Computational Diagnosis of Parkinson's Disease Directly from Natural Speech using Machine Learning Techniques

A. Frid[1], H. Hazan[2], D. Hilu[3], L.M. Manevitz[4], L. Ramig[5] and S. Shapir[6]

Abstract—The human voice signal carries much information in addition to direct linguistic semantic information. This information can be perceived by computational systems. In this work, we show that early diagnosis of Parkinson's disease is possible solely from the voice signal. This is in contrast to earlier work in which we showed that this can be done using hand-calculated features of the speech (such as formants) as annotated by professional speech therapists. In this paper, we review that work and show that a differential diagnosis can be produced directly from the analog speech signal itself. In addition, differentiation can be made between seven different degrees of progression of the disease (including healthy). Such a system can act as an additional stage (or another building block) in a bigger system of natural speech processing. For example it could be used in automatic speech recognition systems that are used as personal assistants (such as Iphones' Siri, Google Voice), or as natural man-machine interfaces. We also conjecture that such systems can be extending to monitoring and classifying additional neurological diseases and speech pathologies. The methods presented here use a combination of signal processing features and machine learning techniques.

Index Terms— Parkinson's disease, Natural Speech Analysis, Classification, Support Vector Machine (SVM), Machine Learning

1 INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder characterized by "masklike" facial features, bradykinesia (i.e. slowness of movement), akinesia, tremor at rest, and muscle rigidity (i.e. resistance to externally imposed movement). PD affects approximately 1% of individuals over the age of 65. The number of patients suffering from PD is increasing; partly in association with increased life expectancy [1]. Typically, by the time the disease is diagnosed, some 60% of nigrostriatal neurons have degenerated, and 80% of striatal dopamine are depleted [2], [3].

Thus, the development of new treatments in this field depends on two main things: (i) early diagnosis of the disease and (ii) correct and constant evaluation of the effectivity of the treatment. Currently, this evaluation is currently usually performed using the Unified Parkinson's Disease Rating Scale (UPDRS) [4]. This tool is based on a score derived from the neurological evaluation that is performed by the physician and thus it is a subjective score which leads to a lack in objectivity and sensitivity of the scale.

It has been suggested that speech is affected very early on [5]–[8] and continues to deteriorate as the disease progresses. Therefore the speech signal is a natural candidate for measuring and quantifying the progress of the disease.

In earlier work, [9] we showed this was possible in a semi-automatic manner. We showed that using specific speech indicators (formants), machine learning techniques could in fact reliably diagnose the disease. This is "semi-automatic" because (i) the choice of these features were selected by speech pathologists and (ii) these specialists also graded these features manually from speech samples.

Specifically, the features chosen were acoustic metrics of vowel articulation that are highly sensitive to changes that occur in the orofacial muscles [10]. 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 [7], [10].

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.

Accordingly, Sapir and colleagues [7], [10] developed 3 acoustic metrics that characterize vowel centralization. These are the Formant Centralization Ratio [7], its inverse, the Vowel Articulation Index [11] and the /i/- /u/ F2 ratio [10], [7].

1. *Edmond J. Safra Brain Research Center for the Study of Learning Disabilities, University of Haifa, Israel Email: alex.frid@gmail.com*
2. *Lorrey L. Lokey Network Biology Laboratories, Technion - Israel Institute of Technology, Haifa, Israel. Email: hananel@hazan.org.il*
3. *Department of Computer Science, University of Haifa, Israel. Email: dan.hilu@hotmail.com*
4. *Department of Computer Science, University of Haifa, Israel. Email: manevitz@cs.haifa.ac.il*
5. *University of Colorado at Boulder and National Center for Voice and Speech Boulder, CO, USA. Email: Lorraine.Ramig@colorado.edu*
6. *Department of Communication Sciences and Disorders, University of Haifa, Israel. Email: sapir@research.haifa.ac.il*

Machine learning techniques then used these basic features in [9] and showed that diagnosis between healthy and diseased could be done at high degrees of accuracy (over 90%).

