Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech

diagnostics

MDPI

Article

Assessing Parkinson's Disease at Scale using Telephone-recorded Speech: Insights from the Parkinson's Voice Initiative

Siddharth Arora[1] and Athanasios Tsanas[2,*]

[1] Somerville College, University of Oxford, Oxford, OX2 6HD, UK; Siddharth.Arora@maths.ox.ac.uk
[2] Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, EH16 4UX, UK; Athanasios.Tsanas@ed.ac.uk
* Correspondence: Athanasios.Tsanas@ed.ac.uk

Abstract (max 200 words): Numerous studies have reported high accuracy using voice tasks for remote detection and monitoring of Parkinson's Disease (PD). Most of these studies, however, report findings on a small number of voice recordings, often collected under acoustically controlled conditions and therefore cannot scale at large. In this study, we aimed to evaluate the potential of using voice as a population-based PD screening tool in resource-constrained settings. Using the standard telephone network, we processed 11942 sustained vowel /a/ phonations from a US-English cohort comprising 1078 PD and 5453 control participants. We characterized each phonation using 304 dysphonia measures to quantify a range of vocal impairments. Given this is a highly unbalanced problem, we used the following strategy: selected a balanced subset (n=3000 samples) for training and testing using 10-fold cross-validation (CV), and the remaining (unbalanced held-out dataset, n=8942) samples for further model validation. Using robust feature selection methods we selected 27 dysphonia measures to present into a radial-basis-function support vector machine and demonstrated differentiating PD participants from controls with 67.43% sensitivity and 67.25% specificity. These findings could help pave the way forward towards the development of an inexpensive, remote, and reliable diagnostic support tool for PD using voice as a digital biomarker.

Keywords: Acoustic measures; biomarker; clinical decision support tool; dysphonia measures; Parkinson's disease; sustained vowel phonations; telemonitoring.

Citation: Arora, S.; Tsanas A. Title. *Diagnostics* **2021**, *11*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

1. Introduction

Neurological diseases strain health systems and pose a considerable ongoing burden on healthcare resources. Parkinson's Disease (PD) has been reported as one of the fastest-growing neurological disorders in terms of prevalence and deaths [1]. A large global burden of disease study identified PD as one of the top 5 leading causes of death from neurological disorders in the US [2]. It is estimated there were approximately 6.1 million people with PD (PwP) globally in 2016 indicating a sharp upward trend compared to 2.5 million PwP in 1990 [1].

Diagnosis of PD requires subjective assessment in-clinic, which incurs logistical costs. Crucially, consultant neurologists might misdiagnose PD up to around 20% of the total cases, while the symptom monitoring accuracy is inherently limited from the intra- and inter-rater variations in the standard clinical scales used to assess PD symptoms severity [3, 4]. Given the current objective constraints and limitations with subjective assessments, there is an urgent and unmet need for developing diagnostic support tools for the objective detection and monitoring of PD.

Parkinson's disease is a neurodegenerative disease that is characterized by four cardinal signs: tremor, bradykinesia, rigor, and postural instability [5]. The vast proportion of PwP also report speech performance degradation as one of the PD symptoms [6]. It is due to this reason, that the potential of capitalizing on acoustic analysis of speech signals to develop PD decision support tools has been pursued vigorously with considerable success over the last 10-15 years. Encouragingly, using voice, studies have proposed technologies based on acoustic analyses to: (1) Differentiate PD from controls [7-10], (2) Monitor the symptom severity of PD [11-14], (3) Assess voice rehabilitation in PD [15], (4) Identify at-risk participants (i.e., those with isolated Rapid Eye Movement (REM) sleep Behavior Disorder as confirmed by a polysomnography test)[16], (4) Identify participants with a higher genetic predisposition for developing PD (i.e., those with a mutation in the Leucine-Rich Repeat Kinase 2 (*LRRK2*) gene) [17], and (5) Predict a range of clinical scores that quantify participants' motor symptoms, cognition, daytime sleepiness, depression, and the overall state of health [18]. A limitation of these studies was, however, that they typically rely on using *high-quality* voice recordings for the analyses which are collected under acoustically carefully controlled conditions with high-end specialized equipment.

Recently, to assess the scalability of voice as a population screening tool for PD, we undertook the largest PD characterization study employing telephone-quality voice [19], which we refer to as the Parkinson's Voice Initiative (PVI) study (Arora et al. 2019). PVI is the first of its kind large scale study collecting speech data from PwP and control participants under free-living acoustic conditions. Using sustained vowel phonations (International phonetic alphabet /a:/) collected from participants in 7 countries, [19] sought to discriminate PD participants from controls using phonations collected under non-acoustically controlled conditions

The use of sustained phonations for quantifying vocal impairment is well established [20, 21]. However, our understanding of variations in dysphonia measures/sustained phonations from participants with different linguistic backgrounds is still rather limited. Historically, the use of sustained vowels has been motivated by the fact that they can be considered generic (certain vowels such as /a/ are met across different languages) and hence the processing of sustained vowel phonations overcomes linguistic differences [20]. In their analyses, Arora et al. (2019) [19] relied on the underlying assumption that sustained vowel phonations are considered generalizable across people from different linguistic backgrounds pooling together all the data from PVI. Tsanas and Arora (2021) [22] investigated the differences in dysphonia measures between UK- and US-English speaking PwP, and reported that although there is an excellent agreement between classical acoustic measures (such as jitter and shimmer), there are pronounced differences in some of the more advanced acoustic measures between the two cohorts. Given that phonations may be language-dependent, this prompts the further question of whether acoustic analyses should be performed separately for participants from different linguistic backgrounds, along with undertaking cross-cohort comparisons. Therefore, this study is a natural extension of the work undertaken by Arora et al. (2019) [19], whereby we focus on the stratified analysis of the sustained phonation by using voice recordings from participants from one linguistic background, specifically, the US-English cohort.

