A cross-linguistic perspective to classification of healthiness of speech in Parkinson’s disease

Vass Verkhodanova a,b,*, Matt Coler a, Roel Jonkers b,c, Sanne Timmermans d, Natasha Maurits d, Bauke de Jong d, Wander Lowie b,c

a University of Groningen, Campus Fryslân, Wirdumerdijk 34, 8911CE, Leeuwarden, the Netherlands
b University of Groningen, Research School of Behavioural and Cognitive Neurosciences, UMCG, GSMS, BCN-Office, FA 30, Postbox 196, 9700, AD, Groningen, the Netherlands
c University of Groningen, Center for Language and Cognition Groningen, PO Box, 716, 9700, AS, Groningen, the Netherlands
d University of Groningen, University Medical Center Groningen, Department of Neurology, PO Box 30.001, 9700, RB, Groningen, the Netherlands

ARTICLE INFO

Keywords: Parkinson’s disease Dysarthria Speech perception Speech classification Prosody Random forest

ABSTRACT

People with Parkinson’s disease often experience communication problems. The current cross-linguistic study investigates how listeners’ perceptual judgements of speech healthiness are related to the acoustic changes appearing in the speech of people with Parkinson’s disease. Accordingly, we report on an online experiment targeting perceived healthiness of speech. We studied the relations between healthiness perceptual judgements and a set of acoustic characteristics of speech in a cross-sectional design. We recruited 169 participants, who performed a classification task judging speech recordings of Dutch speakers with Parkinson’s disease and of Dutch control speakers as ‘healthy’ or ‘unhealthy’. The groups of listeners differed in their training and expertise in speech language therapy as well as in their native languages. Such group separation allowed us to investigate the acoustic correlates of speech healthiness without influence of the content of the recordings. We used a Random Forest method to predict listeners’ responses. Our findings demonstrate that, independently of expertise and language background, when classifying speech as healthy or unhealthy listeners are more sensitive to speech rate, presence of phonation deficiency reflected by maximum phonation time measurement, and centralization of the vowels. The results indicate that both specifics of the expertise and language background may lead to listeners relying more on the features from either prosody or phonation domains. Our findings demonstrate that more global perceptual judgements of different listeners classifying speech of people with Parkinson’s disease may be predicted with sufficient reliability from conventional acoustic features. This suggests universality of acoustic change in speech of people with Parkinson’s disease. Therefore, we concluded that certain aspects of phonation and prosody serve as prominent markers of speech healthiness for listeners independent of their first language or expertise. Our findings have outcomes for the clinical practice and real-life implications for subjective perception of speech of people with Parkinson’s disease, while information about particular acoustic changes that trigger listeners to classify speech as ‘unhealthy’ can provide specific therapeutic targets in addition to the existing dysarthria treatment in people with Parkinson’s disease.
1. Introduction

1.1. Background

Parkinson’s Disease (PD) often leads to a distinctive motor speech disorder referred to as hypokinetic dysarthria (HD). HD results from disturbances in muscular control over the speech mechanism and manifests itself in all dimensions of human speech production (Darley et al., 1969a; Brabenec et al., 2017). HD manifestations negatively impact spoken communication of people with PD (hereafter, PwPD) (Miller, 2017; Moreau & Pinto, 2019) and have been considered to serve as additional marker useful for early diagnosis of PD (Brabenec et al., 2017).

Studies by Darley et al. (1969a, 1969b) were the first to investigate the most recognizably affected speech characteristics in dysarthria. According to the authors, the top ten deviant speech dimensions in HD are monotonous pitch or monopitch, reduced stress, monotony of loudness or monoloudness, imprecise consonants, inappropriate silences, short rushes of speech, harsh voice, breathy voice, low pitch and variable rate. These ten dimensions are associated with phonatory, articulatory and prosodic speech components (Darley et al., 1969b; Duffy, 2012). However, six out of these ten dimensions pertain to prosody, making prosodic insufficiencies the most severely affected cluster of speech dimensions in HD (Darley et al., 1969a; Martens et al., 2011).

Deviant speech dimensions have been extensively studied on an acoustic level. There are many studies focusing on the acoustic aspect of speech production in PD with attention to monopitch, or in acoustic terms - fundamental frequency ($f_0$) deviations (Galaz et al., 2016; Skodda et al., 2013), distorted rhythm of speech (Skodda & Schlegel, 2008), monoloudness - reduced intensity variability of voice (Galaz et al., 2016; Skodda et al., 2013), reduced stress (Tykalova et al., 2014), imprecise consonants (Fischer & Goberman, 2010; Tykalova et al., 2017) and a hoarse and breathy voice quality (Tsanas et al., 2009). A number of studies have also demonstrated that prosody deficits together with harsh voice and reduced articulation are among the most prominent speech characteristics that are present in acoustics of speech produced by PwPD (Brabenec et al., 2017; Galaz et al., 2016; Rusz et al., 2011; Verkhodanova et al., 2019). Some studies even suggested that the prosodic deficits arising from HD are universal (Pinto et al., 2017).

There is a growing number of studies exploring the efficiency of speech production by PwPD and focusing on the perception of speech produced by PwPD. Many researchers have described prominent changes in prosodic characteristics of speech of PwPD. For example, when compared to control speakers, PwPD are less efficient at producing question-statement intonation contrasts (Basirat et al., 2018; Pell et al., 2006) or at conveying both lexical and contrastive stress (Martens et al., 2016; Pell et al., 2006). Overall, the existing literature monopitch and monoloudness are described as having the greatest influence on the perception of speech affected by HD, and are seen as the most prototypical source of prosodic speech problems for PwPD (De Letter et al., 2007; Kuo et al., 2014; Martens et al., 2016).

Such prosody impairment has crucial consequences for the intelligibility of speech, and for the daily communication of dysarthric speakers with PD (Miller et al., 2006; Anand & Stepp, 2015; Pinto et al., 2017). These speech disturbances affect the quality of life of PwPD, resulting in communication problems and often in a feeling of social isolation. This frequently leads to tension, depression, resignation and withdrawal from a conversation (Miller et al., 2006). For example, Schalling et al. (2017) collected and presented self-reported information of affected communication from 188 Swedish PwPD. Their findings demonstrated that 92.5% of the respondents reported at least one symptom related to communication, and that the speech and communication problems resulted in restricted communicative participation for around a third of all respondents. Even though almost every participant reported speech and communication problems, only 45% reported receiving speech-language therapy, highlighting inadequate access to speech-language pathology services for PwPD (Schalling et al., 2017). Moreover, research demonstrated that speech changes caused by HD affect the daily lives of PwPD long before impairment of intelligibility is apparent (Miller, 2017; Miller et al., 2006).

1.2. Assessment of dysarthria

According to many studies, (for instance, Bunton et al., 2007; Sussman & Tjaden, 2012; Duffy, 2012; Näsström & Schalling, 2020), the auditory-perceptual evaluation of dysarthria continues to be the ‘gold standard’ for clinical decisions. The means of assessment performed by listeners ranges from judging vowels (Sapir et al., 2007), words (Bunton & Keintz, 2008; Smith et al., 2019) to read passages (Tjaden & Wilding, 2004) and spontaneous conversational speech (Bunton et al., 2007; Bunton & Keintz, 2008). One of the standard measures of assessment and management of speakers with dysarthria is speech intelligibility scores. These scores are commonly used as a measure of the severity of speech disorder and as a source of information for treatment planning and for monitoring changes in speech (Yorkston & Beukelman, 1978; Bunton & Keintz, 2008). Another approach to auditory-perceptual assessment of dysarthria is component-specific perceptual judgements based on evaluating lists of deviant dimensions, as described by Darley et al. (1969a). Both approaches have limitations and reliability concerns as summarized in a study by Sussman and Tjaden (2012).

