Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic

Taha Khan
School of Technology and Business Studies, Computer Engineering, Dalarna University, 79188 Falun, Sweden
School of Innovation, Design and Technology, Malardalen University, 72123, Vasteras, Sweden
Email: tkh@du.se, Phone: 0046(0)23778852

Abstract

Background:
The mel-frequency cepstral coefficients (MFCC) are relied for their capability to identify pathological speech. The literature suggests that triangular mel-filters that are used in the MFCC calculation provide an approximation of the human auditory perception. This approximation allows quantifying the clinician’s perception of the intelligibility of the patient’s speech that allows mapping between the clinician’s score of the severity of speech symptoms and the actual symptom severity of the patient’s speech. Previous research on speech impairment in Parkinson’s disease (PD) used sustained-phonation and diadochokinesis tests to score symptoms using the unified Parkinson’s disease rating scale motor speech examination (UPDRS-S).

Objectives:
The paper aims to utilize MFCC computed from the recordings of running speech examination for classification of the severity of speech symptoms based on the UPDRS-S. The secondary aim was to compare the performance of the MFCC from running-speech, and the MFCC from sustained-phonation and diadochokinesis recordings, in classifying the UPDRS-S levels.

Method:
The study involved audio recordings of motor speech examination of 80 subjects, including 60 PD patients and 20 normal controls. Three different running-speech tests, four different sustained-phonation tests and two different diadochokinesis tests were recorded in different occasions from each subject. The vocal performance of each subject was rated by a clinician using the UPDRS-S. A total of 16 MFCC computed separately from the recordings of running-speech, sustained-phonation and diadochokinesis tests were used to train a support vector machine (SVM) for classifying the levels of UPDRS-S severity. The area under the ROC curve (AoC) was used to compare the feasibility of classification models. Additionally, the Guttman correlation coefficient ($\mu_2$) and intra-class correlation coefficient (ICC) were used for feature validation.

Results:
The experiments on the SVM trained using the MFCC from running-speech samples produced higher AoC (84% and 85%) in classifying the severity levels of UPDRS-S as compared to the AoC produced by the MFCC from sustained-phonation (88% and 77%) and diadochokinesis (77% and 77%) samples in 10-fold cross validation and training-testing schemes respectively. The $\mu_2$ between the MFCC from running speech samples and clinical ratings was stronger ($\mu_2$ up to 0.7) than the $\mu_2$ between the clinical ratings and the MFCC from sustained-phonation and diadochokinesis samples. The ICC of the MFCC from the running-speech samples recorded in different test occasions was stronger as compared to the ICC of the MFCC from sustained-phonation and diadochokinesis samples recorded in different test occasions.

Conclusions:
The strong classification ability of running-speech MFCC and SVM, on one hand, supports suitability of this scheme to monitor speech symptoms in PD. Besides, the values of $\mu_2$ and ICC suggest that the MFCC from running-speech signals are more reliable for scoring speech symptoms as compared to the MFCC from sustained-phonation and diadochokinesis signals.

Keywords: Parkinson’s disease; Speech processing; Hypokinetic dysarthria; Support vector machine; Telemonitoring; Cepstral analysis
1. Introduction

Parkinson’s disease (PD) is caused by a progressive deterioration of dopamine producing nerve cells in the mid-brain (Olanow et al., 2009). A lack of dopamine causes a number of motor and non-motor symptoms including reduced muscular movement, tremor and speech dysfunction. PD is an incurable disease. The treatment to alleviate the PD symptoms requires an accurate assessment of the symptom status and a timely adjustment of medication. The patients are to be followed up at regular intervals, which is problematic given the physical restriction of patients to access medical facilities and the established procedures of assessment. This causes an increase in the utilization of hospital resources that result in an elevated cost of managing the disease. According to a survey (Findley, 2007), the annual cost of PD management in the UK is estimated between 449 million to 3.3 billion pounds. The annual cost per patient in the USA is around 10000 USD. Unfavorably, the medication forms 20% of this cost and the major cost is incurred in the clinical management. The number of PD patients is increasing in an aging society that predicts a lift in these figures. One solution for cost reduction is to use automated methods for the quantification of PD symptoms, which allows remote monitoring of the patients from their home environment conditions.

Speech impairment is an early onset indicator of PD and it was estimated that up to 90% of the patients develop speech symptoms with the disease progression (Harel et al., 2004). The most noticeable speech symptoms are related to breathing, phonation, articulation and prosody. These symptoms are manifested in the speech quality of the patients in the form of distorted vowels and consonants, harsh and hoarse voice quality, reduced speaking rate, hypernasality, monoloudness and monopitch. Previous research has reported a progressive deterioration of these symptoms over the course of PD (Harel et al., 2004; Holmes et al., 2000). Some investigations suggested that the degradation in speech quality and the general PD severity are strongly interlinked (Skodda et al. 2009).

Previous methods of estimating vocal impairment typically utilized sustained vowel phonation (SVP) tests, where the speaker is requested to sustain phonation for as long as possible with a steady frequency and amplitude. Admittedly, the SVP can provide symptom estimates related to the vibration of vocal folds, however, a clear picture of impairment related to the movement of vocal tract organs including the lips, nasal tract, jaw etc. cannot be captured (Krom, 1995). In order to assess the vocal tract anomalies, another vocal test called the diadochokinesis (DDK) test is utilized. In this test, the patient is required to utter /pa/-/ta/-/ka/ syllable repeated and as long and constantly as possible. Albeit, the SVP and DDK can provide estimates for screening voice pathology, they are artificial in the context of daily life communication. In both these methods, the dynamic aspects of speech such as co-articulation, onset and offset effects etc. are not present.

Besides, from the perspective of the assessment of voice quality, the text-dependent running speech (TRS) can provide much more impairment-related information as compared to the SVP and DDK as it reflects the dynamic aspects of continuous speech and is a better representative of communication (Klingholtz, 1990). Zraick et al. (2003) supported that the TRS with standard formulation has potential to exploit capacious symptoms in the PD speech, providing a broader perspective of evaluation. Despite of these advantages, a problem in estimating vocal impairment using TRS is the structural complexity in processing the speech signal due to the varying annotation of syllables in continuous speech.

Current technology facilitates remote interaction between patients and medical experts (Goetz et al. 2009). Specifically, speech can be recorded and transferred using mobile networks and can be stored in a centralized server for clinical evaluation and diagnosis. The recording of speech samples does not require additional expertise or equipment. An important advantage is that, the recordings can be processed remotely using acoustic algorithms that allow a non-invasive characterization of speech symptoms.

