

# Detecting Parkinson’s disease using Vocal Data from Patients

Applied Probability and Statistics for Engineers

**Submitted to: Dr. Amar Sabih**

**Team 1:**

**Abdulah Alhoothy: 40075668**

**Muhammad Haris: 40150933**

**Qamar Bani**-**Melhem: 40109566**

**Sourav Patne: 40137125**

**Yaoxin Li: 40092820**

Concordia University

Gina Cody School of Engineering and Computer Science

Montreal, Quebec, Canada

**Table of content**

**List of tables**

**List of figures**

**1.Abstract**

**[Abdullah] - [01/04/2020]**

**2.Introduction**

**Abdullah - 30/03/2020**

There are various medical exams that a person has to go through if he/she wants to be diagnosed in order to identify if they have parkinson's. the level of severity of the parkinson's disease are identified by the doctor who made the tests. By using dysphonia data taken from both healthy and patients with parkinson's disease we hope in our study to be able to discriminate between the two groups. An average of 6 lines were taken from each patient [31 patients]. Using calibrated and verified measurement procedures to detect speech signals is the main method by which the data was collected.

Vocal impairment was the main way in which parkinson’s was exhibited in the patients, (nearly 90% of patients) [3, 4]. At the beginning of the illness vocal impairments are the signs that doctors would look for this makes it particularly convenient because the measurement of vocal impairment is noninvasive and is simple to carry out[5, 6, 7]. and so the significance of tracking the progression of symptoms through voice measurement is very high[8].

For this study, Using a Multi-Dimensional Voice Program (MDVP) as the primary of recording the phonations and as software tool for quantitative acoustic assessment of quality of a patient’s voice, all the recording were carried out in a laboratory and under observation. the MDVP is able to measure up to 22 parameters on a single voice line from a patient [8].

**3.Problem description**

**Sourav - [28/03/2020]**

Parkinson’s disease is a neurodegenerative, progressive disorder of the central nervous system that affects movement and causes tremors and stiffness. This affects dopamine-producing neurons in the brain and every year, it affects more than 10 million individuals. Recently we have begun to utilize the data science to improve healthcare and services – predicting diseases early will have countless advantages on the prognosis. Our project proposal is to build a model that would detect Parkinson’s disease using vocal data taken from patients.

**4.Assumptions and limitations HARIS [28/03/2020]**

1. limited data size
2. Limited access to the patient's medical history, therefore we can’t use any other information

Assumptions **HARIS - [28/03/2020]**

1- we assume that the doctor’s diagnosis are always correct

2- all equipments & testing are in the best working conditions and properly calibrated

3- All standards have been followed during testing procedures

**5.Data analysis**

In this report, we are introducing the Parkinson disease (PD) and the type of parameters that had been tested on the patients. We use real data from source [1], and analyze the data with different ways. We studied 31 patients, 23 of them with PD. Then we select the highly 10 uncorrelated parameters and start working on them.

The period since diagnosis with PD is from 0 to 28 years, and the ages of the subjects are from 46 to 85 [8]. Multiple phonation tests were taken by the subjects (Averages of six phonations were recorded from each subject) ranging from one to 36 seconds in length [8].

The data chosen for the project had several parameters:

* MDVP (FO): Fundamental frequency (Fo) is the vibratory rate of the vocal folds. It can be measured in hertz or cycle per second (CPS). Average fundamental frequency during a conversation for males ranges from 100 to 150 Hz, whereas for females it ranges from 180 to 250 Hz.
* MDVP(FHI): maximum FO.
* MDVP(FLO): minimum FO.
* MDVP (Jitter %):Jitter is a measure of frequency instability. A normal voice has a small amount of instability during sustained vowel production Normal instabilities are influences by tissue and muscle properties. It is measured in %.
* MDVP (Jitter abs): Absolute jitter.
* MDVP (RAP): Relative measure of the pitch disturbance.
* MDVP (PPQ): Pitch perturbation quotient.
* MDVP (Shimmer): Shimmer is a measure of amplitude instability.
* MDVP (Shimmer db): Shimmer in db.
* Shimmer (APQ 3-5): Six measures of variation in amplitude perturbation quotient (APQ).
* (NHR): Noise-to-harmonics Ratio.
* (DFA): Signal fractal scaling exponent.
* Spread 1-2: Two nonlinear measures of fundamental frequency variation.
* (RPDE): recurrence period density entropy.
* (DFA): detrended fluctuation analysis.
* (PPE): Pitch period entropy.

