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

Verkhodanova, Vass

DOI:
10.33612/diss.183425053

CHAPTER 6

ACOUSTIC CUES IN THE CROSS-LINGUISTIC RECOGNITION OF HEALTHINESS OF SPEECH IN PARKINSON'S DISEASE

ABSTRACT

This chapter builds upon the experiments and results of cross-linguistic recognition of healthiness of speech produced by PwPD. Using the same data as in chapter 5, this chapter investigates how listeners' recognition of speech healthiness is related to the acoustic changes appearing in the speech of PwPD. The study in this chapter investigates how the speech healthiness recognition of different listener groups can be predicted from conventional acoustic features. A set of 18 features included in the Random Forest analyses were from the domains of voice quality, prosody and articulation. The listeners who participated in the recognition experiment differed in their training and expertise in speech and language therapy as well as in their native languages providing a cross-linguistic perspective to the discussion of different effects on speech healthiness recognition in PD speech.

6.1. INTRODUCTION

Speech disturbances due to hypokinetic dysarthria negatively impact spoken communication of people with Parkinson's disease (PwPD), and have been commonly described

This chapter is adapted from:

Verkhodanova, V., Coler, M., Jonkers, R., Timmermans, S., Maurits, M., de Jong, B., and Lowie, W. (submitted). A cross-linguistic perspective to classification of healthiness of speech in Parkinson's disease. *Journal of Neurolinguistics*.

in terms of deviant speech dimensions from seminal studies by Darley et al. (1969a,b). These deviant speech dimensions have been extensively investigated on the acoustic level. Many studies have focused on the acoustic aspect of speech production in PD with attention to monopitch, or in other terms – lack of fundamental frequency (f_0) variations (Skodda et al., 2013; Galaz et al., 2016), distorted rhythm of speech (Skodda and Schlegel, 2008), monoloudness – reduced intensity variability of voice (Skodda et al., 2013; Galaz et al., 2016), reduced stress (Tykalova et al., 2014), imprecise consonants (Fischer and Goberman, 2010; Tykalova et al., 2017) and a hoarse and breathy voice quality (Tsanas et al., 2010a). A number of studies have also demonstrated that prosody deficits together with harsh voice and reduced articulation are among the most prominent speech characteristics present in the acoustics of speech produced by PwPD (Rusz et al., 2011; Galaz et al., 2016; Brabenec et al., 2017; Verkhodanova et al., 2019). Some studies also suggest that the prosodic deficits arising from HD are universal (Pinto et al., 2017).

There is a growing number of studies exploring the efficiency of speech production by PwPD and focusing on the perception and recognition of speech produced by PwPD. Many researchers have described prominent changes in prosodic characteristics of speech of PwPD. For example, when compared to control speakers, PwPD are less efficient at producing question-statement intonation contrasts (Pell et al., 2006; Basirat et al., 2018) or at conveying both lexical and contrastive stress (Pell et al., 2006; Martens et al., 2016). Overall, in the literature, monopitch and monoloudness are described as having the greatest influence on the perception and recognition of speech affected by HD and are seen as the most prototypical source of prosodic speech problems for PwPD (see chapter 3 for details). As discussed in the previous chapters, these speech disturbances affect the quality of life of PwPD, resulting in communication problems and often in a feeling of social isolation. This frequently leads to tension, depression, resignation and withdrawal from a conversation (Miller et al., 2006). For example, Schalling et al. (2017) collected and presented self-reported information of affected communication from 188 Swedish PwPD. Their findings demonstrated that 92.5% of the respondents reported at least one symptom related to communication, and that the speech and communication problems resulted in restricted communicative participation for roughly one third of all respondents. Even though almost every participant reported speech and communication problems, only 45% reported receiving speech and language therapy, highlighting inadequate access to speech-language pathology services for PwPD (Schalling et al., 2017). Moreover, research has demonstrated that speech changes caused by HD affect the daily lives of PwPD long before impairment of intelligibility is apparent (Miller et al., 2006, 2007).

6.1.1. ASSESSMENT OF DYSARTHRIA

Chapter 2 highlighted that there are various ways of assessing dysarthric speech in literature. Thus, many studies rely on the auditory-perceptual evaluation of dysarthria, which continues to be referred to as the "gold standard" for clinical decisions (Bunton et al., 2007; Sussman and Tjaden, 2012; Duffy, 2013; Näsström and Schalling, 2020). There are various means of assessment performed by listeners, ranging from judging vowels (Sapir et al., 2007a) and words (Bunton and Keintz, 2008; Smith et al., 2019) in isolation to assessing read passages (Tjaden and Wilding, 2004) and spontaneous conversational speech (Bun-

ton et al., 2007; Bunton and Keintz, 2008). One of the common measures of assessment and management of speakers with dysarthria is speech intelligibility scores (Sussman and Tjaden, 2012). These scores are commonly used as a measure of the severity of a speech disorder and as a source of information for treatment planning and for monitoring changes in speech (Yorkston and Beukelman, 1978; Bunton and Keintz, 2008). Another approach to auditory-perceptual assessment of dysarthria is using component-specific perceptual judgements, based on evaluating lists of deviant dimensions, as described by Darley et al. (1969b). Both approaches have limitations and reliability concerns as summarized in a study by Sussman and Tjaden (2012).

