A Method for Automatic and Objective Scoring of Bradykinesia Using Orientation Sensors and Classification Algorithms

O. Martinez-Manzanera*, E. Roosma, M. Beudel, R. W. K. Borgemeester, T. van Laar, and N. M. Maurits

***Abstract*—Correct assessment of bradykinesia is a key element in the diagnosis and monitoring of Parkinson's disease. Its evaluation is based on a careful assessment of symptoms and it is quantified using rating scales, where the Movement Disorders Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is the gold standard. Regardless of their importance, the bradykinesia-related items show low agreement between different evaluators. In this study, we design an applicable tool that provides an objective quantification of bradykinesia and that evaluates all characteristics described in the MDS-UPDRS. Twenty-five patients with Parkinson's disease performed three of the five bradykinesia-related items of the MDS-UPDRS. Their movements were assessed by four evaluators and were recorded with a nine degrees-of-freedom sensor. Sensor fusion was employed to obtain a 3-D representation of movements. Based on the resulting signals, a set of features related to the characteristics described in the MDS-UPDRS was defined. Feature selection methods were employed to determine the most important features to quantify bradykinesia. The features selected were used to train support vector machine classifiers to obtain an automatic score of the movements of each patient. The best results were obtained when seven features were included in the classifiers. The classification errors for finger tapping, diadochokinesis and toe tapping were 15–16.5%, 9.3–9.8%, and 18.2–20.2% smaller than the average interrater scoring error, respectively. The introduction of objective scoring in the assessment of bradykinesia might eliminate inconsistencies within evaluators and interrater assessment disagreements and might improve the monitoring of movement disorders.**

***Index Terms*—Bradykinesia, clinical diagnosis, computer-aided diagnosis, sensor fusion, supervised learning, support vector machines (SVMs).**

I. Introduction

BRADYKINESIA is defined as slowness of movement [1] and is one of the main symptoms of Parkinson's disease (PD) [2]. Its accurate evaluation is essential for correct diagnosis and monitoring of PD. The gold standard for assessing its severity and that of other movement disorder's symptoms is the evaluation by a well-trained clinician using standard clinical rating scales [3]. While a physical examination of the patient and a careful evaluation of the symptoms are required for the assessment of bradykinesia, rating scales are employed to express its severity as a quantity. The most widely used for PD is the Movement Disorders Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [4]. In its motor evaluation section, it defines a series of tasks that are performed by the patient and the movement performance characteristics that should be assessed. The assessment is represented by a sum score that summarizes movement performance. However, in spite of its ubiquitous use, the evaluation of the bradykinesia-related items of the MDS-UPDRS shows low interrater agreement between movement disorders specialists [1]. This limitation hampers the evaluation of bradykinesia and the diagnosis and monitoring of PD. An objective, unbiased scoring of these items of the MDS-UPDRS could improve the evaluation of bradykinesia.

The characteristics that are evaluated for the bradykinesia-related items of the MDS-UPDRS include amplitude, speed, hesitations, halts, and any variability or changes in these features over time [4]. The objective measurement and analysis of these characteristics (or very similar ones) has been the goal of previous studies [1], [2], [5]–[10]. Different sensors or combinations of sensors have been employed, such as accelerometers [1], [2], [6], gyroscopes [1], [7], [8], magnetic sensors [9], and tactile screens [10]. This resulted in a wide variety of measurement systems and methodologies that have allowed for bradykinesia assessments to be extended to even outside the hospital [8], [11]. In recent years, the modified bradykinesia rating scale (MBRS), which assesses amplitude, speed, and rhythm of movements with individual scores was introduced [12]. Its reliability has been evaluated with motion sensors in different tasks [1], [13], [14]. While it provides increased sensitivity in identifying different components of bradykinesia, it shares some of the limitations of the UPDRS because it also relies on subjective clinical judgment [1]. In spite of good results, these objective assessments are still not commonly used and the MDS-UPDRS remains the gold standard for the quantification of bradykinesia [15]. In order to bring objective and unbiased assessment tools to clinical practice, the gap between current subjective clinical rating scales and the wide variety of sensors and methods used for objective assessment of bradykinesia needs to be closed.

Current assessment of bradykinesia is impaired by two inherent inconveniences: the evaluator's individual bias and

Manuscript received May 1, 2015; revised July 16, 2015; accepted September 15, 2015. Date of publication September 18, 2015; date of current version May 19, 2016. This work was supported in part by the Mexican National Council for Science and Technology (CONACYT).Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org. *Asterisk indicates corresponding author.*

*O. Martinez-Manzanera is with the Department of Neurology, University Medical Center Groningen (UMCG), 9713 GZ Groningen, Netherlands (e-mail: o.e.martinez.manzanera@umcg.nl).

E. Roosma is with the Martini Hospital.

M. Beudel, R. W. K. Borgemeester, T. van Laar, and N. M. Maurits are with the Department of Neurology, University Medical Center Groningen (UMCG).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2015.2480242

inconsistency and scale limitations due to the limited number of categories of the scale [16]. In a separate study [16], we propose a solution to the problem of the limited number of categories. Here, we propose an automatic and objective method for assessment of the bradykinesia-related items of the MDS-UPDRS that uses a supervised classification algorithm (support vector machine (SVM) based) to reproduce the evaluators' classification results. Specifically, to bridge the gap between the current quantification of bradykinesia and automatic measurement and assessment tools, we base our analysis on data that are highly comparable to what an evaluator can observe and define features that are very similar to the characteristics that are evaluated for the bradykinesia-related items of the MDS-UPDRS.

The most accurate technique to monitor human movements in a research setting is by using optical motion analysis systems [3]. However, such systems impose many restrictions that make them unsuitable for routine clinical assessment. Instead, to obtain an accurate description of movement, a nine degrees-of-freedom (9DoF) sensor (Shimmer [17], Dublin, Ireland, version 2r, composed of three accelerometers, three gyroscopes, and three magnetic sensors) was employed to capture movement performance. By integrating the information of each individual signal using a sensor fusion algorithm an accurate estimate of 3-D movement was obtained. The result of this algorithm, in the form of quaternions, was transformed to Euler angles. By selecting the Euler angle that best represented the observed movement for the specific MDS-UPDRS item, and subsequently, extracting features that are very similar to the characteristics defined in the MDS-UPDRS, we ensured that the automatic measurement and assessment method was highly comparable to the current quantification of bradykinesia.