**Tables and Figures**

The following table shows the list of subjects with sex, age, Parkinson’s stage and the number of years since diagnosis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject code | Sex | Age | Stage (H&Y) | Years since diagnosis |
| S01 | M | 78 | 3.0 | 0 |
| S34 | F | 79 | 2.5 | ¼ |
| S44 | M | 67 | 1.5 | 1 |
| S20 | M | 70 | 3.0 | 1 |
| S24 | M | 73 | 2.5 | 1 |
| S26 | F | 53 | 2.0 | 1½ |
| S08 | F | 48 | 2.0 | 2 |
| S39 | M | 64 | 2.0 | 2 |
| S33 | M | 68 | 2.0 | 3 |
| S32 | M | 50 | 1.0 | 4 |
| S02 | M | 60 | 2.0 | 4 |
| S22 | M | 60 | 1.5 | 4½ |
| S37 | M | 76 | 1.0 | 5 |
| S21 | F | 81 | 1.5 | 5 |
| S04 | M | 70 | 2.5 | 5½ |
| S19 | M | 73 | 1.0 | 7 |
| S35 | F | 85 | 4.0 | 7 |
| S05 | F | 72 | 3.0 | 8 |
| S18 | M | 61 | 2.5 | 11 |
| S16 | M | 62 | 2.5 | 14 |
| S27 | M | 72 | 2.5 | 15 |
| S25 | M | 74 | 3.0 | 23 |
| S06 | F | 63 | 2.5 | 28 |
| S10 (healthy) | F | 46 | n/a | n/a |
| S07 (healthy) | F | 48 | n/a | n/a |
| S13 (healthy) | M | 61 | n/a | n/a |
| S43 (healthy) | M | 62 | n/a | n/a |
| S17 (healthy) | F | 64 | n/a | n/a |
| S42 (healthy) | F | 66 | n/a | n/a |
| S50 (healthy) | F | 66 | n/a | n/a |
| S49 (healthy) | M | 69 | n/a | n/a |

**Table 1: List of subjects with sex, age, Parkinson’s stage and the number of years since diagnosis.**

Table 2: : sample of Test Dataset

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| name | MDVP:Fo(Hz) | MDVP:Fhi(Hz) | MDVP:Flo(Hz) | MDVP:Shimmer | HNR | RPDE | DFA | spread2 | D2 | spread1 |
| phon\_R01\_S01\_1 | 119.992 | 157.302 | 74.997 | 0.04374 | 21.033 | 0.414783 | 0.815285 | 0.266482 | 2.301442 | -4.813031 |
| phon\_R01\_S01\_2 | 122.4 | 148.65 | 113.819 | 0.06134 | 19.085 | 0.458359 | 0.819521 | 0.33559 | 2.486855 | -4.075192 |
| phon\_R01\_S01\_3 | 116.682 | 131.111 | 111.555 | 0.05233 | 20.651 | 0.429895 | 0.825288 | 0.311173 | 2.342259 | -4.443179 |
| phon\_R01\_S01\_4 | 116.676 | 137.871 | 111.366 | 0.05492 | 20.644 | 0.434969 | 0.819235 | 0.334147 | 2.405554 | -4.117501 |
| phon\_R01\_S01\_5 | 116.014 | 141.781 | 110.655 | 0.06425 | 19.649 | 0.417356 | 0.823484 | 0.234513 | 2.33218 | -3.747787 |
| phon\_R01\_S01\_6 | 120.552 | 131.162 | 113.787 | 0.04701 | 21.378 | 0.415564 | 0.825069 | 0.299111 | 2.18756 | -4.242867 |
| phon\_R01\_S02\_1 | 120.267 | 137.244 | 114.82 | 0.01608 | 24.886 | 0.59604 | 0.764112 | 0.257682 | 1.854785 | -5.634322 |
| phon\_R01\_S02\_2 | 107.332 | 113.84 | 104.315 | 0.01567 | 26.892 | 0.63742 | 0.763262 | 0.183721 | 2.064693 | -6.167603 |
| phon\_R01\_S02\_3 | 95.73 | 132.068 | 91.754 | 0.02093 | 21.812 | 0.615551 | 0.773587 | 0.327769 | 2.322511 | -5.498678 |
| phon\_R01\_S02\_4 | 95.056 | 120.103 | 91.226 | 0.02838 | 21.862 | 0.547037 | 0.798463 | 0.325996 | 2.432792 | -5.011879 |
| phon\_R01\_S02\_5 | 88.333 | 112.24 | 84.072 | 0.02143 | 21.118 | 0.611137 | 0.776156 | 0.391002 | 2.407313 | -5.24977 |