As an alternative to the standard means of assessing speech of people with dysarthria, Sussman and Tjaden (2012) suggest exploring more ‘global’ perceptual judgments of speech disorder severity for individuals with multiple sclerosis and PD. This idea is related to the approach proposed by Weismer et al. (2001) and is in line with early suggestions by Kreiman et al. (1993) who recommended using more global ratings of overall speech competence, such as ‘good/poor voice’ or ‘not impaired/severely impaired’. Sussman and Tjaden’s (2012) results demonstrate that scaled estimates of speech severity appear to be sensitive to aspects of speech impairment in both multiple sclerosis and PD that are not captured by word and sentence intelligibility scores.

However, intelligibility scores, component-specific perceptual judgements and scaled estimates of speech disorder severity are
developed for a specific language. This approach is dependent on language specific differences. For example, findings of a comparative study (Kim & Choi, 2017) show that even though PwPD of different languages, American English and Korean, demonstrate similar acoustic patterns of articulation insufficiencies compared to control groups, the degree to which the same acoustic features contribute to the intelligibility scores is language-dependent. Nevertheless, with the global trend towards migration and growing ageing population, there is an emerging need for dysarthria assessment in a language unfamiliar to the assessor (Näström & Schalling, 2020). Nevertheless, there are few studies that explore cross-linguistic assessment of dysarthria. One such study is that of Hartelius et al. (2003), who report results of cross-language assessment of Swedish and Australian speakers with dysarthria secondary to multiple sclerosis. Australian and Swedish speech language therapists (SLT’s) demonstrated high interrater reliability, resulting in similar prevailing sets of dimensions for both languages: imprecise consonants, harshness and glottal fric, reduced speech rate, pitch level, and loudness despite the foreign language. However, some of these dimensions, namely precision of consonant production, pitch and loudness level, general rate and harshness, were associated with higher disagreement values, while there were language specific difficulties in assessment of general stress pattern and phoneme length (Hartelius et al., 2003). Another study by Näström and Schalling (2020) focuses on developing a systematic assessment method for SLT’s who do not speak the native language of an individual with dysarthria. Authors observed that despite being unfamiliar to the Arabic language, a Swedish SLT was sensitive to respiration, phonation and some articulation changes in dysarthric speech (Näström & Schalling, 2020). Their results demonstrated that an SLT who is performing the assessment according to the method in collaboration with an interpreter shows comparable results as a SLT who speaks the language of an individual with dysarthria (Näström & Schalling, 2020). These findings demonstrate that, irrespective of a listener’s familiarity with the language, the way dysarthria affects phonation and some aspects of articulation might be largely universal and accessible for the assessment of the trained listener than prosody (Hartelius et al., 2003; Näström & Schalling, 2020).

In addition to familiarity with the language, an increasing body of evidence suggests that listeners’ experience and training can also matter (Krieman et al., 1990; Edie & Baylor, 2006; Walsh et al., 2008; Smith et al., 2019; Carvalho et al., 2020). Many studies dedicated to perception of speech affected with HD take one group of listeners as a source of assessments. Researchers often focus on either untrained (Anand & Stepp, 2015; Pell et al., 2006), trained non-expert (Klopfenstein, 2015; Ma et al., 2010; Whitehill et al., 2003) or trained expert listeners (De Letter et al., 2007; Harris et al., 2016; Martens et al., 2011; Plowman-Prine et al., 2009). Relatively few studies have compared perception of speech affected by HD, or any other dysarthria, in groups of listeners with different levels of familiarity with it. Nevertheless, there is conflicting evidence regarding the role of experience (expert versus untrained general population) in the assessment of dysarthric speech.

A study by Walsh et al. (2008) compares, among others, the perception and rating of (Irish) English speech affected by dysarthria from the point of view of dysarthric speakers, SLT’s, and untrained listeners. The authors found no significant differences between the groups, but the intra-rater reliability was lower for the trained listener group suggesting that their perceptual judgements could have changed during the task. This lower intra-rater reliability for the SLT group also suggests that training might influence speech perception. Walsh et al. (2008) confirmed the findings of Kreiman et al. (1990), who demonstrated that untrained listeners employ comparable strategies when listening to dysphonic populations, while trained listeners differ on an individual basis. A similar conclusion regarding a trained group’s varying strategies was also reached by Wolfe et al. (2000), who found that trained listeners can become more sensitive to ‘high frequency noise components as dimensions of dysphonic voice quality’ (p.703).

In line with the findings of Walsh et al. (2008), Smith et al. (2019) reported no significant differences between the groups in their investigation into how trained and untrained listeners rated intelligibility of speech affected with HD. Some studies, however, demonstrate that groups of listeners with different expertise tend to rate speech of PwPD differently (Verkhodanova et al., 2019, 2020). In a longitudinal case study by Verkhodanova et al. (2019) both trained and untrained listeners assess global ‘healthiness’ of a single speaker with PD similarly: both groups rated the recordings made at a later stage as less healthy than the earlier ones. However, trained listeners’ ratings showed a steeper trend towards the ‘less healthy’ scores for recordings made at a later stage. In another study, Verkhodanova et al. (2020) explored the perception of PD speech by Dutch and Czech trained and untrained listeners. They investigated the effect of experience and familiarity with the language on identification of healthiness of speech and of intended sentence type intonation in speech of PwPD. The findings demonstrate that both expertise and familiarity with the speakers’ language act as important factors in listeners’ perception of PD speech. Additionally, identification accuracy depends on a task type: untrained listeners outperformed trained listeners in the task of speech healthiness identification, while for the identification of sentence type intonation trained listeners were more accurate (Verkhodanova et al., 2020).

Carvalho et al. (2020) demonstrate that regarding intelligibility ratings, expertise with dysarthria and experience specifically with PD are important. The authors show that neurologists working with PD give slightly higher intelligibility scores than SLT’s working with adult dysarthria, and higher than other tested listeners groups: listeners familiar with PD, general untrained population, and PwPD themselves. Interestingly, Carvalho et al. (2020) found homogeneity of the ratings across the untrained listeners with no difference between PwPD, relatives of PwPD, and a general population group unfamiliar with PD. The authors conclude that healthcare professionals who work with dysarthria are more likely to understand the speech of PwPD than the groups of untrained listeners (Carvalho et al., 2020).

1.3. Focus of this study

The current study explores the global perceptual judgements of speech ‘healthiness’ similar to the studies by Verkhodanova et al. (2019, 2020) and following the ideas of Kreiman et al. (1993), Weismer et al. (2001), Sussman and Tjaden (2012), Maryn and Debo (2015). However, previous research on speech ‘healthiness’ does not provide any insights into acoustic nature of such perceptual
judgements of speech. To address this gap, the current study focuses on three research questions. First, we investigate how listeners’ perception of speech healthiness correlates with acoustic cues that reflect speech production issues of PwPD. Second, we explore which acoustic features in speech of PwPD serve as better predictors of listeners’ perception of speech healthiness. Third, we investigate whether the set of acoustic cues best predicting the perception of healthiness differs between listeners with different experience in treating speech disorders (hereafter, SLT experience) and between listeners with different familiarity language spoken of PwPD (hereafter, language familiarity). Knowing whether the acoustic aspects of speech are predictive of perceived healthiness of speech can potentially help with the early diagnosis of dysarthria. At the same time, understanding the specific acoustic cues that affect listeners’ perception of healthiness would be beneficial adjuncts for the existing dysarthria treatments in PD.

To address these questions, we performed an online experiment with three groups of listeners: Trained listeners who speak Dutch (hereafter, trained-Dutch group), untrained listeners who speak Dutch (hereafter, untrained-Dutch group), and untrained listeners who do not speak Dutch (hereafter, untrained-non-Dutch group). Therefore, we converted the acoustic cues to a set of features that can be measured directly from the speech signal and reduce listeners’ perceptual judgements about healthiness of speech to yes/no responses in a speech classification task. Regarding the first research question, we hypothesized that listeners’ responses can be predicted from conventional acoustic features that are clinically interpretable (Brabenec et al., 2017). Secondly, since monopitch is described as one of the most ‘deviant’ dimensions in speech of PwPD contributing to many aspects of successful communication (Bunton et al., 2007; Ma et al., 2010; Kuo et al., 2014; Anand & Stepp, 2015; Verkhodanova 2019), we expected that $f_0$ variance would be among the main predictors of listeners’ responses to the classification task. Thirdly, we hypothesized that, in line with observations by Näström and Schalling (2020) and Hartelius et al. (2003), listeners with SLT experience would rely primarily on voice quality and phonation features when classifying dysarthric speech. We also expected that listeners with no training but with the similar language background would rely on the mix of phonation, articulation and prosody related features (DeBodt, Huici, & Van DeHeyning, 2002; Hartelius et al., 2003), while listeners unfamiliar with the speakers’ language would rely more on universal components of prosodic insufficiencies such as monopitch (Whitehill, 2010; Pinto et al., 2017).

