*Proceedings of the RaPID-4 @LREC 2022*, pages 1–8
Marseille, 25 June 2022


# The effect of eHealth training on dysarthric speech

**Chiara Pesenti**[1], **Loes van Bemmel**[2,3], **Roeland van Hout**[4], **Helmer Strik**[4,5,6]
Department of Humanities, Department of Artificial Intelligence,
Institute for Computing and Information Sciences, Centre for Language Studies,
Centre for Language and Speech Technology, Donders Institute for Brain, Cognition and Behaviour
University of Turin
Radboud University Nijmegen
chiara.pesenti@edu.unito.it, {loes.vanbemmel, roeland.vanhout, helmer.strik}@ru.nl

**Abstract**

In the current study on dysarthric speech, we investigate the effect of web-based treatment, and whether there is a difference between content and function words. Since the goal of the treatment is to speak louder, without raising pitch, we focus on acoustic-phonetic features related to loudness, intensity, and pitch. We analyse dysarthric read speech from eight speakers at word level. We also investigate whether there are differences between content words and function words, and whether the treatment has a different impact on these two classes of words. Linear Mixed-Effects models show that there are differences before and after treatment, that for some speakers the treatment has the desired effect, but not for all speakers, and that the effect of the treatment on words for the two categories does not seem to be different. To a large extent, our results are in line with the results of a previous study in which the same data were analyzed in a different way, i.e. by studying intelligibility scores.

**Keywords:** eHealth, Parkinson's Disease, dysarthric speech, POS tagging, function and content words

## 1. Introduction

The automatic acoustic-phonetic analysis of atypical speech is a promising pathway in pathological speech assessment. Automatically identifying the most relevant characteristics of pathological speech could lead to a reliable, accurate and non-invasive assessment method, able to distinguish typical speech from atypical speech, as well as measuring the extent of speech problems and diagnosing different types of atypical speech.

We focus in this study on dysarthria caused by Parkinson's disease (PD). PD is a chronic and progressive neurodegenerative disorder that significantly affects the use and cost of societal resources. More than 90% of patients with PD suffer from speech disorders (De Swart et al., 2003), collectively referred as dysarthria. Such disorders are typically characterized by increased acoustic noise, reduced voice intensity, harsh and breathy voice quality, lack of emotional expression and tonal changes, disturbances of speech rate, imprecise articulation of consonants, involuntary introduction of pauses, rapid repetitions of words and syllables, and sudden deceleration or acceleration in speech (Yang et al., 2020). These symptoms often have serious repercussions on speech intelligibility and daily communication. Moreover, some of them, such as the lack of emotional expression, characterize dysarthria caused by PD, and do not arise in other types of dysarthria.

Speech training with a serious game was given to eight PD patients. Especially focusing on acoustic features related to loudness, intensity and pitch, the game aimed to improve the intelligibility of people with dysarthric speech. Ganzeboom et al. (2022) collected human ratings of the speakers' intelligibility scores of utterances in a pre and post test and concluded that there was a significant speaker-specific improvement. Furthermore, the positive effect of a web-based treatment is thoroughly investigated and confirmed in Ganzeboom et al. (2022). We aimed to investigate its positive effects on acoustic-phonetic features related to loudness, intensity, and pitch.

In this study, we investigated whether this improvement is directly reflected in the acoustic features of loudness, intensity and pitch at the word level, using NLP parsing tools. We also wanted to explore the differences of these three types of acoustic features among two global word categories, namely content words and function words.

## 2. Background

### 2.1. *Treasure Hunters*: a web-based treatment

The serious speech training game Treasure Hunters was developed in the project CHASING: 'CHAllenging Speech training In Neurological patients by interactive Gaming'. Additional information about CHASING project is given in (Ganzeboom et al., 2022) (`http://waag.org/project/chasing`, `http://hstrik.ruhosting.nl/CHASING`).

The Treasure Hunters game is based on the Pitch Limiting Voice Treatment (PLVT), where the goal is to improve speech intelligibility by speaking louder, without raising the pitch. Treasure Hunters gives automatic feedback on the users' voice loudness and pitch, encouraging them to speak loud and low.

