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

Verkhodanova, Vass

DOI:
10.33612/diss.183425053

CHAPTER 4

HOW DYSARTHRIC PROSODY IMPACTS UNTRAINED LISTENERS' RECOGNITION

ABSTRACT

The class of speech disorders known as dysarthria arise from disturbances in muscular control over the speech mechanism caused by damage to the central or peripheral nervous system. Dysarthria is typically classified into one of five classes, each corresponding to a different neurological disorder with distinct prosodic cues (Darley et al., 1969b). The assumption in this classification is that dysarthric speech can be classified based on perception. In this study, we investigated how accurately untrained listeners can recognize stress and intonation in dysarthric speech, and if different neurological disorders impact the ability to convey meaning with these same two cues. To those ends, we collected speech data from Dutch speakers diagnosed with cerebellar lesions (ataxic dysarthria), Parkinson's disease (hypokinetic dysarthria), Multiple Sclerosis (mixed classes of dysarthria) and from a healthy control group. Thirteen untrained Dutch listeners participated in the perceptual experiment which targeted recognition of intended realization of four prosodic functions: lexical stress, sentence type, boundary marking and focus. We analyzed recognition accuracy for different groups and performed acoustic analyses to check for fundamental frequency trajectories. Results attest to different accuracy recognition results for different speaker groups. The sentence type recognition task was the most sensitive of all tasks for differentiating diseases both by acoustic analysis and by analysis of recognition results.

This chapter is adapted from:

Verkhodanova V., Timmermans S., Coler M., Jonkers R., de Jong B., Lowie W. (2019) How Dysarthric Prosody Impacts Naïve Listeners' Recognition. In *Salah A., Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2019. Lecture Notes in Computer Science*, vol 11658. Springer publishing, (pp. 510-519). The paper was slightly adapted for this dissertation.

4.1. INTRODUCTION

Dysarthria is a condition which is caused both by weakness of muscles used in speech and by difficulties in controlling these muscles. The most common and simple description of this speech disorder is "slow speech that can be difficult to understand" (Mayo Clinic, 2020, p.1). Common causes of dysarthria arise from cerebral dysfunction at the level of brainstem nuclei, supra nuclear brain dysfunction or neuromuscular impairment. Neurological conditions that may lead to dysarthria include Parkinson's disease (PD), Amyotrophic lateral sclerosis (ALS), Multiple sclerosis (MS), head injury, Spinocerebellar ataxia (SCA) and a number of others. Since dysarthria causes communication difficulties, it may lead to social deprivation and depression (Mayo Clinic, 2020).

The seminal contribution to understanding dysarthria was made by Darley et al. (1969a,b), who introduced a classification system of dysarthrias. Since then this system (hereafter, the *Mayo System*) has been widely used for research and clinical purposes. The Mayo System links brain pathology based on the lesion site to prominent speech characteristics, united in clusters of deviant speech dimensions. However, despite the wide use of the system, there are doubts about its suitability for clinical purposes. For example, two independent studies tested the classification accuracy in the groups of neurologists and neurology trainees (Fonville et al., 2008), and in the groups of neurologists, residents in neurology, and speech therapists (Van der Graaff et al., 2009). Both studies have reported accuracy of correct classification between 35% and 40%, concluding that "perceptual judgements" alone are not reliable, and clinicians should always take other sources of information into consideration.

Since then, researchers have been trying to classify dysarthria classes using acoustic cues to support the Mayo System. For example, in the study by Guerra and Lovey (2003), authors matched the acoustic measurements of dysarthric speech to the dimensions used by clinicians and compared performance of two different classifiers with the clinicians' assessments of the speech from the speech corpus with different dysarthria classes linked to eight neurological disorders. The combined feature set of clinicians' assessments and objective acoustic measurements provided more accurate information about the speech disturbances, while the best classifier proved to be self-organising maps (SOM), which improved the accuracy of clinicians' judgements by nearly 20% (Guerra and Lovey, 2003). These findings indicate the value of acoustic analysis as an additional tool for clinical purposes.