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

However, intelligibility scores, component-specific perceptual judgements, and scaled estimates of speech disorder severity have been developed for specific languages, as this approach is dependent on language-specific differences. Therefore, such ways of dysarthric speech assessment are dependent on language-specific differences. For example, findings of a comparative study (Kim and Choi, 2017) show that even though PwPD who speak American English and Korean, demonstrate similar acoustic patterns of articulation deviances compared to control groups, the degree to which the same acoustic features contribute to the intelligibility scores is language-dependent. That said, in light of the the upward trend for international migration coupled with increases in life expectancy, there is a need for techniques to assess HD in languages unfamiliar to the assessor (Nässtrom and Schalling, 2020). Nevertheless, few studies explore cross-linguistic assessment of dysarthria. One such study is that of Hartelius et al. (2003), who reported results of cross-language assessment of Swedish and Australian speakers with dysarthria secondary to multiple sclerosis. Australian and Swedish speech and language therapists (SLTs) demonstrated high inter-rater reliability, resulting in similar prevailing sets of dimensions for both languages: imprecise consonants, harshness and glottal fry, reduced speech rate, pitch level, and loudness despite the foreign language. However, some of these dimensions, namely precision of consonant production, pitch and loudness level, and general rate and harshness, were associated with higher disagreement values, while there were language-specific difficulties in the assessment of general stress pattern and phoneme length (Hartelius et al., 2003). Another study, by Nässtrom and Schalling (2020), focuses on developing a systematic assessment method for SLTs who do not speak the native language of an individual with dysarthria. The authors observed that despite being unfamiliar with Arabic, a Swedish SLT was sensitive to respiration, phonation and some articulation changes in dysarthric speech (Nässtrom and Schalling, 2020). Nässtrom and Schalling (2020) found that an SLT who performs an 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. These findings demonstrate that, irrespective of a listener's familiarity with the language, the way dysarthria affects phonation and some aspects of articulation might be universal and accessible for the assessment of a trained listener (Hartelius et al., 2003; Näsström and Schalling, 2020).

As discussed in previous chapters, in addition to familiarity with the language, an increasing body of evidence suggests that listeners' experience and training can also matter (Kreiman et al., 1990; Eadie and Baylor, 2006; Walshe et al., 2008; Smith et al., 2019; Carvalho et al., 2020). Many studies dedicated to perception and recognition of speech affected by HD take one group of listeners as a source of assessment. Researchers often focus on either untrained, trained non-expert or trained expert listeners (also discussed in chapter 3). Relatively few studies have compared perception or recognition of speech affected by HD, or any other dysarthria, in groups of listeners with different levels of familiarity with it. Nevertheless, there is conflicting evidence regarding the role of experience (trained experts versus the untrained general population) in the assessment of dysarthric speech.

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

In line with the findings of Walshe et al. (2008), Smith et al. (2019) reported no significant differences between the groups in their investigation into how trained and untrained listeners rated intelligibility of speech affected with HD. Some studies, however, demonstrate that groups of listeners with different expertise tend to rate speech of PwPD differently (Verkhodanova et al., 2019, 2020). In a longitudinal case study by Verkhodanova et al. (2019), both trained and untrained listeners assess global "healthiness" of a single speaker with PD similarly: both groups rated the recordings made at a later stage as less healthy than the earlier ones. However, trained listeners' ratings showed a steeper trend towards the "less healthy" scores for recordings made at a later stage (for details, see chapter 7). In another study, Verkhodanova et al. (2020) explored the recognition of PD speech by Dutch and Czech trained and untrained listeners. They investigated the effect of experience and familiarity with the language on recognition of healthiness of speech and of intended sentence type intonation in speech of PwPD. The findings demonstrate that both expertise and familiarity with the speakers' language act as important factors in listeners' recognition of PD speech (for details, see chapter 5). Additionally, recognition accuracy depends on a task type: untrained listeners outperformed trained listeners in the speech healthiness recognition task, while trained listeners were more accurate in recognizing sentence type intonation (Verkhodanova et al., 2020).

Carvalho et al. (2020) found that regarding intelligibility ratings, expertise with dysarthria and experience specifically with PD are important. The authors showed that neurologists working with PD gave slightly higher intelligibility scores than SLTs working with adult dysarthria, and higher than other listeners groups in their study (listeners familiar with PD, the general untrained population, and PwPD themselves). The authors concluded that healthcare professionals who work with dysarthria are more likely to understand the speech of PwPD than the groups of untrained listeners (Carvalho et al., 2020).

6.2. FOCUS OF THE CURRENT STUDY

This study explores the global assessment of speech "healthiness" similar to the studies by Verkhodanova et al. (2019, 2020) and following the ideas of Kreiman et al. (1993), Weismer et al. (2001), Sussman and Tjaden (2012), and Maryn and Debo (2014). The main focus of the study is to investigate how listeners' recognition of speech healthiness correlates with acoustic cues that reflect speech production issues of PwPD. We convert the acoustic cues to a set of features that can be measured directly from the speech signal. We also convert listeners' impressions about healthiness of speech into yes/no responses in a speech recognition task. Therefore, we address three research questions. Our first research question is whether listeners' responses about speech healthiness can be predicted from a set of acoustic features. Secondly, we want to find out which acoustic features are more important when predicting listeners' responses about speech healthiness. Our third research question is if the relevance of acoustic features that are predictors of the responses about speech healthiness depends on listeners' SLT experience and language background.