2. Materials and methods

We conducted an experiment to investigate which cues in speech lead listeners with different language backgrounds and expertise to classify healthiness of speech. In this section we describe the three groups of participants (2.1), the creation of stimuli and collection of data (2.2), the experimental procedure (2.3), and data analysis (2.4).

All data was anonymized at the stage of data collection, with researchers being blind to any personal information of the participants. The data collection was approved by the Medical Ethics Committee of the University Medical Center Groningen.

2.1. Participants

One hundred ninety-three listeners were recruited via convenience sampling and participated in the online experiment. We had to exclude 24 participants who reported hearing loss, were familiar with speech and language disorders but received no SLT training, or who had finished the experiment faster than the time of our test run (18 min) performed by one trained phonetician experienced in speech and hearing pathologies. The remaining 169 listeners with different education and language background were split into three groups:

1) trained-Dutch group: 18 trained listeners who speak Dutch, have completed four years of university level SLT training and had SLT working experience in Dutch. Out of 18 trained-Dutch listeners, seven have experience working with neurodegenerative disorders with three listeners experienced specifically with PD (>12 years).
2) untrained-Dutch group: 27 untrained listeners, native Dutch speakers who reported no prior professional experience with speech disorders;
3) untrained-non-Dutch group: 124 untrained listeners, who reported no prior professional experience with speech disorders or working knowledge of Dutch. Among the diverse linguistic backgrounds, the biggest subgroups of untrained-non-Dutch listeners were listeners who are native speakers of Germanic languages ($n = 14$) or Slavic languages ($n = 101$). 112 listeners were recruited outside of the Netherlands, 12 were recruited in the Netherlands within their first months of their stay in the Netherlands.

Table 1
Participant (listeners) demographics. Age is given in years. DT stands for Dutch trained listeners, DU - for Dutch untrained listeners, nDU - for non-Dutch untrained listeners.

| Group name | Age: mean/SD (range) | Gender (F:M) | Native languages |
|------------|----------------------|--------------|-----------------|
| DT group, n = 18 | 34/13.2 (21–64) | F:M - 17:1 | Dutch or/and Frisian/German |
| DU group, n = 27 | 25.2/8.2 (18–60) | F:M - 17:9 and 1 nonbinary | Dutch and: Arabic, English, Frisian, Polish, Spanish, Turkish |
| nDU group, n = 124 | 36.1/13.7 (18–79) | F:M - 90: 34 | Adyghe, English, Estonian, Finnish, German, Hebrew, Hungarian, Indonesian, Italian, Kabardian, Lithuanian, Portuguese, Russian, Turkish, Ukrainian |
Demographics of the participants are presented in Table 1.

The consent form for the listeners was accompanied by a short questionnaire, which inquired into the demographic background and language and expertise backgrounds.

2.2. Data collection and stimuli

We collected speech data from 60 Dutch native speakers: 30 PwPD and 30 control speakers. Prior to the onset of the listening experiment, the severity of each speaker’s dysarthria was assessed by four experienced SLT’s. Each listened to a short sample of spontaneous speech from each speaker and independently assigned an estimate of severity to each speaker on a scale of mild, moderate or severe (Klopfenstein, 2015). Assessors were in excellent agreement, kappa = 0.76 (Fleiss, 2003).

All speakers, both PwPD and control speakers, reported (corrected-to) normal vision and hearing. They all provided informed consent. Exclusion criteria for patients were cognitive problems as assessed by the Minimal Mental State Examination (MMSE <26), brain damage caused by (a) stroke(s) that inflicted aphasia and/or apraxia of speech, and language and/or (motor) speech disorders other than dysarthria. Exclusion criteria for the control speakers were cognitive problems (MMSE <26), brain damage, and language and/or (motor) speech disorders. One inclusion exception was made for a speaker with PD whose MMSE score was 25 due to difficulty in the drawing part of the task. All PwPD were on stable dopamine replacement medication and were recorded in the medication state ON, none had deep brain stimulation. The demographics of all speakers can be found in Table 2.

The recording protocol included prolonged phonation, free speech elicitation (interviews with open questions), picture and short video clip descriptions, reading of ‘The North Wind and the Sun’ passage, a diadochokinesis test, and prosody elicitation tasks targeting production of lexical stress, boundary marking, sentence type and focus intonations (Martens et al., 2011). The recording sessions took place in quiet rooms with the TASCAM DR-100 recorder and an external Sennheiser-e865 microphone placed at a distance of approximately 40 cm from the participant.

Stimuli were created from the recordings of the interview and reading tasks. Decision on including these tasks was based on previous findings: Kempler and Van Lancker (2002) demonstrated that perception of HD speech differs depending on the speech task with which it was elicited. We used one fragment of 3–4 s per speaker taken from the spontaneous speech material and one fragment of 2–3 s per speaker taken from the reading material. The stimuli from interviews and reading we extracted from declarative sentences, did not include artefacts or stuttering and consisted of at least four words.

There were 117 stimuli for the healthiness classification task: 58 phrases from the interviews, 59 phrases from the reading task. There were fewer stimuli from the interviews (58 instead of 60) due to two cases of technical issues in the beginning of the protocol leading to two damaged recordings. The lower number of stimuli from the reading task (59 instead of 60) was a result of reading difficulties of one patient with PD. All (fully anonymized) speech samples that were used as stimuli in the perception experiment were intensity normalized in Praat (Boersma & Weenink, 2020) and did not contain any personal information.

2.3. Procedure

Participants completed the online experiment implemented in JavaScript using jsPsych library (deLeeuw, 2015) and running on the JATOS platform, version 3.5.3 (Lange et al., 2015). The procedure for the task of healthiness classification consisted of a main part preceded by a short practice session meant to acquaint participants with the task. Stimuli were presented in a randomized order, with each stimulus appearing only once. The language of instruction was either Dutch for the Dutch-speaking participants or English or Russian for participants who did not report working knowledge of Dutch. Participants were blind to the diagnosis of the speakers. The instruction for the classification task was as follows: ‘You will listen to sentences spoken by different people. You will judge how healthy they sound to you.’ After listening to each recording, listeners answered the question ‘Did this voice sound healthy to you?’.

The concept of healthiness was not defined in the experiment and listeners had to rely on their own impressions of it. There were three possible answers: ‘Yes’, ‘No’ and ‘I don’t know’. In addition, for every response listeners were requested to specify whether they felt ‘rather sure’ or ‘rather unsure’ about their assessment. Results of the experiments were stored in a JSON format file with participants’ responses assigned numerical values, that was converted to CSV by means of a Python script for further analyses. All subsequent analyses were conducted only on definitive responses, with all ‘I don’t know’ responses removed from the dataset (6.6% of all the answers, for trained-Dutch group it was 2.6% of all the answers, for untrained-Dutch group 2.3%, for untrained-non-Dutch group 8.5%).

Fleiss’ Kappa interrater reliability for multiple categorical variables and multiple raters was calculated for different listener groups. The resulting values were between 0.48 and 0.53 representing fair to good agreement beyond chance (Fleiss et al., 2003) for all answer types. Exceptions were poor agreement for “I don’t know” answers in every listener group.

