INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Do Phonatory Features Display Robustness to Characterize Parkinsonian Speech Across Corpora?

Anna Favaro[1], *Tianyu Cao*[1], *Thomas Thebaud*[1], *Jesús Villalba*[1], *Ankur Butala*[2], *Najim Dehak*[1], *Laureano Moro-Velázquez*[1]

[1] Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

[2] Department of Neurology, The Johns Hopkins University, Baltimore, MD, USA

afavaro1@jhu.edu

Abstract

Sustained vowels have been largely used to quantify vocal impairment in Parkinson's disease (PD), with most studies focusing on a single corpus. Presumably, features obtained from sustained vowels are language-independent, but how findings generalize across cohorts is unclear. This work analyzes 61 phonatory features from 5 corpora in American English, Italian, Castilian Spanish, Colombian Spanish, and German, respectively, by conducting a statistical and correlation analysis. We use *robustness* as a criterion in which a feature displays the same behavior across corpora. The statistical analysis showed that the features provided good separability between PD and controls in only two out of five corpora, and none of the features displayed robustness. However, experiments report significant correlations between feature values and clinical scores. These findings provide valuable insights into the acoustic corpora-based dissimilarities, which should be considered when generalizing findings.

Index Terms: Parkinson's Disease, Sustained vowels, Statistical analysis, Correlation Analysis, Robustness

1. Introduction

Sustained phonations have been extensively used for quantifying vocal impairment occurring in Parkinson's Disease (PD) as this task can help to circumvent part of the articulatory and linguistic confounds of *running speech* [1]. Phonatory approaches working with sustained vowels could be considered language-independent, even though a vowel might not always have the same phonetic realization across languages.

Traditional phonatory features are perturbations of fundamental frequency (jitter), perturbations of amplitude (shimmer), and noise (e.g., Harmonic to Noise Ratio (HNR), Noise to Harmonics Ratio (NHR)). Previous studies reported significant differences between groups for these measurements, with jitter, shimmer, and NHR higher and HNR lower in people with PD [2, 3, 4]. More recently, non-linear features have been introduced to better characterize abnormal vocal fold vibration and non-linear pressure flow in the glottis, leading to parametrization sets that include Recurrence Period Density Entropy, Detrended Fluctuation Analysis, correlation dimension, Hurst Exponent, and Largest Lyapunov Exponent [5, 6].

Even though many studies analyzed phonatory features, only a few considered several corpora simultaneously. Rusz et al. [7] performed a speech analysis of Czech, English, German, French, and Italian speakers in the early phase of PD. In that study, as phonatory features, they only analyzed HNR via a sustained phonation paradigm, but this feature did not reach significance in any corpus. Tsanas and Arora [8] investigated the differences in 307 dysphonia measures between UK- and US-English-speaking subjects with PD. Even though the classical acoustic measures, such as jitter and shimmer, were very similar in the two cohorts, there were pronounced differences in the behaviors and trends of more complex metrics, such as Vocal Fold Excitation Ratio and Mel Frequency Cepstral Coefficients (MFCCs). The authors conducted the same analysis on a larger cohort of participants from seven countries whose speech samples were collected under uncontrolled acoustic environments [9]. They reported that the majority of dysphonia measures did not differentiate PDs from healthy controls (CNs) sufficiently well, probably because of the reduced signal bandwidth. Kovac et al. [10] considered recordings from Czech, American English, Israeli, Colombian Spanish, and Italian-speaking subjects. With respect to the phonatory features analyzed, harmonic richness factor, mean normalized amplitude quotient, mean quasi-open quotient, and jitter reached significance in one speech corpus only.

Overall, it is unclear how findings generalize across corpora, as corpora-based dissimilarities such as different background noises, sampling frequency, or microphones (i.e., the channel characteristics) can influence the estimation of phonatory measurements [11, 12]. Hence, as phonations may be cohort-dependent, it is crucial to understand whether phonatory analysis should be conducted separately for participants from different cohorts, undertaking cross-cohort comparisons [8]. Previous studies on phonation performing cross-corpora analysis displayed some limitations. First, they typically investigated only the behaviors of the most typical phonatory features (e.g., jitter, HNR). Second, they only considered a few speech corpora at a time without assessing whether the features display homogeneous behaviors across corpora. Third, the criteria adopted to probe the robustness of the features were too weak, increasing the probability of falsely reporting as robust features that display different behaviors across cohorts and corpora.