The paper is organized as follows. Section 2 presents the data, followed by the methodology used for acoustic analysis comprising data pre-processing, feature extraction, feature selection, classification, and evaluation strategy. Section 3 presents the results, focusing on describing the most salient dysphonia measures that differentiate PwP from controls, along with the out-of-sample classification results. Discussions and directions for future research are provided in Section 4. Conclusions are provided in Section 5.

2. Data and Methods

2.1 *Data characteristics*

We processed sustained vowel (/a/) phonations collected as part of the PVI. The recordings were sampled at 8kHz with 16 bits resolution and were collected via telephone digital audio lines. The participants were instructed to say 'aaah' for as long and as steadily as possible. All calls were non-identifiable, and participants were entirely self-selected. During the call, participants were asked to provide basic demographics (age, gender) and whether they have received a clinical PD diagnosis. For further details on the data collection protocol, please see Arora et al. (2019) [19]. As mentioned previously, here we focus on the cohort where we have the largest participation (US) in the PVI study, and aim to progressively explore further differences in follow-up work.

Table 1 presents the data details and participant characteristics of the US PVI-cohort that is used hereafter. A total of 12675 phonations from 6942 participants were originally collected. We used an automated algorithm to exclude phonations that had excessive background noise, erroneous recordings, or otherwise missing information following the methodology we had previously described [19, 23]. Specifically, 1987 phonations from 1078 PD participants and 9955 phonations from 5453 controls were further processed.

Table 1. Data details and participant characteristics.

Characteristics	PD participants	Controls
No. of phonations	1987	9955
No. of participants	1078	5453
Age (years)	62.65 (12.03)	49.19 (15.89)
Male/Female	566/512	2976/2477

Note: Age is reported as mean and standard deviation (in brackets)

2.2 Dysphonia measures

We acoustically characterized each sustained vowel /a/ phonation using speech signal processing algorithms to extract 304 dysphonia measures. These dysphonia measures have been developed specifically to characterize sustained vowel /a/ phonations in the context of PD voice assessment, quantifying physiological patterns including deviation from vocal fold periodicity (jitter and shimmer variants), acoustic/turbulent noise, and articulator placement. For the rationale, background and detailed algorithmic expressions for the computation of the dysphonia measures we refer interested readers to our previous work [12, 21, 24, 25]. The MATLAB source code for the computation of the dysphonia measures is freely from the last author's website: https://www.darth-group.com/software. For completeness, we succinctly summarize these algorithms in Table 2 categorized in algorithmic families along with a brief description.

The fundamental frequency (F0) is a critical component in speech signal analysis and is often used as a pre-processing step for many of the dysphonia measures such as jitter [20, 21]. Strictly speaking, F0 is only defined for strictly periodic signals. In practical speech signal processing, we use the concept of F0 to refer to the vibrating pattern of the vocal folds in the short term and typically compute the F0 contour in short pre-specific segments (typically every 10 msec) [12, 26, 27]. This is therefore a practically applicable approach even in speech signals which are not periodic [12, 26]. Here, we computed F0

using the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) algorithm [28] which we had previously reported is one the most accurate F0 estimators in the context of sustained vowels [29]. We clarify that we processed only the most stationary 2-second signal segment from each phonation, which was determined by identifying the least fluctuating 2-second continuous F0 contour segment (in 10 msec steps) as determined using SWIPE: this circumvents problems with highly fluctuating signals. Applying the speech signal processing algorithms gave rise to a 11,942×304 feature matrix that was subsequently processed to map onto the binary outcome (0 was used to denote controls and 1 to denote PwP).

Table 2: Breakdown of the dysphonia measures used in the study

Family of acoustic measures	Brief description	Number of measures
Jitter variants	F0 perturbation	28
Shimmer variants	Amplitude perturbation	21
Harmonics to Noise Ratio (HNR) and Noise to Harmonics Ratio (NHR)	Signal to noise, and noise to signal ratios computed using standard approaches relying on autocorrelation	4
Glottis Quotient (GQ)	Vocal fold cycle duration changes	3
Glottal to Noise Excitation (GNE)	Extent of noise in speech using energy and nonlinear energy concepts	6
Vocal Fold Excitation Ratio (VFER)	Extent of noise in speech using energy, nonlinear energy, and entropy concepts	9
Empirical Mode Decomposition Excitation Ratio (EMD-ER)	Signal to noise ratios using EMD-based energy, nonlinear energy, and entropy	6
Mel Frequency Cepstral Coefficients (MFCC)	Amplitude and spectral fluctuations on the Mel scale quantifying envelope and high frequency aspects	39
F0 related	Comparisons of F0 against age and gender matched controls, inclduing probabilistic variabilities	3
Wavelet-based coefficients	Amplitude, scale, and envelope fluctuations quantified using wavelet coefficients	182
Pitch Period Entropy (PPE)	Variability of F0 expressing inefficiency of F0 stability over and above controls	1
Detrended Fluctuation Analysis (DFA)	Stochastic self-similarity of turbulent noise	1
Recurrence Period Density Entropy (RPDE)	Uncertainty in estimation of F0	1

Algorithmic expressions for the dysphonia measures summarized above are described in detail in [12, 21, 24, 25] . The MATLAB source code for the computation of the dysphonia measures is freely available from the last author's group website: https://www.darth-group.com/software. F0 refers to fundamental frequency estimates, here computed using SWIPE [28].