The mel-frequency cepstral coefficients (MFCC) are relied for their capability to estimate anomalies in pathological speech (Tsanas et al., 2012; Londono et al., 2011; Llorente et al., 2009; Gelzinis et al., 2008). Previous research on the PD speech (Jafari, 2013; Tsanas et al., 2012) used MFCC from SVP signals to classify the severity of speech symptoms according to the unified Parkinson’s disease rating scale (UPDRS) (Fahn et al., 1987). However, these methods did not evaluate MFCC from the TRS, possibly due to complexities in processing the speech signal. On the other hand, the research on vocal disorders, other than PD, used MFCC from the TRS signals to model vocal tract aberrations. For example, Llorente et al. (2009) parameterized MFCC from 140 recorded TRS samples that classified between 117 dysarthric and 23 normal samples with an accuracy of 96%. Similarly, Paja et al. (2012) used MFCC computed from the TRS samples to evaluate spastic dysarthria. Their experiments indicated a strong
correlation between the MFCC and the 2-level (‘Low-Mid’ and ‘Mid-High’) subjective ratings of speech intelligibility. In another experiment to estimate speech depression (Cummins et al., 2013), the MFCC from the TRS samples displayed the strongest discriminatory characteristics compared to other acoustic features, when classifying the presence and absence of depression.

This paper aims to utilize MFCC computed from the TRS test recordings for classification of the severity of speech symptoms according to the UPDRS motor examination of speech (UPDRS-S). A secondary objective is to perform a comparative analysis between the classification performance of MFCC, computed from the recorded TRS tests and the MFCC computed from the recorded SVP and DDK tests, in discriminating between the UPDRS-S levels. The study suggests that MFCC from the TRS tests are better classifiers of speech symptoms and that they remain more consistent and reliable on different test occasions as compared to the MFCC from SVP and DDK tests.

2. **Material and methods**

2.1Patients and data

The speech recordings in this study were obtained from a feasibility study of an at-home treatment device through a procedure described by Goetz et al. (2009). The data acquisition was conducted at the University of California, San Francisco in collaboration with Parkinson’s Institute. One-year (i.e. from June 2009 till June 2010) speech data was acquired from a total of 80 subjects, 48 males and 32 females with an average age of 63.8 years, using a computer-based test battery called QMAT. 60 subjects (40 males and 20 females) had PD duration of 75.4 weeks. 20 other subjects were normal controls.

The recorded speech samples consisted of four different types of SVP, two different types of DDK and three different types of TRS tests. In the SVP tests, the vocal breathiness of patients in keeping the pitch (e.g. ‘aaaah...’) constant for 12 seconds was examined. Four types of the SVP tests, one at the comfortable level of loudness, second with twice the level of initial loudness, third with thrice the level of initial loudness, and fourth with quadrice the level of initial loudness, were recorded. In the DDK tests, the ability of patients to produce rapid alternating speech (e.g. ‘puh-tuh-kuh...puh-tuh-kuh...’) was assessed. Two types of the DDK tests, one at the comfortable level of loudness and other with twice the level of initial loudness were recorded. In the TRS tests, the subjects were asked to recite standard phonetic paragraphs (No, 1999) displayed on the QMAT screen. The paragraphs included ‘the north wind and the sun’, ‘the rainbow passage’ and ‘the grandfather passage’ in test 1, 2 and 3 respectively. These paragraphs were structured in a way such that the level of textual difficulty in reading the paragraphs increased from test 1 to test 3 (Khan et al., 2013). The paragraph ‘the north wind and the sun’ used in the TRS test-1 consisted of 5 sentences and was recorded for 40 seconds from each subject. The paragraph ‘the rainbow passage’ used in the TRS test-2 was more difficult to read than the TRS test-1 paragraph, and consisted of 6 sentences recorded for 50 seconds. The paragraph ‘the grandfather passage’ used in the TRS test-3 was the most difficult to read. It consisted of 8 sentences and was recorded for 60 seconds.

A clinician examined the speech samples and rated the severity of symptoms using the UPDRS-S. The UPDRS-S is item 18 of part-III of the UPDRS. The levels of symptom severity in the UPDRS-S are ranged between ‘0’ and ‘4’, where 0=’normal’, 1=’slight loss of expression, diction and volume’, 2=’monotone, slurred but understandable; moderately impaired’, 3=’marked impairment, difficult to understand, and 4=’unintelligible’. Out of the 80 subjects, 24 subjects were rated ‘0’, 25 subjects were rated ‘1’, 28 subjects were rated ‘2’ and 3 subjects were rated ‘3’. The sampling rate of speech samples was 48 kHz with 16 bit resolution. In total 320 SVP, 240 TRS and 160 DDK samples were used in feature analysis and classification.

2.2 Mel-frequency cepstral coefficients

The use of MFCC for speech analysis stems from the field of speaker identification (Davis and Mermelstein, 1978). Presently, the MFCC are relied for their capability to assess pathological speech (Tsanas et al., 2012; Londono et al., 2011). The selection of MFCC for analyzing impairment in speech is supported more by empirical evidences rather
than by theoretical reasoning. The three main reasons that support this selection are: 1) the computation of MFCC does not require pitch detection, 2) the MFCC parameters are robust to frequency distortions, 3) the cepstral analysis in computing the MFCC allows the removal of noise information in the speech signal as a fact that the MFCC give a compression of this information on the first part of speech cepstrum. This provides some dimensionality reduction that eases the task of pattern classifiers.

The MFCC work by partitioning the speech frequency into overlapping mel-frequency filter banks. This is followed by the application of cepstral and cosine transformations on each bank (Davis and Mermelstein, 1978). For their ability to partition the speech frequency, the MFCC are suited to effectively quantify the potential problems of speech articulation. For instance, the speech signal is characterized not only by the vibration of vocal folds but also by the resonating frequencies formed by the controlled movement of articulators (tongue, jaw, lips etc). Besides, it is known that PD affects the movement of articulatory muscles in addition to the movement of vocal folds (Ho et al., 1998). The subtle dislocation in the movement of articulators deteriorates the intelligibility of speech signal. The MFCC compute the energy difference between the bands of speech frequency which can be used to discriminate between the varying energy levels of disturbed resonances.