To objectively evaluate movement performance, we employed SVMs. An SVM classifier can include every feature available, but this might result in overfitting and poor classification performance due to the curse of dimensionality [18]. Alternatively, the classifier can only include the features that produce an improvement on the classification. However, this can result in the exclusion of some important features. We took a middle way between these two methods and included two features *a priori* that can be related to two important characteristics described in the MDS-UPDRS (amplitude and speed), and then, included additional features into the classifier based on their performance. An alternative approach to reduce data dimensionality, and thereby, avoid the curse of dimensionality is principal component analysis (PCA). We explored this alternative approach as well, as features based on principal components will express more of the variance recorded by the sensors and may, therefore, result in a better classifier. These two approaches were adopted to evaluate whether features from expert knowledge obtain a better performance over features from dimensionality reduction.

To determine which features should be included in the classifier, an iterative method (forward-selection wrapper [19]) was used. In each iteration, an extra feature was included in the classifier, based on the classifier performance. This process was repeated until there was no improvement in the performance of the classifier.

The SVM is a supervised classifier that learns from given labels. The scores from evaluators were used as labels to train the classifier. The performance of the classifier was obtained using leave-one-out cross validation [20] (LOOCV), which is a technique used to estimate the classification error on new data.

In this study, we aim to obtain an objective evaluation of bradykinesia that eliminates inconsistency of an evaluator. Different evaluators might weight movement characteristics differently. Therefore, the features selection procedure was performed using the scores of four clinical evaluators, separately. This resulted in four different classifiers (for each MDS-UPDRS bradykinesia-related item) that learned from different labels and that might include different features. The classification error of these classifiers was averaged for each iteration and these averages were compared against the interrater scoring error to assess the performance of our automatic measurement and assessment methods.

II. Methods

Twenty-five patients with mild to moderate PD (age: 64.4 ± 1.7 y, 13 male, 12 female, SCOPA-COG cognition test: 30.0 ± 1.0) and ten age-matched controls (age: 65.2 ± 3.2 y, 6 male, 4 female, SCOPA-COG cognition test: 28.5 ± 1.4). Every participant performed items 3.4 (finger tapping), 3.6 (diadochokinesis), and 3.7 (toe tapping) of the motor examination section of the MDS-UPDRS with both right and left limbs. All participants were asked to perform the tasks as fast and accurately as possible. Controls were included to evaluate the relevance of features included *a priori* in the classifier. For every patient, each task was videoed and later scored by four well-trained clinicians according to the guidelines of the MDS-UPDRS. The study was conducted according to the principles of the Declaration of Helsinki (2008) with prior approval of the Ethics committee of the University Medical Center Groningen (UMCG).

A. Signal acquisition

Before each task was performed, a 9DoF orientation sensor was placed on the specific body part of interest. For finger tapping, the sensor was placed on the dorsal side of the proximal phalange of the index finger. For diadochokinesis, the sensor was placed on the dorsal side of the forearm close to the wrist. Finally, for toe tapping, the sensor was placed on the instep of the foot over the shoe of the participant. Each 9DoF sensor incorporates nine internal sensors (three accelerometers, three gyroscopes, and three magnetic sensors), where sensors of the same type are orthogonally aligned to each other. Before every measurement, each sensor was calibrated using the Shimmer 9DoF Calibration v2.3 [17] application. This prevented misalignment of the electronic board containing the internal sensors with the outer case and ensured proper recording of the magnetic sensors. All signals were recorded at a sampling rate of 51.2 Hz and streamed via Bluetooth to a computer.

B. Sensor Fusion

To reduce the effects of noise and to obtain a more accurate estimate of movement, all signals from each sensor were

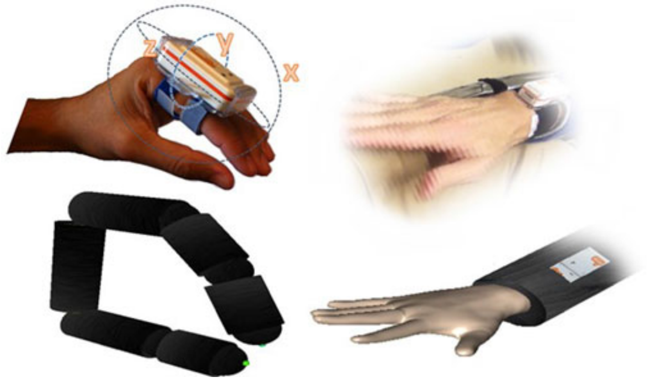


Fig. 1. Left: Orientation sensor on index finger for finger tapping task (top) and its corresponding model (bottom). Right: Orientation sensor on the wrist for diadochokinesis task (top) and its corresponding model (bottom).

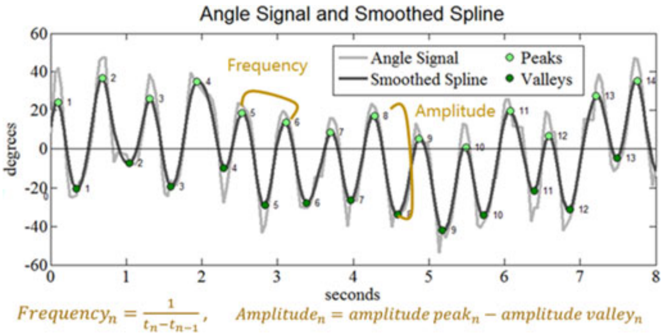


Fig. 2. Example of raw angle signal (gray) and smoothing spline angle (black) signal for diadochokinesis. The frequency of each tap is obtained as the inverse of the time (*t*) between consecutive peaks. In the figure, the frequency of tap six was defined as the inverse of the time difference between the sixth and the fifth peaks. The amplitude of each tap is obtained as the difference in amplitude from a peak to the next valley. In the figure, the amplitude of tap eight was defined as the amplitude difference between the eighth peak and the eighth valley.