The majority of the research focusing on finding a method for the reliable differentiation of dysarthria classes rely on acoustic metrics. In one such study, Liss et al. (2009) investigated the rhythm metrics, addressing dimension of prosody in the corpus of five different dysarthrias with different prosodic profiles. The results show that rhythm metrics can be used to distinguish neurologically healthy speech from moderate and severe dysarthric speech as well as to discriminate dysarthria classes with accuracy up to 80%. A follow up study by Liss et al. (2010) explored whether speech envelope modulation spectra, which quantifies the rhythmicity of speech within specified frequency bands, could be used for automatic analysis. Discriminant function analysis showed 84%–100% accuracy for different dysarthria classes compared to all others, with HD scoring at 100% (Liss et al., 2009).

There are two studies on acoustic metrics that investigate the dimension of articulation focusing on the vowel metrics (Lansford and Liss, 2014a,b). In one study, Lansford and Liss (2014b) explored whether such metrics could be used to distinguish neurologically healthy from dysarthric speech as well as to differentiate among four different classes of dysarthria (ataxic, hypokinetic dysarthria, hyperkinetic and mixed flaccid-spastic dysarthria). All vowel metrics explored, particularly metrics that capture vowel distinctiveness, demonstrated the significant differences between dysarthric and control speakers. However, only the slope of the second formant (F2) feature yielded between-group differences across the groups of speakers with different dysarthrias. In the second study, Lansford and Liss (2014a) investigated whether vowel metrics reflect the listeners' assessment of the intelligibility of dysarthric speech. The results showed a correlation between classification by disordered vowels metrics and intelligibility assessments.

A study by Kim et al. (2011b) explored both dimensions of articulation and prosody simultaneously, using eight acoustic features as predictors for classification of different classes of dysarthria occurring from PD, stroke, multiple system atrophy or traumatic brain injury. Interestingly, the reported results have shown that classification accuracy into dysarthria types was typically worse than by disease type or severity, while the best classification was achieved when disease type was the grouping variable. Regarding intelligibility, F2 slope showed significance for each disease group, serving as a potential universal predictor. Articulation rate, however, was not a significant predictor of speech intelligibility for speakers with PD, although it showed significance in the pooled analysis (Kim et al., 2011b).

In the present study, we explore if different dysarthria classes affect the ability of speakers to convey intended prosody. To this end, we collected recordings of the speakers with diseases and divided them into three groups based on the diseases: PD group, SpinoCerebellar Ataxia group (SCA) and Multiple Sclerosis (MS) group. These diseases are frequent causes of different dysarthrias, namely hypokinetic dysarthria, ataxic dysarthria and either spatic, flaccid or mixed dysarthria. Many studies have indicated that such dysarthria classes have different prosodic deficit profiles (Darley et al., 1969a; Liss et al., 2010; Miller, 2018), which, among other cues, is reflected by different patterns of disturbances of fundamental frequency (f_0).

To determine if untrained listeners could recognize intended intonation and stress patterns produced by speakers of different disease groups, we approached the question from two perspectives. The first perspective is related to listeners' recognition of prosody. The second perspective is related to the acoustics analysis. For listeners' recognition of prosody, we hypothesized that if there is a correlation between disease groups and accuracy of recognition, it would be most prominent in the PD group. For the acoustic analysis, we hypothesized that f_0 would hinder listeners' accuracy of recognition at least for the PD group. To test these hypotheses we collected data (4.2), designed a perception experiment (4.2), and performed an acoustic and recognition accuracy analyses (4.3). Section 4.4 summarizes and discusses the findings of both the acoustic analysis and the recognition experiment.

Table 4.1 | Participants demographics. Age and duration of disease are given in years

Group name	Mean age	Gender (F:M)	Diagnoses	Disease duration
PD	53.9	4:4	Idiopathic PD	mean: 11.5, range: 20
SCA	55.3	5:3	Spinocerebellar ataxia, adult form of Alexander disease, idiopathic late onset cerebellar ataxia, multiple system atrophy with cerebellar ataxia	mean: 6, range: 10
MS	51.9	4:4	Primary progressive MS, secondary progressive MS, relapsing-remitting MS	mean: 13.5, range: 21
HC	56.2	4:4	-	-

4.2. METHODS

DATA COLLECTION

Speech recordings were collected from 32 Dutch native speakers, 24 patients (eight per disease group) and eight control speakers. The demographics can be seen in Table 4.1.