Table 2
Speaker demographics. Age and duration of disease are given in years.

| Group name | Age: mean/SD (range) | Gender (F: M) | Diagnosis | Disease duration mean, range | MMSE score mean, SD |
|------------|----------------------|---------------|-----------|-----------------------------|---------------------|
| PD         | 65.6/9.8 (45–87)     | F:M - 9:21    | Idiopathic PD; mild to severe dysarthria with one speaker with no dysarthria | Mean: 10.7 Range: 31 none | 27.8 (1.4) |
| Control    | 65.9/7.9 (45–79)     | F:M - 12:18   | none      | none                        | 29.0 (1.1) |
2.4. Data analysis

In the current study we explore whether a set of conventional acoustic features can be predictive of the participants' responses in the speech healthiness classification task. We focus on conventional acoustic features because they are clinically interpretable and can be correlated with auditory-perceptual assessments of dysarthria (Brabenec et al., 2017). We used two different feature sets: demographic information features and acoustic features for the classification of the answers related to the perceived healthiness. As a means of predictive analysis we used the Random decision forest ensemble learning method to classify listeners' responses because of its high accuracy, suitability for smaller and imbalanced datasets, and robustness against correlated predictors (Breiman, 2001).

2.4.1. Demographic information features

Demographic information features were used to evaluate the performance of the participants and test whether the model can predict listeners' responses based on the demographic information about the speakers. This was done to examine whether listeners are more sensitive to the presence/absence of the diagnosis and to the duration of the disease (since diagnosis), or rather to the other changes related to speakers' gender, age, or self-reported dialectal pronunciation. This feature set included seven features: speaker participant number, listener participant number, presence or absence of the diagnosis, duration of the disease, speaker age and gender, as well as whether speakers consider themselves speaking a dialect of Dutch. All features were obtained from the questionnaire. The last feature was derived from the answers to the question about other people's perception of participants' speech ('Spreekt u thuis dialect? Kunnen mensen horen waar u vandaan komt als u praat?' - 'Do you speak in a dialect at home? Can people hear where you are from when you talk?'). For the trained-Dutch group we added two features to explore whether information about experience with neurodegenerative disorders or experience specifically with PD will allow our model to classify the responses of trained-Dutch group more accurately. The full list of demographic features is presented in Appendix A.1.

2.4.2. Acoustic features

The second feature set consisted of acoustic measurements to test if the model can reliably predict listeners' answers based on the acoustic information extracted from the speech signal. Acoustic analysis was performed on the prolonged phonation of /a:/, the stimuli set, and on the stimuli extracted with the surrounding context of 1/3 of the length of the stimulus.

Selection of the acoustic features was based on the previous research that listed the affected characteristics in speech of people with HD (Darley et al., 1969a; Brabenec et al., 2017) and other acoustic research concerned with voices affected by dysarthria or dysphonia (Muhammad et al., 2011; Kim et al., 2011; de Keyser et al., 2016). In the study by Brabenec et al. (2017) the authors provided conventional tasks and the feature set they recommend for exploratory HD dimensions analysis, which included measurements of phonation, articulation and prosody. We included several conventional acoustic features characterizing all three aspects of speech production - phonation, articulation and prosody - to test whether listeners are more sensitive to the HD-associated articulation, prosody and/or phonation changes (Brabenec et al., 2017). Another motivation to include the selected acoustic features was the possibility to automatically compute them, which allows for replicability of the research and minimizes the researcher assessment bias.

In this study we included features related to the prosody domain: fundamental frequency variances, articulation and speech rates, ‘inappropriate silences’ (Darley et al., 1969), articulation: vowel space area, vowel articulation index, and phonation and voice quality: means and standard deviations of first two formants, jitter, harmonics-to-noise ratio, maximum phonation time, fundamental frequency variance of prolonged phonation. Because the recordings in the current study were done without strict supervision over the distance between speaker and the microphone, the intensity measurements had to be excluded from the analysis. The details of the 18 acoustic features are described below.

Fundamental frequency ($f_0$) variance calculation was based on the $f_0$ tracking in the stimuli and stimuli with the context taken from the recordings of the interview and of the read passage. The calculation was done by means of a Python script and the Speech Signal Toolkit (SPTK) (SPTK, 2017) based on the robust algorithm for pitch tracking (RAPT) (Talkin, 1995) and included normalization to factor out individual and gender pitch differences. Two other prosodic features, speech rate and articulation rate, were calculated by means of a Praat script (De Jong & Wempe, 2009). Speech rate was measured as the number of syllables divided by the total time of the recording, articulation rate - as a number of syllables divided by phonation time in that recording. The rate measurements were performed on the same stimuli and on stimuli with fixed context. Another feature belonging to the prosody domain, inappropriate silences, was calculated from the results of the same Praat script by De Jong and Wempe (2009) as the number of pauses relative to total speech time after removing periods of silence lasting less than 60 ms. (Brabenec et al., 2017). The measurements were performed on the same stimuli and stimuli with the context set. Concerning phonation/voice quality measurements of means and standard deviations of the first two formants (F1 and F2), maximum phonation time (MPT), jitter and harmonics-to-noise ratio (HNR), all were calculated from the prolonged phonation (/a:/) recordings and by means of a Praat script. Next, two articulation features included in the analysis are conventional measures capturing vowel centralization. First, the Vowel Space Area (VSA) was constructed by the Euclidean distances between the F1 and F2 coordinates of the vowels /i/, /u/, and /a/ (triangular VSA) and was expressed as follows (Liu et al., 2005):

$$0.5 | (F1/i*/(F2/a/ - F2/u/) + F1/a*/ (F2/a/ - F2/i/) + F1/u*/ (F2/i/ - F2/a/)|$$

Second, the Vowel Articulation Index (VAI) is based on the description in the study by Roy et al. (2009), where it is described as maximally sensitive to vowel centralization and decentralization. VAI is expressed as:

$$(F2/i/ + F1/a/)/F2/a/ + F2/a/ + F1/a/)$$
Table 3
Results of the model with demographic predictors for every listener group. DT - Dutch trained listeners, DU - Dutch untrained listeners, nDU - non-Dutch untrained listeners. In the confusion matrix level 0 refers to ‘healthy’ and level 1 – to ‘unhealthy’.

| Group/results | Train set |       |       |       |       | Test set |       |       |       |       |
|---------------|-----------|-------|-------|-------|-------|----------|-------|-------|-------|-------|
|               |           | DT    | DU    | nDU   |       | DT       | DU    | nDU   |       |       |
| Confusion matrix: reference by prediction |           | 0 737 | 122   | 0 1182| 217   | 0 4769   | 996   | 1 423 | 2910 |       |
| Accuracy      |           | 87.3% | 86.3% | 84.4% |       | 84.3%    | 84.4% | 85.2% |       |       |
| 95% CI        |           | [0.8544, 0.8894] | [0.8479, 0.8771] | [0.8364, 0.8514] |       | [0.8118, 0.8713] | [0.8191, 0.8668] | [0.8404, 0.863] |       |       |
| NIR           |           | 0.55  | 0.59  | 0.57  |       | 0.58     | 0.58  | 0.56  |       |       |
| p-value [Acc > NIR] |           | <2.2 × 10^{-16} | <2.2 × 10^{-16} | <2.2 × 10^{-16} |       | <2 × 10^{-16} | <2.2 × 10^{-16} | <2.2 × 10^{-16} |       |       |
| Kappa         |           | 0.74  | 0.71  | 0.68  |       | 0.68     | 0.67  | 0.70  |       |       |
| McNemar’s test p-value |           | 1.36 × 10^{-5} | 9.38 × 10^{-16} | <2.2 × 10^{-16} |       | 0.014    | 3.04 × 10^{-12} | <2.2 × 10^{-16} |       |       |
| Sensitivity   |           | 0.92  | 0.94  | 0.92  |       | 0.90     | 0.95  | 0.92  |       |       |
| Specificity   |           | 0.81  | 0.76  | 0.75  |       | 0.77     | 0.70  | 0.77  |       |       |
The vowels were extracted from the fourth sentence of the ‘North Wind and the Sun’ text, which contained all three vowels. The full list of acoustic features with the feature labels and descriptions is presented in Appendix A.2.

