HIT 140   
FOUNDATIONS OF DATA SCIENCE  
Group Number 106

Shishiranjan Thakur (371465)

Rishika Shrestha (371778)

Sheetu Prakash Nakarmi (369887)

Ashma Bhandari (370896)

# Project Objective 1

(NIH National Institute on Aging (NIA), 2022) Parkinson’s Disease (PD) is a brain disorder that causes unintended or uncontrollable movements, such as shaking, stiffness, and difficulty with balance and coordination. It is the second most ubiquitous neurodegenerative disorder that affects approx. 1% of the population over 60 (Reeve, 2014). When the neurons die or become impaired, they produce less dopamine, which causes the movement problems associated with the disease. (Dellwo, 2023) Dopamine is a neurotransmitter, a brain chemical that transmits messages between neurons (brain and nerve cells). It plays an important role in mood regulation, movement, and how humans experience pleasure and pain. While there is no single definitive set of features that can perfectly distinguish people with PD from those who are healthy, there are several salient variables (features) that researchers and medical professionals commonly consider when diagnosing or studying Parkinson's disease. People with Parkinsonism (PWP) suffer from speech impairments like dysphonia (defective use of the voice), hypophonia (reduced volume), monotone (reduced pitch range), and dysarthia (difficulty with articulation of sounds or syllables).

**From the dataset po1\_data.txt:**

**General Overview:** Parkinson's disease (PD) is a neurological disorder caused by decreasing dopamine in the brain. Speech is one of the first functions that are disrupted. Accordingly, speech features are a promising indicator in PD diagnosis for telemedicine applications. The purpose of this study is to investigate the impact of Parkinson's disease on a minimal set of Jitter and Shimmer voice indicators and studying the difference between male and female speech features in noisy/noiseless environments. Variety of voice samples, including sustained vowels, words, and sentences compiled from a set of speaking exercises for people with Parkinson’s disease. There are two main issues in learning from such a dataset that consists of multiple speech recordings per subject:

1) How predictive these various types, e.g., sustained vowels versus words, of voice samples are in Parkinson’s disease (PD) diagnosis?

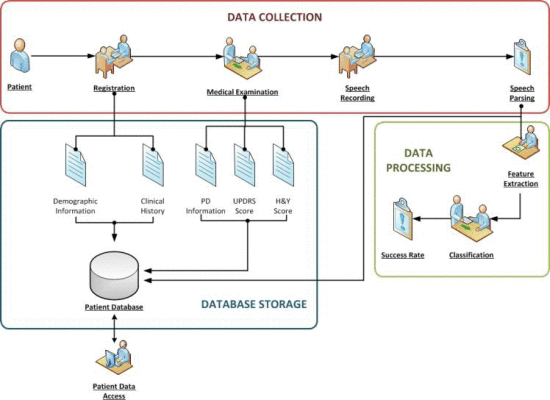
2) How well the central tendency and dispersion metrics serve as representatives of all sample recordings of a subject?

In this study, using speech data from subjects is expected to help the development of a noninvasive diagnostic. There are important examples of these kinds of Alzheimer and PD studies all around the world. The studies based on the PD focus on symptoms like slowness in movement, poor balance, trembling, or stiffness of some body parts but especially voice problems. The main reason behind the popularity of PD diagnosis from speech impairments is that telediagnosis and telemonitoring systems based on speech signals are low in cost and easy to self-use.

**Methodology**

**A. Overall Structure of PD Diagnosis System**

During this study, data are collected via a physician's examination process. The overall structure of the PD diagnosis system is shown in Figure 1. When the patient arrives at the hospital, his/her demographic information including gender, age, profession, educational status, and a brief health history including the chronic diseases, smoking rate, permanently used drugs, and symptoms of diseases are recorded. Demographic and clinical history information collected in the context of this study are not used in the PD-diagnosis system but only to design a computer-aided data storage system in the hospital. Then the patient is taken to a medical examination. During the examination, the physician asks the patient to read or repeat a predefined text including voice samples, write his/her name and draw some loops while conjointly the physician determines an UPDRS score. The speech of each patient is recorded during this process. Subsequently, the speech is parsed to be split into voice samples, and time-frequency based features are extracted from the voice samples using Praat (Weenink, 2022) acoustic analysis software.