In this study, the behaviors of 61 interpretable phonatory features were examined using five corpora of speakers who have different mother tongues: American English, Castilian Spanish, German, Italian, and Colombian Spanish, respectively. Besides considering the most classic amplitude and frequency perturbation parameters, more recent approaches based on complexity parameters, modulation spectra, noise, and fluctuation (or tremor) were analyzed. The proposed analysis leverages *robustness* [13] as a criterion in which a feature behaves the same, independently of the corpus considered. Even though the robustness of different prosodic and linguistic features for the detection of PD has already been assessed [14, 13], the robustness of phonological measurements still needs further investigation. In this respect, identifying which tasks and features are robust represents a step forward in the development of a *universal* framework for PD evaluation and monitoring. Moreover, given their interpretable meaning, if these features would display robustness, they might be employed in clinical scenarios as PD biomarkers.

10.21437/Interspeech.2023-1784

2. Materials

Five different corpora were used in this study: Neurological Signals (NLS) [15], Neurovoz [16], Italian Parkinson's Voice and Speech [17], GITA [18], and GermanPD [19].[1] Table 1 summarizes corpora demographics and disease severity statistics.

2.1. American English

NeuroLogical Signals (NLS) is a data set collected at Johns Hopkins Medicine (JHM) by the authors of this study. It contains recordings from individuals with neurological disorders (NDs) and CNs. The Johns Hopkins Medical Institutional Review Board approved the data collection, and all participants signed informed consent. Participants with PD received dopaminergic medication before the recording session. Speech signals were recorded with a headset at 24 kHz in a quiet room. In this study, we considered 23 participants with clinically established PD and 27 CN participants matched in age. None of the participants in the CN group had a history of symptoms related to PD or any other NDs.

2.2. Castilian Spanish

Neurovoz is a data set that contains speech samples from 32 CNs and 47 participants with PD whose native language is Castilian Spanish. The data collection was performed in compliance with the Helsinki Declaration and was approved by the Ethics Committee of Hospital General Universitario Gregorio Marañón in Madrid (Spain). All participants involved in the study signed informed consent. Participants with PD received dopaminergic medication before the recording session. Recordings were originally sampled at 44.1 kHz and collected in a quiet room.

2.3. Colombian Spanish

GITA is a data set collected by Universidad de Antioquia in Medellín (Colombia). It contains recordings from 50 participants with PD and 50 CNs whose native language is Colombian Spanish. The data collection was performed in compliance with the Helsinki Declaration and was approved by the Ethics Committee of the Clínica Noel, in Medellín. All participants signed informed consent. Participants with PD received dopaminergic medication before the recording session. None of the CN participants reported symptoms associated with PD or other NDs. Recordings were originally sampled at 44.1 kHz and collected in a quiet room.

2.4. German

GermanPD is a data set collected in the hospital of Bochum (Germany). It contains speech recordings from 88 PD and 88 CN participants whose native language is German. The ethics committee of the Ruhr-University Bochum approved the study. All participants signed informed consent. Participants with PD received dopaminergic medication before the recording session. Speech samples were collected in a quiet room using a headset microphone, located approximately 5 cm from the participant's mouth. Recordings were originally sampled at 16 kHz.

2.5. Italian

The Italian Parkinson's Voice and Speech (ItalianPVS)[2] is a corpus containing recordings from 22 elderly CNs and 28 participants with PD. The recordings employed in this study are publicly available. The information concerning the participant's informed consent is detailed in their reference papers and dissemination platforms. Participants with PD received dopaminergic medication before the recording session. Speech samples were collected in a quiet, echo-free room using an external condenser and sampled at 16 kHz.

Table 1: Demographic and disease severity statistics of the study population. When available, gender, age distribution, and scores on the Unified Parkinson's Disease Rating Scale Part III (UPDRS-III) and the Hoen & Yar Scale are reported. Abbreviations: M, Male; F, Female; std: standard deviation.

