2012 IEEE 27-th Convention of Electrical and Electronics Engineers in Israel

Early Diagnosis of Parkinson's Disease via Machine Learning on Speech Data

Hananel Hazan
Department of Computer Science
University of Haifa, Israel
hananel.hazan@gmail.com

Dan Hilu
Department of Computer Science
University of Haifa, Israel
danram4@zahav.net.il

Larry Manevitz
Department of Computer Science
University of Haifa, Israel
The Rotman Research Institute at Baycrest
Toronto, Canada
manevitz@cs.haifa.ac.il

Lorraine O. Ramig
University of Colorado at Boulder and National Center for Voice and Speech
Boulder, CO, USA
Lorraine.Ramig@colorado.edu

Shimon Sapir
Department of Communication Sciences and Disorders, University of Haifa, Israel
sapir@research.haifa.ac.il

***Abstract*— Using two distinct data sets (from the USA and Germany) of healthy controls and patients with early or mild stages of Parkinson's disease, we show that machine learning tools can be used for the early diagnosis of Parkinson's disease from speech data. This could potentially be applicable before physical symptoms appear. In addition, we show that while the training phase of machine learning process from one country can be reused in the other; different features dominate in each country; presumably because of languages differences. Three results are presented: (i) separate training and testing by each country (close to 85% range); (ii) pooled training and testing (about 80% range) and (iii) cross-country (training in one and testing in the other) (about 75% ranges). We discovered that different feature sets were needed for each country (language).**

***Index Terms*- Parkinson Disease, Early Diagnosis, Classification, Machine Learning, Speech Data, Pattern Matching, SVM.**

I. INTRODUCTION

Parkinson's disease (PD) is a slowly progressing and highly debilitating disease, affecting 1.0-1.5% of individuals 60 years old and older. The disease also affects young people in their 30-50s. By the time the disease is diagnosed, some 60% of nigrostriatal neurons are degenerated, and 80% of striatal dopamine is depleted [1], [2].

Thus, there is an urgent need to find biomarkers that can help detect the disease in its early stages, as well as to monitor its progression and severity. The rationale is that the earlier the disease is diagnosed and treated, the less damage will be accrued.

It has been suggested that speech abnormalities might be present at early stages of the disease, before the classical symptoms of the disease appear (muscle rigidity, rest tremor, bradykinesia or hypokinesia, and postural instability), and that by acoustic and classification analyses the individuals at risk for PD could be identified from the population at large [3–6]

Recently, researchers have developed acoustic metrics of vowel articulation that are highly sensitive to changes that occur in the orofacial muscles [7]. The acoustic metrics include the first (F1) and second (F2) formants of the corner vowels /i/, /u/, and /a/ , and various ratios of these vowel formants. The reason for using such acoustic analysis is that the F1 and F2 of these vowels reflect the movements of the tongue, lips, and jaw [5], [7].

Specifically, F2 increases and F1 decreases in formant frequency as the tongue moves forward (as in the case of the vowel /i/), and, respectively, F2 decreases and F1 increases in frequency when the tongue moves backwards (as in the vowel /u/). Also, F1 formant frequency decreases when the tongue goes up (e.g., for the vowels /i/ and /u/) and increases as the tongue goes down, along with the jaw (e.g., for the vowel /a/). Furthermore, Both F1 and F2 decrease when the lips are rounded (e.g., for the vowel /u/) and increase when the lips are unrounded or detracted (as in the vowels /i/ and /a/).

In PD the movements of the speech articulators (lips, tongue, jaw) are restricted in range (hypokinetic), and as a result the vowels become centralized, i.e., formants that normally have high frequency tend to have lower frequency, and formants that normally have low frequency tend to have higher frequencies.

