More than words: Recognizing speech of people with Parkinson's disease
Verkhodanova, Vass

DOI:
10.33612/diss.183425053

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:
2021

Citation for published version (APA):
CHAPTER 7
RECOGNITION OF LONGITUDINAL CHANGES IN SPEECH

ABSTRACT

This chapter provides a longitudinal perspective on how PD impacts both speech production and listeners’ recognition of PD speech. It reports on a study that analyzes time series of monthly recordings of the same individual with Parkinson’s disease over one year, investigating linear trends in both acoustic features belonging to prosody and phonation domains and listeners’ ratings of healthiness of speech. The findings demonstrate the benefit of listeners’ training for speech healthiness recognition with regards to the disease progression.

7.1. INTRODUCTION

There are several longitudinal studies focusing on speech of people with PD (PwPD) (Skodda and Schlegel, 2008; Skodda et al., 2009). Most of these studies employ pre- and posttest designs and analyze recordings collected at two time points with intervals between them ranging from seven months (Skodda and Schlegel, 2008) to 3.7 years (Huber and Darling-White, 2017). The results of these studies demonstrate that the disease progression is reflected in speech. Pitch variability reduces (Skodda and Schlegel, 2008), steady syllable repetition becomes unstable (Skodda et al., 2011a), speech rate increases (Huber and Darling-White, 2017), and quality of voice and articulatory velocity and precision decrease (Skodda et al., 2011a). Another study provides a longitudinal analysis of speech in a single PD speaker over an 11-year period (seven years prior to diagnosis of PD, and three years post-diagnosis), based on archives of national television

This chapter is adapted from:
(Harel et al., 2004). Their findings suggest that changes in $f_0$ variability can be detected as early as five years prior to diagnosis (Harel et al., 2004).

In this study, we explore whether the effects of PD progression are reflected in the speech of a single speaker with PD without the diagnosis of dysarthria. We also investigate if such effects influence listeners’ recognition of speech healthiness. To those ends, we made monthly measures of the same speaker with PD over a year.

In this study, we focus on several speech characteristics related to prosody and phonation that are most indicative of PD speech and which easily allow for automatic measurements (see Table 7.1).

Prosodic changes in speech due to PD are described to demonstrate similar patterns in different languages (Ma et al., 2010b; Rusz et al., 2011; Skodda et al., 2009), with changes in $f_0$ being the most prominent and most studied. At the same time, literature on recognition of prosody in speech affected by PD usually involves people already diagnosed with dysarthria. Changes in phonation and voice quality, namely harsh or rough voice, were identified as being among the most severely affected speech dimensions by several studies (Ma et al., 2010b; Whitehill et al., 2003), but received less attention than changes in prosody.

To determine if longitudinal changes in speech of a single PD speaker without diagnosis of dysarthria can be detected, we approached the question from two perspectives. The first was related to objective acoustic measurements. The second perspective was related to the listeners’ subjective assessments of speech. For the former, we hypothesized if any acoustic changes are to manifest themselves within one year, monopitch would be one of them. For the latter, we hypothesized that listeners would be able to recognize a difference in the speaker’s voice, provided there were sufficient acoustic differences including changes in prosody and voice quality. Additionally, based on the findings of the existing studies (Eadie and Baylor, 2006; Harris et al., 2016), we expected that listeners with training in speech and language pathology to be more sensitive to such changes relative to listeners who lack such training.

### 7.2. Methods

To test our hypotheses we collected data (subsection 7.2.1), designed a speech recognition experiment (subsections 7.2.2 – 7.2.4), and performed an analysis of the collected data: both of the acoustic speech signal (subsection 7.2.5), and listeners’ ratings (subsection 7.3.2). Although the recognition experiment was conducted with both Dutch and English stimuli, we analyzed only the Dutch part of the data relative to the hypotheses of the current study.

The collection and analysis of the material was approved by the Ethics Committee of the faculty Campus Fryslân, University of Groningen.

#### 7.2.1. Data Collection

We recorded speech from one male native Dutch speaker who is fluent in English, and who uses both languages daily. At the start of the recording sessions, the participant was 66 years old. He was diagnosed with PD six years prior to the beginning of the recording
sessions. At the time of writing this chapter, he has not been diagnosed with hypokinetic
dysarthria. He has a history of stuttering.

The recording protocol included five speech tasks: sustained phonation of the vowel
\(/a:/\), an interview with an open question, picture and video descriptions, and reading.
For the picture description task we used one of the Heaton pictures per session (Heaton,
1972), for the video description, we used a short video clip from one of the Charlie Chaplin
silent films, and for reading, we used the “North Wind and the Sun” passage. All tasks
were performed first in English and subsequently in Dutch. The recordings were collected
every month to the extent possible (mean interval was 5.2 weeks, SD = 2.2), one to three
hours after medication intake. The recording sessions took place in quiet rooms at the
university with a Zoom H2 recorder placed at around a 40 cm distance.

