

University of Groningen

## How expertise and language familiarity influence perception of speech of people with Parkinson's disease

Verkhodanova, V.; Coler, M.; Jonkers, R.; Lowie, W.

*Published in:*  
Clinical Linguistics & Phonetics

*DOI:*  
[10.1080/02699206.2021.2003433](https://doi.org/10.1080/02699206.2021.2003433)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2022

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*  
Verkhodanova, V., Coler, M., Jonkers, R., & Lowie, W. (2022). How expertise and language familiarity influence perception of speech of people with Parkinson's disease. *Clinical Linguistics & Phonetics*, 36(2-3), 165-182. <https://doi.org/10.1080/02699206.2021.2003433>

### Copyright

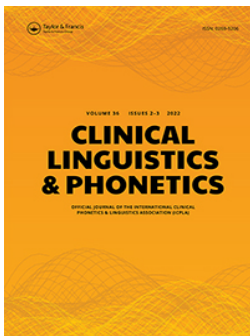
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



## How expertise and language familiarity influence perception of speech of people with Parkinson's disease

V. Verkhodanova, M. Coler, R. Jonkers & W. Lowie

To cite this article: V. Verkhodanova, M. Coler, R. Jonkers & W. Lowie (2022) How expertise and language familiarity influence perception of speech of people with Parkinson's disease, Clinical Linguistics & Phonetics, 36:2-3, 165-182, DOI: [10.1080/02699206.2021.2003433](https://doi.org/10.1080/02699206.2021.2003433)

To link to this article: <https://doi.org/10.1080/02699206.2021.2003433>



© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 22 Nov 2021.



Submit your article to this journal [↗](#)



Article views: 605



View related articles [↗](#)



View Crossmark data [↗](#)

# How expertise and language familiarity influence perception of speech of people with Parkinson's disease

V. Verkhodanova <sup>a,b</sup>, M. Coler<sup>a</sup>, R. Jonkers<sup>b,c</sup>, and W. Lowie<sup>b,c</sup>

<sup>a</sup>Campus Fryslân, University of Groningen, Leeuwarden, The Netherlands; <sup>b</sup>Research School of Behavioural and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands; <sup>c</sup>The Center for Language and Cognition Groningen, University of Groningen, Groningen, The Netherlands

## ABSTRACT

Parkinson's disease (PD) is a progressive neurological disorder characterized by several motor and non-motor manifestations. PD frequently leads to hypokinetic dysarthria, which affects speech production and often has a detrimental impact on everyday communication. Among the typical manifestations of hypokinetic dysarthria, speech and language therapists (SLTs) identify prosody as the most affected cluster of speech characteristics. However, less is known about how untrained listeners perceive PD speech and how affected prosody influences their assessments of speech. This study explores the perception of sentence type intonation and healthiness of PD speech by listeners with different levels of familiarity with speech disorders in Dutch. We investigated assessments and classification accuracy differences between Dutch-speaking SLTs ( $n = 18$ ) and Dutch/non-Dutch speaking untrained listeners ( $n = 27$  and  $n = 124$ , respectively). We collected speech data from 30 Dutch speakers diagnosed with PD and 30 Dutch healthy controls. The stimuli set consisted of short phrases from spontaneous and read speech and of phrases produced with different sentence type intonation. Listeners participated in an online experiment targeting classification of sentence type intonation and perceived healthiness of speech. Results indicate that both familiarity with speech disorders and with speakers' language are significant and have different effects depending on the task type, as different listener groups demonstrate different classification accuracy. There is evidence that untrained Dutch listeners classify PD speech as unhealthy more accurately than both trained Dutch and untrained non-Dutch listeners, while trained Dutch listeners outperform the other two groups in sentence type classification.

## ARTICLE HISTORY

Received 29 October 2020  
Revised 2 November 2021  
Accepted 2 November 2021

## KEYWORDS

Parkinson's disease; dysarthria; speech perception; prosody; speech healthiness

## Introduction

Since the late 1960s there has been an increasing amount of research targeting speech production and speech intelligibility in people diagnosed with Parkinson's Disease (PD). Hypokinetic dysarthria (HD), a speech disorder that is associated with parkinsonism, results from disturbances in muscular control over the speech mechanism (Darley et al., 1969a). It is considered as an additional marker useful for an early diagnosis of PD (Brabenec et al., 2017). It is common to assess and investigate HD based on intelligibility ratings or by means of

**CONTACT** V. Verkhodanova  [v.verkhodanova@rug.nl](mailto:v.verkhodanova@rug.nl)  Campus Fryslân, University of Groningen, Wirdumerdijk 34, Leeuwarden 8911CE, The Netherlands

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

component-specific auditory perceptual assessments of speech (Sussman & Tjaden, 2012). Both methodologies are usually designed for a particular language and utilize language-specific protocols. However, studies have demonstrated that people with HD experience speech changes and communication difficulties long before impairment of intelligibility becomes apparent (Miller et al., 2006). This underscores the need for more ‘global’ perceptual assessment that would better reflect such speech changes (Sussman & Tjaden, 2012). Moreover, with prolonged longevity and migration, there is an increasing need to assess dysarthria in a language unfamiliar to a speech therapist (Näsström & Schalling, 2020).

According to Darley et al. (1969a), the 10 most recognizably affected characteristics of speech changes, so-called ‘deviant’ speech dimensions, in speech affected by 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. Out of these ten, six are attributed to prosodic insufficiency – monopitch, reduced stress, monoloudness, inappropriate silences, short rushes of speech and variable rate (Darley et al., 1969a; Martens et al., 2011).<sup>1</sup> A number of studies have also demonstrated that prosody deficits together with harsh voice and reduced articulation are among the most prominent changes present in the acoustics of speech affected by HD (Anand & Stepp, 2015; Brabenec et al., 2017; Verkhodanova et al., 2019a). Some studies even suggested that the prosodic deficits arising from HD are universal for all languages (Pinto et al., 2017).

Such affected speech production in PD is related to multiple communication issues and changes in how listeners perceive dysarthric speech, with prosody impairment having crucial consequences for the speech intelligibility and for daily communication of speakers with HD (Anand & Stepp, 2015; Carvalho et al., 2020; Pinto et al., 2017). These speech disturbances may affect the quality of life of people with PD. It is reflected in their communication problems and a feeling of social isolation. This may result in tension, depression, resignation and withdrawal from conversation (Miller et al., 2006).