The mel-frequency filter banks are triangular in shape and compute the energy spectrum around the center frequency in each individual band of speech frequency. The boundary frequencies of filter banks are uniformly spaced using the mel-scale given in equation 1.

\[ m = 1127 \ln \left(1 + \frac{f}{700}\right), \quad 0 \leq f \leq F_s \]  

(1)

Where \( F_s \) is the sampling rate of frequency \( f \) in hertz. The mel-frequency cepstrum is computed by applying the mel-frequency filter banks to the logarithm of the DFT of speech signal \( S \). These filter banks divide the frequency bands of speech log-spectrum into mel-spaces. This division between the frequency bands provides an approximation of the human auditory system response that simulates the human ear behavior in separating sounds of different frequencies (Stevens and Volkman, 1940).

In order to compute MFCC, the log-energy at the output of each filter is computed. The MFCC is the discrete cosine transform of filter energy outputs, given in equation 2.

\[ MFCC_n = \sum_{k=1}^{K} \hat{E}_k \cos[n(k - 0.5) \pi / K], \quad n = 0 \ldots L \]  

(2)

Where \( L \) is the number of MFCC, \( K \) is the number of filter banks and \( \hat{E}_k \) is the log energy of the \( k \)th filter. Typically, a value of \( L \) between 10 and 16 is used (Davis and Mermelstein, 1978). The 0th MFC coefficient represents the original signal energy and is generally ignored. The value of \( K \) is chosen between 20 and 40. The block diagram of MFCC computation is shown in figure 1.

Fig. 1. The block diagram of MFCC extraction.
2.3 Sequential minimal optimization algorithm for support vector machines

Support vector machines (SVM) have gained immense popularity in biomedical decision support systems for their ability to produce optimal training results by implementing flexible decision boundaries in high dimensional feature space (Guan, 2011). Previous classifiers used to separate classes using hyper-planes. The SVM widened this idea of separating the hyper-planes to data that cannot be separated linearly by mapping predictors (support vectors) onto a new higher-dimensional space in which the data can be separated linearly. This non-linear classification and mapping of data into high-dimensional feature space is performed by SVM using a trick called kernel. Computationally, finding the best location of hyper-planes to facilitate a kernel function, to create linear boundaries through non-linear transformation, may lead to a convex quadratic programming (QP) optimization problem. This QP optimization problem can be solved using the sequential minimal optimization (SMO) algorithm (Scholkopf et al., 2001).

Considering an n-class classification problem and a set of training vectors \( \{V_i\}_{i=1...M} \) with corresponding label \( S_i \), the SVM classifier assigns a new label \( \hat{S} \) to a test vector \( T \) by evaluating

\[
\hat{S} = \sum_{i} \alpha_i S_i K(T, V_i) + b
\]

(3)

Where, the weights \( \alpha_i \) and bias \( b \) are SVM parameters which are maximized during the SVM training. \( K(\cdot, \cdot) \) is the SVM kernel function. The maximization of weights \( \alpha_i \) in training the SVM leads to a QP optimization problem which can be expressed in a dual form as:

\[
\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j S_i S_j K(V_i, V_j)
\]

(4)

Subject to

\[
0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_i \alpha_i S_i = 0 \quad \text{for } i=1,2,...n
\]

(5)

Where, \( C \) is a positive constant called the SVM hyper-parameter that weights the influence of training errors.

SMO is an iterative algorithm that breaks the optimization problem expressed in equations 4 and 5 into a series of smallest possible sub-problems which are then solved analytically in each SMO iteration. The SMO treats SVM weights \( \alpha_i \) as Lagrange multipliers. The idea is that, for the smallest possible optimization problem involving two Lagrange multipliers \( \alpha_1 \) and \( \alpha_2 \), the linear equality constraints between the two multipliers should be reduced to

\[
0 \leq \alpha_1, \alpha_2 \leq C
\]

(6)

And

\[
S_1 \alpha_1 + S_2 \alpha_2 = k
\]

(7)

Where, \( k \) is the equality constraint variable. The SMO finds a global optimal solution by following these steps: 1) the algorithm finds a Lagrange multiplier \( \alpha_1 \) that violates Karush-Kuhn-Tucker (Kjeldsen, 2000) conditions, 2) picks a second multiplier \( \alpha_2 \) and optimize the pair \( (\alpha_1, \alpha_2) \), and 3) repeat steps 1 and 2 until convergence. The QP optimization problem is solved when all the Lagrange multipliers satisfy the Karush-Kuhn-Tucker conditions.
Albeit the SMO guarantees the convergence of SVM function, the choice of kernel function is important for transforming non-linear feature space into a straight line classification solution. The choice of function is based on the nature of feature space that can be linear, polynomial or radial basis. In our case the underlying specificity regarding the qualitative nature of data could not be determined due to a high variability of speech signals in different speech tests. To circumvent this limitation, a universal kernel function based on Pearson VII function (PUK) (Ustun et al., 2006) was utilized. The PUK is generally used to solve curve-fitting problems and has a general form given in equation 8.

\[
f(x) = \frac{H}{1 + \left[ \frac{2(x-x_0) \sqrt{2^{1/\omega} - 1}}{\sigma} \right]^{2^{1/\omega}}} \tag{8}
\]

Here \(H\) is the peak height at centre \(x_0\) of the peak and \(x\) is an independent variable. The variables \(\omega\) and \(\sigma\) control the tailing factor and half-width of the peak respectively. A curve with \(\omega\) equals to 3 and \(\sigma\) equals to 1, is comparable to a sigmoid function used in neural network modeling (Ustun et al., 2006).

Ustun et al. (2006) modified equation 8 to formulate a kernel function for SVM given in equation 9. For a given set of training vectors \(\{V_i\}_{i=1}^{M}\), the single variable \(x\) in equation 8 is replaced by two training vectors \(V_i\) and \(V_j\). A Euclidean distance between these vectors is introduced so that two identical training vectors would have a zero distance. The peak height \(H\) is replaced by 1 and the peak-offset \(x_0\) is removed. The SVM configured with the PUK kernel function and optimized by the SMO algorithm was trained using the MFCC to classify the severity of speech symptoms based on the UPDRS-S.

\[
k(V_i, V_j) = \frac{1}{1 + \left( \frac{2\sqrt{|V_i - V_j|}}{\sqrt{2^{1/\omega} - 1}} \right)^{2^{1/\omega}}} \tag{9}
\]

3. Experiments, results and analysis

It can be speculated that a higher demand of linguistic stress required to perform a vocal test may result in a higher precision of estimating speech symptoms (Khan et al., 2013) as a fact that the vocal muscles in PD may not cope with the demanding level of stress. Among the different speech tests used, the level of linguistic stress in TRS test-3 was higher than TRS test 1 and 2. Likewise, the level of stress in SVP test-4 was higher than SVP tests 1, 2 and 3. The level of stress in DDK test-2 was higher than DDK test-1. For the sake of comparison, the representative samples of TRS test-3, SVP test-4 and DDK test-2, rated ‘0’ (normal) and ‘3’ (severely impaired) by a clinician, are shown in figure 2.