2.3 Dimensionality reduction using feature selection and feature transformation

High dimensional datasets often lead to well-known problems broadly known as the *curse of dimensionality*. In short, the presence of a large number of noisy and redundant features may affect the predictive performance of the statistical learning algorithm [30]. To address this problem, traditionally feature selection or feature transformation approaches are used, aiming to reduce the dimensionality of the dataset before presenting it

to the statistical learner. We indicatively used three feature selection methods and one feature transformation method to explore different approaches to the problem of optimizing the out of sample performance of the subsequent statistical learner. Specifically, we applied the following feature selection methods: (1) GSO [31], (2) RELIEF [32], and SIMBA [33]. Each of those feature selection methods provides a ranking of the features. In each case, we used the feature selection voting strategy we had previously introduced [10, 15] to robustly determine the final feature subset for each feature selection algorithm. In all cases, we restricted the search to the top-30 features selected using each algorithm. Finally, we explored feature transformation using standard principal component analysis (we extracted the first 30 principal components).

2.4 Statistical Mapping

We have used three state-of-the-art statistical mapping algorithms: (1) Random Forests (RF) [34], (2) Support Vector Machines (SVM) [35], (3) Adaptive Boosting (AdaBoost) [36] to tackle the binary differentiation problem in the study. We chose these methods as they are commonly used off-the-shelf classifiers that have been shown to be accurate in diverse supervised learning problems and in particular in a similar context differentiating PwP from controls using voice [18, 19, 25]. For the RF we explored optimizing performance using Breiman's recommendation with half and twice the default recommended number of features over which to select features for each node, and explored findings using 500 trees and 1000 trees. For the SVM we used the LIBSVM implementation with a MATLAB wrapper [37] and followed the suggestions of the developers of that implementation for optimizing hyper-parameters [38]: we linearly scaled each of the features to lie in the range [-1, 1], and used a Gaussian, radial basis function kernel. We clarify that for the scaling of the features in both the training and the testing subsets only the information from the training subset was used. The penalty parameter C and the kernel bandwidth w were determined using a standard grid search (C, w) defined by the product of the sets $C = [2^{-5}, 2^{-13}, \dots, 2^{15}]$, and $w = [2^{-15}, 2^{-13}, \dots, 2^{3}]$. The optimal parameter pair (C, w) was determined using the highest balanced accuracy. For the Adaboost, the learning rate hyper-parameter was optimized in the range 0.01 to 0.5 (we searched the following possible values: 0.01, 0.03, 0.05, 0.1, 0.3, 0.5) and the number of trees used as weak base learners of the boosted classifier was set to 1000. We refer to the original papers and Hastie et al. [30] for an authoritative description of the methods and further details on parameter fine-tuning and optimization.

Given the dataset is highly unbalanced (9955/11942 samples are from controls and 1987/11642 samples from PwP, i.e. >80% samples in the dominant class) and this setting is known to be particularly challenging for statistical learning models [39], we wanted to explore a different strategy to mitigate potential problems due to a class dominating the performance of the classifiers. The strategy we followed for training and testing the model comprises two steps.

In the first step, we randomly selected 1500 samples from PwP and 1500 samples from controls to create a balanced binary classification dataset (n=3000 samples) which we will use to train, explore, optimize, and validate the classifiers used using a standard 10-fold with 100 iterations model validation approach, following the standard methodology we had previously used in similar applications in this field [10, 12, 18]. The aim is to use this first step to decide on the final model, by optimizing and setting any hyper-parameters so that it can be finalized and used externally in new datasets. We clarify that the feature selection and feature transformation approaches were applied using only the balanced dataset. We report performance on the out-of-sample CV data. The second step is used as a final model validation assessment where we have used the remaining data that has not been already used in step 1. In this case, we have an unbalanced dataset with the

remaining samples (8942 samples, 8455 recordings from controls and 487 recordings from PwP). This is used to provide further evidence of the model generalization performance with samples that have not been used for any of the preceding steps with feature selection/transformation and statistical mapping.

Throughout we report performance in terms of the accuracy, along with sensitivity and specificity. In the final model validation step, we provide the full confusion matrix to facilitate understanding of the classifier's output. The full methodology of the study is concisely summarized in Figure 1.