Data set	Category	Sample size	Gender	Age (std)	UPDRS-III (std)	H&Y (std)
NLS	CN	27	M=16; F=11	64.70 (13.3)	–	–
	PD	23	M=14; F=9	67.31 (15.6)	26.5 (11.4)	2.3 (0.4)
Neurovoz	CN	32	M=14; F=18	67.50 (6.2)	–	–
	PD	47	M=29; F=18	71.40 (10.3)	12.8 (11.3)	2.3 (0.7)
GermanPD	CN	88	M=44; F=44	64.60 (13.7)	–	–
	PD	88	M=47; F=41	67.00 (10.5)	22.7 (10.5)	2.4 (0.7)
ItalianPVS	CN	22	M=10; F=12	67.30 (4.8)	–	–
	PD	28	M=19; F=9	66.40 (9.4)	–	–
GITA	CN	50	M=25; F=25	60.90 (9.6)	–	–
	PD	50	M=25; F=25	61.10 (9.4)	37.6 (18)	2.3 (0.5)

3. Methods

3.1. Phonatory Feature Extraction

Table 2 summarizes the phonatory features extracted. Features were extracted from the sustained phonation of different vowels: /a:/, /e:/, /i:/, /o:/, /u:/. NLS only contains recordings for the vowel /e:/, while GermanPD for the vowel /a:/. When multiple phonations were recorded, we averaged feature values across phonations for each speaker separately. On ItalianPVS, GermanPD, and Neurovoz, three phonations were recorded for each vowel, two on NLS and one on GermanPD. The Automatic Voice Condition Analysis (AVCA)[3] library in MATLAB was used to perform the extraction [5]. This library computes 261 coefficients per recording, calculated per frame, representing the mean and standard deviation of four main feature families: amplitude and frequency perturbation and fluctuation, spectral-cepstral, complexity, and modulation spectra. However, the spectral-cepstral features, which include MFCC, perceptual linear predictive coefficients (PLP), and some Modulation Spectra features such as Modulation Spectra Centroids (MSCents), and dynamic range (MSDR), were not considered as their physical interpretation is unclear. In this sense, although features like MFCCs can contain valuable information about phonation, these values will only be used in machine-learning scenarios and not as biomarkers in clinical environments. In the end, 61 features were extracted, including statistics such as mean and standard deviation for several of them. Before the feature extraction, all recordings were resampled at 16 kHz as required by the AVCA library.

3.2. Statistical and Correlation Analysis

The non-parametric Kruskal–Wallis H-test [30] was used to conduct pair-wise statistical tests to determine any significant differences between the feature distributions of PD and CN participants[4]. The analysis was conducted in each corpus separately. To control the False Discovery Rate (FDR), we applied the Benjamini–Hochberg correction[5]. As family-wise error rate, α was set to 0.05. Moreover, the correlations of the feature values with

[1] We considered only these five data sets as they were the only ones to which we had or received access.

[2] `https://ieee-dataport.org/open-access/italian-parkinsons-voice-and-speech`

[3] https://github.com/jorgomezga/AVCA-ByO

[4] To perform the pair-wise Kruskal-Wallis H-tests, we used *scipy.stats.kruskal* library in Python.

[5] To perform the Benjamini–Hochberg correction, we used *statsmodels.stats.multitest.fdrcorrection*, with default method.

Table 2: For each feature family, the feature names with their corresponding abbreviation in parenthesis, the expected behavior (EB) of the features in individuals with PD, references to previous studies supporting the hypothesis behind the features' EB, and the number of coefficients in each family are reported. Abbreviations: ↑, increasing; ↓, decreasing; N°: number.

Feature Family	Coefficients	EB	Related Works	N°
Amplitude, frequency perturbation and fluctuation	Absolute and relative jitter (rJitta) and shimmer (rShim), Relative Average Perturbation Quotient (rAPQ), Pitch Period Perturbation Quotient (PPQ), Three-point Amplitude Perturbation Quotient (APQ3), Five-point Amplitude Perturbation Quotient (APQ5), F0-Tremor Intensity Index (FTRI), Amplitude-Tremor Intensity Index (ATRI), Noise-to-Harmonics Ratio (NHR), Normalised Noise Energy (NNE), Glottal-to-Noise Excitation Ratio (GNE)	↑	[20], [21], [22] [2], [3], [4],	21
	Statistics about Harmonics-to-Noise Ratio (HNR), Cepstral-HNR (CHNR)	↓	[2], [5], [23], [24]	
Complexity	Correlation dimension (D2), Lempel-Zip Complexity (LZC), statistics about Largest Lyapunov Exponent (LLE), Hurst Exponent (HE), Approximate Entropy (ApEn), Sample Entropy (SampEn), Gaussian kernel Approximate Entropy (GApEn), Fuzzy Entropy (FuzzyEn), Modified SampEn (mSampEn), Permutation Entropy (PE), Detrended Fluctuations Analysis (DFA) Recurrence Period Density Entropy (RPDE), Markovian Entropy (MarkEnt)	↑	[25], [6], [26] [27], [5]	26
Modulation Spectra	Modulation Spectrum Percentiles (MSP), and statistics about Modulation Spectra Homogeneity (MSH), Cumulative Intersection Point (CIL), Rate of Points Above Linear Average (RALA), and Ratio Above Linear Percentiles (RALP)	↑	[5], [28], [29]	14