2.4.3. Random decision forests method

We used Random decision forests or Random Forest (RF) method. This is a technique for predictive modelling, which is based on a collection of unpruned classification or regression trees that are induced from the training data using random feature selection in the process of the tree induction (Thambi et al., 2014; Breiman, 2001). The RF method is known for its high accuracy, suitability for smaller and imbalanced datasets, and for its robustness against correlated predictors and multicollinearity issues (Breiman, 2001). RF has been used for diverse purposes, including models for speech recognition (Su et al., 2007), for background noise classification (Saki & Kehtarnavaz, 2014), or for speech detection (Thambi et al., 2014). In the current study, RF was used to classify the collected responses from the experimental task on recognition of a set of speech stimuli as healthy or unhealthy.

For the purposes of the analysis, we used two RF models. The first model predicts listeners’ responses based on seven demographic features (predictors) to examine whether listeners are more sensitive to speech in the presence/absence of the diagnosis and to the duration of the disease (since diagnosis), or rather to the other changes related to speakers’ gender, age, or self-reported dialectal pronunciation. The seven predictors used for the first model were: presence or absence of the diagnosis, speaker age and gender, self-reported dialect pronunciation, listener ID, speaker ID, and duration of the disease represented as a vector with four values corresponding to control speakers, and three disease duration periods calculated as 3 quantiles from the ordered vector of the disease duration values. We have added listener ID and speaker ID to see how important their contribution is to the model, to understand whether the responses are more dependent on individual rating patterns of each listener or some individual characteristics of the speakers rather than on predictors that are constant across speakers. Therefore, the formula for the first RF model was: response ~ diagnosis + disease duration + speaker age + speaker gender + dialect + listener ID + speaker ID. Two additional predictors for the DT group were experience with neurodegenerative disorders (+ neurodegenerative exp) and specifically experience with HD caused by PD (+ PD exp).

The second model predicts listeners’ answers based on the acoustic information from the speech signal, examining whether conventional and objective acoustic measurements are representative of subjective (perceptual) responses of listeners. The second model used 18 conventional acoustic features that were described earlier. Therefore, the formula for the second model was: response ~ 8 prosodic features + 8 voice quality/phonation features + 2 vowel articulation features.

All the RF analyses were performed in R with the package RandomForest (Liaw & Wiener, 2002). For both models we set the number of input predictors randomly selected at each node of a given tree in the forest (mtry) to be a square root of the number of predictors as suggested in Breiman (2001), thus mtry was 3 and 4 for the two models correspondingly. Due to relatively small datasets (on average 6000 data points per group) we set the total number of trees to grow in the forest (ntree) to 500. For every group (trained-Dutch, untrained-Dutch, untrained-non-Dutch) the model was trained on 70% of the dataset and tested on 30% of the dataset. Since our goal was not to achieve the highest possible predictive accuracy, we intentionally did not include any optimization methods to avoid possibility of additional bias.

For each model we report accuracy results and the ‘out-of-the-bag’ (OOB) error, prediction estimate, the average estimate of prediction errors across all trees in the forest for those observations that are left out of the randomly sampled ‘bag’ of data used to train the trees. We estimate the importance of the prediction used in the models as reflected by two measures: Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI) index, or Mean Decrease Gini. MDA is a measure of decrease in accuracy depending on the presence or absence of specific variables, its value ranges from 0 to 100. MDI is a measure of decrease in data partition impurity during

![Fig. 1(a).](image-url) Variable importance for the model trained on demographic predictors from the dataset of trained-Dutch group’s responses.
the classification, its value ranges from 0 and there is no set maximum: depending on the data it weighs the impurities by the raw counts, not by the proportions (Liaw & Wiener, 2002; Hur et al., 2017).

3. Results

3.1. Demographic predictors

The run of the RF method on the DT training dataset showed an OOB error of 16.9%. The results of the model with demographic predictors are summarized in Table 3 including confusion matrices and statistics: namely, accuracy and confidence interval, kappa coefficient, no-information-rate (NIR) and the results of a binomial test of whether the accuracy of the model is significantly greater than NIR (p-value [Acc > NIR]), as well as McNemar’s test for model bias, the model specificity (rate of True Positives captured by the algorithm) and sensitivity (rate of True Negatives captured by the algorithm). The importance of the predictors used by the model and expressed by MDA and MDI variable importance measures is depicted in Fig. 1(a). We evaluated the variable importance measures by means of the rfPermute library that allows the estimation of statistical significance of variable importance measures by permuting the response variable and then fitting the same RF model again (Archer, 2021), the results of the estimation (number of permutations n = 100) are depicted on Fig. 1(b).

On the untrained-Dutch training set, the RF model yielded an OOB error rate of 14.7%. The results of the model are summarized in Table 3, and variable importance is plotted in Fig. 2(a) with evaluation by permutation (n = 100) plotted in Fig. 2(b).

The RF model run on the untrained-non-Dutch train set demonstrated an OOB error rate of 15.8%. The results of the model are summarized in Table 3, and variable importance is plotted in Fig. 3(a) with evaluation by permutation (n = 100) plotted in Fig. 3(b).

‘Disease duration’ was the most important predictor for the demographic model independent of the listener group. ‘Speaker ID’ and ‘Speaker age’ were consistently in the top three predictors significantly contributing to the accuracy as measured by MDA. ‘Diagnosis’ was the lowest contributing predictor to accuracy, while being ranked higher by the MDI impurity measure.

3.2. Acoustic predictors

The results of the model with acoustic predictors for each listener group are summarized in Table 4 including confusion matrices and statistics. The model trained with acoustic parameters showed an OOB error rate of 16.9% on the trained set of the responses from the trained-Dutch group. Variable importance with colour-coded feature domains is presented in Fig. 4(a). The evaluation of variable importance scores by permutation (n = 100) showed that all acoustic predictors were found to be statistically significant (p ≤ 0.05) for each listener group.

On the train set of untrained-Dutch group’s responses, the acoustic model showed an OOB error rate of 16.6%. Variable importance of the model trained on untrained-Dutch dataset with colour-coded feature domains is presented in Fig. 5.

The model trained on the train set of untrained-non-Dutch group responses showed a 14.7% OOB error rate. Variable importance of the model trained on untrained-non-Dutch dataset with colour-coded feature domains is presented in Fig. 6.

‘Speech rate’ measured in stimuli with context and maximum phonation time, and ‘MPT’, appeared to be the most important predictors for the acoustic model for all groups contributing to the accuracy. The articulation-related predictor ‘VAI’ also appeared to

![Variable Importance with significance](image-url)

Fig. 1(b). Variable importance for the model trained on demographic predictors from the dataset of trained-Dutch group’s responses, with significant scores (p ≤ 0.05) coloured red.
be very important, being third for the trained-Dutch and untrained-Dutch groups, and fourth for the untrained-non-Dutch group. Inappropriate ‘Silences’ measured both in stimuli and in stimuli with context were the least contributing predictors to accuracy for the Dutch speaking groups, while being ranked higher by the MDI measure.

### 3.3. First language background influence on recognition predictors

In determining if language background has any influence on importance of certain demographic and conventional acoustic features for the model, we ran the analysis separately on the two largest subgroups of the untrained-non-Dutch group: listeners with Germanic (non-Dutch) languages as their first languages, and listeners with Slavic languages as their first languages.

#### 3.3.1. Germanic language background

Results of the model with demographic predictors trained on the subset of data from Germanic listeners yielded a 15.6% OOB error estimate. Demographic model results for the Germanic and Slavic listeners are summarized in Table 5. From the variable importance analysis, it appeared that ‘Disease duration’ was once again the most important predictor, followed by ‘Speaker ID’, ‘Speaker age’,
Diagnosis' and 'Speaker gender.' Dialect' and 'Listener ID' had the lowest contributions. Evaluation of variable importance scores by permutations (n = 100) demonstrated that for MDA only 'Listener ID' and 'Dialect' predictors did not show statistical significance, while for MDI there were four predictors with scores that did not reach significance: 'Listener ID', 'Dialect', 'Speaker gender' and 'Speaker age'.

Training the model with acoustic predictors demonstrated a 18.6% OOB error estimate, results are presented in Table 6. Variable importance is depicted in Fig. 7. Evaluation of variable importance scores by permutations (n = 100) demonstrated that for MDA all acoustic predictors showed significant scores (p ≤ 0.05).