The target group for the Treasure Hunters game are older patients suffering from dysarthria due to PD. Previous studies by Ganzeboom et al. (Ganzeboom et al.,

2018) (Ganzeboom et al., 2022) showed that the effect of the game on intelligibility varied between speakers. For some speakers the game seemed to have to desired effect, while for others this is not the case.

### 2.2. Loudness, intensity and pitch

PLVT is the standard treatment in dysarthria therapy in the Netherlands (Kalf et al., 2011). In this treatment, patients are encouraged to speak 'loud and low', implying that they should try to increase voice intensity, while avoiding to raise their pitch, which easily happens when intensity increases. Increasing voice intensity often also results in better articulation. The purpose of avoiding the side effect of pitch increase distinguishes PLVT from Lee Silverman Voice Treatment (LSVT) (Ramig et al., 1995), which focuses only on increasing intensity. Therefore, in the current study, we focus on loudness, intensity, and pitch.

To study the effects of this therapy, we selected features from eGeMAPS (Eyben et al., 2015) and features extracted with our own Praat script that are related to loudness, intensity, and pitch. Note that the entire set of acoustic features was already analyzed at the utterance level by van Bemmel et al. (2021) for the same dysarthric speakers. Several individual features related to loudness and pitch were found to be relevant in classifying before and after treatment recordings. Here, instead of at utterance level, we want to analyse those features at word level. Moreover, we want to identify more general factors or components in the many features related to intensity, loudness, and pitch.

### 2.3. Content and function words

One of the aims of this study is to investigate whether there are meaningful acoustic-phonetic features for the distinction between content words and function words. These two classes of words are based on syntactic-semantic criteria. Content words are members of open word classes with a clear lexical meaning, such as names, nouns, lexical verbs, adverbs or adjectives. They are phonologically and morphologically independent. Function words belong to closed word classes that do not carry a full lexical meaning, and determine the grammatical relations between content words. They are often phonologically and morphologically dependent, have reduced scrambling possibilities, and usually have a shorter length and high frequency of occurrence.

Importantly, content and function words have been studied in relation to atypical speech. Several studies on stuttering (Howell et al., 1999), aphasic speech (Bird et al., 2002), and amyotrophic lateral sclerosis (Turner and Tjaden, 2000), observed differences in the production and perception of these two word classes. In particular, Turner and Tjaden (2000) focused on acoustic differences between content and function words, finding no statistical difference between healthy and pathological speech for vowels acoustic features of formants, space area and duration. Their work also highlights that these features values were generally larger for content words and that the difference of vowel space area for content and function words, although not statistically relevant, tended to be smaller in pathological speech than healthy speech. Bird et al. (2002) instead, studied the production and the comprehension of these two categories of words, finding discrepancies between content and function words only for reading tasks, but not when the imageability was controlled.

To determine how Dutch parts of speech are distributed among these two categories, a literature study on Dutch content and function words was carried in order to classify Dutch POS tags into the two word categories, as explained in section 3.3.

### 2.4. Research Questions

1. Does the web-based gaming treatment have an impact at word level with respect to loudness, intensity, and pitch?

2. (a) Is the impact general or dependent on the speaker involved, and (b), if speaker dependent, do loudness, intensity, and pitch values improve in line with the intelligibility score improvements found in Ganzeboom et al. (Ganzeboom et al., 2022)?

3. Does the treatment have the same impact on content and function words?

## 3. Material and methods

### 3.1. Data and Participants

The speech data were recorded from eight native Dutch speakers with Parkinson's Disease (PD) who underwent web-based treatment with the "Treasure Hunters" serious game for speech training (Ganzeboom et al., 2022). During the four weeks of training the speakers were instructed to speak loud and low, following the concept of Pitch Limiting Voice Treatment used in the Treasure Hunters game. Each speaker was recorded twice, pre-treatment (T2) and four weeks after continuous treatment (T3). During both times, each speaker was asked to read seven Dutch sentences out loud after reading it silently for themselves.

We selected all 32 read sentences from the phonetically balanced story "Papa en Marloes" (11 sentences; (Van de Weijer and Slis, 1991)) and the text of apple pie recipes (21 sentences), both used by Ganzeboom et al. (2022). These sentences vary between 4 and 14 words with a total of 251 words (143 content words, 108 function words).

Table 1 shows general information about the speakers.