the Hoehn and Yahr Scale (H&Y) and the Unified Parkinson's Disease Rating Scale (UPDRS-III) were assessed.[6] Before the correlation analysis, features were normalized by subtracting the mean and dividing by the standard deviation.

3.3. Feature Robustness

We considered *robust* a feature if these three conditions are met:

1. If it reports a significant difference between the medians of the PD and CN group distributions in at least *two* data sets.
2. If it displays the same expected behavior (EB) in *all* the data sets where it is significant.
3. If the correlation between the feature values and the clinical scores is significant, the trend of the correlation follows the EB postulated for the feature.

Similar criteria for robustness were previously introduced by Kovac et al. and Favaro et al. [14, 13]. The EB of a given feature is grounded in the previous literature documenting the speech dysfunctions connected to PD. An ideal robust feature is a feature displaying a homogeneous behavior *across corpora*. The features analyzed and their EBs are summarized in Table 2.

4. Results and Discussion

The upper section of Table 3 reports the pairwise Kruskal–Wallis $H-$test results for the features that were significant ($p < 0.05$) in the statistical analysis. With respect to the first family of phonatory features (see Table 2), the values of the features representing rJitta, rShim, rAPQ, relative PPQ, NNE, and GNE were significantly higher for PDs than CNs on GITA, but significantly lower on ItalianPVS. A similar behavior on ItalianPVS can be found in a previous study using a different feature extraction library [31]. Differently, features related to HNR ratio and CHNR were significantly lower for the PD than for the CN group on GITA (as expected) but significantly higher for participants with PD than for CNs on ItalianPVS. A similar result for CHNR was obtained on Neurovoz using the vowel /u:/, where the median value of the PD group was significantly higher than that of the CN group. Thus, since the observed behaviors of amplitude, frequency perturbation, and fluctuation features contradicted the assumptions on their EBs on ItalianPVS and Neurovoz, they violated the second criterion of robustness. As such, these features cannot be considered robust in our framework. Moreover, as the Hurst Exponent (mean) was significant only on GITA, it violates the first criterion of robustness, which requires a feature to show significance in at least two data sets.

[6]To perform the correlation analysis, we used *scipy.stats.spearmanr* library in Python.

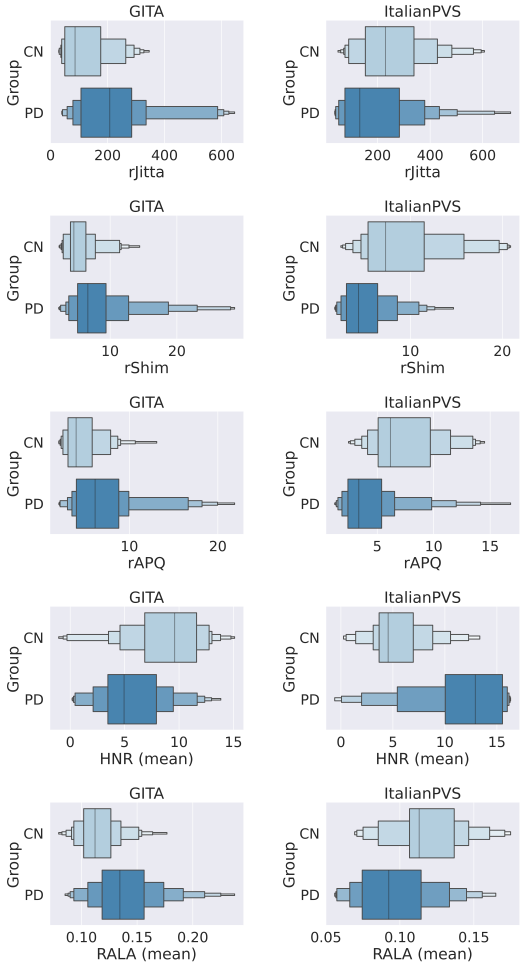

Figure 1: Box plots showing some of the significant features ($p < 0.05$) extracted from the sustained phonation of the vowel /a:/ on GITA and ItalianPVS.