An increasing number of studies explore the efficiency of speech production in people with PD focusing on listeners’ perception. Many researchers have described prominent changes in prosodic characteristics of speech affected with HD. For example, when compared to healthy controls, speakers with PD are less efficient at producing question-statement intonation contrasts (Basirat et al., 2018; Pell et al., 2006) or at conveying lexical and contrastive stress (Martens et al., 2016; Pell et al., 2006). Overall, in the literature, monopitch and monoloudness are said to have the greatest influence on the perception of speech affected by HD and are seen as the most prototypical source of prosodic speech problems for speakers with PD (Anand & Stepp, 2015; Duffy, 2012).

Since the publication of the seminal work by Darley et al. (1969a) on deviant speech dimensions and perceptual classification of different dysarthrias (Darley et al., 1969b), interest in HD production and perception has rapidly grown. According to many studies (for instance, Bunton et al., 2007; Näsström & Schalling, 2020; Sussman & Tjaden, 2012), 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) to spontaneous conversational speech (Bunton & Keintz, 2008;

---

<sup>1</sup>dimension of *imprecise consonants* is sometimes considered as a component of prosodic insufficiency in hypokinetic dysarthria (Darley et al., 1969b; Duffy, 2012).

Bunton et al., 2007). One of the common measures of assessment and management of speakers with dysarthria is the use of speech intelligibility scores. These scores are commonly used as a measure of the severity of the speech disorder (Bunton & Keintz, 2008). Another approach to auditory-perceptual assessment of dysarthria is the use of component-specific perceptual judgements, first described by Darley et al. (1969a). Both approaches have limitations and reliability concerns as summarized in a study by Sussman and Tjaden (2012). These authors suggest exploring global perceptual judgments of speech disorder severity for individuals with multiple sclerosis and PD. Their idea is related to the approach proposed by Weismer et al. (2001) and is in line with the early suggestion 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”.

Methodologies dedicated to obtaining intelligibility scores, component-specific perceptual judgements, and scaled estimates of speech disorder severity are usually developed for a specific language. However, there is an increasing need for dysarthria assessment in a language unfamiliar to the assessor (Näsström & Schalling, 2020). Näsström and Schalling (2020) focused on developing and testing a systematic dysarthria assessment method for SLTs who do not speak the language of an individual with dysarthria. Their results indicate that an SLT who does not speak the target language and performs the assessment according to the method in collaboration with an interpreter shows comparable results to an SLT who speaks the language of an individual with dysarthria (Näsström & Schalling, 2020). This indicates the potential for generalizing assessment methods to many languages providing access to speech-language pathology services to a broader group of people.

In addition to the familiarity with the language, an increasing body of evidence suggests that listeners’ experience and training can also matter (Carvalho et al., 2020; Kreiman et al., 1993; Smith et al., 2019; Walshe et al., 2008). There is conflicting evidence regarding the role of experience (expert versus the untrained general population) in the assessment of dysarthric speech. Work by Walshe et al. (2008) compares the intelligibility rating of (Irish) English speech affected by dysarthria from the point of view of dysarthric speakers, speech and language therapists (SLTs), and untrained listeners. The authors reported mixed results: there were no significant differences between the three groups. However, the intra-rater reliability was lower for the trained listeners. This suggests that the way they assessed speech could have changed during the task. Similarly, a study by Smith et al. (2019) contributed to the understanding of perception of intelligibility of speech affected by PD by comparing ratings performed by trained and untrained listeners. Smith et al. (2019) reported no significant differences between the trained and untrained groups. However, different results can be found in recent studies demonstrating that groups of listeners with different expertise rate speech produced by people with PD differently (Verkhodanova et al., 2019a, 2020). In the longitudinal case study by Verkhodanova et al., (2019a) 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 despite the absence of HD diagnosis. 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. The authors demonstrated that both expertise and familiarity with the speakers’ language are important factors in listeners’ perception of PD

speech. The importance of expertise and experience with PD speech is even clearer in the study by Carvalho et al. (2020) who focused on the intelligibility ratings. The authors showed that neurologists working with PD demonstrate slightly higher intelligibility scores than SLTs working with adult dysarthria, and significantly higher intelligibility scores than other listeners familiar with PD, listeners from the general untrained population, and listeners with PD. Interestingly, the authors found homogeneity of ratings across the untrained listener group with no difference between listeners with PD, relatives of people with PD, and the general population group unfamiliar with PD (Carvalho et al., 2020).

This study investigates the ability of listeners to recognize speech of people with PD as ‘unhealthy’ based on the idea of a more global assessment of dysarthric speech which was proposed and discussed in a number of studies (Kreiman et al., 1993; Maryn & Debo, 2015; Sussman & Tjaden, 2012; Weismer et al., 2001). We investigate whether the classification of speech as ‘unhealthy’ is related to the experience with speech and language disorders as is the case for intelligibility in the study by Carvalho et al. (2020). We are interested in whether healthiness can be related to the changes in the acoustic characteristics of speech without any influence of the semantic content of an utterance. We also explore whether familiarity with speech disorders and with speakers’ language affect listeners’ classification of prosody in PD speech.

To address these issues, we performed an experiment with three groups of listeners: Trained listeners who speak Dutch (hereafter, DT group), Untrained listeners who speak Dutch (hereafter, DU group), and Untrained listeners who do not speak Dutch (hereafter, nDU group). Following the results of Verkhodanova et al., (2019a), we hypothesized that the DT group would most accurately classify PD speech as ‘unhealthy’, and that DT listeners would be more accurate than DU listeners (Martens et al., 2011; Pell et al., 2006). We expected that the DT group would outperform other listener groups at classifying question/statement intonations, similar to the finding that trained listeners understand speakers with PD better than untrained listeners (Carvalho et al., 2020). Given their familiarity with the prosodic system of the Dutch language, we expected DU listeners to more accurately classify prosodic differences in PD speech relative to nDU listeners. We expected that nDU listeners would classify PD speech as ‘unhealthy’ on the basis of their intuition about healthiness – though we supposed they would perform with less accuracy than the DT group. This expectation follows from our assumption that listeners assessing speech in an unfamiliar language are not distracted by the meaning of the speech and accordingly are able to resort exclusively to acoustic impressions. This hypothesis is also in line with the observation that trained listeners are sensitive to changes in speech of people with PD in an unfamiliar language (Näsström & Schalling, 2020). We also hypothesized that the accuracy of responses would differ between nDU listeners with Germanic and non-Germanic language backgrounds due to the differences in the phonetic systems of the languages (Best & Tyler, 2007).

## Materials and methods

We conducted an experiment with three groups of listeners: the DT group, the DU group, and the nDU group with different language backgrounds. We investigated whether the level of familiarity with both speech and language disorders and with speakers’ language affects listeners’ ability to correctly classify PD speech and speech of healthy controls. We also

examined the listener's ability to correctly identify question/statement intonations from audio recordings of control speakers and speakers with PD. The results were subjected to cross-comparisons in the subsequent analysis.