### 3.2. Extracting and POS tagging OTs

Part-of-speech (POS) tags were created for each word in the orthographic transcriptions (OTs) using Alpino (Bouma et al., 2000), a dependency parser for Dutch. The OTs were obtained for each recording through manual transcription by students at Radboud University. The differences between the standard written text

Table 1: *Speakers' general data*

| Speaker | Gender | Age (years) | Time since diagnosis (years) |
|---|---|---|---|
| 01 | M | 73 | 4.5 |
| 02 | M | 56 | 8.0 |
| 03 | M | 60 | 4.5 |
| 04 | M | 63 | 5.5 |
| 05 | F | 53 | 9.0 |
| 06 | M | 75 | 2.0 |
| 07 | F | 67 | 3.0 |
| 08 | F | 62 | 3.0 |

of the prompts and uttered words of the OTs are due to the intrinsic nature of the read speech. Indeed, read speech has a syntactic structure rather consistent and not fragmented, similarly to written language, but at the same time shares some typical elements of spontaneous speech such as stuttering, repetitions, fragmented words, filled pauses, elongated vowels and no punctuation.

By comparing the POS tags of the prompts with the POS tags of the OTs, it turned out that the absence of punctuation marks prevented the correct functioning of Alpino for the OTs. Therefore, a Python code was created with the aim of locating and identifying the punctuation of the prompts and reinserting it into the OTs. To avoid errors, the spot of insertion of the punctuation was determined by the first three words preceding and the first three words following the prompt punctuation. In order to predict the accuracy of the OTs POS tagging, all POS tags of the prompts were manually checked, and it turned out that out of a total of 252 words, 24 were incorrectly tagged by Alpino. Only four of these errors concerned the tagging of a content words instead of a function words or vice versa, while the others were wrong tags assigned to words belonging to the same word category.

### 3.3. Words labelling into content words and function words

The ultimate goal of POS tagging was to label all uttered words as content words and function words. All POS tags used by Alpino have been matched with an additional tag indicating the membership to one of the two global word classes. This step was accomplished through a Python script.

According to the literature research on Dutch content words and function words, a model of tags matching between Alpino POS tags and the two words categories was created. The matching model, shown in Tables 2 and 3, was designed with the aim of making the matches compatible with both the POS tags available and the outcomes of the literature studies.

### 3.4. Acoustic features, outliers detection, data normalization

A total of 103 acoustic features were automatically extracted using Praat (Boersma and Weenik, 2020)

Table 2: *Matching model for content words*

| POS tags | Content words |
|---|---|
| N, SPEC | nouns (+ proper nouns) |
| WW | main verbs |
| ADJ | adjectives |
| BW | adverbs with semantic meaning |

Table 3: *Matching model for function words*.
*: + 300 adverbs (adverbial grammatical function)

| POS tags | Function words |
|---|---|
| LID | determiners (articles, prenominal pronouns) |
| VNW | pronouns |
| VG | conjunctions |
| VG | subordinate conjunctions |
| TSW | interjections |
| VZ | adpositions |
| TW | (cardinal) numerals |
| WW | auxiliary and copula verbs |
| BW | conjunctive adverbs* |

and openSMILE (Eyben et al., 2010). The 15 features extracted by Praat are duration, the four formants, pitch variance, gravity center and the mean, minimum, maximum and standard deviation of pitch and intensity. Using the python package openSMILE, the 88 extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016) features were extracted. These 103 features were extracted on word-level after the speech was force aligned using an in-house forced aligner (`http://webservices.cls.ru.nl`). Note that in (van Bemmel et al., 2021) the same features were extracted with the same method at phoneme, word and utterance level.

Afterwards, outlier detection was carried out among the values of the extracted feature. Thus, a percentage of outliers of about 20% was detected. It turned out that most outliers were due to a worse quality of the audio at the end and beginning of recordings.

A further pre-processing step was the data normalization. Our data were standardized calculating the z-scores. Since there were multiple recordings per speakers, it was possible to calculate z-scores per speaker, thus minimizing the inter-speaker differences and the identify-confounding (Chaibub Neto et al., 2019).