With respect to the modulation spectra features, RALA (mean) was significant on both GITA and ItalianPVS. However, although this feature displayed the EB on GITA, where it was significantly higher in the PD than in the CN group, it shows the opposite behavior on ItalianPVS (see Figure 1). Similarly, even though CIL, PE, and LLE were significant on ItalianPVS, they behaved against the EBs postulated for these features. It

Table 3: Summary of the significant results from the statistical and correlation analysis. For the statistical analysis, corpus name, feature name, p-value, observed behavior (OB), and AUROC are reported for each significant feature. For the correlation analysis, corpus name, clinical score, feature name, p-value, OB, and Spearman's rank correlation coefficient (ρ) are reported for each significant correlation. In square brackets, next to the corpus name, we specify the vowel(s) on which the results were obtained. The AUROC is reported for the vowel /aː/. If the OB matches the EB of a given feature, the arrow representing the EB of the feature is bolded. Abbreviations: ↑, increasing; ↓, decreasing.

Statistical Analysis

Corpus	Feature	*p*-value	OB	AUROC
GITA [/aː/,/eː/, /iː/, /oː/, /uː/]	rJitta	< 0.001	**↑**	0.75
	rRrRAP	< 0.001	**↑**	0.78
	rPPQ	< 0.001	**↑**	0.80
	rSPPQ	< 0.001	**↑**	0.75
	rShdB	< 0.001	**↑**	0.75
	rShim	< 0.001	**↑**	0.74
	rAPQ	< 0.001	**↑**	0.75
	NNE (mean)	< 0.001	**↑**	0.76
	HNR (mean)	< 0.001	**↓**	0.75
	CHNR (mean)	0.03	**↓**	0.65
	GNE (mean)	< 0.001	**↓**	0.66
	RALA (mean)	< 0.001	**↑**	0.74
Neurovoz [/uː/]	CHNR (mean)	0.03	↑	0.66
ItalianPVS [/aː/,/eː/, /iː/, /oː/, /uː/]	rJitta	0.04	↓	0.70
	rPPQ	0.04	↓	0.69
	rShdB	0.07	↓	0.77
	rShim	0.08	↓	0.76
	rAPQ	0.007	↓	0.77
	NNE (std)	0.02	↓	0.62
	HNR (mean)	< 0.001	↑	0.84
	CHNR (mean)	0.003	↑	0.79
	GNE (mean)	0.008	↑	0.75
	LLE (mean)	< 0.001	↓	0.86
	PE (mean)	< 0.001	↓	0.86
	RALA (mean)	0.04	↓	0.70
	MSHphase (mean)	0.04	↓	0.68
	CIL (mean)	0.03	↓	0.71

Correlation Analysis

Corpus	Score	Feature	*p*-value	OB	ρ
GITA [/aː/]	H&Y	Hurst (mean)	0.01	**↑**	0.35
		rApEn (mean)	0.04	**↑**	0.28
		CIL (mean)	0.03	**↑**	0.30
		RALP25 (mean)	0.04	**↑**	0.29
	UPDRS-III	Hurst (mean)	< 0.001	**↑**	0.48
		DFA (mean)	0.01	**↑**	0.36
Neurovoz [/aː/]	H&Y	rJitt	0.004	**↑**	0.51
		rPPQ	< 0.001	**↑**	0.57
		rSPPQ	0.001	**↑**	0.56
		rShdB	0.008	**↑**	0.47
		rShim	0.004	**↑**	0.51
		rAPQ	0.003	**↑**	0.52
		HNR (mean)	0.003	**↓**	0.53
		CHNR (mean)	0.01	**↓**	0.44
		RALA (mean)	0.04	**↑**	0.38
		RALP75 (mean)	0.03	**↑**	0.40
	UPDRS-III	rSPPQ	0.04	**↑**	0.37
GermanPD [/aː/]	H&Y	RALP75 (mean)	0.04	**↑**	0.22
		RALP95 (mean)	0.01	**↑**	0.27
NLS [/eː/]	UPDRS-III	HNR (mean)	0.006	**↓**	0.63

follows that even the modulation spectra features did not satisfy the robustness conditions and, as such, did not pass the test of robustness.