Here we proceed to deepen and extend this work in two directions: (1) we show that the information can be extracted directly from the voice signal without human intervention or analysis on the data. That is, without any measurements of the formants. (2) The system can also successfully distinguish between degrees of severity of PD.

The features chosen are standard auditory signal processing features that are typically used in speech analysis tasks such as speech-to-text [12]–[14] and speaker recognition [15]. This means that this system can be easily embedded in existing perceptual computing control devices.

2 Systems and Methods

Our algorithm consists of the following steps that will be explained in detail below: speech data acquisition, windowing, preprocessing, feature extraction, classification and an optional decision scheme. The whole process is depicted in Figure 22.

A. Data Acquisition: For the data acquisition step, the patients were tested by the physician, and evaluated by the Unified Parkinson's Disease Rating Scale (UPDRS) [4] using the 0-5 grades with 0.5 resolution. Then, the patients were directed to read the "Rainbow" passage described in Figure 1, in natural manner while being recorded. The "Rainbow Passage" is a public domain text, and a phonetically or phonemically balanced passage [16], commonly used both in research and in clinical settings for speech language pathologies tests and treatments [17].

B. *Windowing*: Then a windowing process was applied. The signal was divided into consecutive windows of 20ms length with 50% overlaps. This setup is considered to be suitable for real time analysis applications of speech, while, on the other hand it contains sufficient data on the phonetic level [18].

C. *Preprocessing*: During the preprocessing stage, the signal is offset to be set to a mean of zero (i.e. "removing the DC"), the amplitude of the signal in the windows are then normalized and this is then filtered to remove frequencies not in the speech range.

D. *Feature Selection and Extraction*: Special thought was given to the selection of features. On the one hand, those features need to evaluate the degree of anomalous fluctuations in speech, but on the other hand, we wished to use standard features that are usually used in applications of speech analysis and recognition systems. Besides making the extraction process simpler, this also allows the proposed scheme to be easily integrated in existing systems. The ones chosen were as follows:

a. *Pitch value and its power*: This feature represents the vibration rate of audio signals, which can be represented by the fundamental frequency and multiples thereof. The average pitch frequency time pattern, gain, and fluctuations change from one individual speaker to another. The values were calculated using an auto-correlation algorithm similar to that described in [19].

b. *Short-time Energy*: The short-time energy (E_n) of speech signals reflects the amplitude variation, and is defined by the following equation:

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m) \qquad (1)$$

where $h(n)$ is chosen to be a hamming window. In voiced (periodic) speech the short time energy values are much higher than during the unvoiced speech.

c. Zero Crossing Rate (ZCR): The zero-crossing rate of a short time window defined as a number of times the audio waveform changes its sign in the duration of the frame:

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1}\left|\operatorname{sgn}\left(x[n]\right) - \operatorname{sgn}\left(x[n-1]\right)\right| \qquad (2)$$

where $x(n)$ is the time domain signal for window t. This feature can indicate regarding the amount of noise in the speech signal, i.e. the periodicity of the signal.

d. *Mean and Standard Deviation values of Zero Crossing Rates*: These values are computed using the statistics of the time intervals between consecutive zero crossings. Together with the ZCR feature these values can indicate about the speech abnormalities or 'noisiness' in different levels during the production of voiced and unvoiced sounds.

e. *Mel Frequency Cepstral Coefficients (MFCC)*: After computing the logarithm of the magnitude spectrum (computed by the Short Time Fourier Transform), and grouping the Discrete Fourier Transform (DFT) bins according to a Mel frequency scale (a logarithmic scale which approximates the response of the human auditory system), a discrete cosine transform is performed on the result. The first three coefficients (out of 26 coefficients) were used in this work.