### 3.5. Features reduction

From the 103 acoustic features, 24 features relating to loudness, intensity, and pitch were selected for further dimensionality reduction with Principal Component Analysis (PCA) (Pearson, 1901). Based on the Eigenvalue being larger than one, six principal components were formed, as shown in Table 4. Three of the principal components were related to loudness and intensity features (*li1, li2, li3*), and the other three were related to formant and pitch features from Praat (*F0.1, F0.2, Ppitch*). Inspecting the subset of features grouped

in each of the six components and the component loadings, it was possible to notice that some components were more representative for loudness, intensity and pitch than others. Among the principal components related to loudness and intensity, *li1* grouped mean and higher values of loudness and intensity, *li2* grouped lower values, whereas *li3* grouped values expressing the variation in loudness. Among the principal components related to pitch, *F0.1* grouped static values of pitch obtained with eGeMAPS, *F0.2* grouped pitch dynamic values and range obtained with eGeMAPS, and *Ppitch* grouped pitch range and variation obtained with Praat.

Note that interpretation of principal components is not as straightforward as interpretation of acoustic features, as components are a combination of features. However, we can state that *li1* and *F0.1* seem to be the most complete and exhaustive components for the representation of loudness and intensity and pitch respectively. For this reason, although we have analyzed and reported the values obtained with all six components, we focused more on *li1* and *F0.1*.

### 3.6. Statistical analysis: Linear Mixed Regression Models

Linear Mixed-Effects Models (lmer; package lme4 (Bates et al., 2007)) in R (R Core Team, 2020) was used for the statistical analysis, in combination with the packages lmerTest and SjPlot. The analysis contained two fixed variables, Time (pre vs. post treatment) and Wordclass (function vs. content words), plus their interaction. We included three random effects: Speaker, Word, and Speaker-by-Time. The last effect is a random slope that enables the analysis to capture speaker specific treatment effects. The criterion variables in these analyses were the scores on the six PCA components.

## 4. Results

### 4.1. The fixed effects

No significant effects ($p<.05$) were found for Time or its interaction with Wordclass for any of the six PCA components.

The variable Wordclass has a significant effect on the values of *li1, li2, F0.2* and *li3*. As the boxplots for these components show, *li1* and *F0.2* are lower for function words, while *li2* and *li3* are larger for function words. Given the absence of significant interactions between Wordclass and Time, speakers thus show similar differences between content and function words in pre- and post-treatment.

The boxplots 3 and 2 show the normalized values of content and function words recorded for the eight speakers for the components *li1* and *F0.1*.

### 4.2. The random effects related to Speaker

Inspecting the plots of the random effects, the observation was made that the component values shown by the speaker intercepts may differ between speakers. More interesting are the slopes of Speaker-by-Time.

The slope values of *li3* and *Ppitch* do not deviate from 0 for any speaker, indicating that there is no significant difference between pre- and post-treatment for any speaker. The speaker results of the random slopes of the other four components can be seen in Table 5. The "-" symbol indicates a negative slope, meaning that the value of this component decreased from T2 to T3 for

Table 4: *PCA groupings obtained with Praat and eGeMAPS features related to intensity, loudness and pitch*

**PC1: li1**
loudness_sma3_percentile80.0
loudness_sma3_amean
intensity_max
intensity_mean
loudness_sma3_pctlrange0-2
loudness_sma3_percentile50.0
loudness_sma3_meanRisingSlope

**PC2: F0.1**
F0semitoneFrom27.5Hz_sma3nz_percentile20.0
F0semitoneFrom27.5Hz_sma3nz_percentile50.0
F0semitoneFrom27.5Hz_sma3nz_amean
F0semitoneFrom27.5Hz_sma3nz_percentile80.0
HNRdBACF_sma3nz_ameanpitch_minpitch_mean
pitch_min
pitch_mean

**PC3: li2**
loudness_sma3_amean
loudness_sma3_pctlrange0-2
loudness_sma3_percentile50.0
intensity_min
loudness_sma3_percentile20.0

**PC4: F0.2**
F0semitoneFrom27.5Hz_sma3nz_amean
F0semitoneFrom27.5Hz_sma3nz_percentile80.0
HNRdBACF_sma3nz_amean
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope
jitterLocal_sma3nz_amean

**PC5: Ppitch**
pitch_min
pitch_maxpitch_var
pitch_mean

**PC6: li3**
loudness_sma3_meanRisingSlope
loudnessPeaksPerSec
loudness_sma3_meanFallingSlope
jitterLocal_sma3nz_amean

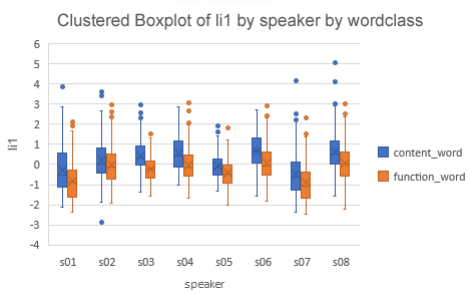


Figure 1: Boxplots of the normalized component *li1* (relating to loudness and intensity) per Speaker per WordClass.

