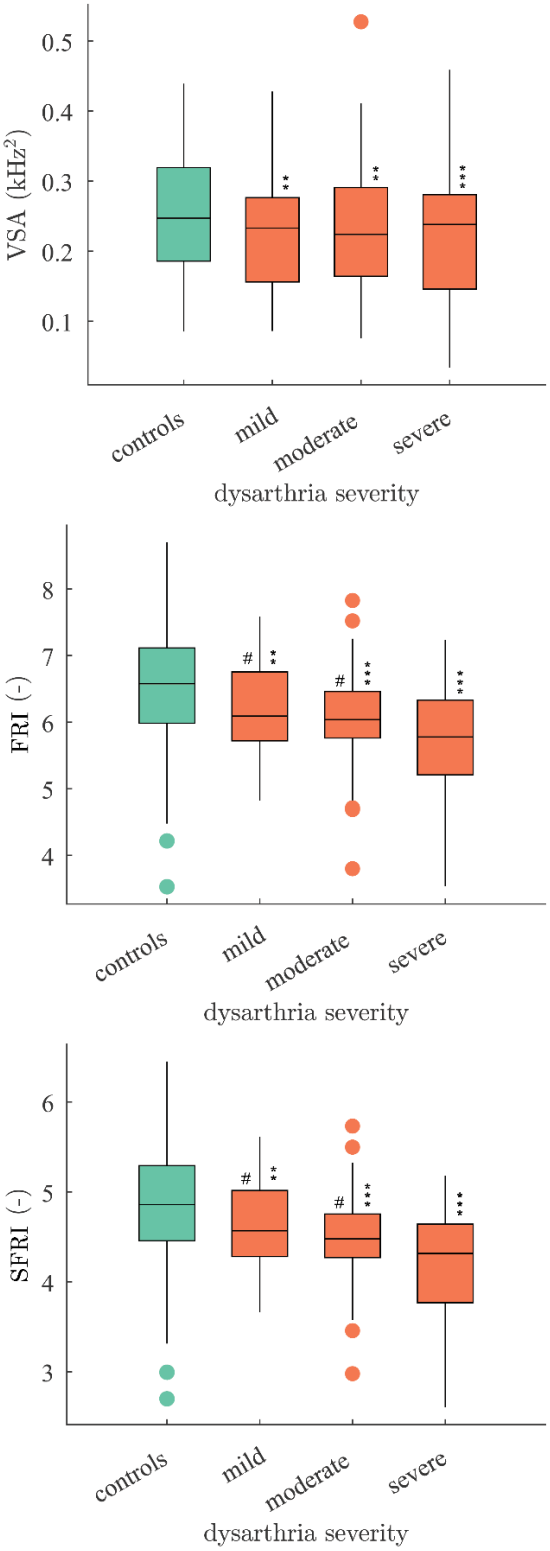


S3. Formant frequencies of corner vowels estimated from monologues for different dysarthria severities compared to healthy controls. To minimize the effects of sex between individual speakers, the estimated formant frequencies were converted into a logarithmic tonal scale (semitones).

	/a/ mean (SD) semitones		**/i/** mean (SD) semitones		**/u/** mean (SD) semitones	
Dysarthria severity	F_1	F_2	F_1	F_2	F_1	F_2
controls	41.44 (2.4)	55.26 (1.9)	28.79 (2.2)	63.48 (1.5)	30.42 (1.6)	44.68 (1.6)
Mild	40.97 (2,6)	55.04 (1.7)	28.77 (2.0)	63.00 (1.6)	30.32 (1.5)	45.21 (1.7)
Moderate	41.04 (3.1)	55.04 (1.9)	28.27 (2.5)	62.93 (1.5)	29.81 (2.3)	45.59 (2.0)
Severe	41.33 (3.2)	54.96 (1.8)	28.48 (2.6)	62.38 (1.5)	29.74 (2.3)	46.68 (2.5)

Captions: SD = standard deviation, F_1 = first formant frequency, F_2 = second formant frequency.

Figure S9. Statistically significant group differences for estimated articulation features in reading passages among the different dysarthria severities compared to healthy controls adjusted by age and sex with ***, **, * referring to $p < 0.001$, $p < 0.01$, and $p < 0.05$. # indicates significant differences to severe dysarthria ($p < 0.05$) after adjusting for age and sex. Middle bars represent the median, and rectangles represent the interquartile range. Maximum and minimum values are by error bars. Outliers are marked as dots. VSA = vowel space area, FRI = formant ratio index, SFRI = second formant ratio index.



S5. Classification analysis

The classification analysis among vowel articulation features manifested accuracy of up to 41.0% for disease type, 39.3% for dysarthria type and 47.4% for dysarthria (**Table S4**). Acoustic metrics of FRI and SFRI were more sensitive to capturing the change of vowel articulation than VSA.

Table S4. Classification analysis for the formant features in reading passages. The numbers indicate the percentage of subjects correctly identified by the discriminant analysis as original groups. Bold numbers indicate the best accuracy across neurological disease type, dysarthria type, and dysarthria severity.

% (monologue / reading passage)	VSA	FRI	SFRI
Neurological disease type	4.4	**41.0**	40.2
Dysarthria type	25.8	**39.3**	**39.3**
Dysarthria severity	39.5	46.7	**47.4**

Captions: VSA = vowel space area, FRI = formant ratio index, SFRI = second formant ratio index.