To address these questions, we performed an online experiment with three groups of listeners (see chapter 5, subsection 5.3 for details). Regarding the first research question, we hypothesized that listeners' responses can be predicted from conventional acoustic features that are clinically interpretable (Brabenec et al., 2017). Secondly, since monopitch is described as one of the most *deviant* dimensions in speech of PwPD contributing to many aspects of successful communication (Bunton et al., 2001; Ma et al., 2010b; Kuo et al., 2014; Anand and Stepp, 2015; Verkhodanova et al., 2020), we expected that f_0 variance would be among the main predictors of listeners' responses to the recognition task. Thirdly, we hypothesized that, in line with observations by Näsström and Schalling (2020) and Hartelius et al. (2003), listeners with SLT experience would rely primarily on voice quality and phonation features when recognizing dysarthric speech. We also expected that listeners with no training but with a similar language background would rely on the mix of phonation, articulation and prosody related features (De Bodt et al., 2002; Hartelius et al., 2003), while listeners unfamiliar with the speakers' language would rely more on universal components of prosodic deviances such as monopitch (Whitehill et al., 2003; Pinto et al., 2017).

6.3. METHODS

Data collection, exclusion and inclusion criteria, stimuli creation, experimental procedure and participants of the online experiment were previously described in chapter 5, subsection 5.3. The current study explores the responses of different groups of listeners – Dutch trained (DT), Dutch untrained (DU), non-Dutch untrained (nDU) – from a different perspective by predicting these responses from a set of features.

6.3.1. DATA ANALYSIS

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

DEMOGRAPHIC INFORMATION FEATURES

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

ACOUSTIC FEATURES

The second feature set consisted of acoustic measurements to test if the model can reliably predict listeners' answers based on the acoustic characteristics of speech signal. Acoustic analysis was performed on the prolonged phonation of /a:/, the stimuli set, and on the stimuli extracted with the surrounding context of 1/3 of the length of the stimulus. Selection of the acoustic features was based on previous research that listed the affected characteristics in speech of people with HD (Darley et al., 1969b; Brabenec

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

In this study we included features related to:

- prosody: fundamental frequency variability, articulation and speech rates, "inappropriate silences" (Darley et al., 1969a),
- articulation: vowel space area, vowel articulation index (see chapter 2 for details),
- phonation and voice quality: means and standard deviations of first two formants, jitter, harmonics-to-noise ratio, maximum phonation time, fundamental frequency variance of prolonged phonation.

Because the recordings in this study were done without strict supervision over the distance between speaker and the microphone, intensity measurements coud not be included in the analysis. The details of 18 acoustic features are described below.

Fundamental frequency (f_0) variability calculation was based on the f_0 tracking in the stimuli and stimuli with the context taken from the recordings of the interview and of the read passage. The calculation was done by means of a Python script and the Speech Signal Toolkit (SPTK) (Imai et al., 2017) based on the robust algorithm for pitch tracking (RAPT) (Talkin, 1995). Two other prosodic features, speech rate and articulation rate, were calculated by means of a Praat script (De Jong and Wempe, 2009). Speech rate was measured as the number of syllables divided by the total time of the recording. Articulation rate was measured as the number of syllables divided by phonation time in that recording. The rate measurements were performed on the same stimuli and on stimuli with the fixed context. Inappropriate silences feature was calculated from the results of the same Praat script by De Jong and Wempe (2009) as the number of pauses relative to total speech time after removing periods of silence lasting less than 60 ms. (Brabenec et al., 2017). The measurements were performed on the same stimuli set and on the stimuli with the context set. Phonation / voice quality measurements of means and standard deviations of the first formant (F1) and second formant (F2), maximum phonation time (MPT), jitter and harmonics-to-noise ratio (HNR), all were calculated from the prolonged phonation (/a:/) recordings and by means of a Praat script. Two articulation features included in the analysis are conventional measures capturing vowel centralization. First, the vowel space area (VSA) is constructed by the Euclidean distances between the F1 and F2 coordinates of the vowels /i/, /u/, and /a/ (triangular VSA, Liu et al. (2005)). Second, the vowel articulation index (VAI) is based on the description in the study by Roy et al. (2009), where it is described as maximally sensitive to vowel centralization and decentralization. Equations for VSA (2.1) and VAI (2.2) can be found in

chapter 2. The vowels were extracted from the fourth sentence of the "North Wind and the Sun" text, which contained all three vowels. The full list of acoustic features with the feature labels and descriptions is presented in Appendix E.2.

6.4. RANDOM DECISION FOREST METHOD

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

For the purposes of the analysis, we used two RF models. The first model predicts listeners' responses based on seven demographic features (predictors) to examine whether listeners are more sensitive to speech in the presence/absence of the diagnosis and to the duration of the disease, or rather to other changes related to speakers' gender, age, or self-reported dialectal pronunciation. The seven predictors used for the first model were: 1) presence or absence of the diagnosis, 2) speaker age and 3) gender, 4) self-reported dialect pronunciation, 5) listener ID, 6) speaker ID, and 7) duration of the disease represented as a vector with four values corresponding to control speakers, and three disease duration periods calculated as 3 quantiles from the ordered vector of the disease duration values. We have added listener ID and speaker ID to see how important their contribution is to the model to understand whether the responses are more dependent on individual rating patterns of each listener or some individual characteristics of the speakers rather than on predictors that are constant across speakers. Therefore, the formula for the first RF model was:

```
response ~ diagnosis + disease duration + speaker age +
           speaker gender + dialect + listener ID + speaker ID
```

Two additional predictors for the DT group were experience with neurodegenerative disorders (+ *neurodegenerative exp*) and specifically experience with HD caused by PD (+ *PD exp*).

The second model predicts listeners' answers based on the acoustic information from the speech signal, examining whether conventional and objective acoustic measurements are representative of subjective (perceptual) responses of listeners. The second model used 18 conventional acoustic features that were described earlier. Therefore, the formula for the second model was:

```
response ~ 8 prosodic features +
           8 voice quality/phonation features +
           2 vowel articulation features
```

Table 6.1 | Results of the model with demographic predictors with accuracy and confidence intervals for every listener group.