Notable difference can be observed in the acoustic properties of samples rated ‘normal’ and ‘severely impaired’. A reduction in amplitude can be seen in the waveforms of ‘severely impaired’ TRS and DDK samples as compared to the ‘normal’ TRS and DDK samples. The articulation rate was low in the impaired TRS waveform, i.e. the speaker paused 19 times between utterances as compared to the normal speaker who paused only 7 times. Also, the pauses between utterances were much longer in the impaired TRS, suggesting that the speaker had to rest on several occasions to re-initiate recitation. In comparison to TRS and DDK, the waveform of SVP sample rated ‘severely impaired’ depicts that the speaker indicated some difficulty in initiating phonation, but later on, she phonated normally. Although, slightly higher amplitude was noticed in the initial part of the impaired SVP waveform, this amplitude reduced with the continued phonation.

The MFCC were computed from the recorded speech tests using equation 2. In order to compute the MFCC, the speech signals were divided into the frames of 50 ms each. A filter bank of \(K=24\) was applied to extract up to \(16^{th}\) order \((L=16)\) MFCC from each frame. The mean of each order of the MFCC between the frames was computed and used for analysis. A comparison between the MFCC, computed from the recorded samples of TRS test-3, SVP test-4...
and DDK test-2, rated ‘0’ (normal) and ‘3’ (severely impaired) by a clinician, is shown in figure 2. A marked difference can be noticed between the amplitudes of MFCC computed from the waveforms of TRS tests rated ‘normal’ and ‘severely impaired’. By contrast, the amplitudes of MFCC computed from the waveforms of the SVP and the DDK tests, rated ‘normal’ and ‘severely impaired’, were not very dissimilar, although there were some differences in the sign of amplitudes from positive to negative.

<table>
<thead>
<tr>
<th>Clinical rating</th>
<th>1. TRS</th>
<th>2. SVP</th>
<th>3. DDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Normal</td>
<td>[Waveform Image]</td>
<td>[Waveform Image]</td>
<td>[Waveform Image]</td>
</tr>
<tr>
<td></td>
<td>[MFCC Image]</td>
<td>[MFCC Image]</td>
<td>[MFCC Image]</td>
</tr>
<tr>
<td>b. Severely impaired</td>
<td>[Waveform Image]</td>
<td>[Waveform Image]</td>
<td>[Waveform Image]</td>
</tr>
<tr>
<td></td>
<td>[MFCC Image]</td>
<td>[MFCC Image]</td>
<td>[MFCC Image]</td>
</tr>
</tbody>
</table>

Fig. 2. A comparison between the MFCC computed from the representative samples of TRS, SVP and DDK tests is shown. A marked difference can be noticed between the MFCC from the TRS samples marked ‘normal’ (1.A.ii) and ‘severely impaired’ (1.B.ii) by the clinician.

3.1 Correlation analysis

The intelligibility of speech can be disturbed by dysfunctions in a number of speech components including respiration, phonation, articulation and prosody. Previous research suggests that impairment of phonation is the most common manifestation in PD speech (Midi et al., 2008), albeit PD causes disruption in the speech production system.
as a whole. For instance, hoarseness and harshness are common symptoms of phonation in the PD speech. Besides, the PD symptoms of articulation involve short rushes of speech, articulation blurring, improper consonant articulation etc. (Ho et al., 1998). Sometimes it is possible that two speech samples are given a similar severity rating of symptoms by a clinician, yet these samples have anomalies in different speech components. For example, a strong symptom of vocal harshness in one sample may exist weakly in another sample rated similar but which has strong articulation blurring. Another issue is that some symptoms may likely remain unnoticed by the clinician as these symptoms may not interfere so much with speech intelligibility. For example, hoarseness (‘soft voice’) may remain a consistent attribute in the PD speech, but harshness (‘rough voice’) may cause more of a disruption in the intelligibility of speech perceived by the clinician (Khan et al., 2013). All these symptoms are estimated using different acoustic features. For instance, the features of ‘cepstral separation difference’ (Khan et al., 2013) can be used to estimate the harsh and hoarse quality of speech. On the other hand, the MFCC can be used to measure the symptoms of articulation (Tsanas et al., 2012). In the given situation, it is possible that feature quantities do not follow a monotonic trend relative to their corresponding symptom severity levels 0’ (normal) to ‘4’ (unintelligible) in the UPDRS-S. This complicates the choice of correlation model.

In the given situation, the one-to-one mapping between a computerized feature (representing an individual speech symptom) and the corresponding clinical rating (based on multiple symptoms) is not possible through the use of rank-order or Likert scale. One may choose Spearman rho which utilizes the rank-order scale to correlate between two variables in an ordered dataset. Then a restriction with the rank-order scale is that, an agreement between two variables on one class level is strictly based upon the agreement between these variables on the former class level. If an agreement between a rater and a computerized feature has to be made in a succeeding class level (say level ‘3’) based on a class property (speech symptom) which does not exist in the former class level (say level ‘2’) but exists in levels preceding to the former level (say levels ‘0’ and ‘1’). In that case the Spearman rho would penalize the correlation value since the agreement between the rater and the feature value in the succeeding class level (level ‘3’) could not be reached due to the absence of this agreement in the former class level (level ‘2’).

This problem of monotonicity in feature quantities can be solved using the Guttman scale (Guttman, 1944). The Guttman correlation model matches the ranked nature of speech dataset where a human rater scores between UPDRS-S levels depending upon the perceived intelligibility of speech that is affected by the proportions of different speech symptoms each estimated by a different acoustic feature. The Guttman monotonicity coefficient ($\mu_2$) expresses an increase in variable x (say a computed speech feature) relative to an increase in variable y (say UPDRS-S levels) without assuming that the intervals between the values of y are perfectly scaled. This is advantageous because the ties between x and y can be untied in the same order without penalty giving an adequate measure of correlation. A $\mu_2$ value equals to +1 or -1 depicts perfect correlation or anti-correlation between x and y respectively. The $\mu_2$ can be computed using equation 10.

$$\mu_2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_h - x_i)(y_h - y_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_h - x_i)(y_h - y_j)}$$

(10)

Where, $h$ is the order of UPDRS-S levels and $i$ is the corresponding order of feature values relative to $h$.