Except for the neurologically healthy control speakers (HC), participants exhibited dysarthric symptoms due to a neurological disorder according to the clinical judgement of a neurologist. All speakers reported (corrected-to) normal vision and hearing and signed informed consent. Exclusion criteria for speakers with dysarthria were cognitive problems assessed by the Minimal Mental State Examination (MMSE < 26), brain damage caused by stroke that inflicted aphasia and/or apraxia of speech, and language and/or (motor) speech disorders other than dysarthria. Exclusion criteria for control speakers were cognitive problems (MMSE < 26), brain damage, language and/or (motor) speech disorders. The recording sessions took place in quiet rooms at the University Medical Centre Groningen or at participants' homes with a TASCAM-DR100 recorder and an external Senheiser e865 microphone placed at around a 40 cm distance from a participant.

The data collection was approved by the Medical Ethics Committee of the University Medical Center Groningen.

LISTENERS

Thirteen native Dutch listeners were recruited via convenience sampling and had diverse work experience (from catering to graduate studies) participated in the prosody recognition experiment (mean age 29). All 13 were untrained and reported no prior experience with speech disorders. All participants reported normal hearing.

Table 4.2 | Prosodic functions and their "perceptual correlates" based on Martens et al. (2011) and Rietveld and Van Heuven (2009). Prominent correlates according to Rietveld and Van Heuven (2009) are marked bold.

Function name	**Description**	**Perceptual correlates (for undisturbed speech)**	**Name used in the current study**
Lexical function	Discriminates between words	***f0* change**, (vowel) duration, intensity	Lexical stress
Phrasing function	Segments the speech stream in information units	**preboundary lengthening pauses**, f_0 change	Boundary Marking
Attentional marking	Highlights the most important elements in a unit	***f0* change**, (vowel) duration, intensity	Focus
Intentional marking	Nuances meaning	f_0 change	-
Sentence type	Discriminates between questions and statements	**general *f0* rise (question)**, high initial f_0 (question)	Sentence Typing
Emotional prosody	Discriminates between different emotional states	**general *f0* , *f0* span, speech rate**	-

STIMULI

Stimuli for this study were created from a prosody task, that included exercises on four Dutch prosody functions: lexical stress, sentence type, boundary marking, and focus intonation (Martens et al., 2011). Table 4.3 summarizes Dutch prosody functions and their "perceptual correlates" based on Martens et al. (2011) and Rietveld and Van Heuven (2009).

Four exercises targeting elicitation of four prosodic functions included sentence completion (to elicit lexical stress and boundary intonation), repetition (for boundary intonation) and the production of negative/affirmative and questions and statements (for sentence type and focus intonation). As result, from these exercises we have created pairs of stimuli for every linguistic prosody function:

- Words segmented from the completed sentences that differed in stress placement: first or second syllable (e.g., *dóorlopen* - 'to continue'and *doorlópen* - 'to complete');
- Phrases syntactically identical but different in question or statement intonation (e.g.*de toets gehaald?* - '<he> passed the test?'and *de toets gehaald.* - '<he> passed the test');
- Phrases syntactically identical but different in complete/statement or incomplete/iteration intonation (e.g., *Andre houdt van honden,* <...> - 'Andre likes dogs, <...>'and *Andre houdt van honden.* - 'Andre likes dogs ');

- Phrases syntactically identical but different in prosodically emphasized words - focus intonation (e.g., *ik ken haar van **dansles**.* - 'I know her from the **dancing class**.'(as opposed to another class) and *ik ken haar van dansles.* - 'I know her from a dancing class.'.

The total number of stimuli was 1233, 320 for the stress and for sentence type, and 310 and 283 for boundary marking and focus respectively. There were fewer stimuli for the two latter functions because some participants decided to quit during the last part of the protocol and some participants made mistakes in the exercise parts.

PROCEDURE

Participants of the recognition experiment completed a recognition task in which they listened to the stimuli in four blocks corresponding to four prosody functions. Participants were told that they would hear words and phrases that were different either in stress or intonation and were asked to answer a simple question by picking one option from a list (e.g., "Was the phrase a question or a statement?" – "1) question, 2) statement, 3) impossible to decide"). For every prosody task there were always three options with one being "impossible to decide". The experiment was built within the OpenSesame programme (Mathôt et al., 2012)[1].