combined with a sensor fusion algorithm [21] that allows the estimation of the spatial orientation parameters of the 9DoF sensor. This algorithm, based on quaternions, achieves the level of accuracy of a Kalman filter (which is considered the most popular probabilistic fusion algorithm [22]) without the computational expense that the latter requires [21]. The quaternion representation of an orientation vector used in this algorithm has the advantage that it is not affected by singularities (gimbal lock) associated with Euler angles [21] that affect other algorithms. The output of the algorithm in the form of quaternions was converted to Euler angles. Since for each task, most of the movement can be described by a single Euler angle (for finger tapping by the angle that describes the flexion and extension of the index finger, for diadochokinesis by the angle that describes the pronation and supination of the wrist and for toe tapping by the angle that describes the dorsiflexion and plantar flexion of the foot) the analysis of each tasks was performed on the corresponding Euler angle signal that explained most of movement.

C. User Interface

Shimmer provides a basic acquisition program in LabView [23] (Austin, Texas, USA.) that includes a three dimensional representation of a 9DoF sensor. This program was modified to display a 3-D model that represents the body parts involved in each movement (see Fig. 1 for two examples). This representation allowed visual identification of improper calibration as indicated by false rotational movements in the model for motionless sensors and verification that the recording procedure was correctly performed (sensors placed incorrectly would be indicated by unusual movements of the model).

D. Signal Processing

By nature, human body movements are limited to a maximum frequency of 20 Hz [24]. Therefore, to decrease artifacts such as drift and the noise produced by the main electrical power line, the signals were band-pass filtered between 0.3 and 20 Hz (second-order Butterworth filter). Then, to obtain a smoother version of the signals for feature extraction, spline interpolation was used to fit each signal (see Fig. 2) using a smoothing parameter [25] $\rho = 0.1$ (1). With this approach, the fitted spline does not go through every single point of the original signal but only represents the general pattern of the signal. The function that was minimized to obtain the smoothing spline is given as

$$\rho \sum_{i} w_i \left(y_i + s\left(x_i\right)\right)^2 + (1 - \rho) \int \left(\frac{d^2 s}{dx_2}\right)^2. \tag{1}$$

Here, ρ is the smoothing parameter and w_i is the specified weight of data point *i*. The first term is the mean squared error (MSE) when the curve *s,* which is a function of *x*, is used to predict *y*. The second term is an added penalty function that limits the curvature of *s* [26]. Two versions of each signal were thus obtained: one with more detail (raw angle (RA) signal) and one with less detail (smoothing spline angle (SSA) signal). Features were subsequently extracted from these two signals.

E. Determination of Features

To obtain features related to the characteristics defined in the MDS-UPDRS (e.g., amplitude, speed, and their variability), we first identified each movement repetition in the signal by distinguishing the peaks and valleys in the SSA signal. Then, we defined the amplitude of a single movement as the difference in amplitude from a peak to the next valley and the frequency of each movement as the inverse of the time between consecutive peaks (see Fig. 2). To represent amplitude and frequency (representing speed of movement) as mentioned in the MDS-UPDRS, the mean amplitude and mean frequency across all identified movements were calculated. Due to its smoothness, the SSA signal more closely resembles the observed oscillation pattern associated with the type of movements studied in the MDS-UPDRS than the RA signal. On the other hand, the low-pass filtering effect of the spline interpolation reduces the amplitude of each individual movement repetition in the SSA signal. We, therefore, decided to calculate features for both the RA and SSA signals.

Other characteristics that are evaluated according to the MDS-UPDRS are decrement of movement amplitude, and slowing of movement. To capture decrement of movement amplitude, the slope of the straight line fitted through all movement amplitudes as a function of movement repetition number was taken (slope amplitude). To capture slowing of movement, a similar procedure was performed for the movement frequencies, resulting in the feature slope frequency. These procedures were performed for both the RA and SSA signals.

Rhythm is another characteristic mentioned in the MDS-UPDRS and can be defined as any sequence of regularly recurring events. To account for this characteristic, we estimated features based on its reciprocal, movement variability. We estimated amplitude and frequency variability by calculating the standard deviations (std) of all individual movement amplitudes and frequencies, respectively. This resulted in the features std amplitude and std frequency.

Another characteristic mentioned in the MDS-UPDRS is regularity. The expected regular signal of a healthy subject describes a smooth pattern. To account for regularity, we obtained features related to the smoothness of the signal. Compared to their corresponding RA signals, SSA signals are much smoother. The goodness of fit of SSA signals to their corresponding RA signals thus provides an indication of the smoothness of movement. The discrepancy between these two signals is summarized in the following additional features: sum of squares due to error (SSE), which is the total deviation of the SSA signal from the RA signal, the coefficient of determination (R2) and the root mean squared error (RMSE) [27].

We additionally included features describing maximum velocity during movement initiation and termination. First, to estimate the velocity of each movement, the first derivative of the SSA signal was calculated. According to Shima *et al.* [9], we determined the maximum velocity during initiation of each movement (extension for finger tapping, dorsiflexion for foot tapping, and supination for diadochokinesis) and during termination of each movement (flexion for finger tapping, plantar flexion for foot tapping, and pronation for diadochokinesis) and used their mean and coefficient of variation (CV) resulting in the features mean and CV maxV during movement initiation and mean and CV maxV during movement termination.

Finally, hesitations were quantified according to Shima *et al.* [9], employing zero crossings in the acceleration signal. The acceleration signal was calculated as the second derivative of the SSA signal. An individual movement was considered to contain hesitations if its corresponding acceleration signal contained more than two zero crossings. The percentage of individual movements containing hesitations (percentage of hesitations) and the CV of the number of zero crossings of each individual movement (CV of zero crossings) were determined as features related to hesitations.