The $\mu_2$ was utilized to map between the MFCC computed from TRS, SVP and DDK samples, and the clinical ratings based on UPDRS-S. As there were only three patients rated ‘3’ (severely impaired) by the clinician, the samples related to this group of patients were merged into the patient group rated ‘2’ (moderately impaired). The jackknifing (leave-one out) cross validation (CV) (Berger, 2007) was used to stratify correlation estimates. The jackknife estimates of $\mu_2$ between MFCC and UPDRS-S are listed in table 1.

The 3rd, 4th, 7th, 8th, 9th, 10th and 12th order MFCC showed strong ($\mu_2 > 0.5$) statistically significant ($p<0.05$) correlation with the UPDRS-S ratings in all the TRS tests. Specifically, the 4th MFCC produced the highest correlation estimates in TRS test-1 ($\mu_2 = 0.60$, $p = 0.004$), test-2 ($\mu_2 = 0.67$, $p = 0.001$) and test-3 ($\mu_2 =0.70$, $p < 0.0001$) respectively, as compared to the other TRS MFCC and the MFCC from SVP and DDK. Importantly, the value of $\mu_2$ between the 4th MFCC and UPDRS-S increased and the p-value decreased from TRS test-1 to test-3,
suggesting that a higher demand of linguistic stress required to perform the vocal test may result in a higher precision of estimating speech symptoms.

Likewise, the jackknife estimates of $\mu_2$ between the 4th order MFCC from the SVP samples and the UPDRS-S ratings were strong in SVP test-3 ($\mu_2 = 0.73, p = 0.001$) and test-4 ($\mu_2 = 0.70, p = 0.002$). However, this correlation did not show a monotonic increase relative to the increasing level of loudness from SVP test 1 to test 4. On the other hand, except for the 6th and the 16th order MFCC, no significant correlation was found between the other MFCC and the UPDRS-S ratings. Similarly, the MFCC computed from the DDK samples did not show a correlation with the UPDRS-S, except for the 3rd order MFCC that showed a strong correlation in DDK test-1 ($\mu_2 = 0.61, p = 0.001$) and test-2 ($\mu_2 = 0.57, p = 0.003$) respectively.

Table 1. The jackknife estimates of $\mu_2$ between the UPDRS-S ratings and the MFCC from SVP, TRS and DDK tests. Estimates in bold represent strong (>0.5) statistically significant (p<0.05) correlation. Estimates in italic represent improving correlation relative to the increasing difficulty of vocal test.

<table>
<thead>
<tr>
<th>MFCC</th>
<th>SVP test-1</th>
<th>SVP test-2</th>
<th>SVP test-3</th>
<th>SVP test-4</th>
<th>DDK test-1</th>
<th>DDK test-2</th>
<th>TRS test-1</th>
<th>TRS test-2</th>
<th>TRS test-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC-1</td>
<td>0.066</td>
<td>-0.083</td>
<td>-0.309</td>
<td>-0.31</td>
<td>-0.042</td>
<td>0.078</td>
<td>0.074</td>
<td>0.115</td>
<td>-0.046</td>
</tr>
<tr>
<td>MFCC-2</td>
<td>0.222</td>
<td>0.067</td>
<td>-0.024</td>
<td>0.079</td>
<td>-0.166</td>
<td>-0.059</td>
<td>0.114</td>
<td>0.21</td>
<td>0.221</td>
</tr>
<tr>
<td>MFCC-3</td>
<td>-0.097</td>
<td>-0.016</td>
<td>0.335</td>
<td>0.368</td>
<td><strong>0.609</strong></td>
<td><strong>0.572</strong></td>
<td>0.484</td>
<td><strong>0.635</strong></td>
<td><strong>0.609</strong></td>
</tr>
<tr>
<td>MFCC-4</td>
<td>0.081</td>
<td>0.317</td>
<td><strong>0.732</strong></td>
<td><strong>0.701</strong></td>
<td>0.366</td>
<td>0.177</td>
<td><strong>0.6</strong></td>
<td><strong>0.672</strong></td>
<td><strong>0.698</strong></td>
</tr>
<tr>
<td>MFCC-5</td>
<td>-0.352</td>
<td>-0.256</td>
<td>-0.21</td>
<td>-0.173</td>
<td>0.165</td>
<td>0.123</td>
<td>0.05</td>
<td>0.148</td>
<td>0.346</td>
</tr>
<tr>
<td>MFCC-6</td>
<td>0.03</td>
<td>-0.144</td>
<td><strong>-0.526</strong></td>
<td><strong>-0.528</strong></td>
<td>-0.471</td>
<td>-0.289</td>
<td>-0.209</td>
<td>-0.226</td>
<td>-0.17</td>
</tr>
<tr>
<td>MFCC-7</td>
<td>0.372</td>
<td>0.391</td>
<td>0.301</td>
<td>0.362</td>
<td>-0.463</td>
<td>-0.328</td>
<td>-0.432</td>
<td><strong>0.596</strong></td>
<td><strong>-0.537</strong></td>
</tr>
<tr>
<td>MFCC-8</td>
<td>-0.119</td>
<td>-0.068</td>
<td>0.098</td>
<td>-0.066</td>
<td>-0.299</td>
<td>-0.014</td>
<td><strong>-0.61</strong></td>
<td><strong>-0.661</strong></td>
<td><strong>-0.616</strong></td>
</tr>
<tr>
<td>MFCC-9</td>
<td>-0.189</td>
<td>0.189</td>
<td>0.115</td>
<td>0.126</td>
<td>-0.425</td>
<td>-0.2</td>
<td><strong>-0.614</strong></td>
<td><strong>-0.581</strong></td>
<td><strong>-0.598</strong></td>
</tr>
<tr>
<td>MFCC-10</td>
<td>-0.018</td>
<td>0.196</td>
<td>0.176</td>
<td>0.211</td>
<td>0.214</td>
<td>0.16</td>
<td><strong>-0.583</strong></td>
<td><strong>-0.575</strong></td>
<td><strong>-0.636</strong></td>
</tr>
<tr>
<td>MFCC-11</td>
<td>-0.199</td>
<td>0.125</td>
<td>-0.146</td>
<td>-0.302</td>
<td>0.069</td>
<td>-0.018</td>
<td>-0.19</td>
<td>-0.216</td>
<td>0.101</td>
</tr>
<tr>
<td>MFCC-12</td>
<td>-0.244</td>
<td>-0.136</td>
<td>-0.445</td>
<td>-0.33</td>
<td>-0.148</td>
<td>-0.172</td>
<td>-0.427</td>
<td>-0.426</td>
<td>-0.405</td>
</tr>
<tr>
<td>MFCC-13</td>
<td>-0.006</td>
<td>0.065</td>
<td>0.023</td>
<td>-0.117</td>
<td>-0.168</td>
<td>-0.198</td>
<td>-0.111</td>
<td>-0.183</td>
<td>0.027</td>
</tr>
<tr>
<td>MFCC-14</td>
<td>-0.067</td>
<td>-0.156</td>
<td>-0.117</td>
<td>-0.116</td>
<td>-0.498</td>
<td>-0.195</td>
<td>-0.208</td>
<td>-0.323</td>
<td>-0.237</td>
</tr>
<tr>
<td>MFCC-15</td>
<td>0.331</td>
<td>0.104</td>
<td>0.194</td>
<td>0.083</td>
<td>-0.019</td>
<td>0.083</td>
<td>-0.275</td>
<td>-0.34</td>
<td>-0.497</td>
</tr>
<tr>
<td>MFCC-16</td>
<td>-0.388</td>
<td><strong>-0.502</strong></td>
<td>-0.466</td>
<td><strong>-0.511</strong></td>
<td>0.12</td>
<td>0.013</td>
<td>-0.016</td>
<td>-0.077</td>
<td>-0.218</td>
</tr>
</tbody>
</table>