For each block the procedure consisted of a short practice session and a main part. The practice session was designed to acquaint participants with the task. They were asked to assess two stimuli of two different voices. For the main part there were 192 stimuli randomly pooled from the set representing current prosody function in such a way that there were six stimuli per speaker in each block. The speech samples were intensity normalized and presented over Koss Pro4S headphones. Participants could listen to each sample only once.

ANALYSES

To analyse listeners' accuracy of dysarthric prosody recognition we calculated percentage of correct, incorrect and unspecified ("impossible to decide") answers along with the confidence interval (CI) estimation for the particular answers using Normal Approximation Method of the Binomial Confidence Interval.

To analyse pitch trajectories of different disease groups and healthy speakers, we assessed f_0 slopes within each stimulus. To do so, we divided each stimulus recording in two parts (the ratio between parts was 1:1 for stimuli of the lexical stress function, for other stimuli it was 7:3). For each part we calculated f_0 average derivative and calculated the difference between the parts of the recording. Pitch tracking was performed with the Talkin's RAPT algorithm (Talkin, 1995) implemented in the SPTK toolkit in 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). RAPT results have been shown to be informative for Dutch dysarthric speech in an earlier study (Verkhodanova and Coler, 2018).

[1] See the link to the source code in Appendix H and the screenshots in the Appendix G

Table 4.3 | Recognition accuracy for different disease groups and healthy speakers presented as percentage of correct/incorrect/unspecified answers with CI.

Group	Correct	Incorrect	Unspecified
HC	67 ± 1.8	27 ± 1.8	4 ± 0.8
MS	60 ± 2.0	28 ± 1.8	11 ± 1.3
SCA	56 ± 2.0	28 ± 1.8	14 ± 1.4
PD	55 ± 2.0	25 ± 1.7	18 ± 1.5

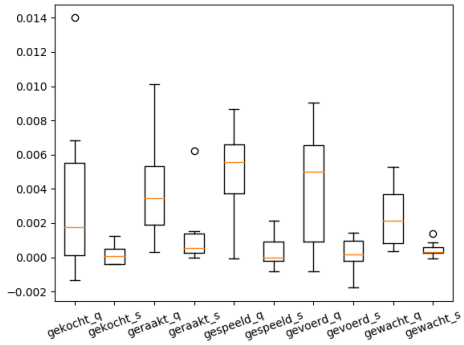
4.3. RESULTS

General accuracy calculation for different disease groups did not show any striking differences, though predictably the stimuli produced by the HC group were recognized most accurately of all, and the stimuli produced by the PD group least accurately of all, with the highest percentage of unspecified answers. The percentage of unspecified answers was also very small for the stimuli from the HC group as compared to the other groups (see Table 4.3)

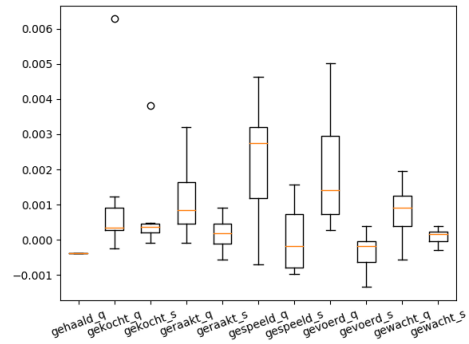
When assessing the differences for listeners' performance depending on the target prosodic function, stimuli from different speaker groups yielded different accuracy results. Overall, boundary and focus tasks were the most difficult prosodic functions for listeners to recognize, especially focus intonation, where the percentage of the unspecified answers was the highest (up to 23 ± 3.4), but even those functions showed a difference between dysarthric and non-pathological speech. Lexical stress appeared to be a relatively successful task for listeners when they recognized stimuli from HC and MS speakers. However, the same task with stimuli produced by the SCA and PD groups proved to be less successful, as listeners demonstrated lower accuracy recognition results. Sentence type was the best recognized function for all speaker groups, with the smallest numbers of unspecified answers. Sentence type was also the only function where stimuli produced by the PD group did not yield the lowest scores of recognition accuracy.