*Figure 1: Overall Structure of the System* (Anon., n.d.)

**B. Feature Extraction**

There are 40 individuals in total, in which 20 of them have the Parkinson’s disease and the other 20 were healthy individuals, ranging between 36 and 85. Furthermore, A group of 26 linear and time-frequency based features (Table I) are extracted from each voice sample considering the previous works held on this field of study. The optimal feature for each category is studied separately for the men's and women's samples. Dysphonia is a well-known speech problem to diagnose and follow the condition of a people with Parkinson (PWP). Dysphonia leads to reduced loudness, breathiness, roughness, decreased, and exaggerated vocal tremor in voice. These indications can be detected by analyzing various frequencies in voice. In this context, during medical examinations, each subject is asked to read or say predetermined 26 voice samples containing numbers from 1 to 10, four rhymed sentences, nine words in Turkish language along with sustained vowels “a”, “o”, and “u.” To extract features from voice samples, Praat (Weenink, 2022) acoustic analysis software is used. These features include jitter, shimmer, harmonicity, pitch, and pulse. And to measure that feature, a rating scale was used called as the Unified Parkinson’s Disease Rating Scale (UPDRS).

The **jitter** variables measure the variation in the frequency of the sound. It is defined as the parameter of frequency variation from cycle to cycle, and shimmer relates to the amplitude variation of the sound wave. (João Paulo Teixeira\*, 2013) To determine jitter, which reflects the variation of the successive periods, and contains the peaks levels corresponding to the beginning of the glottal pulse signal, this means, this function returns a vector of the same size but only with the peaks. The methods used for determining the **Shimmer** are identical to jitter, the main difference is that the jitter considers periods and shimmer considers the maximum peak amplitude of the signal. The **shimmer** variables measure the variation in the amplitude of the sound.

*A typical jitter during sustained voiced sound in adults ranges between 0.50% and 1.00% and shimmer ranges between 0.05 dB and 0.22 dB*

The **harmonicity** variables are related to vocal quality and assess the noise in the sound. The **pitch** variables measure how high or low is the sound based on the frequency of vibration of the sound waves produced. The **pulse** variables measure the glottal pulse, which are variances in voice quality affected by manipulating the vocal cords when speaking. The **voice** variables measure the extent to which a subject has trouble maintaining vocal cord vibration when saying a sustained vowel. (Denison, 2023)

As given in dataset, we have 26 features extracted from each individual. Let’s first analyze each of the features for the **second** individual, which is further applicable for all of the remaining individuals. First one is the **Jitter** percentage (%), followed by absolute jitter, which is measured in microseconds, and then jitter which is measured in the form of relative amplitude perturbation (rap), the last one measures the average absolute difference of differences of periods (ddp) between jitter cycles.

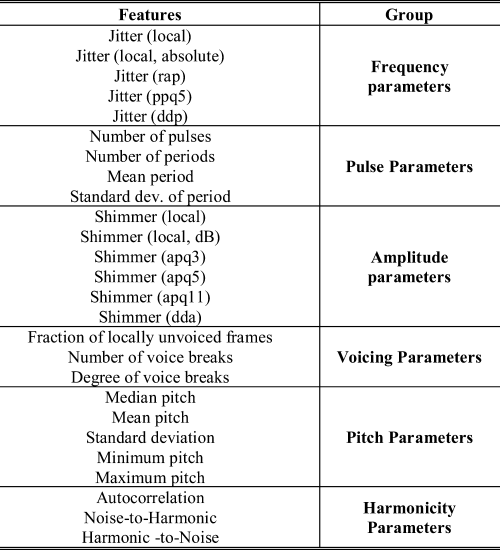
Talking about **Shimmer**, we measure this feature in percentage as well as in absolute shimmer in decibels (db). Next up, we measure shimmer as 3-point, 5-point, and 11-point amplitude perturbation quotient. Eventually, we also measure shimmer as an average absolute difference between consecutive differences between the amplitudes of shimmer cycles (d.d.a.). Similarly, the next feature is **Harmonicity,** in which we can measure Noise-to-Harmonic Ratio (NHR), and Harmonic-to-Noise Ratio (HNR). Finally, we need to find the autocorrelation between these two.