Group	Model results on the training set		Model results on the test set	
	Accuracy	**95% CI**	**Accuracy**	**95% CI**
DT	87.3%	[0.8544, 0.8894]	84.3%	[0.8118, 0.8713]
DU	86.3%	[0.8479, 0.8771]	84.4%	[0.8191, 0.8668]
nDU	84.4%	[0.8364, 0.8514]	85.2%	[0.8404, 0.863]

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

6.5. RESULTS

For each model we report accuracy results and the "out-of-the-bag" (OOB) error, prediction estimate, and the average estimate of prediction errors across all trees in the forest for observations left out of the randomly sampled "bag" of data used to train the trees. We estimate the importance of the prediction used in the models as reflected by two measures: Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI) index, or Mean Decrease Gini. MDA is a measure of the decrease in accuracy depending on the presence or absence of specific variables, and MDI is a measure of the decrease in data partition impurity during the classification (Hur et al., 2017).

6.5.1. DEMOGRAPHIC PREDICTORS

The RF method on the DT training dataset showed an OOB error of 16.9%. Accuracy results of the model with demographic predictors are summarized in Table 6.1. The importance of the predictors used by the model, and expressed by MDA and MDI variable importance measures, are depicted in Figure 6.1.

On the DU training set, the RF model yielded an OOB error rate of 14.7%. Accuracy results of the model with demographic predictors are presented in Table 6.1. The importance of the predictors is plotted in Figure 6.2.

The RF model ran on the nDU training set demonstrated an OOB error rate of 15.8%. All accuracy results of the model are summarized in Table 6.1, and variable importance is

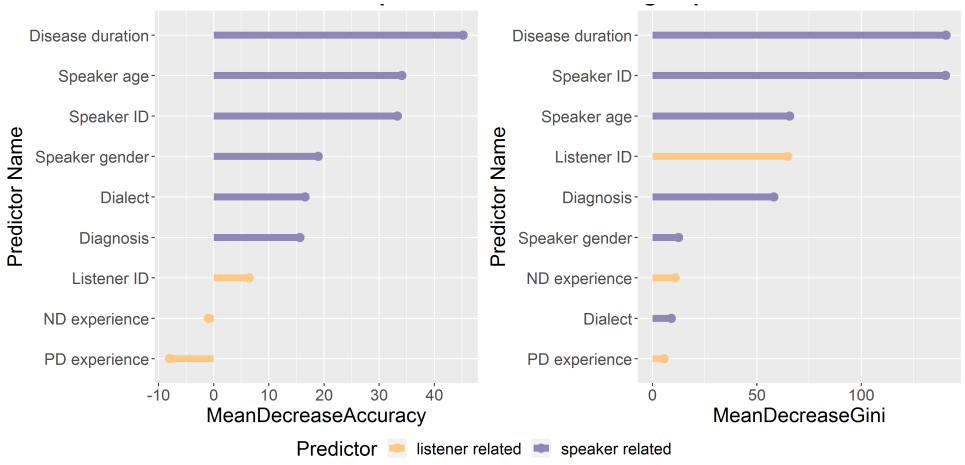

Figure 6.1 | Variable importance for the model trained on demographic predictors from the dataset of responses from the Dutch trained listeners.

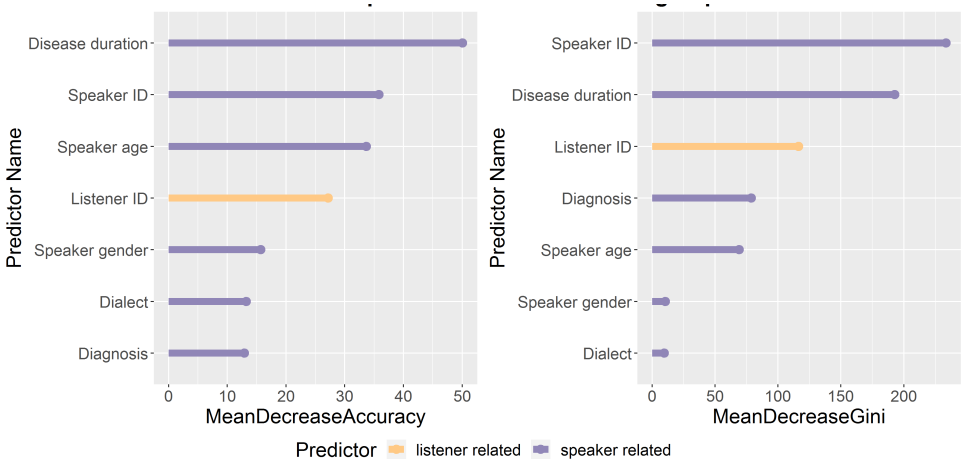

Figure 6.2 | Variable importance for the model trained on demographic predictors from the dataset of responses from the Dutch untrained group.