### 3.2 Test-retest reliability

In order to consider a feature to be reliable, the values of the feature must be consistent in different test settings and test occasions for a similar class level (i.e. in our case, the level of symptom severity). The intra-class correlation coefficient (ICC) (McGraw and Wong, 1996) is generally used to assess the consistency by quantifying the degree to which the feature values resemble each other at similar class levels in different test occasions. The ICC is computed using single and average measurements. In the single measurements, a class value of feature in one test is mapped to a corresponding class value of that feature in another test to estimate consistency. In the average measurements, a class value of feature in one test is mapped with all the class values of that feature in the other test to estimate consistency.

The single and average ICC measures were computed for the MFCC computed from the samples of TRS, SVP and DDK tests recorded at different test occasions. In order to compute these measures, the MFCC in the order 1 to 16 were computed separately from the samples of TRS test 1, 2 and 3 respectively and were merged to form 16
different TRS-MFCC groups, such that each group consisted of an order of MFCC computed from the three different test occasions (test 1, 2 and 3) of TRS. Likewise, the MFCC in the order 1 to 16 were computed separately from the samples of SVP test 1, 2, 3 and 4, and were merged to form 16 SVP-MFCC groups. Similarly, the MFCC in the order 1 to 16 were computed separately from the samples of DDK test 1 and 2, and were merged to form 16 DDK- MFCC groups. A comparison of single and average ICC measures between each MFCC group is given in figure 3.

![Intra-class correlation coefficients](image)

**a. Single measures**

**b. Average measures**

---

Fig. 3. Intra-class correlation coefficients. A comparison of ICC between the MFCC groups computed from SVP, DDK and TRS tests, is given. Notice that the ICC of MFCC groups from TRS tests are higher than the ICC of MFCC groups from SVP and DDK tests.

The MFCC groups computed from the TRS test occasions showed the highest ICC, both in the single and average measurements, as compared to the MFCC groups computed from the SVP and DDK test occasions. The 8th order MFCC group from TRS showed the highest ICC, in both single (ICC = 0.94) and average (ICC = 0.98) measurements. The mean between the single ICC measures of MFCC groups were 0.64, 0.80 and 0.87 for the test occasions of SVP, DDK and TRS respectively. Similarly, the mean between the average ICC measures of MFCC groups were 0.87, 0.89 and 0.95 for the test occasions of SVP, DDK and TRS respectively.

The MFCC groups from the test occasions of SVP were the most inconsistent. One reason for this inconsistency is that, the change in the level of loudness and pitch during sustained phonation is inter-dependent (Gramming et al., 1988). The change in the level of loudness on different test occasions of SVP, as well as the change in the level of loudness due to voice impairment, may shift the fundamental frequency in the frequency spectrum (figure 4ci and 4cii) resulting in varying quantities of MFCC among different bands of mel-frequency filters. Also, the fundamental frequency alters between speakers of different gender (Gelfer and Bennett, 2013). By contrast, the TRS does not encounter this problem, because the frequencies of different sounds are uniformly spread along the frequency spectrum (figure 4ai) and any disturbance related to the impairment in these frequencies can be captured using the MFCC. Likewise, the frequencies of DDK resulting from the repeated utterance of ‘puh-tuh-kuh’ are quasi uniformly spread on the frequency spectrum (figure 4bi); however they are not as uniform as the frequencies of the TRS (figure 4ai). A comparison between the log-spectrums of TRS, SVP and DDK, rated ‘0’ and ‘3’ is shown in figure 4.
Clinical rating | a.TRS | b.DDK | c.SVP
---|---|---|---
‘0’ Normal | ![Normal TRS log-spectrum](image) | ![Normal DDK log-spectrum](image) | ![Normal SVP log-spectrum](image)
a. i. Normal TRS (test-3) log-spectrum | b. i. Normal DDK (test-2) log-spectrum | c. i. Normal SVP (test-4) log-spectrum
‘3’ Severely impaired | ![Impaired TRS log-spectrum](image) | ![Impaired DDK log-spectrum](image) | ![Impaired SVP log-spectrum](image)
a. ii. Impaired TRS (test-3) log-spectrum | b. ii. Impaired DDK (test-2) log-spectrum | c. ii. Impaired SVP (test-4) log-spectrum

Fig. 4. A comparison between the log power spectrums of representative TRS, DDK and SVP samples, rated ‘0’ and ‘3’ is shown. The irregular power levels of TRS (a. ii) and DDK (b. ii) spectrums rated ‘3’ can be compared to the smoother power levels of TRS (a. i) and DDK (b. i) spectrums rated ‘0’. By contrast, this comparison is difficult between the SVP samples rated ‘0’ (c. i) and ‘3’ (c. ii) due to power level shifts in SVP spectrum rated ‘0’.