Sapir and colleagues [5], [7] developed 3 acoustic metrics that characterize vowel centralization. These are the Formant Centralization Ratio see Eq. (1), [5], its inverse, the Vowel Articulation Index is Eq. (2), [8] and the Eq.(3) ratio [5], [7].

$$FCR = (F2u+F2a+F1i+F1u)/(F2i+F1a) \quad (1)$$
$$VAI = (F2i+F1a)/(F2u+F2a+F1i+F1u) \quad (2)$$
$$F2i/F2u \quad (3)$$

Acoustic analyses with the FCR, VAI, and F2i/F2u (see Eq. (1), Eq. (2), Eq. (3)) have been shown to differentiate between PD and Healthy controls (HC) groups and monitor changes (deterioration or improvement) in speech in studies in the USA [5], [7] and in Germany [6]. The study in 2007 [7] in the USA was done on 29 individuals with PD and 14 HCs and in 2010 study [5] in the USA was done on 38 individuals with PD and 14 HCs. The 2011 study [2] in Germany was done on 98 individuals (34 men & 34 women with PD and 15 men & 15 women who served as HCs). The studies in the USA showed statistically significant differences between the PD and HC groups ([7] study:, Eq. (3), p=0.0015, with a large effect size, ES=1.345; 2010 study: Eq. (1), p=0.0006, Eq. (3), p=0.0002, with large effect sizes >0.95). The study in Germany showed significant differences (p<0.0002, ES>1.11) between the PD and HCs. Note that the present work relied on a new composite from the raw data (i.e. not from the acoustic metrics developed by Sapir and his colleagues).

In this work we pursue a slightly different direction. We use machine learning on the gathered data, to classify the subjects as Parkinson's or not. This does not use human judgment on the metric; although there was, of course, human judgment as to which basic features to record. The rationale was that the machine learning should be able to discover the appropriate combination of features by itself. We applied this to the data of the two studies mentioned above (33 individuals from the USA – 8 men and 8 women with PD and 7 men and 7 women who served as HCs – and the 98 individuals from Germany). Using all the features in each study we reached satisfactory results (~88% accuracy for the American data set and ~80% for the German data set).

II. Methods

We used the Support Vectors Machine (SVM) with a radial basis function (RBF) kernel as the machine learning tool for the classification. (See [9] for basic information on SVM.)

The SVM is a tool that does a 2-class classification; that is by receiving labeled (training) data of two classes it finds an optimal hyper plane that divides the two classes. Once this has been determined, new data points are classified depending on which side of the hyper-plane they lay.

The data used by us was the data extracted from the recordings done on two distinct data sets (the American data set and the German data set as described above). Each set had its points labeled as to whether they were in PD or healthy.

The American data set's data file contained the following features: the health condition (PD vs. HC), gender, age, F1a, F1u, F1i, F2a, F2u, F2i of each person who was recorded and the German data set's data file contained the features: health condition, gender, F1a, F1u, F1i, F2a, F2u, F2i of each person who was recorded.

We did the classification under three versions of representation of the data:

I. The raw numerical data values.

II. Unit normalization – the sum of the data values would equal 1.

III. Log representation – we applied the ln function on each of the values. Such logarithmic transformation has been used to reduce inter-speaker variability that is related to anatomical differences, thus reducing noise, statistically speaking).)

Since, from a machine learning perspective, the total number of features was rather small, we decided to do an exhaustive search for the best subset of features. It is well established in machine learning that feature selection can have a big effect on the quality of the results. (See e.g. [10]). Note that a subset of features can do a better classification due to the fact that not all the features are essential for the classification process. (Some features can harm the classification process since they just add noise to the system.)

The search was done by rerunning the SVM classification algorithm on each set separately.

The first (Country-wise) was to examine each country's data set on its own; that is search for the best feature set that maximizes the generalization results on the American data set using the American data to train and to test' similarly for the German data set.

We produced generalization results by using cross validation using the “leave one out” method. (That is, for each combination of features we reran the SVM leaving one individual out in the training; and then seeing if that individual was classified correctly. We then report the statistics over these runs.)

Using this method we calculated the accuracy of each of the features combinations.

Our second approach (Cross-Country) used both of the data sets, one data set is used for the SVM's training process and the second data set is used for testing the accuracy. (Because the American data set contains more features we used only the German data set's features for this test). First, we trained the SVM using the data of the American data set. Then, we checked how many data points of the German data set were identified correctly. Secondly, we trained the SVM with the data of the German data set. Then, we checked how many data points of the American data set were identified correctly.

Interestingly, the features selected for the cross country were typically quite poor for the training country. We discuss this further below.

Our third approach (Pooled data sets) also uses the two data sets, but this time we combine the two data sets into a one large data set. We gave the SVM this data set to train on and then checked the accuracy.