## Materials

### Data collection

Speech recordings were collected from 60 Dutch native speakers: 30 individuals diagnosed with PD and 30 healthy controls (HC). The demographics appear in Table 1.

The severity of each speaker's dysarthria was assessed by four experienced SLTs. 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 absent, mild, moderate or severe dysarthria (Klopfenstein, 2015). Assessors were in excellent agreement, kappa = 0.76 (Fleiss, 2003).

Speakers reported (corrected-to) normal vision and hearing and signed informed consent. Exclusion criteria for speakers with PD were cognitive problems assessed by 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 unrelated to PD. Exclusion criteria for HCs were cognitive problems (MMSE < 26), brain damage, language and/or (motor) speech disorders. One inclusion exception was made for a speaker with PD whose MMSE score was 25 due to the difficulty in the drawing part of the MMSE assessment.

The recording protocol included several language tasks: prolonged phonation, free speech elicitation (interviews with open questions), picture and short video descriptions, reading, 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 e86 microphone placed at a distance of approximately 40 cm from the participant.

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.

### Stimuli

For stimuli creation, we used the recordings of the interviews and reading and the recordings of the prosody elicitation exercise on sentence type. We decided on including both reading and spontaneous speech as research demonstrated that perception of speech

**Table 1.** Speaker demographics. Age and duration of disease are given in years.

Group name	Age: Mean/SD (range)	Gender: %F	Diagnosis	Disease duration: Mean/SD (range)	MMSE score: Mean/SD (range)
PD, n = 30	65.6/9.8 (45– 87)	30%	Idiopathic PD; mild to severe dysarthria except for one speaker	10.8/7.2 (1– 32)	27.8/1.4 (25– 30)
HC, n = 30)	65.9/7.9 (45– 79)	40%	None	none	29.0/1.1 (26– 30)

affected by HD differs depending on the speech task with which it was elicited (Kempster & Van Lancker, 2002). Sentence type intonation was included because previous studies showed that listeners are most sensitive to sentence type intonation compared to other linguistic prosody types in the experiment on differentiating between HC and PD speech (Verkhodanova et al., 2019b).

For each speaker, we used short fragments of 3–4 seconds from the spontaneous speech task and of 2–3 seconds from the reading task. The stimuli from interviews and reading were selected according to three criteria: they should not include artefacts or stuttering, they should consist of at least four words, and they should be extracted from declarative statements. For the stimuli from the sentence type task, we selected one out of five pairs of phrases per speaker. The phrases were syntactically identical but different in question or statement intonation (e.g., [*Heeft hij de toets gehaald?*] – ‘[Has he] passed the test?’ and [*Hij heeft de toets gehaald.*] – ‘[He has] passed the test’).

There were 245 stimuli: 58 phrases from the interviews, 59 phrases from the reading task, and 120 phrases from the exercise targeting question/statement intonation elicitation. 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 interviews. The lower number of stimuli from the reading task (59 instead of 60) was a result of reading difficulties of one speaker 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 sensitive information.

### **Participants in the perception experiment**

In total, 193 listeners found through convenience sampling took part in the experiment. We excluded 24 people who reported hearing loss, were familiar with speech language disorders but did not receive SLT training, or who finished the experiment faster than the time of our pilot run (18 minutes). The remaining 169 listeners belonged to one of three groups:

- (1) The DT group: 18 Dutch-speaking listeners with four years of university level SLT training and working experience. Out of these, seven had experience with neurodegenerative disorders, while three listeners had specific experience with PD (>12 years).
- (2) The DU group: 27 Dutch-speaking listeners, reporting no prior professional experience with speech disorders;
- (3) The nDU group: 124 listeners, reporting no prior professional experience with speech disorders or working knowledge of Dutch. Among the diverse linguistic backgrounds, the biggest subgroups of nDU listeners were native speakers of Germanic ( $n = 14$ ) or Slavic languages ( $n = 101$ ).

Participant demographics appear in [Table 2](#).

Before the experiment, all participants signed an informed consent accompanied by a short questionnaire on their demographic, language and expertise background.



**Table 2.** Participants demographics. Age is given in years.

Group name	Age: mean/ SD (range)	Gender: %F	Native languages
Trained group, n = 18	34/13.2 (21– 64)	94%	Dutch or/and Frisian/German
Dutch untrained group, n = 27	25.2/8.2 (18– 60)	63%	Dutch and: Arabic, English, Frisian, Polish, Spanish, Turkish
Non-Dutch untrained group, n = 124	36.1/13.7 (18– 79)	73%	Adyghe, English, Estonian, Finnish, German, Hebrew, Hungarian, Indonesian, Italian, Kabardian, Lithuanian, Portuguese, Russian, Turkish, Ukrainian

## Procedure

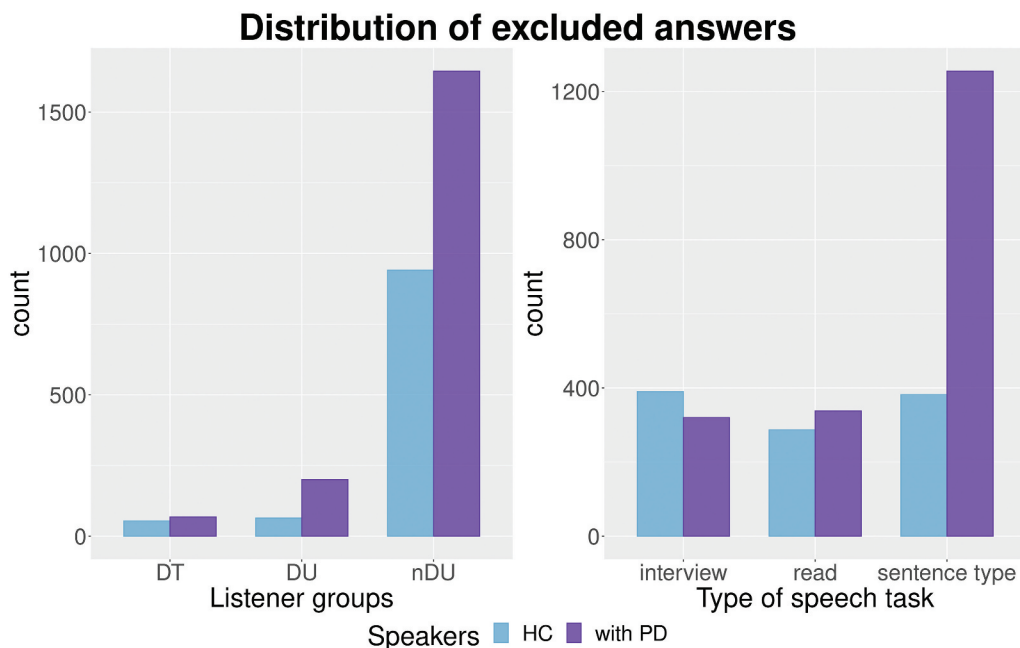
Participants completed two classification tasks in the experiment implemented in JavaScript using jsPsych library (De Leeuw, 2015) and running on the JATOS platform, version 3.5.3 (Lange et al., 2015), which enabled the online testing procedure. The tasks were organised into two blocks. For each block, the procedure consisted of a main part, preceded by a short practice session. 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.