### 3.3 Classification

The SVM classifier optimized by the SMO algorithm (section 2.3) and configured with the PUK kernel function (of $\sigma=1$ and $\omega=3$) was used for the classification of speech data. As there were only 3 patients rated ‘3’ (severely impaired) by the clinician; in order to avoid a high standard error pertaining to a low sample size in class ‘3’ and to balance the overall class distribution, the samples related to this group of patients were merged into the patient group rated ‘2’ (moderately impaired). This left behind 3 levels of symptom severity for classification, where class ‘0’ (normal) consisted of 24 subjects, class ‘1’ (mild impairment) consisted of 25 subjects and class ‘2’ (moderate-severe impairment) consisted of 31 subjects.

The MFCC computed from the samples of TRS, DDK and SVP tests were used separately in different classification experiments to discriminate between the levels of symptom severity. The clinical UPDRS-S ratings were used as classification targets. Two different classification experiments were performed. In the first experiment, data were stratified using 10 fold CV to obtain unbiased generalization estimates of MFCC in different speech tests. In the second experiment, data were separated between training and testing sets to validate the generalization performance of MFCC in different speech tests, with a statistical assumption that the MFCC used in the testing data will have a similar distribution to MFCC used in training the classifier. For optimal results, the SVM hyper-parameter $C$ was tuned between 1 and 20. A constant value of $C=15$ produced the best generalization results and was maintained in all classification experiments. The results are given in table 2.

In the first experiment, the MFCC computed from the samples of TRS test 1, 2 and 3 were merged together to form a classification matrix with the dimensions of 16 MFCC x 240 samples (80 patients x 3 TRS tests). Data stratification with 10-fold CV on this input vector in SVM produced an overall classification accuracy of 78% with the true positive rate (TPR) of 80%, 60% and 90% for class ‘0’, ‘1’ and ‘2’ respectively (table 2). In a similar test on the MFCC computed from the samples of DDK tests, the 10-fold CV on a classification matrix of the dimensions of
16 MFCC x 160 samples (80 patients x 2 DDK tests) produced a classification rate of 66% with the TPR of 54%, 75% and 68% for class ‘0’, ‘1’ and ‘2’ respectively. Likewise, in a further classification test on the MFCC computed from the samples of SVP tests, the 10-fold CV on an input vector of the dimensions of 16 MFCC x 320 samples (80 patients x 4 SVP tests) produced a classification rate of 83% with the TPR of 77%, 80% and 90% for class ‘0’, ‘1’ and ‘2’ respectively.

Table 2. A comparison between the performance of SVM in classifying the severity levels of UPDRS-S when using the MFCC from the samples of TRS, SVP and DDK tests separately.

<table>
<thead>
<tr>
<th>Classification Type</th>
<th>Vocal Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TRS</td>
</tr>
<tr>
<td></td>
<td>Total samples = 240 (samples from TRS test 1, 2 &amp; 3)</td>
</tr>
<tr>
<td>10 fold cross validation</td>
<td>TPR class ‘0’=80%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘1’=60%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘2’=90%</td>
</tr>
<tr>
<td></td>
<td>Classification accuracy = 78%</td>
</tr>
<tr>
<td>Training</td>
<td>Total samples = 160 (samples from TRS test 1 &amp; 2)</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘0’=82%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘1’=83%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘2’=97%</td>
</tr>
<tr>
<td></td>
<td>Classification accuracy = 88%</td>
</tr>
<tr>
<td>Testing</td>
<td>Total samples = 80 (samples from TRS test 3)</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘0’=64%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘1’=88%</td>
</tr>
<tr>
<td></td>
<td>TPR class ‘2’=71%</td>
</tr>
<tr>
<td></td>
<td>Classification accuracy = 74%</td>
</tr>
</tbody>
</table>

In the second experiment, the MFCC from the samples of TRS test 1 and 2 were selected to form a training set matrix of the dimensions of 16 MFCC x 160 TRS samples. The SVM was trained using this matrix against the UPDRS-S targets ‘0’, ‘1’ and ‘2’. Another set of MFCC computed from the samples of TRS test-3 was used to form a test set matrix of the dimensions of 16 MFCC x 80 TRS samples. This test set matrix was then used for testing the trained classifier. An accuracy of 74% was achieved by this scheme in classifying the test set between the three levels of severity in UPDRS-S with the TPR of 64%, 88% and 71% for class ‘0’, ‘1’ and ‘2’ respectively.

In the same way, the MFCC from the samples of DDK test-1 were selected to form a training set matrix of the dimensions of 16 MFCC x 80 DDK samples. This matrix was used to train the SVM classifier. The MFCC from the samples of DDK test-2 were used to form a test set matrix of the dimensions of 16 MFCC x 80 DDK samples. Testing the classifier on the test set produced a classification accuracy of 66% with the TPR of 60%, 71% and 68% for class ‘0’, ‘1’ and ‘2’ respectively.

Likewise, the MFCC from the samples of SVP test-1 and test-2 were selected to form a training set matrix of the dimensions of 16 MFCC x 160 samples. This matrix was used to train the SVM classifier. The MFCC from the samples of SVP test-3 and test-4 were selected to form a test set matrix of the dimensions of 16 MFCC x 160 samples. Testing the classifier on the test set produced a classification rate of 68% with the TPR of 62%, 67% and 74% for class ‘0’, ‘1’ and ‘2’ respectively.
In both these classification experiments, first using the 10-fold CV and second using the training and testing sets, the MFCC from the samples of TRS tests on average produced the highest accuracy (76%) in classifying the three levels of UPDRS-S as compared to the MFCC from the SVP and the DDK samples. Importantly, the classification rates produced by the MFCC from the TRS samples in both experiments were consistent. On the other hand, the 10 fold CV on the MFCC from SVP samples although produced a high classification rate (83%). But this rate drastically reduced (68%) when the MFCC from the SVP tests with varying levels of loudness were used to train and test the classifier, supporting the fact that the change in loudness levels and fundamental frequency in SVP causes inconsistency in the quantities of the MFCC that makes them less suitable to be used for voice assessment. Besides, the classification rates produced by the MFCC from the DDK samples were the lowest.