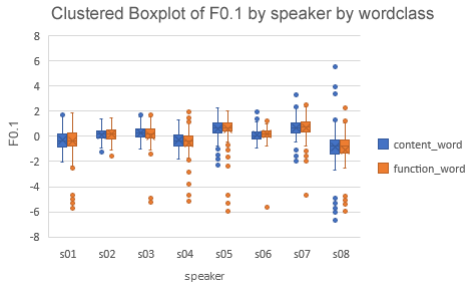


Figure 2: Boxplots of the normalized component ıF0.1 (relating to formants) per Speaker per WordClass.

that particular speaker. The "+" indicates a positive slope, meaning an increase in component value instead. The confidence interval of specific slopes did not cross the 0, and those slopes are therefore considered significant. These slopes are marked in the Table 5 with *s. Other slopes were too small to be considered significant.

Given their component loadings, the *li1* and *F0.1* components are most closely linked to the mean values of loudness, intensity and pitch. The other principal components are related to, e.g., lower values [minimum, 20-percentile] and variation in the feature values. Interpreting the outcome of the statistical analyses for these components is thus more complex, but also less relevant for the current research. We therefore focus on the *li1* and *F0.1* components.

Figures 3 and 5 show the normalized values for the components ıli1 and *F0.1* before (T2) and after (T3) treatment per speaker. Even if the $p$ values of the intercept and the variable Time are not $< .05$, speaker 03, 05 and 07 show a clear increase for the *li1* component and decrease or keep approximately stable for the *F0.1* component. Figures 4 and 6 show the slopes of the random effects (intercept) speaker and Time for the components *li1* and *F0.1*, respectively. Note that the left part of Figures 3 and 5 shows the speaker intercept, indicating the between-speaker variance of the component values, whereas the right part of the figure shows the speaker slope indicating the treatment effect, i.e. the difference between the pre and post test per speaker.

Speakers 01, 02, and 06, had a negative slope for both *li1* and *F0.1*. For speakers 04, 05, and 07, a positive slope was found for both *li1* and *F0.1*. Speaker 03 had a positive slope for *li1* but a negative slope for *F0.1*, while speaker 08 had the opposite.

Table 5: Random effects on Time(T3). Symbols - and + indicate a negative or positive slope. *: $|\text{value}| > 0.5$, **: $|\text{value}| > 1$, ***: $|\text{value}| > 1.5$.
*li3* and *Ppitch* components are not included since their values do not deviate from zero.

| Speaker | F0.1 | F0.2 | li1 | li2 |
|---|---|---|---|---|
| 01 | - | - | -*** | -* |
| 02 | - | - | - | - |
| 03 | - | - | + | + |
| 04 | + | - | + | + |
| 05 | + | + | + | + |
| 06 | - | + | - | - |
| 07 | + | + | +** | + |
| 08 | + | - | - | + |

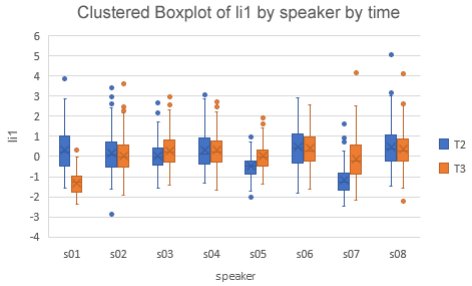


Figure 3: Boxplot of the normalized component *li1* (relating to loudness and intensity) per speaker per time point.