In the first block, participants listened to the stimuli created from the interviews and reading material. 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 understanding of it. Listeners had three answer options to choose from: “Yes”, “No” and “I don’t know”, and for every response they specified whether they felt “rather sure” or “rather unsure” about their answer.

In the second block, participants listened to the stimuli created from the recordings of the sentence type exercise. Participants were asked to answer the question “was the phrase a question or a statement?” by choosing from options “question”, “statement”, or “I don’t know”. Just as in the first block, they specified how confident they were of their answer by selecting “rather sure” or “rather unsure”.

Results of the experiment were stored in JSON format with participants’ responses assigned numerical values. The format was converted to CSV with a Python script. All subsequent analyses were conducted only on definitive responses, with all “I don’t know” (IDK) responses removed from the dataset (6.6% of all the responses, for the DT group it was 2.6% of the responses, for the DU group 2.3%, and for the nDU group 8.5%). The distribution of the excluded answers is presented in Figure 1.

Fleiss’ Kappa interrater reliability for multiple categorical variables and multiple raters was calculated for different listener groups. The resulting values were between 0.40 and 0.75 representing fair to good agreement beyond chance (Fleiss et al., 2003) for all answer types. Exceptions were poor agreement for classifying statement intonation in case of the nDU group. Excellent agreement was found in the DT group when DT listeners classified question intonations (see Table 3).



**Figure 1.** IDK answers distribution by listener group and by type of speech task.

**Table 3.** The Fleiss' Kappa interrater agreement for different listener.

Experiment part/answers		DT group	DU group	nDU group		
				Germanic	Slavic	All
Part 1	Healthy	0.511	0.522	0.504	0.489	0.480
	Unhealthy	0.533	0.502	0.526	0.511	0.504
Part 2	Healthy	0.755	0.618	0.425	0.502	0.473
	Unhealthy	0.732	0.557	0.325	0.414	0.385

### Statistical analyses

To analyze how the experience with speech and language disorders and familiarity with speakers' language affect accuracy of listeners' classification of PD speech, we applied bootstrap resampling ( $R = 1000$ ). To estimate the classification accuracy between PD and HC speech groups and between sentence types we compared distribution of answers expressed by the statistics (1):

$$|\text{mean}(\text{HC\_scores}) - \text{mean}(\text{PD\_scores})| / \text{mean}(\text{SD}(\text{HC\_scores}), \text{SD}(\text{PD\_scores})) \quad (1)$$

We measured the participants' ability to distinguish speakers with PD from HC speakers as the bootstrapped classification accuracy was based on their answers in the healthiness classification tasks. To take confidence of the listeners into account, each test was run on a second set of scores – with confidence introduced as a weight. The interaction between answers and confidence was calculated using the expression (2):

$$0.5 + (\text{answer} - 0.5) * (1 - \text{conf}/2) \quad (2)$$

A similar procedure was applied to measure the participants' ability to identify question/statement intonations which was based on participants' answers in the sentence type classification task.

The analyses of both healthiness and sentence type classification accuracies were independently compared between the three groups using Welch's F test, a non-parametric alternative of one-way ANOVA. The choice of Welch's F test for statistical analysis was motivated by the significantly different variance of data in the bootstrapped groups. The test was run twice for each task: first, with the scores only, and afterwards with confidence answers used as scores' weight.

To gain more insight into the influence of familiarity with the language on the accuracy of healthiness and sentence type classification, subsequent analyses compared four groups: DT, DU, nDU listeners whose native languages were from Germanic language family, and nDU listeners whose native languages were from Slavic language family. These two language families were better represented in the nDU group, with 14 speakers of German or English and 101 speakers of Russian and/or Ukrainian. The rest of the nDU group, nine listeners, were excluded from this subsequent analysis for the reasons of heterogeneity of the language backgrounds.

## Results

### *Healthiness classification*

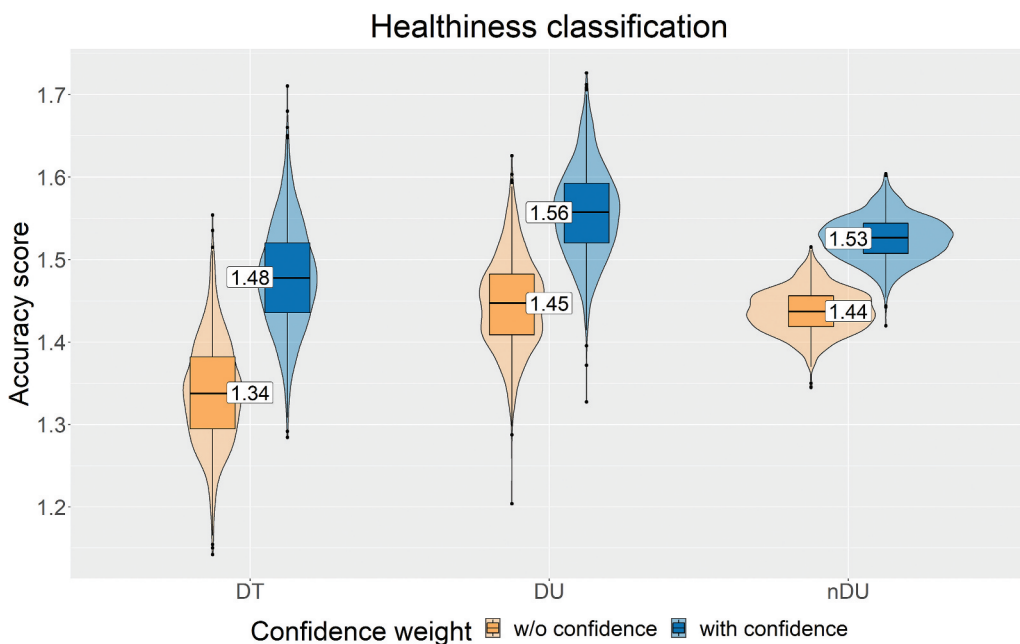
The DU group performed with the most accuracy. The DT group performed with the least. With confidence weights introduced, the accuracy scores became higher for every group. The comparison of the accuracy scores for three groups is depicted in [Figure 2](#).