**5.1.Methodology**

In our study, we have a Dataset of 32 people with 195 records, 23 of them with Parkinson disease. The status attribute shows if the person is an illness with 1 or not with 0. For each patient, we study 10 out of 24 attributes. The following steps are our work on the dataset we have:

1. We plotted a histogram for each attribute(column) and analyzed the distribution for the data.
2. We calculate the cross-correlation for the 24 attributes, then we calculated the correlation for each two attributes that had a high positive correlation with the target (0, 1). Let’s say, spread 1 and ppe their correlation is equal to 0.96 so it’s more than 0.65; we know that they are both correlated with each other. Now, we are correlating them with the target (status). Then, we will drop whichever has a weaker correlation with the target, in this case, is ppe (0.53).
3. We reduced the attributes to 10 according to our results, and this will be explained in detail in the final report.
4. We picked a sample of 150 rows for each attribute, then we calculated the SD, variance, mean, standard error and confidence interval of 95% for each one of them.
5. We choose two attributes and apply the hypothesis test on them with the target (status) and analyze the results depending on p-value and alpha.

**Add more information on the method of reducing the data, also put a picture of the excel sheet cross correlation. [QAMAR] - [28/03/2020]**

**will add all the graphs that are not line graphs.(histograms, box plot,Scatter plots) choose the best looking one [Haris] - [28/03/2020]**

* **comparing Stages of the disease [max and min] [QAMAR] - [28/03/2020]**

**6.Statistical analysis**

**6.1Descriptive Statistical analysis of features/parameters [STD,Variance, Mean, CI,Mode,median] make a table with 6 columns and 11 rows, rows=features columns=STD,Variance, Mean, CI,Mode,median [Yaoxin] -[28/03/2020]**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **95%Ci** | **Standard Error** | **Median** | **Standard Deviation** | **Variance** | **Mode** |
| **MDVP:Fo(Hz)** | **154.445691275168** | **6.47107490785483** | **3.27463248293943** | **151.884** | **39.9719835240883** | **1597.75946684999** |  |
| **MDVP:Fhi(Hz)** | **193.810758389262** | **13.3939327308719** | **6.77788587203334** | **179.139** | **82.7346408540707** | **6845.02079725206** |  |
| **MDVP:Flo(Hz)** | **115.653711409396** | **7.05661618242746** | **3.57094067054669** | **104.773** | **43.5888858955136** | **1899.9909736121** |  |
| **MDVP:PPQ** | **0.00335646666666667** | **0.000395759379043313** | **0.00020028175090721** | **0.00264** | **0.00245294047256934** | **0.00000601691696196868** | **0.00182** |
| **Jitter:DDP** | **0.00951166666666667** | **0.00134055194919445** | **0.000678412454092098** | **0.00727** | **0.00830882173837478** | **0.0000690365186800894** | **0.00496** |
| **MDVP:Shimmer** | **0.0302877181208054** | **0.00297425608423009** | **0.00150509702401649** | **0.02378** | **0.0183720505307326** | **0.00033753224070379** | **0.01608** |
| **MDVP:Shimmer(dB)** | **0.0305638** | **0.0030041470912267** | **0.00152030721512689** | **0.024025** | **0.0186198846466629** | **0.000346700104255033** | **0.01608** |
| **HNR** | **22.0511744966443** | **0.735719873184028** | **0.372304791611685** | **22.244** | **4.54455914481218** | **20.653017820696** |  |
| **RPDE** | **0.497656798657718** | **0.0166189147456896** | **0.00840986062321398** | **0.495954** | **0.10265543141783** | **0.0105381375995809** |  |
| **Spread 1** | **-5.68876318791947** | **0.172502137423024** | **0.0872932411733086** | **-5.657899** | **1.06554980325965** | **1.13539638322667** |  |