3.3.2. Slavic language background
The model with demographic predictors of the Slavic data subset yielded similar results to the model run on the Germanic subset of untrained-non-Dutch group dataset: the OOB error estimate was 15.5%. Results of the demographic model are presented in Table 5. The variable importance analysis provided a similar picture to the model run on Germanic untrained-non-Dutch subset, with predictors of 'Speaker age' and 'Speaker gender' shifted higher in terms of accuracy importance. In terms of the purity index, results of the variable importance analysis for models trained on both Germanic and Slavic subsets showed that 'Disease duration' and 'Speaker ID' were the most important predictors. Evaluation of variable importance scores by permutations (n = 100) demonstrated that for MDA all
Table 4
Results of the model with acoustic predictors for every listener group. DT - Dutch trained listeners, DU - Dutch untrained listeners, nDU - non-Dutch untrained listeners. In the confusion matrix level 0 refers to ‘healthy’ and level 1 – to ‘unhealthy’.

| Group/results | Train set | | | | | Test set | | | | |
|---------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|               | DT        | DU        | nDU       | DT        | DU        | nDU       | DT        | DU        | nDU       | DT        |
| Confusion matrix: reference by prediction | 0  | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Accuracy | 86.1% | 86.1% | 85.4% | 84.2% | 83.6% | 85.8% | 0.55 | 0.59 | 0.57 | 0.58 | 0.58 | 0.56 |
| 95% CI | [0.842, 0.8783] | [0.8452, 0.8749] | [0.8462, 0.8608] | [0.81, 0.8698] | [0.8101, 0.8588] | [0.8465, 0.8688] |
| NIR | 0.55 | 0.59 | 0.57 | 0.58 | 0.58 | 0.56 |
| p-value [Acc > NIR] | <2 × 10⁻¹⁶ | <2.2 × 10⁻¹⁶ | <2.2 × 10⁻¹⁶ | <2 × 10⁻¹⁶ | <2.2 × 10⁻¹⁶ | <2.2 × 10⁻¹⁶ |
| Kappa | 0.72 | 0.71 | 0.70 | 0.67 | 0.65 | 0.71 |
| McNemar's test p-value | 0.017 | 9.23 × 10⁻⁷ | <2.2 × 10⁻¹⁶ | 0.26 | 1.23 × 10⁻⁶ | 9.3 × 10⁻⁸ |
| Sensitivity | 0.90 | 0.92 | 0.90 | 0.88 | 0.92 | 0.90 |
| Specificity | 0.82 | 0.78 | 0.79 | 0.79 | 0.73 | 0.80 |
demographic predictors showed significant scores (p ≤ 0.05), while for MDI there were only two predictors with scores not reaching significance: ‘Listener ID’, ‘Dialect’, and ‘Speaker sex’.

The model with acoustic predictors demonstrated a lower OOB error estimate of 14%. Results for the second model are summarized in Table 6. Variable importance is presented in Fig. 8. Evaluation of variable importance scores by permutations (n = 100) demonstrated that for MDA all acoustic predictors showed significant scores (p ≤ 0.05).

4. Discussion

We explored whether conventional acoustic measurements could be predictive of the way different listener groups classify PD and control speech as healthy or unhealthy. As we discussed in section 1.2, while there exists a body of literature targeting perception of speech of PwPD by listeners with different expertise and experience, the studies are mostly focused on measuring intelligibility scores or on assessing the component-specific perceptual judgements of speech produced by PwPD. A few studies have looked into global perceptual judgements of different listeners assessing speech affected by HD (Weismer et al., 2001; Sussman & Tjaden, 2012), and even fewer provided a cross-linguistic perspective (Näström & Schalling, 2020; Pinto et al., 2017). Therefore, we focused on a more global perceptual judgements of speech of PwPD from two perspectives: influence of SLT experience and influence of language familiarity. To

Fig. 4. Variable importance of the acoustic predictors used by the model on the trained-Dutch group dataset.

Fig. 5. Variable importance of the acoustic predictors used by the model on the untrained-Dutch group dataset.
explore this, we have analyzed two feature sets to see whether listeners' responses can be reliably predicted from any of them. On both acoustic and demographic feature sets, Random Forest analyses demonstrated the accuracy of listeners' responses classification above 84% for every listener group, significantly above guessing level, which demonstrates the importance of the selected features for prediction.

Table 5
Results of the model with demographic predictors for Germanic and Slavic listeners. In the confusion matrix level 0 refers to ‘healthy’ and level 1 – to ‘unhealthy’.

| Group/results | Train set | Test set |
|---------------|-----------|----------|
|               | Germanic  | Slavic   | Germanic | Slavic |
| Accuracy      | 87.4%     | 85.4%    | 87.4%    | 85.1%  |
| 95% CI        | [0.844, 0.8601] | [0.8452, 0.8613] | [0.8386, 0.9043] | [0.8383, 0.8634] |
| p-value [Acc > NIR] | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ |
| Kappa         | 0.74      | 0.69     | 0.74     | 0.69   |
| McNemar's test p-value | $5.12 \times 10^{-5}$ | $<2.2 \times 10^{-16}$ | 3.78 $\times 10^{-5}$ | $<2.2 \times 10^{-16}$ |
| Sensitivity   | 0.93      | 0.94     | 0.95     | 0.93   |
| Specificity   | 0.80      | 0.74     | 0.78     | 0.74   |

Table 6
Results of the model with acoustic predictors for Germanic and Slavic listeners. In the confusion matrix level 0 refers to ‘healthy’ and level 1 – to ‘unhealthy’.

| Group/results | Train set | Test set |
|---------------|-----------|----------|
|               | Germanic  | Slavic   | Germanic | Slavic |
| Accuracy      | 85.4%     | 86.3%    | 86.5%    | 86.4%  |
| 95% CI        | [0.8304, 0.8746] | [0.8522, 0.8679] | [0.8282, 0.8958] | [0.8517, 0.8759] |
| p-value [Acc > NIR] | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ | $<2.2 \times 10^{-16}$ |
| Kappa         | 0.70      | 0.71     | 0.72     | 0.72   |
| McNemar's test p-value | $1.2 \times 10^{-4}$ | 3.06 $\times 10^{-7}$ | 2.1 $\times 10^{-4}$ | 3.6 $\times 10^{-4}$ |
| Sensitivity   | 0.91      | 0.90     | 0.94     | 0.90   |
| Specificity   | 0.77      | 0.81     | 0.77     | 0.81   |

Fig. 6. Variable importance of the acoustic predictors used by the model on the untrained-non-Dutch group dataset.
For the first model based on demographic features of both speakers and listeners, results demonstrated the expected influence of the predictor ‘Disease duration’ on the model’s accuracy for every listener group. The importance of both ‘Listener ID’ and ‘Speaker ID’ predictors suggest the other less random predictors available to the model did not have high predictive power. This indicates the possible importance of unlisted predictors related to both speakers and listeners. It is also clear that in the trained-Dutch and untrained-non-Dutch listeners assessed speech samples more uniformly, as predictor of the ‘Listener ID’ showed very low contribution to the model’s accuracy (and was not found to be significant for trained-Dutch group), while in untrained-Dutch the group mean accuracy would have decreased more prominently if the ‘Listener ID’ predictor with 25% of contribution to the model accuracy were to be randomized. This could have been a sign of a lower inter-rater agreement, but the Fleiss’ Kappa interrater agreement for all listener groups was very similar: from 0.48 to 0.53, which represents good agreement beyond chance (Fleiss et al., 2003). Therefore, given the good agreement for the untrained-Dutch group, this difference in accuracy-related importance for ‘Listener ID’ predictor may be caused by some group characteristic or specific sensitivity of untrained-Dutch listeners to certain unlisted predictor compared to trained-Dutch or untrained-non-Dutch group. This finding is indirectly in contrast with the results of Walshe et al. (2008), where authors demonstrated lesser agreement within their SLT group. This finding of importance of ‘Listener ID’ predictor for the model trained on untrained-Dutch listeners’ responses could also be a result of framing the question in the task (“Did this voice sound healthy to you?”) that was interpreted by the trained group in a more uniform way than by the Dutch untrained group. Further research into
rating patterns of speech healthiness of trained and untrained groups is needed to explore whether the untrained-Dutch group was indeed more sensitive to certain speech aspects.