In order to check how accurate the SVM separation on the large data set is we used cross validation using the "leave two out" method. Each iteration we took out one data point from each data set, one from the American data set and one from the German data set. (i.e. each time the SVM trained on all of the data points both of the sets have, all but two). After the training process was over we saw whether both of the data points were classified correctly.

III. Results

After doing the exhaustive search for the best combination of features we received the accuracy for each of the combinations for each of the three approaches.

A. Investigation 1*: (Separate Data Sets)*

For each of the three investigations, we checked which of the data representations was best. We also did exhaustive search over all choices of subsets of features. Since the two data sets do not have identical features (the American data set has some features that the German data set does not have) we used an exhaustive search for the best combination of features on the German data set features. (The German features - health condition, gender, F1a, F1u, F1i, F2a, F2u, F2i - can be found in both of the data sets.)

The best result for the German data set occurred using three features with the “log representation” and the best results for the American data set is using five features with "log representation". In Table I, we summarize the best results by number of features used under all representations.

Table I *The generalization results under all representations by number of features for the case of training and testing in the separate data-sets. Each row reports the best results for that number of features*

# of features for training	basic data		unit normalization		Log representation	
	American	German	American	German	American	German
1	72.73%	69.39%	69.70%	71.43%	69.70%	75.51%
2	72.73%	71.43%	75.76%	78.57%	75.76%	75.51%
3	72.73%	73.47%	81.82%	84.69%	81.82%	72.45%
4	69.70%	76.53%	84.85%	81.63%	87.88%	74.49%
5	84.85%	76.53%	90.91%	80.61%	93.94%	75.51%
6	84.85%	73.47%	90.91%	81.63%	90.91%	71.43%
7	84.85%	70.41%	84.85%	78.57%	87.88%	66.33%

The best accuracy for the American data set is 94% when using the features: gender, age, F2i, F2a, F1i.

The best accuracy for the German data set is 85% when using the features: F2i, F2a, F2u.

For interest and comparison we ran the SVM with the 3 acoustic metrics mentioned above, the SVM accuracy results were 58% when only using FCR, 69.7% when only using VAI and 73% when only using the F2i/F2u ratio.

B. Investigation 2: (Cross-country)

As can be seen in *Table II*, the best accuracy is about 75% in each direction. (This time, it is best to use log representation – it gives the best accuracy), Thus we see a penalty in cross-country training. This might be expected since there is a language difference under-lying the data. Nevertheless, it is interesting that we are still able to classify in the 75% range.

Table II *The results as in Table I, but for the case where training was done in one country and testing in the other country. The results are listed under the testing country.*

# of features for training	basic data		unit normalization		Log representation	
	American	German	American	German	American	German
1	69.39%	72.73%	69.39%	57.58%	74.49%	69.70%
2	70.41%	66.67%	69.39%	57.58%	75.51%	72.73%
3	72.45%	60.61%	69%	63.64%	74.49%	75.76%
4	72.45%	66.67%	69%	69.70%	74.49%	75.76%
5	72.45%	66.67%	54.08%	69.70%	69.39%	72.73%
6	71.43%	60.61%	37.76%	57.58%	68.37%	69.70%
7	69.39%	54.55%	30.61%	57.58%	66.33%	57.58%

Interestingly, and perhaps counter-intuitively, the features trained on Americans used on the Germans do not give good results on the Americans. The American data set's features: F1u, F2i (who give us ~75% in the German data give us only ~58% accuracy ratio in the American data set. Similarly, the German data set's features: F1i, F1u, F2i, F2u trained on the German data (who give us ~75% on the American data in the cross-country) give us ~71% accuracy ratio on the German data. These results are of course lower than the 94% and 85% accuracy ration we got in the first approach. We hypothesize that this is due to the language difference.

C. Investigation 3: (Pooled Data Sets)

In this approach we put all the data together, and trained on both and tested on both. The testing used a "leave two-out" approach where one point from the American side and one point from the German side were tested in each run and results over all the runs were reported. In Table III we report the generalization for each country under this pooled training.

The best result was 84% accuracy ratio for identifying the American data points and 76% accuracy ratio for identifying the German data points.