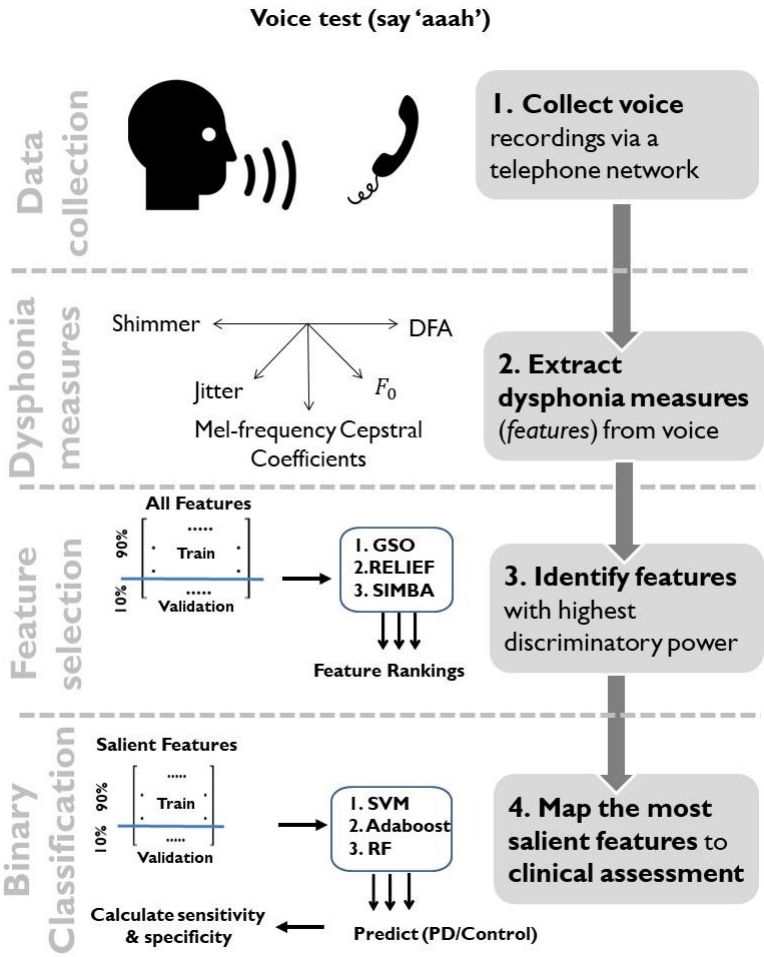

Figure 1. Schematic diagram showing the different stages of this study. Specifically: (**Step 1**) data collection: sustained phonations were collected over a standard telephone line network; (**Step 2**) feature extraction: 304 dysphonia measures were extracted from each phonation to characterize voice impairment; (**Step 3**) feature selection: using a balanced dataset (n=1500 PwP and 1500 control participants), the feature matrix was split into non-overlapping training and test data using a 10-fold cross-validation scheme, and three feature selection techniques (GSO, RELIEF, and SIMBA) were employed for identifying the most salient features on the training data; (**Step 4**) classification: the most salient subset of features were mapped onto clinical assessment (PD/Control) using binary classifiers (SVMs, Adaboost, and Random Forests). The final classification step was on the test data held out as part of the CV; subsequently, once we decided on the final statistical learning model, the trained classifier was also presented with the held-out dataset (8942 samples) as an additional performance assessment approach.

3. Results

Figure 2 illustrates the performance (balanced accuracy) of the model as a function of the features presented into SVM in the standard 10-fold CV setup. Table 3 summarizes

the different performance measures for the three classifiers considered in this study for completeness. The performance was evaluated using only the test data, using a 10-fold CV scheme with 100 iterations. We remark that 27 features with an SVM led to a balanced accuracy of about 67.3% (sensitivity: 67.43%, specificity: 67.25%). Therefore, we selected this trained model with the 27 features to test further how well findings generalize on the out of sample (held-out) unbalanced dataset. The resulting confusion matrix for the unbalanced held-out dataset (n=8942 samples) is provided in Figure 3 (balanced accuracy 66.3%).

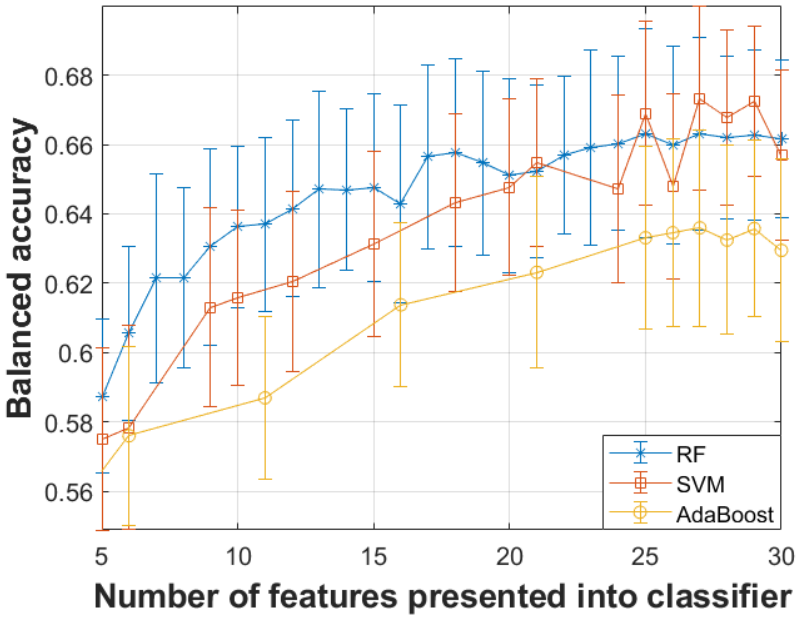

Figure 2. Balanced accuracy as a function of the number of features presented into the three binary classifiers for the validation dataset comprising 3000 samples (1500 controls and 1500 PwP). The bars denote the standard deviation around the quoted mean score. The features presented into the classifiers were selected using SIMBA.

Table 3. Out-of-sample performance measures for the three classifiers (SVM, Adaboost, and Random Forests) using 10-fold CV with 100 iterations on the balanced dataset (n=3000 samples).