3.4 ROC analysis

The area under the receiver operating characteristic (ROC) curves (AoC) is generally used to assess the feasibility of a classification model independent of cost context and class distribution (Metz, 1978). An ROC curve can be plotted by taking the true positive rate of a class on y-axis against the false positive rate of that class on x-axis. Each predicted instance of the class represents one point in the ROC space. The best classification model produces a point in the upper left coordinate (0, 1) of the ROC space, which means that there is no false negative or false positive in classification. A point along a diagonal line between the coordinates (0, 0) and (1, 1) represents a ‘complete random guess’ by the model. This diagonal line divides the ROC space into two halves. An ROC curve above this line represents an AoC greater than 50% of the ROC space area. An AoC of 100% represents a ‘perfect classification model’, whereas an AoC of less than 50% represents a ‘worthless model’. An AoC between 80% and 100% represents an ‘excellent classification model’. 

Table 3. The ROC analysis.

<table>
<thead>
<tr>
<th>Classification type</th>
<th>TRS</th>
<th>DDK</th>
<th>SVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-fold cross validation</td>
<td>ROC Curves: TRS 10-fold CV</td>
<td>ROC Curves: DDK 10-fold CV</td>
<td>ROC Curves: SVP 10-fold CV</td>
</tr>
<tr>
<td>Area under ROC curve</td>
<td>Class 0</td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Class 0</td>
<td>86%</td>
<td>82%</td>
<td>87%</td>
</tr>
<tr>
<td>Area under ROC curve</td>
<td>Class 0</td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Class 0</td>
<td>84%</td>
<td>84%</td>
<td>83%</td>
</tr>
</tbody>
</table>
In the two classification experiments, first using the 10 fold CV and second using the training-testing sets, the MFCC from the TRS samples produced the average AoC of 85% and 84% respectively (table 3), suggesting that this scheme can yield an ‘excellent classification model’ for categorizing the severity of speech symptoms. Importantly, the AoC for all severity classes in both experiments were above 80%, indicating the indubious distinction of samples in each severity class. Moreover, the values of AoC for each severity class were consistent in both experiments.

By contrast, although the MFCC from the SVP samples produced a high average AoC (88%) in 10 fold CV scheme (table 3). This value decreased to 77% when testing and training sets were utilized, suggesting again that the change in the loudness levels of phonation may demote the performance of MFCC in categorizing the severity of speech symptoms. Besides, although the values of average AoC produced by the MFCC from the DDK samples were consistent (77%) in both classification experiments, the results do not suggest that the MFCC from the DDK samples together with the SVM can yield an ‘excellent classification model’ for classifying the severity of speech symptoms.

4. Discussion

Previous research that utilized SVP supported the feasibility of MFCC for scoring voice symptoms in PD. However, the use of mel-filter banks for evaluating phonation frequencies in a relatively broad spectrum of speech is suboptimal. The mel-filters simulate the human auditory system. This suggests that the MFCC from the TRS are more suited for the quantification of speech symptoms as they allow estimating the clinician’s (hearing) perception of the patient’s speech that can be mapped with the clinician’s rating of the severity of speech symptoms. Additionally, the use of MFCC from the TRS opens the possibility for a clinician to make a decision about the presence or absence of pathologies from all voiced sounds present in speech.

Our experiments suggest that the performance of the MFCC from the SVP signals in classifying the severity levels of speech symptoms was demoted due to the varying level of loudness in different SVP tests. The ICC values of the MFCC groups were low, suggesting that the quantities of the MFCC were not consistent between different test occasions. Also, the Guttman correlation was weak between the MFCC and the UPDRS-S. Further, the ROC analysis of classification schemes involving SVM and MFCC from the SVP samples produced varying values of AoC in the 10-fold CV and training-testing experiments. These findings suggest that the recordings of SVP must be standardized in terms of loudness level and the length of phonation to enable an accurate estimation of voice symptoms. Besides, this standardization is difficult because the perception of loudness is different for different speakers. For instance, a vocalist can phonate in a louder voice than a non-vocalist. One possible solution is to use a sound analyzer to maintain a standard level of loudness during recording.

On the other hand, the MFCC from the DDK samples were consistent in classifying the severity levels of UPDRS-S i.e. the value of the average AoC was reproducible in the 10-fold CV and training-testing experiments. Also, the values of ICC measurements suggest that the quantities of these MFCC were consistent between different test occasions. However, the classification performance of these MFCC was the lowest as compared to the MFCC from the TRS and SVP samples. Also, the Guttman correlation between these MFCC and the UPDRS-S was low.

Our experiments on the MFCC computed from the TRS samples suggest that a significant amount of impairment related information can be obtained using the TRS as compared to using SVP and DDK. For instance, the Guttman analysis showed that a total of seven MFCC from the TRS samples (3rd, 4th, 7th, 8th, 9th, 10th and 12th) were strongly significantly correlated ($\mu_2 > 0.5, p<0.05$) with the UPDRS-S ratings, which is a higher number of MFCC as compared to the number of strongly correlated MFCC from the SVP and the DDK samples. Additionally, the strong values of ICC measurements suggest that the quantities of MFCC from the TRS samples were consistent between different test occasions. Importantly, the classification scheme using the SVM and the MFCC generated high levels of AoC (>84%) in the 10-fold CV and training-testing experiments, suggesting that this classification scheme is an ‘excellent model’ for categorizing the severity levels of UPDRS-S.

To our knowledge, prior to this research, the MFCC from the recorded samples of TRS tests were never computed for categorizing the severity levels of speech symptoms in PD. However, it should be bear in mind that for a more precise estimation of speech symptoms, a more comprehensive set of features is required for estimating symptoms in
the sub-systems of speech. For instance, the symptoms of prosody are generally assessed using pause rate or deviation in the fundamental frequency. The MFCC should be combined with other speech features to obtain a complete status of speech pathology.

A limitation associated with this work is that, the speech tests were recorded in a silent room. In a real-life environment, the processing of noisy signals to quantify speech symptoms could be challenging. Another drawback of using the TRS is that, the analysis of text is language dependent. It is possible that the MFCC from the spoken text of a different language may exhibit a different behavior in classifying the severity levels of speech symptoms. There is a need to validate the classification performance of MFCC in languages other than English. Despite of these limitations, the high classification performance obtained using the MFCC from the TRS samples in combination with the SVM classifier supports the suitability of this scheme to be used for categorizing the severity levels of speech symptoms in PD.

Acknowledgment

This research is a module of the project “PAULINA”. This project has been running in Dalarna University in collaboration with Abbott Laboratories and Nordforce Technology and is funded by a grant from Swedish Knowledge Foundation.

References

Guan, W., 2011. New support vector machine formulations and algorithms with application to biomedical data analysis, PhD thesis. Georgia Institute of Technology.
Guttman, L., 1944. A basis for scaling qualitative data. American sociological review 9(2), 139-150.