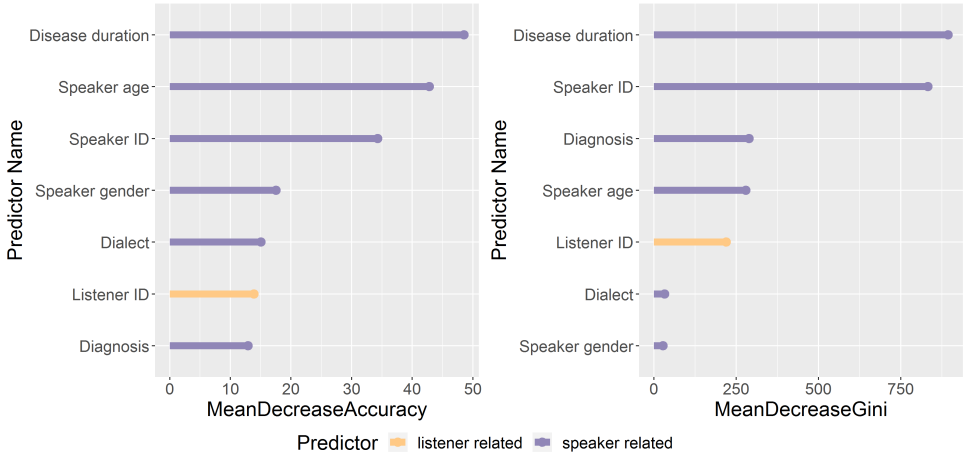

Figure 6.3 | Variable importance for the model trained on demographic predictors from the dataset of responses from the non-Dutch untrained group.

Table 6.2 | Results of the model with acoustic predictors with accuracy and confidence intervals for every listener group.

Group	Model results on the training set		Model results on the test set	
	Accuracy	95% CI	Accuracy	95% CI
DT	86.1%	[0.842 , 0.8783]	84.2%	[0.81 , 0.8698]
DU	86.1%	[0.8452, 0.8749]	83.6%	[0.8101, 0.8588]
nDU	85.4%	[0.8462, 0.8608]	85.8%	[0.8465, 0.8688]

plotted in Figure 6.3.

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

6.5.2. ACOUSTIC PREDICTORS

The model trained with acoustic parameters showed an OOB error rate of 16.9% on the trained set of the responses from the Dutch trained group. Accuracy results for the model are summarized in Table 6.2. Variable importance with colour-coded feature domains is presented in Figure 6.4.

On the training set of the Dutch untrained group's responses, the acoustic model showed an OOB error rate of 16.6%. Variable importance of the model trained on the dataset of responses of untrained Dutch listeners with colour-coded feature domains is

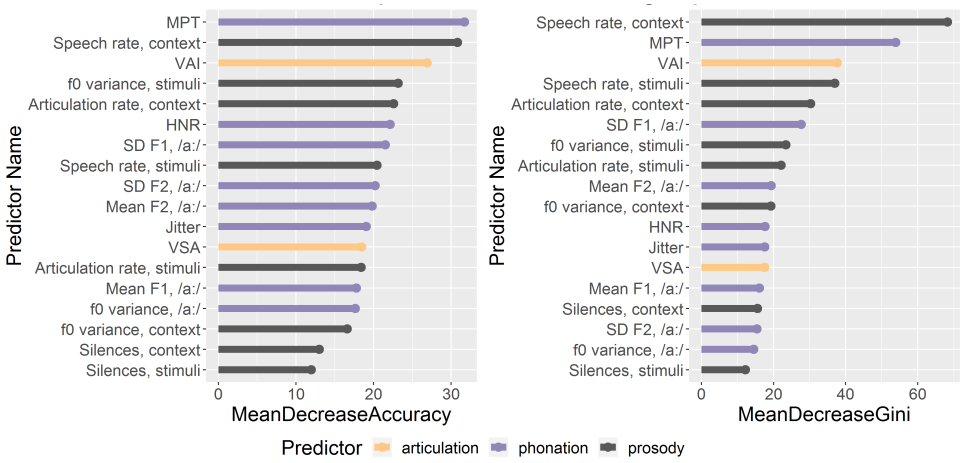

Figure 6.4 | Variable importance of the acoustic predictors used by the model on the dataset of responses from the Dutch trained listeners.

presented in Figure 6.5.

The model trained on the training set of non-Dutch untrained group's responses showed a 14.7% OOB error rate. Variable importance of the model trained on the dataset of responses of non-Dutch untrained listeners with colour-coded feature domains is presented in Figure 6.6.

Speech rate measured in stimuli with context and maximum phonation time (MPT), appeared to be the most important predictors for the acoustic model for all groups contributing to the accuracy. The articulation-related predictor *VAI* also appeared to be very important, being third for the Dutch trained and untrained groups, and fourth for the non-Dutch untrained group. *Inappropriate silences* measured both in stimuli and in stimuli with context were the predictors that contributed to accuracy the least for the Dutch speaking groups, while being ranked higher by the MDI measure.

6.5.3. FIRST LANGUAGE BACKGROUND INFLUENCE ON RECOGNITION PREDICTORS

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

GERMANIC LANGUAGE BACKGROUND

Results of the model with demographic predictors trained on the subset of data from Germanic listeners yielded a 15.6% OOB error estimate. Accuracy results for the Germanic and Slavic listeners are summarized in Table 6.3. From the variable importance analysis, it appeared that *Disease duration* was once again the most important predictor, followed