F. Feature Selection

The features so far described constitute the basic set of features (set 1) that was used in the forward-selection wrapper to select features for the classifier. Since the relationship between the selected features and an evaluator's scores might be better described by nonlinear than by linear relationships, we also formed a set of features (set 2), which was composed of the features of set 1 and their squared values. A summary of all features is given in Table I.

TABLE I
FEATURES' NUMBERS FOR SETS 1 AND 2

Number	FEATURE	Squared Version Number
1	Slope amplitude RA	22
2	Mean amplitude RA	23
3	Standard deviation amplitude RA	24
4	Slope frequency RA	25
5	Mean frequency RA	26
6	Standard deviation frequency SSA	27
7	Slope amplitude SSA	28
8	Mean amplitude SSA	29
9	Standard deviation amplitude SSA	30
10	Slope frequency SSA	31
11	Mean frequency SSA	32
12	Standard deviation frequency SSA	33
13	Filtered signal fit (SSE)	34
14	Filtered signal fit (R2)	35
15	Filtered signal fit (RMSE)	36
16	Percentage of hesitations	37
17	CV of zero crossings	38
18	Mean maxV during movement initiation	39
19	CV maxV during movement initiation	40
20	Mean maxV during movement termination	41
21	CV maxV during movement termination	42

Set of features for subsets 1 and 2. Set 1 includes features 1 to 21. Set 2 is composed from these features plus their squared version (features 22 to 42).

Our goal is to select features such that the characteristics described in the MDS-UPDRS are captured. Amplitude and frequency (representing speed of movement) are two characteristics that can be more easily and more reliably estimated from sensor recordings than the rest of the characteristics (variability, hesitations, halts, etc.). Since these two characteristics are mentioned in the MDS-UPDRS, we decided to include the features that represent them (mean amplitude and mean frequency) in the algorithm. The relevance of these two features to improve the classification performance was evaluated using a *t*-test to compare their values between patients and controls. A *t*-test is a univariate feature importance method [28]. Univariate methods assume feature independence. This assumption is not met by amplitude and frequency (higher amplitudes can only be obtained at the cost of speed and vice versa). Therefore, feature importance was tested on the product of amplitude and frequency. The results of the *t*-test indicate that this feature is significantly different between patients and controls. Therefore, we decided to include the two features mean amplitude and mean frequency of the SSA signal in the classifiers *a priori,* before the first iteration of the feature selection algorithm.

Wrappers are multivariate methods that take into account feature dependences. They potentially achieve better results because they do not make simplifying assumptions regarding feature independence. The forward-selection wrapper approach is an iterative method that includes one feature into the classification algorithm with each iteration [19]. It allows to observe the performance of the classifier (in terms of number of tasks correctly classified) as features are added to the classifier. The inclusion of features with meaningless variance in terms of classification will only confound learning methods [18]. Thus, instead of entering every feature into the classifier, a subset of features must be used. There are different feature selection methods. The forward-selection wrapper approach [16] was selected for this study, because it is easier to interpret the incorporation of each feature into the model than in the backward-selection approach where relations between variables are taken into account. To select features, wrappers use the same evaluation criterion as employed by the classifier itself (in this case, the classification error). To determine which feature should be added, in each iteration, the performance of the classifier is evaluated with all already included features plus each candidate feature individually. The method will select the feature that results in the largest performance improvement. This method was used separately on sets 1 and 2 resulting in a subset of features for each (subsets 1 and 2).

We also explored classification performance when features are obtained by dimensionality reduction using PCA. PCA was applied to set 2 only, as it contains more features than set 1. The resulting principal components (PCs) represent the (combined) features from expert knowledge in order of explained variance. However, more explained variance does not necessarily imply better classification performance. We built two classifiers on the basis of the resulting PCs (set 3): subset 3 was built by adding the PC to the classifier (with each iteration) that explained most of the remaining variance. Finally, subset 4 was built using the forward-selection wrapper approach on the PCs. For a fair comparison across methods, the first two PCs that explained most variance were also included *a priori* before the first iteration of the feature selection method (see Fig. 3 for an overview of the four classification approaches).

G. Classification

All classification approaches employed the classification error as evaluation criterion. The classification error was defined as the percentage of patients that were incorrectly classified (according to the evaluator's classification) using LOOCV [29], which was used instead of other less computationally expensive algorithms in view of the relatively low number of participants [30].

The automatic classification was done using SVMs. SVMs employ kernels to map the data into a higher dimensional feature space where data can be separated by a hyperplane [31]. Originally, SVMs were designed for binary classification. In this study, the performance on each task was scored between zero and four according to the MDS-UPDRS criteria by each of the four evaluators. From the several methods that extend SVMs use to multiclass classification [32], a one-versus-all strategy, which employs binary classifiers (e.g., score-zero class versus the rest of the classes) was selected (illustrated in Fig. 4). In each binary classifier, the features derived from every performance but one (according to LOOCV) were used as training samples to construct a hyperplane. The remaining performance is used as a test sample. Its location in the feature space determines the confidence value, which can be interpreted as the Euclidian

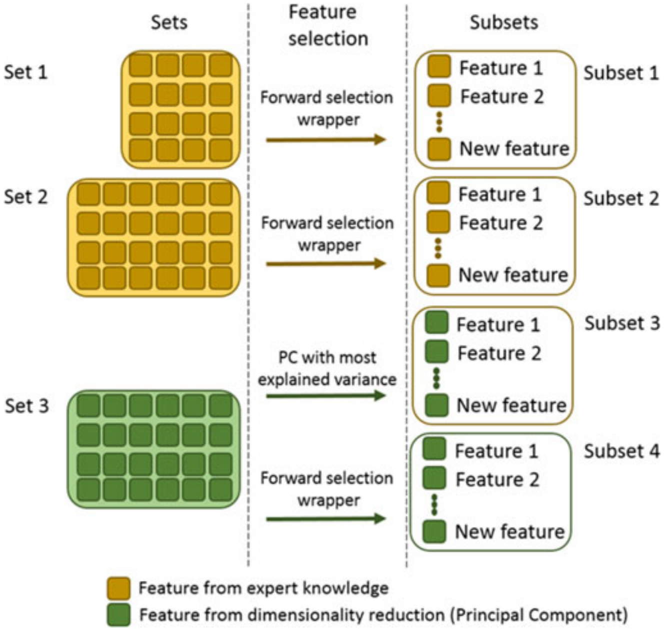

Sets
Feature selection
Subsets
Set 1
Set 2
Set 3
Forward selection wrapper
Forward selection wrapper
PC with most explained variance
Forward selection wrapper
Feature 1
Feature 2
New feature
Subset 1
Subset 2
Subset 3
Subset 4
Feature from expert knowledge
Feature from dimensionality reduction (Principal Component)