Classifier	# of optimal features	Sensitivity	Specificity	Balanced accuracy
SVM	27	**67.43%**	**67.25%**	67.34%
Random Forests	27	66.38%	66.20%	66.29%
Adaboost	27	63.11%	63.60%	63.36%

Note: highest scores are highlighted in **bold**.

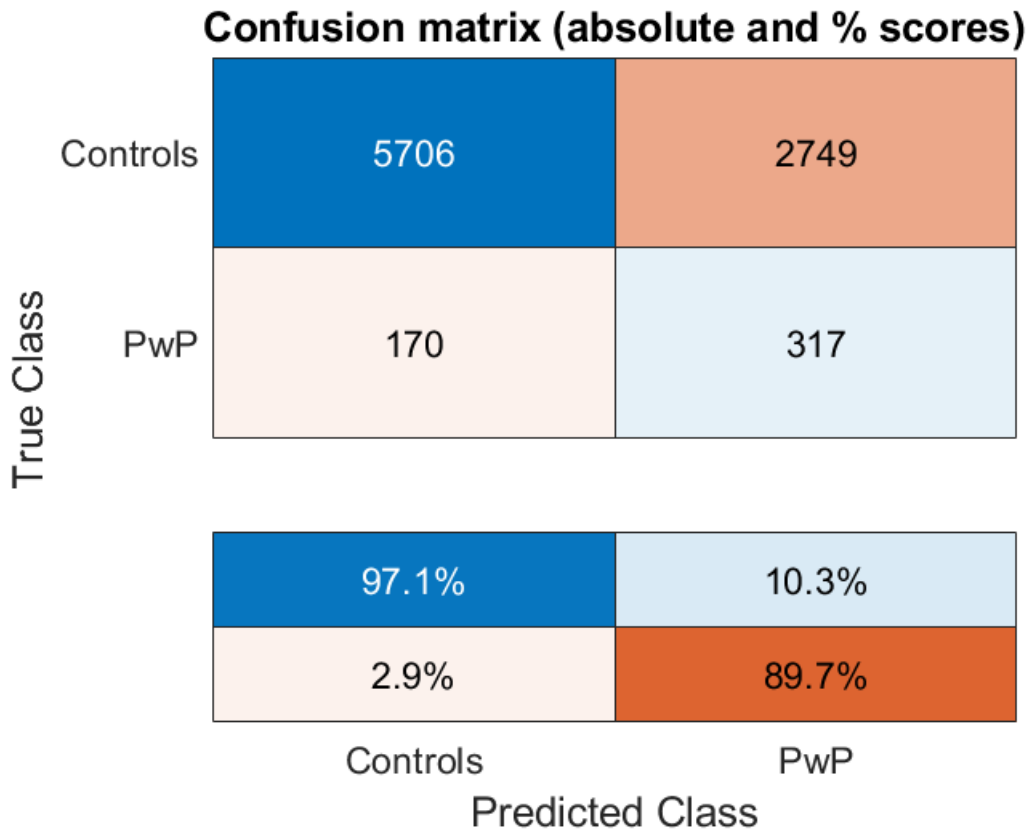

Figure 3. Confusion matrix denoting performance on the held-out unbalanced dataset (n=8942 samples) when using the best performing model selected from the results presented in Fig. 2 (SVM with 27 features selected using SIMBA).

The results in Fig. 3 suggest that we can indeed correctly identify the vast majority of PwP in the held-out (unbalanced) dataset, and hence this supports the presented methodology as a potentially useful biomarker that could be further explored.

4. Discussion

We investigated the potential of differentiating PwP and controls using telephone-recorded speech collected under acoustically non-controlled conditions exploring different statistical machine learning techniques and strategies. This study is part of our wider goal to explore whether we can develop a PD screening tool that is readily accessible, accurate, ideally free-of-charge, and is the underlying reason we set up the PVI study where data for this study were drawn from. We demonstrated 67.34% balanced accuracy using 27 acoustic features presented into an SVM with a standard 10-fold CV approach. This finding was further verified on an additional out of sample unbalanced dataset where we found a balanced accuracy of 66.3% (sensitivity: 65.09%, specificity: 67.49%). Overall, this is very similar performance to what we had previously reported in Arora et al. (2019) (66.4% balanced accuracy), however this has now been achieved using 27 acoustic features compared to the need to include 100 features that we had reported in the afore-mentioned study and hence is a more parsimonious result.

Compared to our previous exploration of the PVI dataset to differentiating PwP from controls, here we had used only the US cohort. This was motivated by some of our earlier investigations that some of the feature distributions are different across the PVI cohorts [22], which suggests we should carefully consider stratifying the PVI data and investigating cohorts independently. We aim to explore transfer learning approaches [40] to account

for covariate shifting between the different datasets in the PVI study (given data has been collected across 7 countries and participants between countries may come from different linguistic backgrounds e.g. English or Spanish).