Another interesting result appeared in the variable importance analysis for the trained-Dutch group: features of listeners’ experience with neurodegenerative diseases and of specific experience with PD negatively influenced the model’s accuracy. Apparently, these features had little predictive power to set apart listeners with more ‘suitable expertise profile’ to more accurately classify their responses. One possible explanation is that with a small subgroup of experienced people these predictors introduced more noise than information into the model: these predictors’ importance did not reach significance. However, there is another possible explanation that it could be potential similarity in strategies applied by everyone in the trained-Dutch group because of the task specifics that is general question about healthiness of voice rather than component-specific perceptual assessments or intelligibility judgements. Based on the findings in Verkhodanova et al. (2021), there is a possibility that given the assessment of only short fragments, when classifying speech as ‘unhealthy’, trained listeners were distracted by their attentive efforts of using specific criteria imposed by their training, while other groups relied on their intuitive impressions of what represents ‘unhealthy’ speech. Additional research with a larger group of listeners with specific experiences with PD is required to reliably test the importance of such predictor as well as to gain more insights into specificity of trained listeners’ perceptual strategies.

Regarding the second model, and the first research question, we found that the subjective responses of different listener groups can be predicted by the conventional objective acoustic features that are used to describe speech and voice of speakers with HD. The accuracy of the model trained on the test sets of responses from different listener groups ranged from 83.6% for the untrained-Dutch group to 85.8% for the untrained-non-Dutch group.

Answering the second research question, in the top three predictors for each listener group there always appear, in different order, both the ‘maximum phonation time’ predictor and the ‘speech rate’ predictor measured in stimuli extracted with the context. For the two Dutch listener groups, the top three predictors also included ‘vowel articulation index’. The ‘maximum phonation time’ predictor is a measurement of glottis efficiency calculated from a separate task of sustained phonation. Listeners have not heard prolonged phonation recordings, therefore the importance of the ‘maximum phonation time’ predictor may be indicative of other important acoustic phenomena present in the stimuli and related to the phonation issues, such as breathing or glottalization.

Another interesting finding is that ‘speech rate’ calculated from the stimuli extracted with their contexts was an important predictor for every group independent of the expertise or language background. Based on the rankings of deviant dimensions in speech of speakers with HD (Bunton et al., 2007; Darley et al., 1969a), rate appears to be much less prominent than monopitch or inappropriate silences. This unexpected finding highlights the need for additional investigations into the influence of rate and its possible correlations with listener’s judgements of healthiness of speech. It is also noteworthy that ‘speech’ and ‘articulation rate’ were often more important for reliable model prediction when calculated from the stimuli taken with the context, while ‘f0 variance’ calculated from the stimuli was in general more important than ‘f0 variance’ calculated from the stimuli with the context. This brings to light that some speech rhythm patterns manifest themselves on longer durations than 3–4 s and that they are still detectable by listeners even when not presented in their entirety. This could also be attributed to short rushes of speech dimension (Darley et al., 1969a) possibly present in speech of people with HD, that might be more reliably measured by the script when it is given a longer speech sample.

Concerning the third research question, whether listeners with different experience with speech disorders are sensitive to different acoustic cues in speech of PwPD, there was no clear tendency of a model relying mostly on prosody or phonation and voice quality for any listener groups. For Dutch speaking listener groups, the top three predictors were the same: ‘maximum phonation time’, ‘speech rate’ in stimuli extracted with context, and ‘vowel articulation index’. However, contrary to our hypothesis, among predictors with the higher contribution to accuracy (that is if randomized, there will be mean decrease in accuracy over 20%) (see Figs. 4–6), predictors from the domain of phonation and voice quality appear to be more important for predicting untrained-Dutch group’s responses. Taking into account a previous study (Verkhodanova et al., 2021), a group of untrained Dutch listeners was more successful at classifying speech of PwPD as unhealthy, these features related to phonation and voice quality can be a valuable source of information for the listeners. It is also in line with observations by Näström and Schalling (2020).

The third research question also focused on whether listeners with different degrees of familiarity with a speaker’s language rely on different acoustic cues when classifying speech sounds as healthy or unhealthy. According to the model, and in line with our original expectations, the Germanic listeners’ responses similar to the Dutch listeners’ responses correlated better with the features related to phonation and voice quality, while for Slavic listeners our model favoured prosodic features. This was another interesting observation regarding demographic features, which highlighted a possible difference in assessment strategies employed by listeners from different language backgrounds. Visible in the variable importance analysis, ‘speakers’ age’ and ‘gender’ appeared to be more important predictors for Slavic listeners’ than for Germanic listeners’ responses, which could be attributed to cultural differences and calls for additional research. However, contrary to our expectations, we found a pattern similar to Slavic listeners in the trained-Dutch group, which is visible in the model’s variable importance (see Fig. 1).

In summary, our findings demonstrate that more global perceptual judgements of different listeners classifying speech of PwPD may be predicted with sufficient reliability from conventional acoustic features. The findings suggest that independent of expertise and language background, when classifying speech as healthy or unhealthy listeners are more sensitive to speech rate, presence of phonation deficiency reflected by maximum phonation time measurement and centralization of the vowels. It is therefore likely that both specifics of the expertise and language background may lead to listeners relying more on the features either from prosody or phonation and voice quality domains. Such findings suggest that these features are more or less representative of universal aspects of acoustic change in speech of PwPD which appear to be prominent for listeners independently of their first language or expertise. This is in line with the finding of Näström and Schalling (2020) who noticed that a Swedish SLT is able, without the interpreter’s help, to distinguish between PwPD with no articulatory impairment and participants with articulatory impairment. This means that
articulatory deficits in speech of PwPD, even though language specific, can be picked up by both trained and untrained listeners who do not speak the language of those PwPD. This also warrants additional research into dependence between the global perception of healthiness and these specific features.

The current study provided evidence that despite expectations Dutch untrained listeners’ responses were better predicted by phonation and voice quality features than trained listeners. Surprisingly, experience with neurodegenerative disorders or specifically with PD had a negative effect on the prediction accuracy of the model. Given that evaluation of the variable importance by permutation showed the lack of significance of these predictors, this result also points to certain limitations of the current study, as enlarged and more balanced groups of listeners including a separate group of trained listeners with PD expertise would provide a clearer perspective on the influence of specific expertise on prediction of the listeners’ responses. Another limitation is that we restricted ourselves to a certain number of predictors, therefore including a broader list of features such as other articulatory and intensity measurements could lead to higher prediction results of the Random Forest model and to a better understanding of acoustic cues important for perceptual judgements of healthiness. Another limitation is a potential bias of Random Forest method in ranking predictors (Strobl et al., 2007).

In the light of the rising risks of PD and growing number of early onset cases (Dorsey et al., 2020), the findings of the current study have similar real-life implications as described in Sussman and Tjaden (2012), where subjective perception of voices as unhealthy may have a negative effect on speakers’ potential employability and/or social activity. The information of specific acoustic changes that trigger listeners to classify speech as ‘unhealthy’ can also provide specific therapeutic targets to enhance communication efficiency of speakers with HD as well as help to work on alleviating the negative attitudes with which speakers with HD can be confronted (Maryn & Debo, 2015; Miller et al., 2006). Such findings also contribute to the growing body of research that recommend both researchers and clinicians to incorporate more global perceptual measures that can help understand and incorporate listeners’ sensitivity to a variety of variables, including voice, prosody, and other speech characteristics (Kent et al., 1989; Sussman & Tjaden, 2012). To further explore subjective perception of healthiness and its acoustic correlates, future research should also investigate additional articulatory measures as well as less conventional acoustic measures. It should also control for the specific HD-related experience in the trained group to be able to differentiate classification patterns in that group and correlate it with the acoustic measures.