Fig. 3. Overview of four classification approaches. From each set of features, a subset with optimal features is constructed. The first two sets are composed of features from expert knowledge and the last two contain features obtained from PCA. Subsets 1, 2, and 4 are built using the forward-selection wrapper as the feature selection algorithm, while subset 3 includes in each iteration the PC that explains most of the variance that has not already been included. Two features are included in each subset *a priori* before the addition of more features.

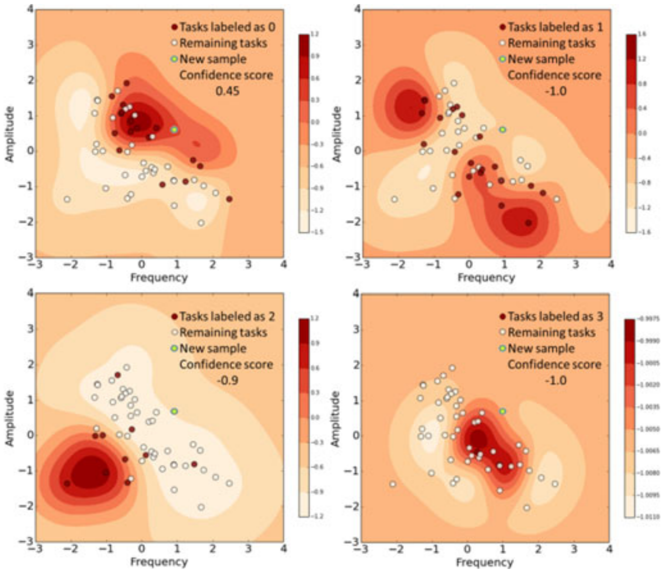


Fig. 4. Multiclass classification using SVM and the RBF kernel in a one-versus-all methodology using LOOCV with only two features. In this example, tasks were evaluated from 0 to 3 (four classes). In a one-versus-all methodology, every single class is evaluated against the rest of the classes (e.g., red dots correspond to the amplitude and frequency of the tasks scored as zero and white dots correspond to the amplitude and frequency of the remaining tasks in the top left figure). The different decision surfaces created using the red and white dots are illustrated with different colors. The confidence of a new sample (green dot representing the task left out by LOOCV) to belong to a certain class is represented by the color of the surface and by the scale on the right of each figure. The new sample is then classified in the class that obtained the highest confidence (score 0 in the example).

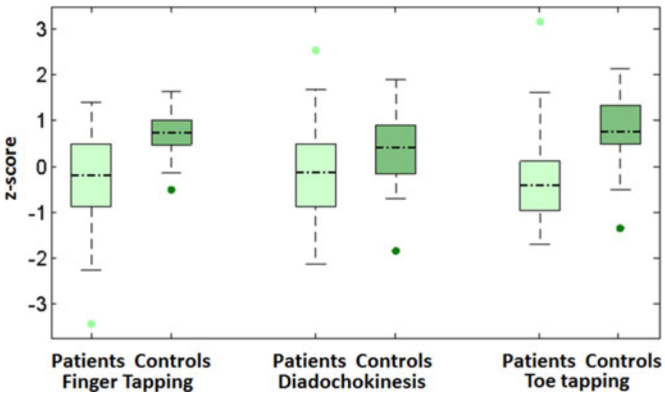


Fig. 5. Boxplots of combined amplitude–frequency feature for finger tapping (left), diadochokinesis (center), and toe tapping (right). On average, for the three tasks controls exhibit a significantly higher combination of amplitude and frequency than patients.

distance of the sample to the separating hyperplane and it expresses the confidence of a sample to belong to a certain class. The remaining sample is then classified in the class that obtained the highest confidence value across all binary classifiers. When using SVMs, the choice of the kernel determines the separation boundaries of the classes. In this study, the radial basis function (RBF) (2) kernel was used, which is generally a reasonable choice [33].

$$K_{\mathrm{RBF}} = (x, x') = \exp(-\gamma \left\| x - x' \right\|^2). \qquad (2)$$

Here, x and x' are two training samples of the feature space and γ determines the influence of the squared Euclidian distance (between the feature vectors x and x') to build the hyperplane. In this study, $\gamma = 1.0$ was selected. To avoid poor performance due to relatively large values of individual features, all features were first normalized using z-scores.

For all methods, the classification error obtained at each iteration of the feature selection wrapper was compared against the average interrater scoring error. This error was defined as the average percentage of tasks that were classified differently by each combination of two evaluators.

III. Results

A. Combined Amplitude–Frequency

The distributions for combined amplitude-frequency were all normally distributed for the three MDS-UPDRS items and for both groups with the exception of toe tapping for patients ($p = 0.04$, Kolmogorov–Smirnov test). T-tests were used for all group comparisons including toe tapping, since its distribution did not show large differences from normality. Combined amplitude–frequency was always higher for controls than for patients (finger tapping: controls $M = 131.85$ deg/s, patients $M = 107.73$ deg/s, $p = 0.0002$; diadochokinesis: controls $M = 170.68$ deg/s, patients $M = 150.78$ deg/s, $p = 0.03$; toe tapping: controls $M = 32.19$ deg/s, patients $M = 12.16$ deg/s, $p < 0.0001$, see Fig. 5).