Further analysis of accuracy focused on specific prosody patterns. It targeted the differences between first or second syllables for lexical stress, between question or statement intonations for sentence type, between finished or unfinished intonation for boundary marking, and between presence or absence of focus intonation for focus. Except for focus, the difference between accuracy for two specific prosody patterns was very clear within each group. Questions were recognized better than statements, words with stressed first syllable were recognized better than words with stressed second syllable, and finished intonation was recognized better than unfinished intonation.

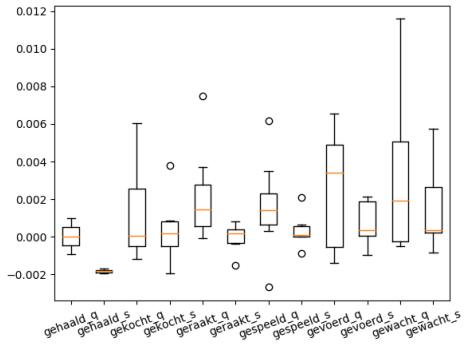
To determine if f_0 trajectories would reflect the listeners' recognition results, we conducted a Kruskal-Wallis rank sum tests for non-parametric data to determine f_0 trajectory differences across the data. We compared differences between the f_0 derivatives for stimuli pairs. For all but one pair, significant results were found in the sentence type task for two speaker groups: HC and PD. Other prosodic functions did not exhibit any



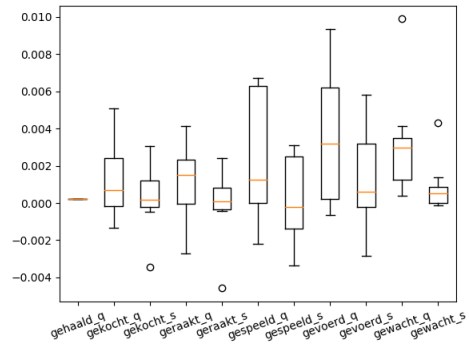
(a) HC



(b) PD



(c) MS



(d) SCA

Figure 4.1 | f_0 derivative differences in sentence typing. Difference between derivatives are placed on the y-axis, stimuli tags are placed on the x-axis: 'q'after each word means question, 's'- statement.

stable significant results within any speaker group. The box plots of f_0 trajectories for sentence type function in different speaker groups are presented on Figure 4.1.

Additionally, we checked for correlations between accuracy of listeners' recognition and speakers' disease duration, and found that there were none.

4.4. DISCUSSION AND CONCLUSIONS

In this study we explored the ability of untrained listeners to recognize intonation and stress patterns produced by speakers with different neurological disorders in comparison with control speakers. We found that different neurological disorders, which lead to different types of dysarthria, affect the recognition of prosodic patterns differently. The HC group was always distinguishable from any dysarthria group based on the listeners' recognition results. As hypothesized, listeners performed the poorest on stimuli produced by the PD group in three out of four prosody function tasks. Sentence type was the function where listeners were more successful in recognizing stimuli from the PD group

than from the SCA group. This can be explained by the specifics of dysarthria for the SCA disorders: the SCA speakers' tendency towards equalized vowel/syllable durations within utterances and unusually large f_0 range across utterances (Kent et al., 2000) interfered with their ability to mark sentence types.

Moreover, not all the tasks were found to be reliable to assess prosody deviances. The focus recognition task appeared to be very difficult for listeners in general, causing high numbers of unspecified ("impossible to decide") answers. The sentence type recognition proved to be the clearest task, and was the only one that showed correlation with f_0 trajectories estimation. However, the results show that f_0 trajectories alone cannot act as a reliable predictor for different dysarthria classes or for the accuracy of listeners' recognition. Nevertheless, it is obviously a meaningful cue for differentiating healthy and dysarthric speech.

Despite the small number of speakers and listeners, and the lack of information about the severity of dysarthria, we showed that assessing how untrained listeners recognize dysarthric speech can shape further research dedicated to exploring the link between acoustic and perceptual cues and, in doing so, shed light on classifying different dysarthria classes. Further research should target other acoustic cues such as temporal cues and formant measurements that might affect listeners' prosody recognition of dysarthric speech.

II

When people listen: speech recognition and acoustic analysis

Between our two lives
there is also the life of
the cherry blossom.

Matsuo Bashō