Welch's F test demonstrated the significance of observed differences with and without confidence weight; that is, there was a significant effect of the listener group on the healthiness classification accuracy (see [Table 4](#)). The Tukey post hoc test revealed that all the differences between the three groups were significant for both weighted and unweighted accuracy scores,  $p < .001$ . The largest differences for both unweighted and weighted accuracy scores was found between the DT and nDU groups (diff = 0.11 and 0.08,  $p < .001$ ). Therefore, results of the healthiness classification task indicate a negative influence of experience and positive influence of familiarity with speakers' language.

### *Sentence type classification*

For both PD and HC speech, the differences for the three groups were similar. The DT group performed most accurately, followed by the DU and nDU groups (see [Figure 3](#)). The introduction of the confidence weight highlighted the differences for both PD and HC speech, showing the rise in the scores boosted by the confidence weight for all the groups ([Figure 3](#)).

Welch's F test demonstrated a significant effect of the listener group on the sentence type classification accuracy scores for both PD and HC speech with and without confidence weight (see [Table 4](#)). The Tukey post hoc test revealed that all the differences between the



**Figure 2.** Accuracy of speech healthiness classification.

**Table 4.** Welch's F test results per task, with and without confidence weight.

Task	Confidence weight	F(2, 2997)	<i>p</i>	$\omega^2$	95% CI	
Healthiness	Without	1230.36	< .001	0.49	[0.462, 0.507]	
	with	516.19	< .001	0.28	[0.26, 0.311]	
Sentence type	<b>PD speech</b>	Without	27 518.98	< .001	0.95	[0.944, 0.949]
		with	33 817.59	< .001	0.96	[0.954, 0.959]
	<b>HC speech</b>	Without	24 052.04	< .001	0.94	[0.938, 0.944]
		with	31 725.97	< .001	0.95	[0.952, 0.957]

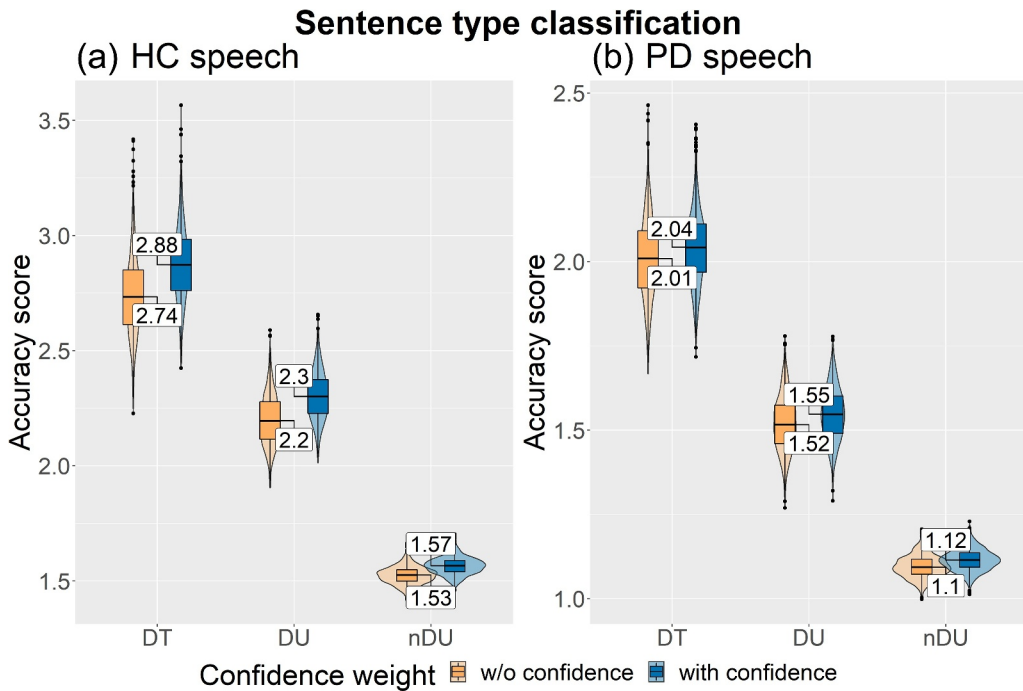
three groups were significant,  $p < .001$ . The biggest differences were found between the DT and nDU groups, for both PD and HC speech with and without weights (PD diff = 0.92 and 0.93, HC diff = 1.21 and 1.31,  $p < .001$ ).

### ***Influence of Germanic and Slavic languages***

Healthiness classification results for the nDU subgroups demonstrated the highest accuracy scores of the nDU listeners with Germanic language backgrounds (see Figure 4).

Welch's F test confirmed that these differences were significant both for weighted and, with a smaller effect, for unweighted accuracy scores (see Table 5). Subsequent Tukey post hoc test showed significance ( $p < .001$ ) for differences between all the groups except for the differences between DU and Slavic nDU ( $p = .116$ ). The highest scores for Germanic nDU subgroup (Figure 4) demonstrate their more accurate classification of PD and HC speech based on perceived healthiness.

The analysis of the sentence type classification showed similar results to the trend that was earlier outlined for the three groups (see Figure 5).



**Figure 3.** Accuracy of sentence type classification in HC speech (a) and in PD speech (b) for three listener groups.

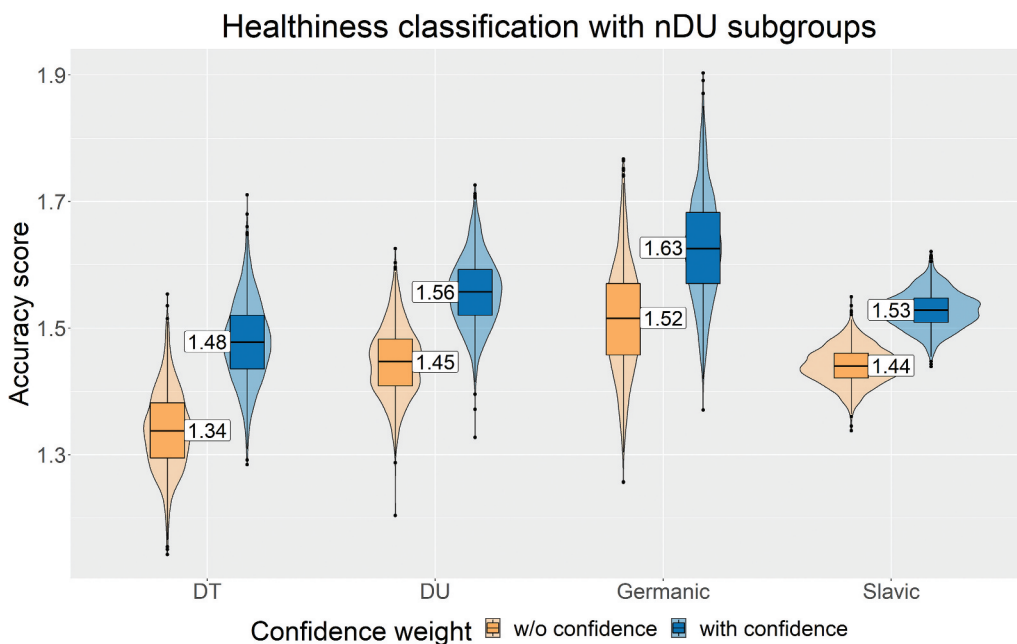
Similar to the results on speech classification based on healthiness perception, the results of Welch's F test for sentence type classification showed that the found differences were statistically significant in both PD and HC speech with and without confidence weight (see Table 5).