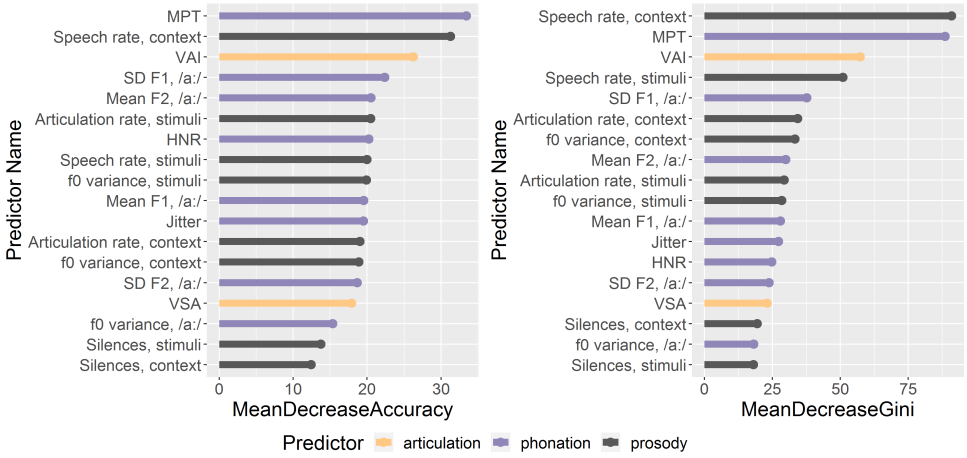

Figure 6.5 | Variable importance of the acoustic predictors used by the model on the dataset of responses from the Dutch untrained listeners.

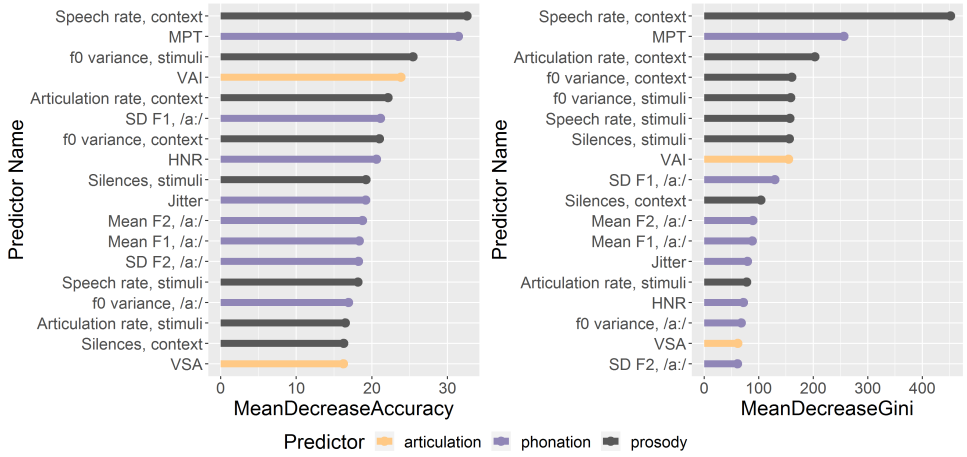

Figure 6.6 | Variable importance of the acoustic predictors used by the model on the dataset of responses from the non-Dutch untrained listeners.

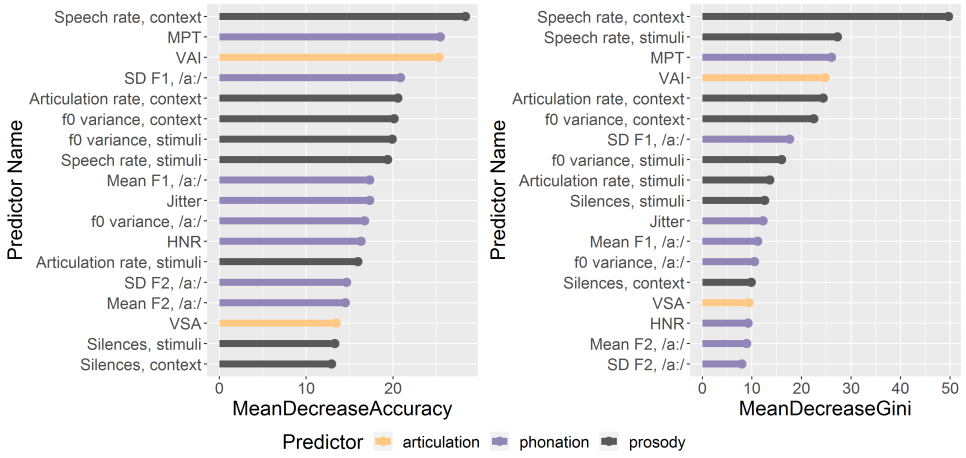

Figure 6.7 | Variable importance of the acoustic predictors used by the model on the Germanic subset of the dataset of responses from the non-Dutch untrained listeners.

by *Speaker ID, Speaker age, Diagnosis* and *Speaker gender. Dialect* and *Listener ID* had the lowest contributions.

Training the model with acoustic predictors demonstrated a 18.6% OOB error estimate, accuracy results are presented in Table 6.3. Variable importance is depicted in Figure 6.7.

SLAVIC LANGUAGE BACKGROUND

The model with demographic predictors of the Slavic data subset yielded similar results to the model ran on the Germanic subset of the dataset of responses from the non-Dutch untrained group: the OOB error estimate was 15.5%. Accuracy results for the demographic model are presented in Table 6.3. The variable importance analysis provided a similar picture to the model ran on the Germanic subset, with predictors of *Speaker age* and *Speaker gender* shifted higher in terms of accuracy importance. In terms of the purity index, results of the variable importance analysis for models trained on both Germanic and Slavic subsets showed that *Disease duration* and *Speaker ID* were the most important predictors.

The model with acoustic predictors demonstrated a lower OOB error estimate of 14%. Accuracy results for the second model are summarized in Table 6.3. Variable importance is presented in Figure 6.8.