E. *Classification*: For the classification stage, a Support Vector Machine (SVM) implementation [20] of the algorithm presented in [21] was used with C-SVC (Support Vector Classification) kernel type and Radial Basis kernel function (a.k.a Gaussian Kernel). SVM was chosen because it is known to perform well on generalization even with small amounts of data. Grid search methodology with a polynomial scale was applied in order to determine the free parameters ('*C*' parameter of the SVC and '*gamma*' parameter of the RBF kernel function). \ Twelve percent of data windows were randomly chosen without repetition for the training procedure each time and the selected data was oversampled in order to avoid learning on imbalanced data. After the best SVM parameters were found a cross validation procedure was applied in order to report on the results.

F. *Decision scheme*: The learning procedure implemented here treats all windows equivalently without regard to their temporal origin. We expect that further analysis using the temporal arrangement could be averaged in some sort of hierarchal fashion (a simple mechanism could be majority voting, but more complex methodologies are also possible).

"When the sunlight strikes rain drops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long, round arch, with its path high above and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but never finds it. When a man looks for something beyond its reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there will be no more universal floods. "

Figure 1. The "Rainbow passage" read by the patients.

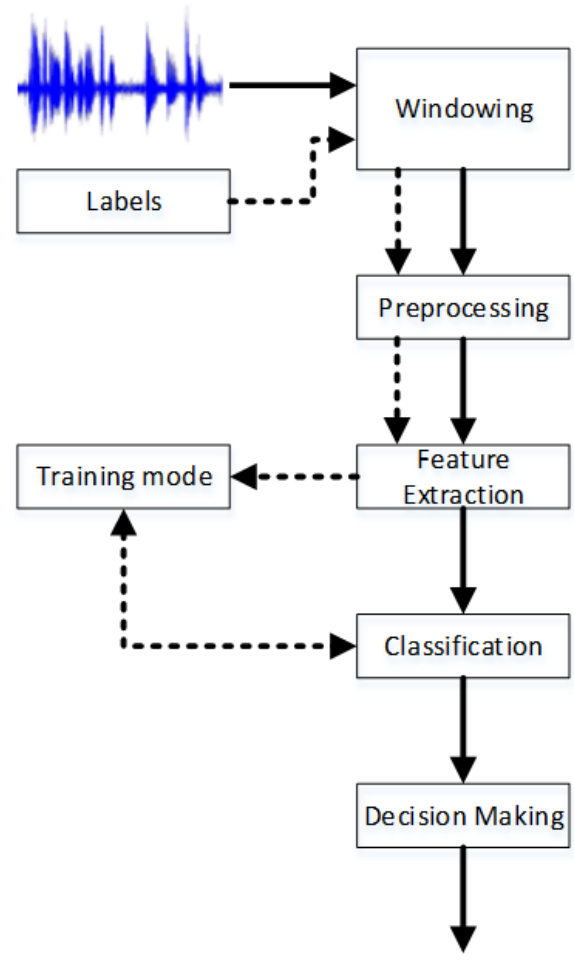

Figure 2. Flow Diagram of the proposed scheme, starting from the signal acquisition through the preprocessing, to the feature extraction, training and classification modes followed by the decision scheme. The full arrows indicate the trained system path and the dashed arrows indicate the machine learning training process.

3 Results

For the system evaluation a total of 43 patients (4 from grade 1, 4 from grade 1.5, 18 from grade 2, 9 from grade 2.5, 7 from grade 3 and 1 from grade 4) and 9 controls were recorded reading the "Rainbow passage". We examined all possible classifications between UPDRS degrees of disease. That is we trained and tested classification on data consisting of two grades of disease severity and repeated this for all such choices The overall accuracy average of these results was 81.8%; the full results are summarized in Table 1. We also checked whether the erroneous results were of type 1 (false positive) or type 2 (false negative), in general these were balanced (refere to Table 1).

Table 1 – The binary classification results between each choise of two grades of disease severity. Correct classification percentage displayed first, within the parenteces false positive (Class 1) and false negative (Class 2) percentages are displayed accordingly.