On the whole, all the significant results reported on ItalianPVS and GITA were replicated across the phonations of the different vowels analyzed. Similarly, in the other data sets, no significance was reported for any vowel, except for Neurovoz when analyzing the vowel /uː/ (see Table 3). Altogether, these results show that, within a certain data set, changing vowels does not exert a particular impact on the experimental results, which confirms the language independence of vowel phonations for most features. It is important to notice that even though these features were most significant in only two corpora and did not display a robust behavior, this fact does not exclude the possibility that, altogether, they could provide some good separability between PD and CN groups in a multivariate analysis or machine learning classification experiments. However, our results suggest that the analyzed features cannot be reliably used in a clinical scenario as biomarkers of PD.

The lower section of Table 3 reports the significant correlations between the feature values and the clinical scores (i.e., H&Y scale, UPDRS-III). The correlation results slightly differed when considering different vowels. Thus, we report and discuss results for the vowel /aː/ as this vowel is contained in all the corpora, except in NLS, for which we report results for the vowel /eː/. For ItalianPVS, no clinical scores were available, so we could not perform any correlation analysis on this data set. Features encoding amplitude, frequency, perturbation, fluctuation, and entropy-related metrics were the most predictive of disorder severity as they show significant weak ($\rho = [0.20 - 0.39]$) or moderate correlations ($\rho = [0.40 - 0.59]$) with clinical scores across corpora. The only two features that showed a significant correlation with the clinical scores in more than one corpus were HNR (mean) and RALP75 (mean). In general, the trends of the significant correlations were aligned with the EBs of the features in each speech corpus analyzed. However, different features were effective in tracking disorder severity in the different corpora analyzed, a phenomenon that further emphasizes the inconsistent behaviors of these measurements.

5. Conclusions and Future Work

This study examines the behavior of a composite set of interpretable features encoding phonatory information. It aims at exploring the effectiveness of these features in modeling characteristic acoustic patterns occurring in PD by assessing their discriminatory power and cross-corpora robustness. Our statistical analysis showed that the phonatory features provided good separability between PD and CN groups in only two out of five corpora. In this respect, none of the features displayed robustness, as their behaviors were not homogeneous across corpora. On the other hand, our experiments highlight the effectiveness of a subset of phonatory features in capturing disorder severity, as they showed significant correlations with the clinical scores, and the trends of the correlations behaved as expected.

On the whole, vowel phonation has been extensively adopted for its intrinsic benefits of being language-independent and eliminating potential linguistic and articulatory confounds originating in the analysis of running speech. However, the suitability of this task for the assessment of PD when considering multiple cohorts and/or corpus at a time has been questioned by the current study. In this regard, it has been recently shown that linguistic and prosodic features extracted from running speech display robust cross-lingual and cross-corpora behaviors [13, 7, 14]. As such, they might be a more valuable candidate for evaluating PD in a universal framework. Even in studies involving a single corpus, the analysis of phonatory features can be problematic as these features require long audio recordings of sustained phonations (more than 3 s), a fine calibration of the recording system, and a controlled acoustic environment to be estimated [9, 11].

In a future study, we plan to validate the results obtained in this work by analyzing a greater sample of cohorts, balancing the classes in terms of the number of individuals, gender, age, medication time, PD disease severity, and disorder phenotype. We will also conduct cross-lingual machine learning experiments following a rigorous methodology [31] to observe to which extent the lack of robustness reported in this study will affect the model performances.

6. References

[1] J. Schoentgen and R. De Guchteneere, "Time series analysis of jitter," *Journal of Phonetics*, vol. 23, no. 1-2, pp. 189–201, 1995.

[2] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[3] J. Jiménez-Monsalve, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and P. Gomez-Vilda, "Phonation and articulation analyses in laryngeal pathologies, cleft lip and palate, and parkinson's disease," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2017, pp. 424–434.