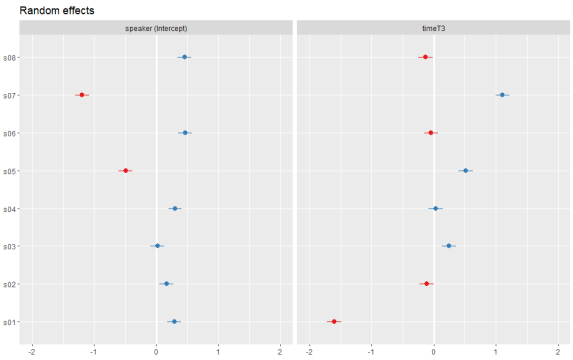


Figure 4: The random effects plot of the component *li1* (relating to loudness and intensity) shown for both intercept (speaker) and time

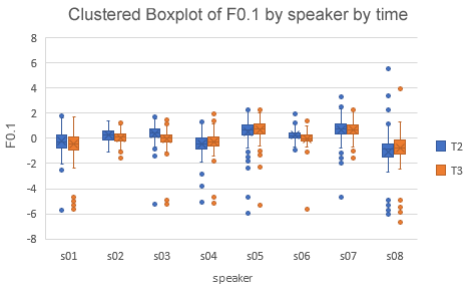


Figure 5: Boxplots of the normalized component *F0.1* (relating to formant features) per speaker per time.

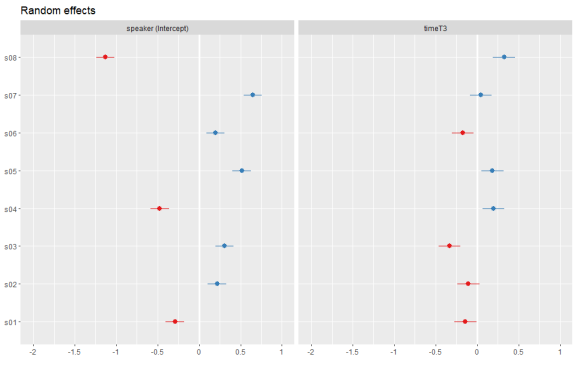


Figure 6: The random effects plot of the component *F0.1* (relating to formant features) shown for both Speaker intercept and Speaker by Time intercept

# 5. Discussion

## 5.1. Trend of loudness, intensity, and pitch compared with intelligibility scores

Our findings are mostly in line with Ganzeboom et al. (2022)'s results, especially with regards to loudness and intensity. Indeed, speakers 03, 04, 05 and 07, who increase in intelligibility scores after the treatment in Ganzeboom et al., show an evident increase of loudness and intensity after the treatment in our research. On the other hand, speaker 08, seemingly in contradiction with their increase in intelligibility score, does not show an increase in loudness and intensity after treatment in our results.

With regards to the pitch, no speaker shows an evident change for the component relating to the Praat pitch features. However, for the other two components relating to pitch (*F0.1* and *F0.2*), the differences per speaker between pre and post treatment recordings also vary. For speakers 03 and 06, who both improved, *F0.1* indeed goes down after treatment. However, for speakers 04, 05, 07 and 08, who also improved, the *F0.1* seems to increase after treatment. Speaker 01 and 2, who did not improve in intelligibility, also show a decrease in *F0.1* and *F0.2* scores after treatment. It seems that pitch is less relevant in the eventual intelligibility scoring than intensity, and even with some increase in pitch, a speaker can still be evaluated as more intelligible after treatment.

The fact that there seem to be multiple significant speaker results but not for the factor Time could be the consequence of our small sample size. With only eight speakers, there are undoubtedly between-speaker effects that interfere with the Time effect. Additionally, not all speakers profit from the treatment, something that came out in inspecting the interaction between speakers and Time.

## 5.2. Treatment impact on words categories

The two words categories show a clear difference in intensity as well among each other. Both in T2 and T3, speakers use greater intensity when pronouncing content words, according to component *li1*. Component *li2* gives us an unexpected result, showing larger values for function words. This is probably due to the aforementioned nature of the two different components. *Li2* indeed, is less representative of the loudness and intensity compared to *li1*, since it groups five acoustic features that mainly indicate low values of loudness and intensity, while *li1* groups features carrying mean and higher values. Even the pitch is slightly higher in the content words according to *F0.2*, while for *F0.1* the pitch of content and function words is approximately the same.