Placing the results in the wider context in the research literature, this study's findings are very modest given we had previously reported more than 98% binary differentiation between PwP and controls using a similar protocol to collect sustained vowel /a/ phonations [10]. Similarly, other research groups had reported accuracies around and over 90% in this binary differentiation application, indicatively [8, 41]. However, we stress that previous work had focused on collecting data under carefully controlled acoustic conditions (e.g. sound-treated booths, using high-quality standardized microphones [10, 15]), whereas in the PVI participants self-enrolled using their own devices which have different specifications in terms of microphone quality and frequency attenuation characteristics, at their own environments which typically had some background noise, whilst using different telephone networks. Moreover, unlike most research studies, participants in the PVI were not screened or clinically assessed for study enrollment and thus we cannot rule out the presence of clinical-pathologic differences in voice within this cohort. Collectively, all these 'degrees of freedom' lead to lower quality data and therefore it is expected there will be considerable performance degradation. For example, some of the most successful nonlinear dysphonia measures in this application rely on the use of high frequencies (2.5-10 KHz) to compute the 'noise' component in the recorded signal, see [10] for details. Given that the sampling rate in PVI is 8 kHz (and hence the useful recorded information is up to 4 kHz according to the Nyquist sampling theorem), this constrains extracting clinically informative features.

Speech impairment is commonly associated with Parkinson's [40] and is characterized by pitch monotonicity, variable rate, imprecise consonants, and breathiness and harshness. As opposed to other types of speech signals that are often used in clinical assessments, such as running speech and reading loud a linguistically rich prespecified text e.g. the Grandfather Passage [20], the use of sustained phonations helps circumvent challenges associated with different accents and linguistic confounds [20]. For example, our previous work has shown that sustained phonations can provide high accuracy in differentiating PwP from controls [10], along with other interesting insights in the speech-PD literature including replicating PD symptom severity and assisting PD rehabilitation [10, 12, 18, 21]. We emphasize also that the methodology adopted in this study for processing sustained vowels had previously also been generalized to analyze different types of speech, e.g. voice fillers [42], and to provide useful insights more widely in different biomedical speech signal processing applications [43]. Therefore, the use of sustained vowels is strongly motivated and has been practically vindicated. A further practical consideration is that this study draws data from PVI, where data was collected across 7 countries with participants coming from different linguistic backgrounds [19]. One of the aims of PVI was to provide cross-linguistic comparisons for the assessment of PD within a short time span of speech samples from a large, self-selected population group. Therefore, for practical reasons and to minimize participant burden, we had decided in PVI to collect exclusively sustained vowels. It is due to these reasons that the focus of this study was on analyzing sustained phonations. Nevertheless, we remark that the use of alternative speech types, e.g. running speech, might be accommodating additional acoustic information which is not captured in sustained vowels (although we stress that the argument goes both ways, the use of sustained vowels may capture information not accounted in running speech). An interesting line of future work would be to evaluate the efficacy of telephone-quality sustained phonations in conjunction with running speech to develop screening tools for PD.

The participants for this study were entirely self-selected, where they were prompted to answer the question – 'Do you have Parkinson's disease?' and their response was treated as the gold standard (or label) for statistical mapping. In the absence of detailed clinical assessments, we cannot rule out clinical-pathologic differences in voice within this cohort, which could be one of the factors contributing to the relatively low discrimination accuracy reported in this study. It is worth noting that diagnosis/monitoring of PD requires in-person subjective assessment, typically by a trained neurologist, which can incur substantial logistical costs in resource-constrained and remote settings. Thus, we deemed it necessary to include only self-reported symptoms. Specifically, the data collection protocol of PVI was designed with the objective to develop a population-based screening (and not monitoring) tool for PD, which would have the potential to transform current practices by reducing logistical costs associated with in-person clinical assessments, while exploring alternate routes to recruiting participants for clinical trials.

This study builds on our previous work on PVI [19] and acoustic analysis [10, 12, 14, 21] to almost completely automate the data processing pipeline. In principle, it may be useful to apply auditory-perceptual analysis relying on human expertise to analyze the data and potentially identify problems, e.g. highly aperiodic/too noisy signals, and also to perceptually characterize the signals (producing additional features). This is indeed often done in studies with a low number of speech samples with speech signals of different nature (e.g. running speech, counting days, reading pre-specified linguistically rich text etc.). Auditory-perceptual analysis is not commonly used when processing sustained vowels, at least in the biomedical speech signal processing literature. Moreover, auditory-perceptual analysis would be practically very challenging and costly for the size of the available data in PVI. Instead, developing automated pattern recognition tools combined with statistical machine learning offers a replicable, objective, automated, and directly scalable approach. This has enabled us to automatically determine, for example, highly aperiodic and noisy signals which were discarded from further analysis (for details on the algorithm see our previous work [19]).

We explored three different feature selection methods and standard feature transformation using PCA to reduce the dimensionality of the dataset. The transformed features using PCA led to consistently worse results and hence these results are not presented in the paper due to space constraints. The three feature selection algorithms led to quite different feature subsets (results not shown), and SIMBA along with SVM provided a somewhat better overall performance in the balanced dataset where we applied the standard CV approach. Therefore, we reported in Figure 2 the performance of classifiers as a function of the number of features progressively selected by SIMBA.

SVMs and RF worked considerably better than Adaboost in this application (see Figure 2). In our experience on this and related PD problems using classification tools, we have observed that generally bagging approaches tend to outperform boosting approaches, although we do not have a theoretical justification for this finding. SVMs led to the best overall result, which is broadly in agreement with our empirical observation in related studies in Parkinson's applications: we have previously reported SVMs slightly outperform RF in binary classification problems, whereas RF generally leads to better outcomes in multiclass classification problems [21]. Again, this should be cautiously considered on the basis of our experience in related applications, and we make no further claims on generalizability of this finding. We remark that the choice of the three classifiers used here is indicative of some commonly used methods, there are many alternative classifiers that could be explored. For example, an interesting line of further research work would be to provide a comparison of different classification methods, including deep learning. Moreover, it would be worth exploring different classifiers in further detail in conjunction with different class balancing schemes and model validation strategies.