6.6. DISCUSSION

We explored whether conventional acoustic measurements could predict how different listener groups recognize PD and control speech as healthy or unhealthy. While there exists a body of literature targeting recognition and assessment of speech of PwPD by listeners with different expertise and experience, the studies are mostly focused on measuring intelligibility scores or on exploring the component-specific perceptual assessment of speech produced by PwPD. A few studies have looked into global assessment and rating

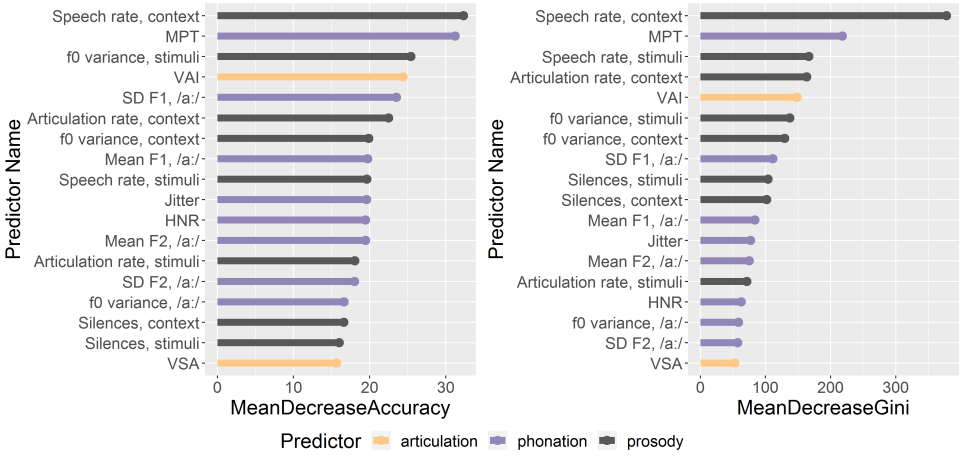

Figure 6.8 | Variable importance of the acoustic predictors used by the model on the Slavic subset of dataset of responses from the non-Dutch untrained listeners.

Table 6.3 | Results of the models with accuracy and confidence intervals for Germanic and Slavic listeners.

Group / type of predictors	**Model results on the training set**		**Model results on the test set**	
	Accuracy	**95% CI**	**Accuracy**	**95% CI**
Germanic / demographic	87.4%	[0.844 , 0.8601]	87.4%	[0.8363, 0.8616]
Germanic / acoustic	85.4%	[0.8304, 0.8746]	86.5%	[0.8282, 0.8958]
Slavic / demographic	85.3%	[0.8452, 0.8613]	85.1%	[0.8383, 0.8634]
Slavic / acoustic	86.3%	[0.8522, 0.8679]	86.4%	[0.8517, 0.8759]

of speech affected by PD by different listeners (Weismer et al., 2001; Sussman and Tjaden, 2012), and even fewer provided a cross-linguistic perspective (Pinto et al., 2017; Näsström and Schalling, 2020). Therefore, we focused on a more global assessment of speech of PwPD from two perspectives: one that focuses on the SLT experience and one that focuses on the influence of language background. Accordingly, we analyzed two feature sets to see whether listeners' responses can be reliably predicted from any of them. On both acoustic and demographic feature sets, Random Forest analyses demonstrated the accuracy of listeners' responses classification above 84% for each listener group, which was significantly above guessing level. This accuracy suggests the importance of the selected features for prediction.

For the first model based on demographic features of both speakers and listeners, results demonstrated the expected influence of the predictor *Disease duration* on the model's accuracy for every listener group. The importance of both *Listener ID* and *Speaker ID* predictors suggest the other less random predictors available to the model did not have high predictive power. This indicates the possible importance of unlisted predictors related to both speakers and listeners. It is also clear that the Dutch trained and non-Dutch untrained listeners assessed speech samples more uniformly: the predictor *Listener ID*'s contribution to the model's accuracy was very small. At the same time, in the Dutch untrained group, the mean accuracy would have decreased by more than 25% if the *Listener ID* predictor were to be randomized. This might be a sign of a lower inter-rater agreement, but the Fleiss' Kappa inter-rater agreement for all listener groups was very similar: from 0.48 to 0.53, which represents good agreement beyond chance (Fleiss et al., 2013). Therefore, given the good agreement for the Dutch untrained group, this surprising difference in accuracy-related importance for the *Listener ID* predictor may be caused by some group characteristic or specific sensitivity of Dutch untrained listeners to certain unlisted predictor compared to both Dutch trained and non-Dutch untrained groups. This finding contrasts with the results of Walshe et al. (2008), where authors found lesser agreement within their SLT group. This finding that the *Listener ID* predictor is important in modelling responses of Dutch untrained listeners could also be a result of framing the question in the task ("Did this voice sound healthy to you?"). It seems possible that this question could have been interpreted by the Dutch trained group in a more uniform way than by the Dutch untrained group. In other words, the trained group could have interpreted the word "healthy" as a term, with a consensual meaning, whereas the untrained groups could have had a more common-sense understanding.

Interestingly, in the variable importance analysis for the Dutch trained group, two features related to listeners' training (experience with neurodegenerative diseases and specific experience with PD) negatively influenced the model's accuracy. Apparently, these features had little predictive power to set apart listeners with a PD-related "expertise profile" to more accurately classify their responses. One possible explanation is that with a small subgroup of experienced people these predictors introduced more noise than information into the model. Another possible explanation could be the potential similarity in strategies applied by everyone in the Dutch trained group because of the task specifics, that is, general question about healthiness of voice, rather than component-specific auditory-perceptual assessments or intelligibility judgements. There is a possibility that, given the assessment of only short fragments, when classifying speech

as “unhealthy”, trained listeners were distracted by their attentive efforts to use specific criteria imposed by their training, while other groups relied on their intuitive impressions of what represents “unhealthy” speech.