A Tukey post hoc test demonstrated that in PD and HC speech, with or without confidence weights, all the differences between the four groups were significant,  $p < .001$ . The differences in mean scores for the Germanic nDU group for PD and HC speech (Figure 5) suggest that the similarity between the Dutch and German phonetic and prosodic inventories is not the only factor which impacts identification of a question/statement intonations.

## Discussion

This study explored the effects of speech and language therapy expertise and familiarity with speakers' language on the classification of PD speech. We investigated the accuracy with which different groups of listeners classified the healthiness of speech and the sentence type intonation. We found that both expertise in speech and language disorders and familiarity with the language have a significant effect on the perception of PD speech.

We found that speakers with PD are perceived as more unhealthy compared to HC speakers by both trained and untrained listeners. This finding is in line with the fact that speech affected by HD exhibits acoustic pathological symptoms due to prosodic and



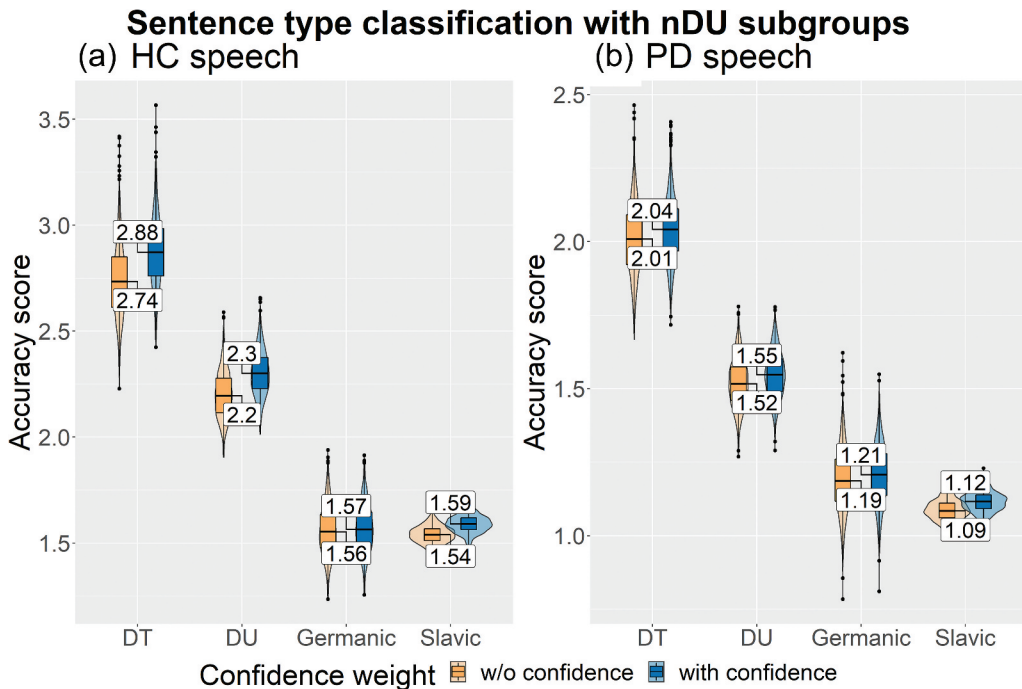
**Figure 4.** Accuracy of speech healthiness classification for four listener groups.

**Table 5.** Welch's F test results for the analysis of language background effect. Results are presented per task and with and without confidence weight.

Task	Confidence weight	F(3, 3996)	<i>p</i>	$\omega^2$	95% CI	
Healthiness	without	1485.27	< .001	0.51	[0.494, 0.534]	
	with	991.44	< .001	0.43	[0.413, 0.454]	
Sentence type	<b>PD speech</b>	without	19 853.92	< .001	0.94	[0.934, 0.938] [0.944, 0.948]
		with	24 320.42	< .001	0.94	
	<b>HC speech</b>	without	21 354.93	< .001	0.94	[0.947, 0.945] [0.954, 0.958]
		with	27 644.19	< .001	0.96	

articulatory deficits (first described by Darley et al. (1969b)). Moreover, listener groups accurately classified unhealthiness in speech elicited by both reading and interview tasks, suggesting that both interview and reading provide enough cues for listeners to identify unhealthiness in PD speech. The accuracy scores differed per speech task for each listener group: DT group showed higher accuracy for stimuli from interviews, while DU and nDU group were more accurate when they assessed stimuli from the recordings of reading. However, the interview elicitation task appears to be potentially preferable, as it is a closer approximation of natural spontaneous speech than reading and does not have any restrictions for speakers to have normal (corrected-to) vision as in the reading exercises.

Regarding the first research question, whether the familiarity with speech disorders affects classification of PD speech into the “unhealthy” category, SLT training proved to be a significant factor. Surprisingly, the DT group did not perform most accurately as expected because of their training, as Carvalho et al. (2020) demonstrated that SLTs without experience in PD outperform the general public and listeners with PD in the intelligibility assessment of words and sentences of speakers with PD. Both DU and nDU groups



**Figure 5.** Accuracy of sentence type classification in HC speech (a) and in PD speech (b) for four listener groups.

performed significantly more accurately than the DT group. These findings can be compared to the results of previous studies by Smith et al. (2019) and Walshe et al. (2008), where trained listeners did not outperform the untrained listeners. Higher sensitivity of the untrained listeners is very likely to arise from a different interpretation of the concept of “healthiness”. In other words, group differences might stem from different task interpretations, as the DT listeners might have approached “healthiness” from a different perspective with varying strategies based on their SLT expertise. A similar conclusion was also reached by Walshe et al. (2008), who reported a large interrater variability in their trained group.

Following our second research question, we investigated if and how familiarity with a particular language relates to the classification of PD speech in that language. Statistical analyses revealed that the differences between DT, DU and nDU groups were significant, and both expertise and familiarity with speakers’ language affect the ability to classify Dutch PD speech into ‘unhealthy’ category. The subsequent analysis of the language backgrounds of untrained listeners resulted in the unanticipated findings.