There are different model validation strategies that could be explored and here it is particularly important because of the highly unbalanced nature of the dataset. In principle, when having a single dataset it is useful to perform CV (e.g. 5-fold or 10-fold CV, along with additional iterations for statistical confidence) rather than leaving a single portion of the data out for testing ('the testing dataset'). This is because often we want to assess the model's robustness with perturbed training/test data, also assessing variability in performance across folds (and iterations) to provide an estimate of the generalization performance including a confidence interval. However, the highly unbalanced nature of the problem given the available dataset in this study poses considerable challenges using a standard CV approach. Therefore, we decided on a strategy where we used both model validation approaches: retaining a completely separate subset of the data for testing at the very end, and using a balanced subset with 3000 randomly selected samples (which overcomes problems with highly unbalanced data) for a standard training/testing scenario using 10-fold CV. This enables us both to assess the model's performance in a 'classifier-friendly' binary classification setting with a balanced dataset where we can also provide a confidence interval on the estimates (see Fig. 2) and also test the model's performance on an additional unbalanced subset (see Fig. 3).

We remark that the developed SVM model was further validated on an unbalanced 'held-out' dataset (see Figure 3), where we observe that most PwP were correctly detected. The false positives rate is still fairly high and there is ample space for improving these results further before it can be meaningfully used as an accurate clinical decision support tool. Nonetheless, the findings in Figure 3 highlight this freely accessible tool towards screening for PD might be a useful direction and could be complemented with additional modalities (e.g. smell [44], smartphone-based tests [11, 16, 18]) to form a more accurate practical tool that people could periodically use for mobile check-up and potentially facilitate referrals for specialized physical neurological assessment.

This study has some key limitations primarily regarding the quality of the speech dataset. The standard recommendation of the speech community is that speech signals should be sampled with at least 20KHz sampling frequency for clinical applications because there is useful information in the higher frequencies of the spectrum [20]. Also, the data in PVI was collected under acoustically non-controlled conditions, which has a clear degradation effect on the data quality of the recorded speech signals. Nevertheless, some exploratory recent work has demonstrated that sustained vowels /a/ transmitted over the simulated standard telephone network (following the typical digital communications process with down-sampling to 8KHz, encoding, transmitting through a noisy channel and decoding) demonstrated that the reduction in voice quality was not prohibitive for replicating the standard PD symptom severity metric [14]. Therefore, there is some justification that the reduced sampling rate used in PVI (8KHz) would still be useful information to be extracted from the sub-optimally recorded data. In principle, a study could be designed these days where people could collect speech samples on a high-end smartphone (which use high-quality microphones) and captured using a dedicated smartphone app at the recommended sample rate. However, that would require people to have access to high-end expensive equipment, and thus such a solution would not be widely available. Instead, PVI was conceptualized as an approach to democratize access to a potentially useful PD screening tool that could be accessible to all at practically no cost. We maintain that if we want to scale up work and deliver responsible innovative solutions to make a meaningful difference in practice with a largely accessible tool, there are some compromises we will likely need to make when collecting data in a practical setting so that it would be as accessible as possible by those who would like to use it.

5. Conclusions

This study further supports the concept of exploring telephone-quality speech towards developing a screening tool to assess PD with an easy-to-use test relying solely on the use of the sustained vowel /a/. Our findings have important implications towards democratizing access to a useful, generalizable, and robust PD tool at practically no cost, which can be easily used remotely and at scale with any telephone device. This study is a part of the broader work that increasingly members of the research community focus on towards developing diagnostic decision support tools in PD which can be adopted at scale. In time, this approach could be expanded to facilitate early diagnosis both for PD and potentially other related conditions. We envisage this tool may be widely applied to provide early probabilistic indication of PD particularly for groups at-risk, potentially facilitating early PD diagnosis which in turn can lead to better longitudinal symptom management.

Author Contributions: Both authors contributed equally.

Funding: The study was funded through an EPSRC award (Co-PIs Tsanas and Arora).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Due to data confidentiality the data cannot be made publicly available. Interested researchers seeking to explore collaborative opportunities can approach the authors.

Acknowledgments: We are grateful to Aculab for the use of their servers to facilitate data collection.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Collaborators, G.B.D.P.s.D., *Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016.* Lancet Neurol, 2018. **17**(11): p. 939-953.
2. Collaborators, G.U.N.D., et al., *Burden of Neurological Disorders Across the US From 1990-2017: A Global Burden of Disease Study.* JAMA Neurol, 2021. **78**(2): p. 165-176.
3. Hughes, A.J., et., *Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases.* J Neurol Neurosurg Psychiatry, 1992. **55**(3): p. 181-4.
4. Post, B., et al., *Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?* Mov Disord, 2005. **20**(12): p. 1577-84.
5. Goetz, C.G., 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, 2008. **23**(15): p. 2129-70.
6. Ho, A.K., et al., *Speech impairment in a large sample of patients with Parkinson's disease.* Behav Neurol, 1998. **11**(3): p. 131-137.
7. Godino-Llorente, J.I., et al., *Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers.* PLoS One, 2017. **12**(12): p. e0189583.
8. Little, M.A., et al., *Suitability of dysphonia measurements for telemonitoring of Parkinson's disease.* IEEE Trans Biomed Eng, 2009. **56**(4): p. 1015.