7.2.2. PARTICIPANTS IN RECOGNITION EXPERIMENT

A group of 61 native Dutch speakers participated in the experiment. Among them were
people with different experiences with speech disorders. Based on their experience and
training, we divided them into two groups: untrained listeners with no prior experience
with speech disorders (hereafter untrained group) and speech therapists and/or students
of a neurolinguistics master programme with experience in listening to disordered speech
(hereafter trained group). The untrained group consisted of 51 people (mean age 27.5, SD
7.7 years). The trained group consisted of ten people (mean age 24.4, SD 1.4 years). All
participants reported normal hearing.

7.2.3. STIMULI

To create a stimuli set we used short fragments of 2-3 seconds taken from the Dutch and
English spontaneous monologues and reading tasks of five sessions out of 12 (namely,
days 0, 107, 204, 286, and 411). From each of the five sessions we selected six samples: four
fragments from spontaneous monologues (two per language) and two fragments from
reading tasks (one per language). The total amount of stimuli (30 phrases) were selected
according to three criteria: they should not include artefacts of stuttering, they should
consist of at least four words, and they should be taken from declarative statements. To
the extent possible, fragments from monologues were extracted from the first and second
halves of the recording.

7.2.4. PROCEDURE

Participants completed a speech recognition experiment in which they listened to the
stimuli in a randomized order. Participants were told that they would hear short phrases
and were asked to rate them on a 7-point Likert scale according to their impressions of
speech healthiness (from “very healthy” to “very unhealthy”). The experiment was built
within the OpenSesame programme, an open-source graphical experiment builder for the
social sciences (Mathôt et al., 2012). The procedure consisted of a short practice session
and the main part. For the practice session, participants were presented with two stimuli

---

1See the link to the source code in Appendix H and the screenshots in the Appendix G
Table 7.1 | Overview of features and their measurement methods.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Parameter</th>
<th>Description</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosody</td>
<td>$f_0$ coefficient of variation</td>
<td>Variance of fundamental frequency ($f_0$), representing the variations of vibration rate of vocal folds</td>
<td>RAPT (Talkin, 1995)</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
<td>The number of syllables per total time</td>
<td>Praat script (De Jong and Wempe, 2009)</td>
</tr>
<tr>
<td></td>
<td>Articulation rate</td>
<td>The number of syllables produced per speaking time</td>
<td>Praat script (De Jong and Wempe, 2009)</td>
</tr>
<tr>
<td>Voice quality</td>
<td>Jitter</td>
<td>Frequency perturbation, representing the extent of variation of the voice range</td>
<td>Praat (Boersma and Weenink, 2017)</td>
</tr>
<tr>
<td>Voice quality</td>
<td>Shimmer</td>
<td>Amplitude perturbation, representing rough speech</td>
<td>Praat (Boersma and Weenink, 2017)</td>
</tr>
<tr>
<td></td>
<td>RPDE</td>
<td>Recurrence period density entropy, representing the inefficiency of voice frequency control</td>
<td>Python script based on algorithm from Little et al. (2007)</td>
</tr>
<tr>
<td>HNR</td>
<td>Harmonics-to noise ratio, representing voice hoarseness. HNR is defined as the amount of noise in the speech</td>
<td>Praat (Boersma and Weenink, 2017)</td>
<td></td>
</tr>
</tbody>
</table>

of female voices: a healthy young voice and an unhealthy one which was elderly and affected by severe dysarthria to create the benchmarks for healthy and unhealthy voices. The main part included our target set of 30 stimuli. All stimuli were presented to the participants using headphones (Koss Pro4S). Participants could listen to each stimulus as many times as they wanted.

7.2.5. Acoustic analyses

To determine whether acoustic changes are reflected in the recorded speech over a one year period, we performed an acoustic analysis of seven speech characteristics in Dutch monologues and reading. The descriptions of all analyzed characteristics and details of their measurements are summarized in Table 7.1 and are described further below.
f₀ COEFFICIENT OF VARIATION
Pitch tracking was performed with the David Talkin’s RAPT algorithm (Talkin, 1995), implemented in the SPTK toolkit for Python (Imai et al., 2017). The RAPT algorithm identifies pitch candidates with the cross-correlation function and then attempts to select the "best fit" at each frame by dynamic programming (Morrison et al., 2007; Talkin, 1995). From the pitch trajectory we calculated the f₀ coefficient of variation (hereafter, variance) to estimate the range of the speaker’s f₀.

SPEECH AND ARTICULATION RATES
Measuring speech and articulation rates commonly requires annotation of phonemes or syllables. However, this procedure is time-consuming and sometimes error-prone. Therefore, these measurements were done automatically by detecting syllable nuclei with a Praat script written by De Jong and Wempe (2009), where syllable nuclei correspond to peaks in intensity preceded and followed by dips in intensity, with unvoiced peaks being discarded. Studies on French (Looze et al., 2012) and Dutch (Verkhodanova and Coler, 2018) dysarthric speech have demonstrated the usefulness of the script. For the current study we have used a -20 dB silence threshold, 4 dB dip and 70 ms as a minimal pause duration. Speech rate was computed as the number of syllables divided by total time of the recording, and articulation rate as the number of syllables divided by phonation time of the recording.