B. Classification

The effects of the curse of dimensionality (the performance does no longer increase (substantially) even though more features were added) are visible for each of the three items approximately after the sixth iteration (seven features used) (illustrated in Fig. 6). Therefore, we focus our analysis on the results obtained before this effect occurs.

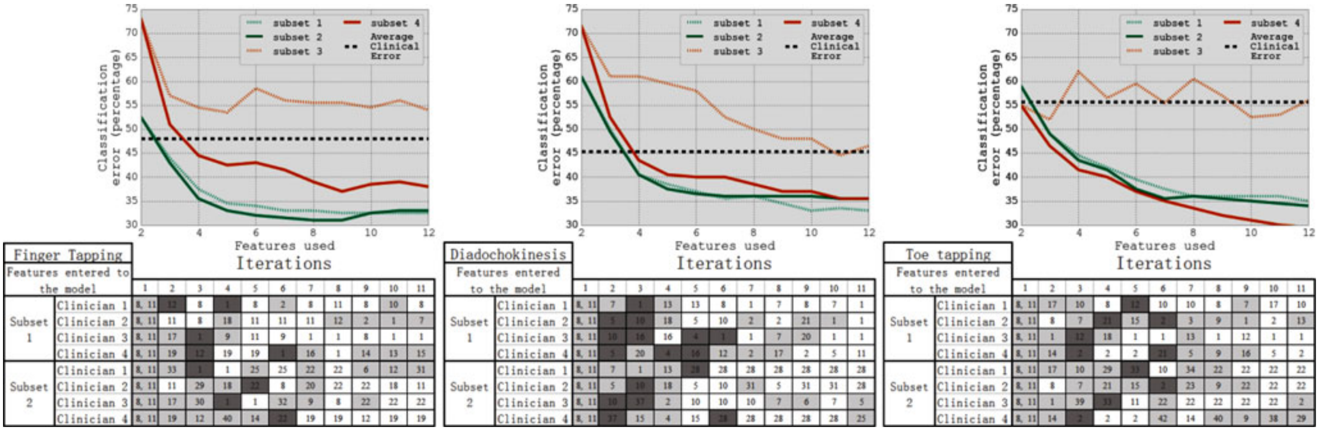
1) Finger Tapping: The average classification error for each subset is illustrated for finger tapping in Fig. 6 (left). Classification employing features in subsets 1 and 2 gave better results than employing features in subsets 3 and 4. When seven features were included (sixth iteration), the classification error for subsets 1 and 2 was 33% and 31.5%, respectively: an improvement of 15–16.5% compared to the average interrater scoring error (48%). These performances were just 0.5–1% lower than the best performance of the classifiers that occurred at the ninth and tenth iterations, respectively. After four iterations the best performance for subset 3 was obtained, resulting in a classification error of 53.5%: 5.5% worse than the average interrater scoring error. After six iterations, the classification error of subset 4 was 41.5%: this performance is 6.5% better than the average interrater scoring error and 4.5% worse than the best performance found at iteration eight.

Until iteration six, the unique features selected by more than one classifier for subsets 1 or 2 (besides features 8 and 11 that were included *a priori*) were the features 12, 1, and 22.

2) Diadochokinesis: The average classification error at each iteration for each subset is illustrated for diadochokinesis in Fig. 6 (center). Overall classification performed better for subsets 1 and 2 than for subsets 3 and 4. After six iterations, the classification errors for subset 1 and 2 were 35.5% and 36%, respectively: an improvement of 9.3–9.8% compared to the average interrater scoring error (45.3%). Classification for subset 3 shows very poor improvement performance. The best performance was obtained after iteration ten, resulting in a classification error of 44.5%: a minor improvement of 0.8% compared to the average interrater scoring error. However, classification for subset 4 shows a similar pattern as for subsets 1 and 2. At the sixth iteration, a classification error of 40% is obtained: an improvement of 5.3% compared to the average interrater scoring error and 9.8% worse than the best performance found at iteration ten.

Until iteration six, the unique features selected by more than one classifier for subsets 1 and 2 (besides features 8 and 11 that were included *a priori*) were the features 1, 4, 5, 10, 16, 28, and 37.

3) Toe Tapping: The average classification error for each subset is illustrated for toe tapping in Fig. 6 (right). After six iterations, the classification error for subsets 1 and 2 was 37.5% and 35.5%, respectively: an improvement of 18.2–20.2% compared to the average interrater scoring error (55.7%) and only 1.5–2.5% worse than the best performance for these subsets. Classification for subset 3 showed very poor performance, obtaining its lowest error at iteration two (52%): only 3.7% better than the average interrater scoring error and not showing improvement afterwards. On the other hand, classification performance



Finger Tapping	Features entered to the model	1	2	3	4	5	6	7	8	9	10	11
Subset 1	Clinician 1	8, 11	12	8	1	8	2	8	11	8	10	8
	Clinician 2	8, 11	11	8	18	11	11	11	12	2	1	7
	Clinician 3	8, 11	17	1	9	11	9	1	1	8	1	1
	Clinician 4	8, 11	19	12	19	19	1	16	1	14	13	15
Subset 2	Clinician 1	8, 11	33	1	1	25	25	22	22	6	12	31
	Clinician 2	8, 11	11	29	18	22	8	20	22	22	18	11
	Clinician 3	8, 11	17	30	1	1	32	9	8	22	22	22
	Clinician 4	8, 11	19	12	40	14	22	19	19	12	19	19

Diadochokinesis	Features entered to the model	1	2	3	4	5	6	7	8	9	10	11
Subset 1	Clinician 1	8, 11	7	1	13	13	8	1	7	8	7	1
	Clinician 2	8, 11	5	10	18	5	10	2	2	21	1	1
	Clinician 3	8, 11	10	16	16	4	1	1	7	20	1	1
	Clinician 4	8, 11	5	20	4	16	12	2	17	2	5	11
Subset 2	Clinician 1	8, 11	7	1	13	28	28	28	28	28	28	28
	Clinician 2	8, 11	5	10	18	5	10	31	5	31	31	28
	Clinician 3	8, 11	10	37	2	10	10	10	7	6	7	5
	Clinician 4	8, 11	37	15	4	15	28	28	28	28	28	25