**6.2Normality Plots [QAMAR] - [28/03/2020]**

**6.3hypothesis test [QAMAR] -[28/03/2020]**

**6.4 Regression model [Abdullah] - [28/03/2020]**

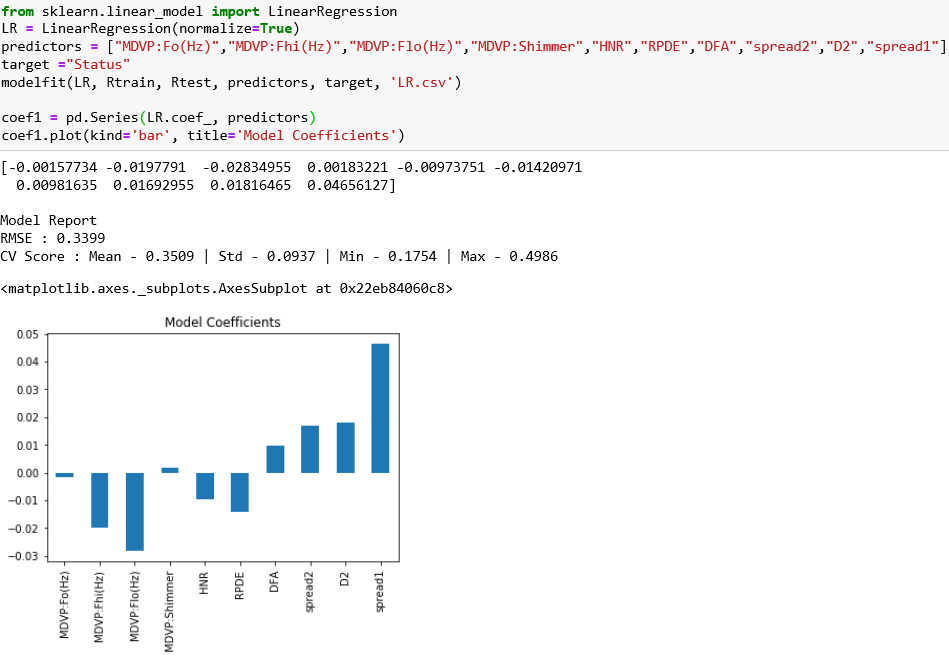
Regression analysis is a way of mathematically sorting out which variables does have an impact. it also investigates the relationship between two or more variables and estimates one variable based on the others.It answers the questions: Which factors matter most?

In this study the target variable that we are trying to predict is the “Status” and the predictors that we are going to use are the 10 selected features shown above.

In the below image you see an implementation of a regression model, it has been implemented through python.

the model is linear regression

the linear equation coefficients are are both plotted in bar chart and shown as an array (the red arrows)



**Calculate the accuracy of the model yields 94% accuracy in its ability to predict parkinson's**

**y\_pred=model.predict(x\_test)**

**print(accuracy\_score(y\_test, y\_pred)\*100)**

**94.87179487179486**

**6.5 Anova**

**7.Conclusions or summery**

**[Sourav & Yaoxin]**

**8.References**

[1] Little, M.A., McSharry, P.E., Roberts, S.J. et al. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. BioMed Eng OnLine 6, 23 (2007). https://doi.org/10.1186/1475-925X-6-23

[2] A. E. Lang and A. M. Lozano, "Parkinson's disease - First of two parts," New England Journal of Medicine, vol. 339, pp. 1044-1053, 1998.

[3] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," Behavioural Neurology, vol. 11, pp. 131-137, 1998.

[4] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and Co-Occurrence of Vocal-Tract Dysfunctions in Speech of a Large Sample of Parkinson Patients," Journal of Speech and Hearing Disorders, vol. 43, pp. 47-57, 1978.

[5] J. R. Duffy, Motor speech disorders : substrates, differential diagnosis, and management, 2nd ed. St. Louis, Mo.: Elsevier Mosby, 2005.

[6] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of Intensive Voice Treatment (the Lee Silverman Voice Treatment [LSVT]) on Vowel Articulation in Dysarthric Individuals With Idiopathic Parkinson Disease: Acoustic and Perceptual Findings," J Speech Lang Hear Res, vol. 50, pp. 899-912, 2007.

[7] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in Parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis," Journal of Voice, vol. 21, pp. 64-71, 2007.

[8] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," in IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015-1022, April 2009.