Next feature is **Pitch**, in which we can measure Mean, Median, Standard Deviation, Minimum and Maximum pitch of vocal sound. As mentioned above for pitch, we also need to measure Mean, Standard Deviation and Number of pulses and number of periods for **Pulse**. Lastly, we also measure how often the **Voice** of each individual breaks in terms of degree, and number. We also measure the fraction of unvoiced frames. For all these measurements mentioned above, we use the rating scale known as UPDRS. The PD indicator indicates 0 as healthy subject and 1 as the subject with PD.

(jagroopofficial, 2023) **Data Wrangling** is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging. It is a crucial topic for Data Science and Data Analysis. **Pandas** Framework of Python is used for Data Wrangling. Pandas is an open-source library in Python specifically developed for Data Analysis and Data Science. It is used for processes like data sorting or filtration, Data grouping, etc, and is imported in Python file as “pd”. Also, we have used **Seaborn**, which is a Python library created for enhanced data visualization and is imported in python file as “sns”. Additionally, we have used **scipy.stats**, which is the SciPy sub-package. It is mainly used for probabilistic distributions and statistical operations. There is a wide range of probability functions. It is imported as “st”. We have also imported “**Math**” library, where Python has a set of built-in math functions, including an extensive math module, that allows you to perform mathematical tasks on numbers.

(geeksforgeeks, 2023) **Inferential statistics** are used for finding inferences on the data and making predictions about the data on a given sample of data. This uses probability to find conclusions.

.



*Table 1:* *Time-Frequency-Based Features Extracted from Voice Samples* (Anon., n.d.)

**CONCLUSION**

**Sustained Vowels Hold Important Information:** When compared to single words and brief phrases, sustained vowels hold more information that is relevant to the discrimination of Parkinson's disease (PD).

**Central Tendency and Dispersion Metrics Analysis:** The study looked at different combinations of central tendency (mean, median, trimmed mean) and dispersion metrics (standard deviation, interquartile range, mean absolute deviation) to see how well they represented the recordings of the subjects.

**Superiority of Mean and Standard Deviation: Among** the analyzed metrics, the capacity of the predictive model to generalize is improved by using the classical mean and standard deviation as representations of a subject's vocal samples.

**Summarizing Representation Approach:** The use of the mean and standard deviation as combining representations for each participant's many recordings seems to be more successful than considering each individual voice recording as a single data sample.

**Efficiency of Prediction Models:** When many recordings from each participant are available for study, using predictive models based on the mean and standard deviation of voice parameters has been shown to be an efficient method for enhancing prediction accuracy.

# References

Dellwo, A., 2023. *What is Dopamine?.* [Online]   
Available at: https://www.verywellhealth.com/dopamine-5086831

Denison, J., 2023. *S223 HIT140 FOUNDATIONS OF DATA SCIENCE.* [Online]   
Available at: https://online.cdu.edu.au/ultra/courses/\_57259\_1/outline/edit/document/\_4961050\_1?courseId=\_57259\_1&view=content

Freepik Company S.L. , 2010-2023. *Freepik.* [Online]   
Available at: https://www.freepik.com/free-vector/parkinson-disease-symptoms-infographic\_9740796.htm

jagroopofficial, 2023. *Data Wrangling in Python.* [Online]   
Available at: https://www.geeksforgeeks.org/data-wrangling-in-python/

João Paulo Teixeira\*, C. O. C. L., 2013. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology 9(5):1112-1122.*

NIH National Institute on Aging (NIA), 2022. *Parkinson’s Disease: Causes, Symptoms, and Treatments.* [Online]   
Available at: https://www.nia.nih.gov/health/parkinsons-disease

Reeve, A. S. E. T. D., 2014. Ageing and Parkinsons Disease: why is advancing age the biggest risk factor?. *Ageing Res. Rev.,* Volume 14, pp. 19-30.

Weenink, P. B. a. D., 2022. *Praat: doing phonetics by computer.* [Online]   
Available at: https://www.fon.hum.uva.nl/praat/