Toe tapping	Features entered to the model	1	2	3	4	5	6	7	8	9	10	11
Subset 1	Clinician 1	8, 11	17	10	8	12	10	10	8	7	17	10
	Clinician 2	8, 11	8	7	21	15	2	3	9	1	2	13
	Clinician 3	8, 11	1	12	18	1	1	13	1	12	1	1
	Clinician 4	8, 11	14	2	2	2	21	5	9	16	5	2
Subset 2	Clinician 1	8, 11	17	10	29	33	10	34	22	22	22	22
	Clinician 2	8, 11	8	7	21	15	2	23	9	22	22	22
	Clinician 3	8, 11	1	39	33	11	22	22	22	22	22	2
	Clinician 4	8, 11	14	2	2	2	42	14	40	9	38	29

Fig. 6. Boxplots of combined amplitude–frequency feature for finger tapping (left), diadochokinesis (center), and toe tapping (right). Below each graph, there is a visualization of the features selected for subsets 1 and 2 for each evaluator on each iteration. The incorporation of a nonrepeated feature is indicated in pale gray. On dark gray, the features that are selected by more than one classifier are indicated.

for subset 4 showed a continuous improvement. After six iterations, it obtained a classification error of 35%: an improvement of 17% compared to the average interrater scoring error and 5.5% worse than its best performance.

Until iteration six, the unique features selected by more than one classifier for subsets 1 or 2 (besides features 8 and 11 that were included *a priori*) were the features 2, 12, 21, and 33.

IV. Discussion

In this study, we showed how objective measurement and assessment of the bradykinesia-related items of the MDS-UPDRS can be achieved using a 9DoF sensor and SVM-based classification. Our approach resulted in a consistent scoring of tasks with a lower classification error than the interrater classification error that occurs when bradykinesia is assessed by different evaluators. Classification based on features that were closely related to the important characteristics assessed in the MDS-UPDRS outperformed classification based on features that resulted from dimensionality reduction (PCA) for two of the three bradykinesia-related items. The importance of selecting appropriate and relevant features is most obvious from the results obtained when, at each iteration, the PC that explained most of the remaining variance in the dataset was added (subset 3); this approach resulted in the worst classification performance among all subsets.

As expected, for all items and before too many features were entered in the algorithm, classification based on subset 2 obtained a slightly better classification more rapidly than for subset 1. This suggests that the relation between the selected features and the clinical evaluation might be nonlinear. Among the many nonlinear transformations of features that could have been used (e.g., logarithmic, square root, etc.), we only investigated the change in classification performance when quadratic features were added to the set of features. It may be that including features derived from other nonlinear transformations of the original features would further improve classification performance.

Amplitude and speed are the two characteristics mentioned in the MDS-UPDRS that can be more directly related with specific features from the recorded signal. After confirming their relevance with a feature importance test, we decided to include them *a priori* into the feature selection algorithm. A different approach would be a feature selection algorithm without *a priori* inclusion of features. However, depending on the scores used to train the classifiers, some features that according to the MDS-UPDRS should be included in the classifier might be left out. The other extreme case would have been to include every feature, which would most likely result in overfitting and problems due to the curse of dimensionality.

A. Feature Selection

For most of the subsets, the best classification performance was obtained around the sixth iteration. In most cases, further inclusion of features did not improve or even declined classifier performance. Our discussion is, therefore, focused on the features selected by the classifiers in subsets 1 and 2 until this iteration. The features selected by the classifiers suggest which features were more relevant for each evaluator.

1) Finger Tapping: From set 1, feature 1 (slope amplitude RA) and feature 12 (std frequency RA) were the only features selected by more than one classifier. This indicates that the variability in movement speed (feature 12) and the decrease in movement amplitude (feature 1) are important characteristics to score this task. A decrease in movement amplitude is typical for patients with PD. Probably both methods selected the slope from the RA signal (feature 1) and not from the SSA signal (feature 7) because the low-pass filter effect of the spline interpolation reduced signal amplitude. From set 2, feature 1 was also selected by more than one classifier. Moreover, the only other feature selected by more than one classifier was its squared version (feature 22). For subset 2, feature 12 was only selected by one classifier. This probably occurred because one classifier included the square of featured 12 (feature 33), instead.

2) Diadochokinesis: From set 1, five features were selected by at least two classifiers for diadochokinesis. Features 1 (slope amplitude RA), 4 (slope frequency RA), 5 (mean frequency RA), 10 (slope frequency SSA), and 15 (percentage of hesitations) were selected. From set 2, only two features were selected by at least two classifiers: feature 15 was substituted by its squared version (feature 37) and the squared version of slope amplitude SSA signal (feature 28) was also included. The inclusion of slope frequency and slope amplitude underlines the importance of the decrease in amplitude and speed of movement to rate this task. Since feature 11 (mean frequency SSA) was one of the *a priori* selected features, it is interesting to notice that two classifiers also included feature 5 (mean frequency RA). This suggests that the information contained in these two features is different. The percentage of hesitations was a feature selected from both sets 1 and 2 for diadochokinesis, while it was not selected for the other two bradykinesia-related items of the MDS-UPDRS. We suggest that this may be explained by the fatigue induced by this task, which may result in short movement halts that can be identified on video.

3) Toe Tapping: From set 1, feature 2 (mean amplitude RA) and feature 12 (std frequency SSA) were the only features selected by more than one classifier for toe tapping. For the classifiers that employed set 2, the std frequency of SSA signal was substituted by its squared version (feature 33). In contrast to the other two bradykinesia-related items of the MDS-UPDRS, feature 1 (slope amplitude RA) was not selected by more than one classifier. This can be explained by the difficulty of evaluating the small amplitude movements involved in toe tapping. Feature 21 (CV maxV during movement termination) was selected by more than one classifier from set 1, but it was not selected anymore from set 2. This probably occurred because one classifier included the square of featured 21 (feature 42).