However, according to our data, there is no particular difference between the treatmen change shown by the content words treatment and the treatment change shown by the function words. Therefore, the web-based treatment seems to impact the two groups of words equally. Since all speakers show significant differences between the two word classes (function and content) in four out of six components, these natural variations between pronunciation of function and content words seems to be unchanged by treatment. Indeed, speech that does not present any tonal variance results is in fact monotonous and unnatural. The fact that the treatment does not flatten the acoustic differences between the two word categories is a positive effect and contributes to corroborate Ganzeboom's intelligibility scores, since some speakers manage both to increase loudness and intensity without changing the pitch, and to keep speech spontaneous and natural even after the treatment.

It is interesting, however, to notice that Turner and Tjaden (2000), comparing healthy speech with speech from speakers with mild to moderate dysarthria associated with amyotrophic lateralsclerosis, did not find any statistically significant difference between the patients and controls, but noticed different trends between function and content words with respect to the two groups of speakers.

## 5.3. Limitations

One of the limitations of our research is its focus on a reduced number of features compared to those avail-

able. This choice was dictated by the need to make the large number of acoustic features extracted for each word compatible with a valid statistical analysis. Therefore, the advantage of avoiding the curse of multi-dimensionality inevitably has the consequence of sacrificing information.

Additionally, the principal components obtained with PCA are less easily interpretable than the acoustic features they are created with. The components are some combination of groups of features with specific component loadings, making it difficult to draw clear conclusions out of increases or decreases of these components.

However, many of the initial 103 features are highly correlated with each other. In fact, we carried out the PCA with the aim to consider the greatest number of features that were linked to the loudness, intensity, and pitch, not limiting the research by selecting a single feature that most represented each of these three traits. Nevertheless, there is a large number of features related to the acoustic spectrum, to the four formants, and to temporal characteristics, which for the aforementioned reasons have been excluded from the research. It would be interesting to study those features in future studies.

Furthermore, we must take into consideration that the extraction of the acoustic features was done at word-level, and that different results perhaps would have come out with an analysis at utterance level as regards loudness, intensity, and pitch. In their work for instance, van Bemmel et al. (2021) detects the most relevant features of the same data used in this research analysing the features at phoneme, words and utterance level. Yet, only a word level analysis would have allowed a distinction between content and function words.

It would be appropriate to interpret our findings also in the light of the type of speech that has been analyzed, that is read speech. The nature of this type of speech has very different characteristics from those of spontaneous speech, and these differences could certainly have repercussions on a phonetic, lexical or syntactic level.

Finally, as regards the method used, we have partially adopted procedures that can only work for small corpora, such as the automatic insertion of punctuation. Also the labeling of words into content words and function words, based on Dutch POS tags, could turn problematic with words as POS tagging does not differentiate sufficiently for a classification in content and function words.

## 6. Conclusions

Using six principle components based on 24 acoustic features related to loudness, intensity, and pitch extracted from speech recordings with Praat and eGeMAPS, we found that some of these components reflect the changes in intelligibility after treatment with a serious game for Parkinson's Disease patients.

While no significant effect was found for the fixed factor Time, providing no proof for treatment impacting loudness, intensity, and pitch that is consistent in all speakers (answering question 1), it was found that the treatment effects differ per speaker.

*Li2* and *F0.2*, components relating to loudness and intensity and eGeMAPS features of pitch respectively, were found to have significant differences between speakers for these two component values. Random effect plots of the intercept also show differences between speakers for other components (answering question 2a).

Random effect plots show the differences between speaker slopes, where the link with previous research into intelligibility improvement after treatment can be made. For the components *li3* and *Ppitch* relating to eGeMAPS features of loudness and jitter and Praat features of pitch respectively, none of the speakers had a significant slope, indicating no difference after treatment for these component values. This is in line with Pitch Limiting Voice Treatment (PLVT), where the pitch is not supposed to increase and loudness does.

The other components (two relating to loudness and intensity and two relating to eGeMAPS features of pitch) and their respective slopes differ per speaker, again showing the speaker-dependent results. Speaker 01 and 02 did not improve in intelligibility, and all four of the other components decreased as well, implying a decrease in articulation quality with the loss of loudness in line with PLVT. Speaker 03 had a large improvement in intelligibility after finishing treatment, and did indeed show an increase in loudness and intensity components while showing a decrease in pitch components, perfectly following the PLVT. Speaker 04 and 06 both had a slight increase in intelligibility and a mix of increase and decrease for both loudness and intensity and pitch. Speaker 05 and 07 both had a large improvement in intelligibility and all four components increased, implying that it is possible that speech is considered more intelligible if both loudness and intensity and pitch are increased. Speaker 08 had a large intelligibility score but a mix of increase and decrease in the four component values (answering question 2b).