Table III *The results for the pooled data sets broken up by country. (The total results of the pooled data set are the average over the two countries.*

# of features for training	basic data		unit normalization		Log representation	
	American	German	American	German	American	German
1	72.70%	69.39%	57.58%	69.39%	70.69%	72.48%
2	66.57%	71.43%	68.27%	69.39%	70.13%	76.16%
3	72.67%	74.77%	72.73%	73.16%	81.32%	75.45%
4	72.67%	75.26%	78.42%	73.13%	84.38%	79.25%
5	69.29%	75.26%	78.60%	74.92%	81.01%	78.48%
6	69.29%	74.37%	84.38%	76.31%	78.70%	71.49%
7	57.64%	72.45%	81.05%	75.39%	75.26%	70.32%

IV. Conclusions

This study shows: (i) early detection of PD from speech data seems to be feasible and accurate with results approaching the 90% mark in two different data sets. (ii) the optimal features seem to be data set dependent which we interpret as language dependent (iii) notwithstanding the previous point, one can successfully use machine learning to train and test different data sets together (iv) not withstanding point ii, one can even use machine learning to train on one country and test in another (v) however, the set of features needed seems to be language (and or text) dependent.

As a last point, in this study we used only vowel formants to classify the PD and HC subjects. There are other acoustic metrics, such as those that measure vocal intensity and vocal decay [11], speech prosody (e.g., inflection in the voice fundamental frequency and/or intensity reflecting emotional or linguistic information) [4], [12], [13], and voice quality [3] that may help in the classification of those individuals with early PD and those who do not have PD. We also point out that Also, the number of subjects in the present study was rather small and the subjects were a combination of early and mild PD. Finally, the acoustic measures in the two languages were based on different text, which may have also affected the differences across countries. Thus, it would be worthwhile to do the acoustic and classification analyses based on a large number of subjects and similar texts (phonetic features) in order to assess their predictive power.

Acknowledgment

We thank the Caesarea Rothschild Institute of the University of Haifa for its support.

Larry Manevitz thanks the Bernard Katz foundation for its support during a visit to the Rotman Institute at Baycrest in Toronto.

Also, this research was partially funded by the R01 NIH DC1150 grant from the USA National Institutes of Deafness and Other Communication Disorders.

References

[1] S. Sapir, L. Ramig, and C. Fox, “Speech and swallowing disorders in Parkinson disease,” *Curr Opin Otolaryngol Head Neck Surg*, vol. 16, no. 3, pp. 205–210, Jun. 2008.

[2] S. Sapir, L. O. Ramig, and C. M. Fox, “Intensive voice treatment in Parkinson’s disease: Lee Silverman Voice Treatment,” *Expert Rev Neurother*, vol. 11, no. 6, pp. 815–830, Jun. 2011.

[3] 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,” *Nature Precedings*, no. 713, Sep. 2008.

[4] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, “Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s disease,” *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 350–367, Jan. 2011.

[5] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, “Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech,” *J Speech Lang Hear Res*, vol. 53, no. 1, p. 114, Feb. 2010.

[6] S. Skodda, W. Visser, and U. Schlegel, “Vowel articulation in Parkinson’s disease,” *J Voice*, vol. 25, no. 4, pp. 467–472, Jul. 2011.

[7] 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, no. 4, pp. 899–912, Aug. 2007.

[8] N. Roy, S. L. Nissen, C. Dromey, and S. Sapir, “Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy,” *J Commun Disord*, vol. 42, no. 2, pp. 124–135, Apr. 2009.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] O. Boehm, D. Hardoon, and L. Manevitz, “Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms,” *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 3, pp. 125–134, Sep. 2011.

[11] K. M. Rosen, R. D. Kent, and J. R. Duffy, “Task-based profile of vocal intensity decline in Parkinson’s disease,” *Folia Phoniatr Logop*, vol. 57, no. 1, pp. 28–37, Feb. 2005.

[12] B. Harel, M. Cannizzaro, and P. J. Snyder, “Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: a longitudinal case study,” *Brain Cogn*, vol. 56, no. 1, pp. 24–29, Oct. 2004.

[13] J. Möbes, G. Joppich, F. Stiebritz, R. Dengler, and C. Schröder, “Emotional speech in Parkinson’s disease,” *Mov. Disord.*, vol. 23, no. 6, pp. 824–829, Apr. 2008.

.