In this study, we allowed the inclusion of repeated features in the feature selection algorithm. The reasons are twofold. First, the kernel employed by the SVM classifier (RBF) defines the shape of the decision boundary. The decision boundary obtained in a larger feature space (with more dimensions) might produce a better classification even if the features included are repeated. Also, limiting the inclusion of features to only nonrepeated features might force the inclusion of nonrelevant features.

V. Conclusion

The objective evaluation based on features eliminates inconsistency within an evaluator. Using a classification algorithm with objective features we were able to score the bradykinesia-related items of the MDS-UPDRS task more accurately than the average interrater scoring error. However, since classifiers learned from labels obtained from evaluators individual bias is still present in each classifier. Following the same methodology with a larger number of evaluators and employing only the tasks where consensus is found could lead to an unbiased objective measuring system. This could lead to an improvement in the assessment and monitoring of movement disorders.

References

[1] D. Heldman *et al.*, "The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures," *Mov. Disord.*, vol. 26, pp. 1859–1863, 2011.

[2] M. Pastorino *et al.*, "Assessment of Bradykinesia in Parkinson's disease patients through a multi-parametric system," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, vol. 2011, Jan. 2011, pp. 1810–1813.

[3] N. L. Keijsers *et al.*, "Online monitoring of dyskinesia in patients with Parkinson's disease," *IEEE Eng. Med. Biol. Mag.*, vol. 22, no. 3, pp. 96–103, May/Jun. 2003.

[4] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[5] P. J. G. Ruiz *et al.*, "Bradykinesia in Huntington's disease," *Clin. Neuropharmacol.*, vol. 23, no. 1, pp. 50–52, Jan. 2000.

[6] J. Cancela *et al.*, "A comprehensive motor symptom monitoring and management system: The bradykinesia case," in *Proc. IEEE Eng. Med. Biol. Soc. Conf.*, Jan. 2010, pp. 1008–1011.

[7] J.-W. Kim *et al.*, "Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease," *Med. Biol. Eng. Comput.*, vol. 49, no. 3, pp. 365–371, Mar. 2011.

[8] A. Salarian and H. Russmann, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 2, pp. 313–322, Feb. 2007.

[9] K. Shima *et al.*, "Measurement and evaluation of finger tapping movements using magnetic sensors," in *Proc. IEEE Eng. Med. Biol. Soc. Conf.*, vol. 2008, Jan. 2008, pp. 5628–5631.

[10] A. L. Taylor Tavares *et al.*, "Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation," *Mov. Disord.*, vol. 20, no. 10, pp. 1286–1298, Oct. 2005.

[11] D. G. M. Zwartjes *et al.*, "Ambulatory monitoring of activities and motor symptoms in Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 11, pp. 2778–2786, Nov. 2010.

[12] A. Kishore *et al.*, Unilateral versus bilateral tasks in early asymmetric Parkinson's disease: Differential effects on bradykinesia," *Mov. Disord.*, vol. 22, pp. 328–333, 2007.

[13] A. J. Espay *et al.*, Differential response of speed, amplitude, and rhythm to dopaminergic medications in Parkinson's disease," *Mov. Disord.*, vol. 26, pp. 2504–2508, 2011.

[14] A. J. Espay *et al.*, Impairments of speed and amplitude of movement in Parkinson's disease: A pilot study," *Mov. Disord.*, vol. 24, pp. 1001–1008, 2009.

[15] G. Pal and C. G. Goetz, "Assessing bradykinesia in Parkinsonian disorders," *Front. Neurol.*, vol. 4, pp. 1–5, Jan. 2013.

[16] O. Martinez-Manzanera *et al.*, "A method for automatic, objective and continuous scoring of bradykinesia," presented at the IEEE Int. Conf. Body Sensor Networks, Cambridge, MA, USA, 2015.

[17] Shimmer, Dublin, Ireland. (2014). [Online]. Available: shimmersensing.com

[18] P. F. Evangelista *et al.*, "Taming the curse of dimensionality in kernels and novelty detection," *Adv. Soft Comput.*, vol. 34, pp. 425–438, 2006.

[19] R. Kohavi and H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 97, pp. 273–324, 2011.

[20] G. James *et al.*, *An Introduction to Statistical Learning*, vol. 103. New York, NY, USA: Springer, 2013.

[21] S. Madgwick *et al.*, "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, 2011, pp. 1–7.

[22] Y. Zheng *et al.*, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1538–1554, May 2014.

[23] Shimmer. (2015). Shimmer LabVIEW development library V0.1". [Online] Available at http://www.shimmersensing.com/support/wireless-sensor-networks-download/

[24] C. V Bouten *et al.*, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 3, pp. 136–47, Mar. 1997.

[25] The MathWorks, (2014). Smoothing Splines. [Online]. Available: mathworks.com/help/curvefit/smoothing-splines.html

[26] C. Shalizi. (2011). Splines: Smoothing by directly penalizing curve flexibility. [Online]. Available: http: www.stat.cmu.edu/∼cshalizi/402/lectures/11-splines/lecture-11.pdf., Accessed: 3-Dec-2014.

[27] The MathWorks. (2015). Evaluating Goodness of Fit. [Online]. Available: nl.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html#bq_5kwr-3, Accessed: 3-Dec-2014.

[28] I. Guyon *et al.*, *Feature Extraction*, vol. 207. Berlin, Germany: Springer, 2006.

[29] P. A. Lachenbruch and M. R. Mickey, “Estimation of error rates in discriminant analysis,” *Technometrics*, vol. 10, no. 1, pp. 1–11, Feb. 1968.

[30] G. C. Cawley and N. L. C. Talbot, “Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters,” *J. Mach. Learn. Res.*, vol. 8, 841–861, 2007.

[31] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 7, pp. 144–152, 1995.

[32] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Jan. 2002.

[33] C. Hsu, C. Chang, and C. Lin. (2010). A practical guide to support vector classification. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf. [Accessed: 10-Apr-2014]

Authors’ photographs and biographies are not available at the time of publication.