9. Orozco-Arroyave, J.R., et al., *Automatic detection of Parkinson's disease in running speech spoken in three different languages.* J Acoust Soc Am, 2016. **139**(1): p. 481-500.
10. Tsanas, A., et al., *Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease.* IEEE Trans Biomed Eng, 2012. **59**(5): p. 1264-71.
11. Arora, S., et al., *Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study.* Parkinsonism Relat Disord, 2015. **21**(6): p. 650-3.
12. Tsanas, A., et al., *Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity.* J R Soc Interface, 2011. **8**(59): p. 842-55.
13. Tsanas, A., et al., *Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests.* IEEE Transactions on Biomedical Engineering, 2010. **57**: p. 884-893.
14. Tsanas, A., M.A. Little, and L.O. Ramig, *Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network.* IEEE Access, 2021. **9**: p. 11024-11036.
15. Tsanas, A., et al., *Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease.* IEEE Trans Neural Syst Rehabil Eng, 2014. **22**(1): p. 181-90.
16. Arora, S., et al., *Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD.* Neurology, 2018. **91**(16): p. e1528-e1538.
17. Arora, S., et al., *Investigating Voice as a Biomarker for Leucine-Rich Repeat Kinase 2-Associated Parkinson's Disease.* Journal of Parkinsons Disease, 2018. **8**(4): p. 503–510.
18. Arora, S. et al., *Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease.* IEEE Access, 2021. **9**: p. 44813-44824.
19. Arora, S., L. Baghai-Ravary, and A. Tsanas, *Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice.* J Acoust Soc Am, 2019. **145**(5): p. 2871.
20. Titze, I.R., *Principles of Voice Production.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1994.
21. Tsanas, A., *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*. PhD thesis, University of Oxford, 2012. xi, 242 pages.
22. Tsanas, A. and S. Arora, *Acoustic analysis of sustained vowels in Parkinson's disease:: New insights into the differences of UK- and US-english speaking participants from the Parkinson's voice initiative.* Models and Analysis of Vocal Emissions for Biomedical Applications (in press), 2021.
23. Tsanas, A. and S. Arora, *Exploring telephone-quality speech signals towards Parkinson's disease assessment in a large acoustically non-controlled study.* 19th IEEE International Conference on BioInformatics and BioEngineering, Athens, Greece, 2019: p. 953-956.
24. Tsanas, A., et al., *New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity.* International Symposium on Nonlinear Theory and its Applications, Krakow, Poland, 2010: p. 457-460.
25. Tsanas, A., *Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms.* 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Florence, Italy, 2013: p. 37-40.
26. Roark, R.M., *Frequency and voice: perspectives in the time domain.* J Voice, 2006. **20**(3): p. 325-54.
27. Tsanas, A., A. Triantafyllidis, and S. Arora. *Automating quasi-stationary speech signal segmentation in sustained vowels: application in the acoustic analysis of Parkinson's disease*. in *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*. 2021. Florence, Italy.
28. Camacho, A. and J.G. Harris, *A sawtooth waveform inspired pitch estimator for speech and music.* J Acoust Soc Am, 2008. **124**(3): p. 1638-52.

29. Tsanas, A., et al., *Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering.* J Acoust Soc Am, 2014. **135**(5): p. 2885-901.
30. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. 2001, New York: Springer. xvi, 533 p.
31. Chen, S., S.A. Billings, and W. Luo, *Orthogonal least squares methods and their application to non-linear system identification.* International Journal of Control, 1989. **50**: p. 1873-1896.
32. Kononenko, I., *Estimating attributes: Analysis and extensions of RELIEF.* Lecture Notes in Artificial Intelligence, Springer, Berlin, Heidelberg, 1994. **784**.
33. Gilad-Bachrach, R., A. Navot, and N. Tishby, *Margin Based Feature Selection - Theory and Algorithms.* Proceedings of the twenty-first international conference on Machine learning (ICML), 2004.
34. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
35. Boser, B., I. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers.* Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh
36. Freund, Y. and R.E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.* Journal of computer and system sciences, 1997. **55**: p. 119-139.
37. Chang, C.C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**: p. 1-27.
38. Hsu, C.-W., C.-C. Chang, and C.-J. Lin, *A practical guide to support vector classification.* Technical report, National Taiwan University, 2010.
39. Bishop, C.M., *Pattern recognition and machine learning*. Information Science and Statistics. 2006, New York: Springer. xx, 738 p.
40. Pan, S.J. and Q. Yang, *A Survey on Transfer Learning.* IEEE Transactions on Knowledge and Data Engineering, 2010. **22**: p. 1345-1359.
41. Sakar, C.O. and O. Kursun, *Telediagnosis of Parkinson's disease using measurements of dysphonia.* Journal of Medical Systems, 2010. **34**: p. 591–599.
42. San Segundo, E., A. Tsanas, and P. Gomez-Vilda, *Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics.* Forensic Science International, 2017. **270**: p. 25-38.
43. Tsanas, A. and P. Gómez-Vilda. *Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels*. in *Multidisciplinary Conference of Users of Voice, Speech and Singing* 2013. Las Palmas de Gran Canaria.
44. Lo, C., et al., *Olfactory Testing in Parkinson Disease and REM Behavior Disorder: A Machine Learning Approach.* Neurology, 2021. **96**(15): p. e2016-e2027.