Surprisingly, listeners with a Germanic language background with no working knowledge of Dutch were more accurate at detecting unhealthiness in Dutch speech than Dutch trained and untrained listeners. This suggests that being unfamiliar with the speakers’ (typologically similar) language, while sharing a number of phonetic and prosodic features, is beneficial for detecting unhealthiness related to PD in speech. One possible explanation of such difference might be the absence of distractions while processing Dutch speech. That is, the influence of the DT group’s experience could have served as a distractor during the classification task, thus making trained listeners more sensitive to expertise-specific cues. The semantics of the utterances could have been another influencing factor in the

classification task for Dutch listeners, whereas the unfamiliarity and salience of some segmental aspects of Dutch phonetic inventory, together with varying prosodic cues, may have served as a distractor for untrained listeners with typologically different language backgrounds (in this case, Slavic). This highlights the need for further exploration, not only of the typological characteristics of the target language spoken by people with PD (Pinto et al., 2017), but also of the relationship between the native languages of speakers and listeners if they are coming from different language backgrounds. Therefore, further research into the acoustic relationships between the native language of listeners and the language spoken by people with HD may shed more light on listeners' perception patterns of non-native dysarthric speech (Alispahic et al., 2017; Kim et al., 2012; Näsström & Schalling, 2020).

Our third research question addressed the effects of familiarity with speech disorders and with speakers' language on the prosody classification in PD speech. Our findings demonstrate that for the sentence type classification task, both language and expertise are important. The assessment by the DT group was more accurate than those of the DU and nDU groups. This shows that expertise plays a key role in listeners' ability to correctly identify question/statement intonations in PD speech. Therefore, the benefit of SLT training for sentence type classification in PD speech goes against findings reported by Smith et al. (2019) and Walshe et al. (2008), where trained listeners performed either less accurately or similar to untrained listeners. However, this finding is in line with the recent study by Carvalho et al. (2020), which demonstrated that healthcare professionals working with HD are more likely to understand PD speech than untrained listeners. The different findings for the healthiness and for the sentence type classifications in the current study also confirm the importance of the task type for assessment of dysarthric speech (Martens et al., 2011).

The expected significance of familiarity with speakers' language is also apparent, as the DU group was more sensitive to the sentence type intonation contrasts than the nDU group. It is interesting to see that in the sentence type classification task in the HC speech, listeners of Germanic and Slavic language backgrounds performed very similarly. At the same time, in the sentence type classification task in PD speech, Germanic listeners outperformed Slavic listeners. It is possible that the lack of distraction allows listeners to recognize the coping strategies that speakers with PD might use to overcome possible prosodic deficits (Pinto et al., 2017). On the other hand, even though many prosodic features are auditory-perceptually salient to listeners cross-linguistically, the use of prosodic cues is language specific and is modulated by the prosodic phonology of a given language (Kim et al., 2012). This could be reflected in language-specific compensatory strategies employed by speakers with PD causing Slavic listeners in this study to be less sensitive to compensatory prosodic strategies employed by Dutch speakers with PD.

Present findings demonstrate the interconnection between listeners' perception of healthiness of speech and the acoustic changes in PD speech. The impressions of speech healthiness and the judgements about question and statement intonations without any influence of semantic content indicate that speech acoustics can be predictive of perceived healthiness. Our findings also confirm that trained listeners understand speakers with PD better than untrained listeners at the level of prosody. The results of the study highlight the benefits of exploring more global perceptual assessments of speech and the importance of training the untrained population, providing them with strategies and tools to understand speakers with PD more easily, not only on the level of intelligible words and phrases as suggested by Carvalho et al. (2020), but also on the level of understanding the linguistic prosody. Our



findings underscore the necessity for further research into the language-specific and language-universal aspects of HD. It also brings into focus the importance of exploring multilingual proficiency of not only dysarthric speakers (Pinto et al., 2017), but also listeners and assessors of their speech. Another outcome of this study is the methodological implications accentuating the need for a broad and rigorous approach towards elicitation methods, task design and, more importantly, the concept of expertise and training.

Future research could expand on these results through experiments with SLT listeners who explicitly work with dysarthria secondary to PD, and by cross-linguistic experiments targeting perception of HD. This will allow one to map which cues and which listener groups could be of help in detecting HD and possibly in contributing to early PD diagnosis. This knowledge will also provide specific therapeutic targets to enhance communication efficiency of speakers with HD and help to work on alleviating the negative attitudes which speakers with HD may face (Maryn & Debo, 2015; Miller et al., 2006).

The results of this study have two important outcomes for clinical practice. First, the finding that especially inexperienced listeners are good at recognizing unhealthiness in speech of speakers with PD 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. Therefore, those who are already close to a speaker at risk will be able to recognize one of the early signs of PD. This might help with an early diagnosis of PD as well as early detection of speech problems, which in turn would allow speakers with PD to start speech therapy at an earlier stage of the disease progression. The second contribution to clinical practice is related to the finding that non-native listeners were better at recognizing unhealthiness in speech affected by PD, as they focused on the acoustics of speech rather than on the content of speakers' message. 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.

## Acknowledgments

We are very grateful to our colleagues from University Medical Center Groningen: Prof. Dr. Natasha Maurits, Dr. Bauke de Jong and Sanne Timmermans for their advice, collaboration and invaluable help with the data collection. We also thank Lea Busweiler, research assistant, for the help in the data collection. We are grateful to all the speakers and listeners who volunteered to participate in our study.

## Data availability

The data that support the findings of this study are openly available in OSF at <https://osf.io/mf4d5>, DOI 10.17605/OSF.IO/MF4D5.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The author(s) reported there is no funding associated with the work featured in this article.

