Analysis of datasets from healthy controls and patients with Parkison's disease

For the following tasks, two distinct datasets, one comprising Czech-speaking participants and the other Spanish-speaking participants, are to be analyzed. For each dataset speech features were extracted from two distinct speech tasks: sustained phonation of the vowel 'a' and a syllable repetition task involving the word 'pataka'. Moreover, demographic and clinical information is available for all participants across both datasets (for further information on the clinical specifics and previous work on both datasets see Orozco et al. (2014) and Rusz et al. (2013)).
The sustained phonation task was performed multiple times in both data sets while the pataka task was just repeated in the Czech dataset. Repetitions of the task are performed to improve validity of the measure, as patients' motivation and performance on those tasks tend to fluctuate. Hence, when analyzing the data, it might be useful to either aggregate the samples or just use the first or the last repetition.

Analytical questions:

1. Which speech features are the best to separate healthy controls from Parkinson's patients?
2. Which speech features are the most different between the two datasets?
3. Do age and gender variations exist within the healthy control group or among the Parkinson's patients?

Open question:

1. What would be your approach to account for age and gender differences when analyzing the speech features?

Please write some Python code to answer the analytical questions. Describe your findings in a short text together with the open question. Please send your results together with the code to Louisa.

References

Orozco, J.R.; Arias-Londoño, J.D.; Vargas-Bonilla, J.; González-Rátiva, M.; Noeth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinsons disease. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 26–31 May 2014; pp. 342–347.

Rusz, J.; Cmejla, R.; Tykalova, T.; Ruzickova, H.; Klempir, J.; Majerova, V.; Picmausova, J.; Roth, J.; Ruzicka, E. Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. J. Acoust. Soc. Am. 2013, 134, 2171–2181.