VOICE QUALITY
To describe the voice quality, we analyzed recordings of the sustained phonation, measuring jitter, shimmer, recurrence period density entropy (RPDE based on the algorithm from Little et al. (2007)) and harmonics-to-noise ratio (HNR). All were measured automatically.

7.3. RESULTS

7.3.1. RESULTS OF ACOUSTIC ANALYSES
The analysis of coefficient of variation for f₀ showed a decline from the first session towards the final session (see Figure 7.1). A linear regression analysis demonstrated that the decline was significant (F = 205.5, p < .000), with an R² of 0.14 and a slope of $-4.41 \times 10^{-5}$. Speech and articulation rate showed no trends, nor did the measurements for RPDE and HNR. Shimmer did not show any significant decline. The decline for jitter was significant: F = 6.2, p < .03, with an R² of 0.18 and a slope $= -3.58 \times 10^{-5}$.

7.3.2. RATING PATTERNS
To assess the rating patterns of the participants, we fitted a simple linear regression model in R, which resulted in a significant regression equation predicting scores depending on time (F = 52.42, p < .000) with an R² of 0.054 and a slope coefficient of 0.0025 (see Figure 7.2). To see if there was a difference between trained and untrained listener groups, we fitted separate linear models. For the untrained group, the regression equation showed significance (F = 36.5, p < .000), with an R² of 0.046; the slope coefficient was 0.0023. For
the trained group, the regression equation was significant as well (F = 18.4, \( p < .000 \)), with an \( R^2 \) of 0.11; the slope coefficient was 0.0039.

Fitting separate models for subsets of stimuli from monologue and reading tasks showed that both groups had steeper slopes for monologues than for reading (0.0026 vs 0.0015 for the untrained group and 0.0041 vs 0.0034 for the trained group). The model for the untrained listener group rating stimuli from the reading task did not reach significance (\( p > .005 \)).

To verify that the results of the linear regression were not random, we applied a Monte Carlo analysis. In the performed simulation, we modelled the probability of different slope outcomes. We randomized the scores 1000 times and calculated the slope for every randomized set of scores. The resulted distribution of slopes had a mean value of \( 1.3 \times 10^{-5} \) and SD of 0.0003 with a standard error of \( 1.14 \times 10^{-5} \). We also used a resampling technique based on the jackknife resampling to evaluate the possibility of bias. We calculated slopes for 1/3 of the data set 1000 times and found that variance for slopes was extremely small: \( 2.78 \times 10^{-7} \).

### 7.4. Discussion

In this study, we explored the question of longitudinal changes in speech of a single speaker with PD and without the diagnosis of dysarthria. Acoustic analysis of his speech showed no significant changes for speech or articulation rates, shimmer, RPDE or HNR with the disease progression. Significant changes were present in \( f_0 \) and jitter. However, a prominent dip in \( f_0 \) variance remains unexplained. We interviewed our speaker at every
session, and he did not report any events that could have affected his speech within the month before the dip. We found neither changes in the recording procedure, nor noise conditions that could have affected the data. Therefore, it is likely that some physical change occurred which triggered the decline. Nevertheless, it is too early to say if it could be an onset of dysarthria, and further research into other speech characteristics such as vowel and consonant articulation might shed some light on this hypothesis.

The results of the recognition experiments validated the acoustic analysis, showing the trend of later recordings being rated as less healthy than earlier recordings. Differences in rating between trained and untrained listeners provide an additional perspective on the importance of listeners’ training for assessment of longitudinal changes in PwPD. Exploring training effects on speech recognition longitudinally requires further analysis and should be addressed in future studies.

The trends resulting from the linear regression analysis were significant, and the $R^2$ values were expected to be lower due to the nature of the data. While the scores are spread quite broadly, the analysis was concerned with the slight changes in the rating patterns. Regarding the prominent dip in $f_0$ variance attested by the acoustic analysis, we found no apparent effect in the dip on the listeners’ rating patterns. Therefore, the $f_0$ difference is not prominent enough on its own to guide listeners’ recognition of speech healthiness.

To summarize, we have found that longitudinal changes in speech of a single speaker with PD are reflected both by the results of acoustic analysis and the results of the recognition experiment, validating our hypotheses on the influence of monopitch. Our findings demonstrate that both trained and untrained listeners are sensitive to the presence of longitudinal speech changes. However, as this is a single case study, it is impossible to
generalize our findings to the larger population of speakers with PD. On the other hand, the results are very promising and show the potential benefit of exploring recognition experiments and prosody tracking for monitoring disease progression in speakers with PD.