## ORCID

V. Verkhodanova  <http://orcid.org/0000-0002-5481-8508>

## References

- Alispahic, S., Mulak, K. E., & Escudero, P. (2017). Acoustic properties predict perception of unfamiliar Dutch vowels by adult Australian English and Peruvian Spanish listeners. *Frontiers in Psychology*, 8, 52. <https://doi.org/10.3389/fpsyg.2017.00052>
- 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](https://doi.org/10.1044/2015_JSLHR-S-14-0243)
- Basirat, A., Patin, C., & Moreau, C. (2018). Relationship between perception and production of intonation of French in Parkinson's disease. *Speech Prosody*, Poznań, Poland. 13-16 June 2018. 809–813. <https://doi.org/10.21437/SpeechProsody.2018-163>
- Best, C. T., & Tyler, M. (2007). Language experience in second language speech learning. In Bohn, O. S., and Munro, M.J. (Eds.). *honour of James Emil Flege*. John Benjamins Publishing. 13–34. <https://doi.org/10.1075/llt.17.07bes>
- Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer [Computer software]*. Version 6.1.14. Retrieved May 2, 2020, from <http://www.praat.org/>
- 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>
- Bunton, K., & Keintz, C. K. (2008). The use of a dual-task paradigm for assessing speech intelligibility in clients with Parkinson disease. *Journal of Medical Speech-Language Pathology*, 16(3), 141–155. <https://pubmed.ncbi.nlm.nih.gov/21637738/>
- 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(6), 1481–1495. [https://doi.org/10.1044/1092-4388\(2007\)102](https://doi.org/10.1044/1092-4388(2007)102)
- Carvalho, J., Cardoso, R., Guimarães, I., & Ferreira, J. J. (2020). Speech intelligibility of Parkinson's disease patients evaluated by different groups of healthcare professionals and naïve listeners. *Logopedics Phoniatrics Vocology*, 46(3), 1–7. <https://doi.org/10.1080/14015439.2020.1785546>
- Darley, F. L., Aronson, A. E., & Brown, J. E. (1969a). Differential diagnostic of patterns of dysarthria. *Journal of Speech and Hearing Research*, 12(2), 246–269. <https://doi.org/10.1044/jshr.1202.246>
- Darley, F. L., Aronson, A. E., & Brown, J. E. (1969b). 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>
- De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Duffy, J. R. (2012). *Motor speech disorders: Substrates, differential diagnosis and management* (3rd ed.). Mosby.
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons, Inc.
- Kempler, D., & Van Lancker, D. (2002). The effect of 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>

- Kim, S., Broersma, M., & Cho, T. (2012). The use of prosodic cues in learning new words in an unfamiliar language. *Studies in Second Language Acquisition*, 34(3), 415–444. <https://doi.org/10.1017/S0272263112000137>
- Klopfenstein, M. (2015). Relationship between acoustic measures and speech naturalness ratings in Parkinson's disease: A within-speaker approach. *Clinical Linguistics & Phonetics*, 29(12), 938–954. <https://doi.org/10.3109/02699206.2015.1081293>
- Kreiman, J., Gerratt, B., Kempster, G., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, 36(1), 21–40. <https://doi.org/10.1044/jshr.3601.21>
- Lange, K., Kühn, S., & Filewich, E. (2015). “Just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, 10(6), 1–14. <https://doi.org/10.1371/journal.pone.0130834>
- Martens, H., Van Nuffelen, G., Cras, P., Pickut, B., De Letter, M., & De Bodt, M. (2011). Assessment of prosodic communicative efficiency in Parkinson's disease as judged by professional listeners. *Parkinson's Disease: Communication Impairments in Parkinson's Disease*, 2011. 129310. <https://doi.org/10.4061/2011/129310>
- Martens, H., Van Nuffelen, G., Wouters, K., & De Bodt, M. (2016). Reception of communicative functions of prosody in hypokinetic dysarthria due to Parkinson's disease. *Journal of Parkinson's Disease*, 6(1), 219–229. <https://doi.org/10.1016/j.neuroimage.2017.02.073>
- Maryn, Y., & Debo, K. (2015). Is perceived dysphonia related to perceived healthiness? *Logopedics Phoniatrics Vocology*, 40(3), 122–128. <https://doi.org/10.3109/14015439.2014.915981>
- Miller, N., Noble, E., Jones, D., & Burn, D. (2006). Life with communication changes in Parkinson's disease. *Age and Ageing*, 35(3), 235–239. <https://doi.org/10.1093/ageing/afj053>
- Näsström, A. K., & Schalling, E. (2020). Development of a method for assessment of dysarthria in a foreign language: A pilot study. *Logopedics Phoniatrics Vocology*, 45(1), 39–48. <https://doi.org/10.1080/14015439.2019.1650392>
- Pell, M. D., Cheang, H. S., & Leonard, C. L. (2006). The impact of Parkinson's disease on vocal-prosodic communication from the perspective of listeners. *Brain and Language*, 97(2), 123–134. <https://doi.org/10.1016/j.bandl.2005.08.010>
- Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., & Sadat, J. (2017). A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. *Journal of Phonetics*, 64, 156–167. <https://doi.org/10.1016/j.wocn.2017.01.009>
- Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., & Fox, C. (2007). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 50(4), 899–912. [https://doi.org/10.1044/1092-4388\(2007\)064](https://doi.org/10.1044/1092-4388(2007)064)
- Smith, C. H., Patel, S., Woolley, R. L., Brady, M. C., Rick, C. E., Halfpenny, R., Rontiris, A., Knox-Smith, L., Dowling, F., Clarke, C. E., Au, P., Ives, N., Wheatley, K., & Sackley, C. M. (2019). Rating the intelligibility of dysarthric speech amongst people with Parkinson's disease: A comparison of trained and untrained listeners. *Clinical Linguistics & Phonetics*, 33(10–11), 1063–1070. <https://doi.org/10.1080/02699206.2019.1604806>
- Sussman, J. E., & Tjaden, K. (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](https://doi.org/10.1044/1092-4388(2011)11-0048)
- Verkhodanova, V., Coler, M., Jonkers, R., De Bot, K., & Lowie, W. (2019a). Prosodic changes in the speech of a single patient with Parkinson's disease. In S. Calhoun, P. Escudero, M. Tabain, and P. Warren, (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia. 5-9 August 2019. (pp. 3046–3050).
- Verkhodanova, V., Timmermans, S., Coler, M., Jonkers, R., De Jong, B., & Lowie, W. (2019b). How dysarthric prosody impacts native listener's recognition. In: A. Salah, A. Karpov, and R. Potapova, (Eds.), *Speech and computer. SPECOM 2019. Lecture notes in computer science*. Springer, Cham (vol. 11658, pp. 510–519). [https://doi.org/10.1007/978-3-030-26061-3\\_52](https://doi.org/10.1007/978-3-030-26061-3_52)

- Verkhodanova, V., Trčková, D., Coler, M., & Lowie, W. (2020). More than words: Cross-Linguistic exploration of Parkinson's disease identification from speech. In A. Karpov, and R. Potapova, (Eds.), *Speech and computer. SPECOM 2020. Lecture notes in computer science*. Springer, Cham (Vol. 12335, pp. 613–623). [https://doi.org/10.1007/978-3-030-60276-5\\_59](https://doi.org/10.1007/978-3-030-60276-5_59)
- Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language and Communication Disorders*, 43(6), 633–648. <https://doi.org/10.1080/13682820801887117>
- Weismer, G., Jeng, J., Laures, J., Kent, R., & Kent, J. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatica Et Logopaedica*, 53(1), 1–18. <https://doi.org/10.1159/000052649>