The results of this study have two important outcomes for clinical practice. The first outcome is the finding that inexperienced listeners are able to recognize ‘unhealthiness’ in speech of PwPD which is important, because it implies that family members of people at risk of developing PD will be able to detect ‘unhealthiness’ in the speech of their relatives. This means that those who are already close to a speaker at risk will be able to recognize changes in their speech - one of the early signs of PD. Together with the findings of the prominent acoustic predictors that allow inexperienced listeners to recognize ‘unhealthiness’ in speech might help with an early diagnosis of PD as well as early detection of speech problems, which consecutively would allow PwPD to start speech therapy at an earlier stage of the disease progression. The second outcome is the higher importance of same speech features for predicting the responses about speech ‘unhealthiness’ in both Dutch and non-Dutch listener groups. This finding furthers and supports the theoretical underpinning of the possibility to develop language independent automatic systems that can detect symptoms of unhealthy speech of a potential speaker with Parkinson’s disease.

Declaration of interest

The authors report no conflict of interest.

Data availability

Acoustic measurements data that support the findings of this study are openly available in OSF, DOI 10.17605/OSF.IO/FRJC2.

Acknowledgements

We are very grateful to Lea Busweiler, research assistant, for the invaluable help in the data collection. We also thank all the speakers and listeners who volunteered to participate in our study.

Appendix A. List and descriptors of all predictors used in the modelling

| Table A.1 |
|---|---|
| **Demographic predictors** | |
| **Label** | **Description** |
| Disease duration | Duration of the disease expressed as four quantiles of disease duration distribution. |
| Speaker ID | Unique ID number assigned to each speaker |
| Speaker age | Age of each speaker at the moment of recording, expressed in years |
| Speaker gender | Speaker gender |
| Dialect | Absence or presence of self-reported dialectal pronunciation |
| Diagnosis | Absence or presence of PD diagnosis |
| Listener ID | Unique ID number assigned to each listener |
Table A.2

Acoustic predictors (based on Brabenec et al., 2017; Muhammad et al., 2011; Kim et al., 2011; de Keyser et al., 2016)

| Label       | Domain       | Description                                                                 |
|-------------|--------------|-----------------------------------------------------------------------------|
| MPT         | Phonation    | Maximum Phonation Time in seconds was calculated from prolonged phonation of vowel /a:/ MPT characterizes the aerodynamic efficiency of the vocal tract reflected by the maximum duration of the prolonged vowel |
| HNR         | Phonation    | Harmonics-to-noise ratio, the amount of noise in the speech signal, mainly due to incomplete vocal fold closure. HNR is defined as the amplitude of noise relative to tonal components in speech |
| SD F1,/a:/  | Phonation    | Standard deviation of the first formant /a:/ measured in Hz. The high standard deviation of formants indicates the unstable nature of the measured vowel. |
| Mean F1,/a:/| Phonation    | Mean of the first formant of /a:/ measured in Hz. Mean values of formants may consistently differ in disordered and healthy speech. |
| SD F2,/a:/  | Phonation    | Standard deviation of the second formant of /a:/ measured in Hz. The high standard deviation of formants indicates the unstable nature of the measured vowel. |
| Mean F2,/a:/| Phonation    | Mean of the second formant of /a:/ measured in Hz. Mean values of formants may consistently differ in disordered and healthy speech. |
| Jitter      | Phonation    | Jitter is a measure of frequency perturbation. It is defined as the variability of the $f_0$ from one cycle to the next. |
| $f_0$ variance,/a:/ | Phonation | Pitch variance estimation was defined as average of the squared deviations from the mean of $f_0$ variation in frequency of vocal fold vibration |
| Speech rate, stimuli | Prosody | Speech rate in stimuli was measured as the number of syllables divided by the total time of the recording for each stimulus. |
| Speech rate, context | Prosody | Speech rate in context was measured as the number of syllables divided by the total time of the recording for each stimulus extracted with its 1/3 of its length context |
| Articulation rate, stimuli | Prosody | Articulation rate in stimuli was measured as a number of syllables divided by phonation time in the recording for each stimulus |
| Articulation rate, context | Prosody | Articulation rate in context was measured as a number of syllables divided by phonation time in the recording for each stimulus extracted with its 1/3 of its length context |
| $f_0$ variance, stimuli | Prosody | Pitch variance estimation was defined as the average of the squared deviations from the mean of $f_0$ variation in frequency of vocal fold vibration. It was measured for each stimulus. |
| $f_0$ variance, context | Prosody | Pitch variance estimation was defined as the average of the squared deviations from the mean of $f_0$ variation in frequency of vocal fold vibration. It was measured for each stimulus extracted with its 1/3 of its length context. |
| Silences, stimuli | Prosody | Inappropriate silences were defined as the number of pauses relative to total speech time after removing periods of silence lasting less than 60 ms. Measurements were performed for each stimulus. |
| Silences, context | Prosody | Inappropriate silences were defined as the number of pauses relative to total speech time after removing periods of silence lasting less than 60 ms. Measurements were performed for each stimulus extracted with its 1/3 of its length context. |
| VAI         | Articulation | Vowel articulation index, based on formant centralization, was defined as $VAI = \frac{(F1a + F2a) - (F1u + F2u)}{2}$. The vowels were extracted from the fourth sentence of the ‘North wind and the sun’ text, which contained all three vowels. |
| VSA         | Articulation | Vowel articulation index, based on formant centralization, was defined as $VSA = 0.5 \times [(F1i + F2i - F2a) + F1a + (F2u - F2i) + (F1u - F2i)]$. The vowels were extracted from the fourth sentence of the ‘North wind and the sun’ text, which contained all three vowels. |

References

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson’s disease. Journal of Speech, Language, and Hearing Research, 58(4), 1134–1144. https://doi.org/10.1044/2015_JSLHR-S-14-0243

Archer, E. (2021). Package r`permute`: Estimate permutation p-values for random forest importance metrics [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rPermute/index.html. (Accessed 25 December 2021).

Basirat, A., Patin, C., & Moreau, C. (2018). Relationship between perception and production of intonation of French in Parkinson’s disease (pp. 809–813). Proc. 9th International Conference on Speech Prosody. https://doi.org/10.21437/SpeechProsody.2018-163, 2018.

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Computer software]. Version 6.1.4. Retrieved from http://www.praat.org/. (Accessed 2 May 2020).

Brabenec, L., Mekyska, J., Galaz, Z., & Rektorova, I. (2017). Speech disorders in Parkinson’s disease: Early diagnostics and effects of medication and brain stimulation. Journal of Neural Transmission, 124(3), 303–334. https://doi.org/10.1007/s00702-017-1676-0

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bunton, K., & Keintz, C. K. (2008). Effects of a concurrent motor task on speech intelligibility in speakers with Parkinson disease. Journal of Medical Speech-Language Pathology, 16, 141–155. PMID: 21627728

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. Journal of Speech, Language, and Hearing Research, 50, 1481–1495. https://doi.org/10.1044/1092-4388(2007/102

Carvalho, J., Cardoso, R., Guimarães, I., & Ferreira, J. J. (2020). Speech intelligibility of Parkinson disease patients evaluated by different groups of healthcare professionals and naive listeners. Logopedics Phoniatrics Vocology, 1–7. https://doi.org/10.1080/14015439.2020.1785546

Darley, F. L., Aronson, A. E., & Brown, J. E. (1969a). Diagnostic differential of patterns of dysarthria. Journal of Speech & Hearing Research, 12(2), 246–259. https://doi.org/10.1044/1092-4388(1969/005

Darley, F. L., Aronson, A. E., & Brown, J. E. (1969b). Clusters of deviant speech dimensions in the dysarthrias. Journal of Speech & Hearing Research, 12(3), 462–496. https://doi.org/10.1044/1092-4388.12.3.462

De Bodt, M. S., Hulic, M. E. H. D., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. Journal of Communication Disorders, 35(3), 283–292. https://doi.org/10.1016/S0021-9924(02)00065-5

De Jong, N. H., & Wempe, V. (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