In this study we addressed three research questions. First, whether listeners' responses about speech healthiness can be predicted from a set of acoustic features. Second, which acoustic features are more important when predicting listeners' responses about speech healthiness. Third, if the relevance of acoustic features that are predictors of the responses about speech healthiness depends on listeners' SLT experience and language background.

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

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

Another interesting finding is that *speech rate* calculated from the stimuli extracted with their contexts was an important predictor for each group, independent of the expertise or language background. Based on the rankings of deviant dimensions in speech of people with HD (Darley et al., 1969b; Bunton et al., 2007), rate appears to be much less prominent than monopitch or inappropriate silences. This unexpected finding highlights the need for additional investigations into the influence of rate and its possible correlations with listener's assessment of speech healthiness. It is also noteworthy that *speech* and *articulation rate* were often more important for reliable model prediction when calculated from the stimuli extracted with the context, while f_0 *variance* calculated from the stimuli was in general more important than f_0 *variance* calculated from the stimuli with the context. This brings to light that some rhythm patterns in speech manifest themselves on durations longer than 3-4 seconds and that they are still detectable by listeners even when not presented in their entirety. This could also be attributed to the dimension of short rushes of speech (Darley et al., 1969b) which may have been present in speech of the speakers with PD and that might be more reliably measured by the script when it is given a longer speech sample.

Concerning the third research question, whether listeners with different experience with speech disorders are sensitive to different acoustic cues in speech of PwPD, there was no clear tendency of a model relying mostly on prosody or phonation and voice quality for any of the listener groups. For the Dutch speaking listener groups, the top three predictors were the same: *maximum phonation time*, *speech rate* in stimuli extracted with context, and *vowel articulation index*. However, contrary to our hypothesis, among

predictors with a higher contribution to accuracy (that is, if randomized, there will be a mean decrease in accuracy over 20%) (see Figures 6.4–6.6), predictors from the domain of phonation and voice quality appear to be more important for predicting responses of the Dutch untrained group. Taking into account the previous study (see chapter 5), in which a group of untrained Dutch listeners was more successful at recognizing speech of PwPD as unhealthy, these features related to phonation and voice quality can be a valuable source of information for the listeners, which is also in line with observations by Näsström and Schalling (2020).

The third research question also focused on whether listeners with different degrees of familiarity with a speaker's language rely on different acoustic cues when recognizing speech as healthy or unhealthy. According to the model, and in line with our expectations, the Germanic listeners' responses, similar to the Dutch listeners' responses, highly correlated with the features related to phonation and voice quality, while for Slavic listeners, our model favoured prosodic features, suggesting different assessment strategies employed by listeners from different language backgrounds. There was another interesting observation regarding demographic features which highlighted a possible difference in assessment strategies. Visible in the variable importance analysis, *speakers' age* and *gender* appeared to be more important predictors for Slavic listeners' than for Germanic listeners' responses, which could be attributed to cultural differences and calls for additional research. However, contrary to our expectations, we found a pattern similar to Slavic listeners in the Dutch trained group which is visible in the model's variable importance (see Figure 6.6).

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

The current study provided evidence that in contrast to our expectations, Dutch untrained listeners' responses were better predicted by phonation and voice quality features than the responses of the Dutch trained listeners. Surprisingly, experience with neurodegenerative disorders or specifically with PD had a negative effect on the prediction accuracy of the model. Such findings calls for additional research with larger and more balanced groups of trained listeners including SLTs working with PD (cf. group division in

Carvalho et al. (2020)) to investigate the resposnse patters, accoustic correlates, and RF model accuracy for listeners with specific expertise in PD.

Such findings also point to certain limitations of the current study, as enlarged and more balanced groups of listeners including a separate group of trained listeners with PD expertise would provide a clearer perspective on the influence of specific expertise on the prediction of the listeners' responses. Another limitation is that we restricted ourselves to a certain number of predictors, therefore, including a broader list of features such as other articulatory and intensity measurements could lead to higher prediction results of the Random Forest model. Exploring additional predictors could also help to better understand the acoustic cues important for recognition of speech healthiness. Another limitation is a potential bias of the Random Forest method in ranking predictors, as it has been shown that variable importance measurements are not always reliable, especially when potential predictor variables vary in their scale of measurement or their number of categories (Strobl et al., 2007).

The findings of the current study have similar real-life implications as described in Sussman and Tjaden (2012), where recognition of voices as unhealthy may have a negative effect on speakers' potential employability and/or social activity. Knowledge of specific acoustic changes that trigger listeners to recognize speech as "unhealthy" can also provide specific therapeutic targets to enhance communication efficiency of speakers with HD as well as help to work on alleviating the negative attitudes speakers with HD can be confronted with (Miller et al., 2006; Maryn and Debo, 2014). Such findings also contribute to the growing body of research that recommend both researchers and clinicians to incorporate more global perceptual measures that can help understand and incorporate listeners' sensitivity to a variety of variables, including voice, prosody, and other speech characteristics (Kent et al., 1989; Sussman and Tjaden, 2012). To further explore recognition of healthiness , future research should also investigate additional articulatory measures as well as less conventional acoustic measures that would provide a more detailed understanding of acoustic correlates of listeners' impressions about speech healthiness. It should also control for the specific HD-related experience in the trained group to be able to differentiate classification patterns in that group and correlate it with the acoustic measures.

III

TIME MATTERS: LONGITUDINAL SINGLE CASE STUDIES

Every day is a journey,
and the journey itself is my home.

Matsuo Bashō