Given that there is no significant effect of the interaction between WordClass and Time, we can conclude that the treatment with the Treasure Hunters game has the same (namely, no) impact on content and function words and any variation between these two groups is consistent pre and post treatment (answering question 3).

Looking at this variation between content and function words, it was found that four out of six components (*li1, li2, li3, F0.2*) had a significant difference between content and function words (answering question 2).

Overall, our research answered some questions and raised many others. Among the topics that would be more interesting to investigate in future studies,

there is certainly the cause of such an evident speaker-dependent result. Many critical aspects of our method could also be further explored, such as the inclusion of more acoustic features or resolving the problems related to the POS tagging of read speech. Lastly, it would be interesting to do a similar analysis with spontaneous speech, focusing in particular on the fact that one of the symptoms of speech disorders in patients with dysarthria is also the lack of emotional expression and tonal changes.

## 7. Acknowledgements

We would like to thank our colleagues Mario Ganzeboom, Marjoke Bakker, Lilian Beijer, and Toni Rietveld, who were involved in acquiring the data within the CHASING project.

## 8. Bibliographical References

Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). The lme4 package. *R package version*, 2(1):74.

Bird, H., Franklin, S., and Howard, D. (2002). 'Little words'—not really: function and content words in normal and aphasic speech. *Journal of Neurolinguistics*, 15(3-5):209–237.

Boersma, P. and Weenik, D. (2020). Praat: doing phonetics by computer (version 6.1.22). `http://www.praat.org`.

Bouma, G., van Noord, G., and Malouf, R. (2000). Alpino: Wide-coverage computational analysis of Dutch. volume 37, pages 45–59, 01.

Chaibub Neto, E., Pratap, A., Perumal, T. M., Tummalacherla, M., Snyder, P., Bot, B. M., Trister, A. D., Friend, S. H., Mangravite, L., and Omberg, L. (2019). Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ digital medicine*, 2(1):1–6.

De Swart, B. J., Willemse, S., Maassen, B., and Horstink, M. (2003). Improvement of voicing in patients with Parkinson's disease by speech therapy. *Neurology*, 60(3):498–500.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Ganzeboom, M., Bakker, M., Beijer, L., Rietveld, T., and Strik, H. (2018). Speech training for neurological patients using a serious game. *British Journal of Educational Technology*, 49(4):761–774.

Ganzeboom, M., Bakker, M., Beijer, L., Strik, H., and Rietveld, T. (2022). A serious game for speech training in dysarthric speakers with Parkinson's disease: Exploring therapeutic efficacy and patient satisfaction. *International Journal of Language & Communication Disorders*.

Howell, P., Au-Yeung, J., and Sackin, S. (1999). Exchange of stuttering from function words to content words with age. *Journal of Speech, Language, and Hearing Research*, 42(2):345–354.

Kalf, J., de Swart, B., Bonnier, M., Hofman, M., Kanters, J., Kocken, J., Miltenburg, M., Bloem, B., and Munneke, M. (2011). Guidelines for speech-language therapy in Parkinson's disease. *Nijmegen, The Netherlands/Miami, FL: ParkinsonNet/NPF*.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

R Core Team, (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramig, L. O., Countryman, S., Thompson, L. L., and Horii, Y. (1995). Comparison of two forms of intensive speech treatment for parkinson disease. *Journal of Speech, Language, and Hearing Research*, 38(6):1232–1251.

Turner, G. S. and Tjaden, K. (2000). Acoustic differences between content and function words in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 43(3):769–781.

van Bemmel, L., Cucchiarini, C., and Strik, H. (2021). Using feature selection to evaluate pathological speech after training with a serious game. *ExLing 2021*, page 245.

Van de Weijer, J. and Slis, I. (1991). Nasaliteitsmeting met de nasometer. *Logopedie en Foniatrie*, 63(97.101).

Yang, S., Wang, F., Yang, L., Xu, F., Luo, M., Chen, X., Feng, X., and Zou, X. (2020). The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson's disease. *Scientific Reports*, 10(1):1–9.