		Class 1					
		1	1.5	2	2.5	3	4
Class 2	0	83.39 (17.47, 15.75)	83.3 (18.46, 14.93)	79.6 (19.77, 21.01)	78.51 (22.54, 20.43)	86.3 (10.57, 17.45)	76.18 (32, 15.63)
	1		78.41 (18.2, 24.97)	79.08 (21.92, 19.92)	81.7 (17.71, 18.89)	85.14 (15.13, 14.58)	85.57 (16.05, 12.80)
	1.5			83.39 (16.93, 16.29)	84.03 (15.04, 16.89)	74.45 (22.78, 28.30)	86.1 (15.42, 12.38)
	2				76.7 (22.35, 24.24)	83.37 (15.64, 17.6)	81.11 (24.76, 13.01)
	2.5					85.81 (13.51, 14.86)	79.49 (23.92, 17.1)
	3						86.57 (17.52, 9.33)

4 Summary and Discussion:

An automatic system for quantification and classification of Parkinson's disease directly from natural speech was developed using the techniques of Machine Learning. This system did not require any human intervention in the analysis.

Besides the application itself, this method shows much promise for general machine perception of human conditions. Interestingly, this method shows that the deep human expertise in choice, selection and combination of speech signals can replaced by an automatic process on the auditory signal itself. While the features from the signal were, in fact, pre-chosen by the state of the art practice in general speech signal processing, it would be interesting in future work, to see if those features or replacements could be automatically discovered.

We also expect that further pre-processing (e.g. removing "silent windows" or adding more global hierarchial windows) should be useful.

As a further point, the general set-up of this work seems appropriate for application to a wide range of neurological diseases and states such as dementias, strokes and speech pathologies. Of course, one can forsee using such modules eventually in telemedicine systems.

References

[1] C. B. Levine, K. R. Fahrbach, A. D. Siderowf, R. P. Estok, V. M. Ludensky, and S. D. Ross, "Diagnosis and treatment of Parkinson's disease: a systematic review of the literature," *Evid. Rep. Technol. Assess. (Summ.)*, no. 57, pp. 1–4, May 2003.

[2] 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.

[3] 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.

[4] S. Fahn, R. Elton, and UPDRS Development Committee, "Unified Parkinson's disease rating scale," *Recent Dev. Park. Dis.*, vol. 2, pp. 153–163, 1987.

[5] 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," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, 2009.

[6] 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.

[7] 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. JSLHR*, vol. 53, no. 1, p. 114, Feb. 2010.

[8] S. Skodda, W. Visser, and U. Schlegel, "Vowel Articulation in Parkinson's Disease," *J. Voice*, vol. 25, no. 4, pp. 467–472, Jul. 2011.

[9] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Early diagnosis of Parkinson's disease via machine learning on speech data," in *2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, 2012, pp. 1–4.

[10] 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. JSLHR*, vol. 50, no. 4, pp. 899–912, Aug. 2007.

[11] 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.

[12] P. Cosi, J.-P. Hosom, and A. Valente, "High performance telephone bandwidth speaker independent continuous digit recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01*, 2001, pp. 405–408.

[13] A. K. V. SaiJayram, V. Ramasubramanian, and T. V. Sreenivas, "Robust parameters for automatic segmentation of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 1, pp. I–513–I–516.

[14] A. Frid and Y. Lavner, "Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm," in *2010*

IEEE 26th Convention of Electrical and Electronics Engineers in Israel (IEEEI), 2010, pp. 000751–000755.

[15] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, 2006.

[16] G. Fairbanks, "The Rainbow Passage," in *Voice and Articulation Drillbook*, 2nd ed., New York: Harper & Row, 1960, p. 127.

[17] S. A. Borrie, M. J. McAuliffe, and J. M. Liss, "Perceptual Learning of Dysarthric Speech: A Review of Experimental Studies," *J. Speech Lang. Hear. Res.*, vol. 55, no. 1, pp. 290–305, Dec. 2011.

[18] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2217–2235, 2001.

[19] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 25, no. 1, pp. 24–33, 1977.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans Intell Syst Technol*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[21] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working Set Selection Using Second Order Information for Training Support Vector Machines," *J Mach Learn Res*, vol. 6, pp. 1889–1918, Dec. 2005.