[4] V. Uloza, V. Saferis, and I. Uloziene, "Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery," *Journal of Voice*, vol. 19, no. 1, pp. 138–145, 2005.

[5] J. Gómez-García, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part iii: Review of acoustic modelling strategies," *Biomedical Signal Processing and Control*, vol. 66, p. 102049, 2021.

[6] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.

[7] J. Rusz, J. Hlavnička, M. Novotnỳ, T. Tykalová, A. Pelletier, J. Montplaisir, J.-F. Gagnon, P. Dušek, A. Galbiati, S. Marelli *et al.*, "Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease," *Annals of neurology*, vol. 90, no. 1, pp. 62–75, 2021.

[8] A. Tsanas 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," *PROCEEDINGS E REPORT*, p. 161, 2021.

[9] ——, "Biomedical speech signal insights from a large scale cohort across seven countries: The parkinson's voice initiative study," *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pp. 45–48, 2019.

[10] D. Kovac, J. Mekyska, V. Aharonson, P. Harar, Z. Galaz, S. Rapcsak, J. R. Orozco-Arroyave, L. Brabenec, and I. Rektorova, "Exploring language-independent digital speech biomarkers of hypokinetic dysarthria," *medRxiv*, 2022.

[11] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, "Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[12] F. Schaeffler, S. Jannetts, and J. M. Beck, "Reliability of clinical voice parameters captured with smartphones–measurements of added noise and spectral tilt," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH, Graz, Austria, 15-19 September 2019*. ISCA, 2019.

[13] A. FAVARO, L. Moro-Velazquez, A. Butala, C. Motley, T. Cao, R. D. Stevens, J. Villalba, and N. Dehak, "Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in parkinson's disease," *Frontiers in Neurology*, vol. 14, p. 317, 2023.

[14] D. Kovac, J. Mekyska, Z. Galaz, L. Brabenec, M. Kostalova, S. Z. Rapcsak, and I. Rektorova, "Multilingual analysis of speech and voice disorders in patients with parkinson's disease," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2021, pp. 273–277.

[15] A. Favaro, C. Motley, T. Cao, M. Iglesias, A. Butala, E. S. Oh, R. D. Stevens, J. Villalba, N. Dehak, and L. Moro-Velázquez, "A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 532–539.

[16] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, and N. Dehak, "A forced gaussians based methodology for the differential evaluation of parkinson's disease by means of speech processing," *Biomedical Signal Processing and Control*, vol. 48, pp. 205–220, 2019.

[17] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, 2017.

[18] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 342–347.

[19] S. Skodda, W. Grönheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199–e205, 2011.

[20] M. Brückl, A. Ghio, and F. Viallet, "Measurement of tremor in the voices of speakers with parkinson's disease," *Procedia Computer Science*, vol. 128, pp. 47–54, 2018.

[21] R. Viswanathan, S. P. Arjunan, A. Bingham, B. Jelfs, P. Kempster, S. Raghav, and D. K. Kumar, "Complexity measures of voice recordings as a discriminative tool for parkinson's disease," *Biosensors*, vol. 10, no. 1, p. 1, 2019.

[22] S. S. Upadhya, A. Cheeran, and J. Nirmal, "Statistical comparison of jitter and shimmer voice features for healthy and parkinson affected persons," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2017, pp. 1–6.

[23] T. Khan, J. Westin, and M. Dougherty, "Cepstral separation difference: A novel approach for speech impairment quantification in parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 25–34, 2014.

[24] L. A. Ramig, I. R. Titze, R. C. Scherer, and S. P. Ringel, "Acoustic analysis of voices of patients with neurologic disease: rationale and preliminary data," *Annals of Otology, Rhinology & Laryngology*, vol. 97, no. 2, pp. 164–172, 1988.

[25] D. A. Rahn III, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis," *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.

[26] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[27] H. Zhang, N. Yan, L. Wang, and M. L. Ng, "Energy distribution analysis and nonlinear dynamical analysis of phonation in patients with parkinson's disease," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 630–635.

[28] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," *BioMed research international*, vol. 2015, 2015.

[29] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in bioengineering and biotechnology*, vol. 4, p. 1, 2016.

[30] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010.

[31] A. S. Ozbolt, L. Moro-Velazquez, I. Lina, A. A. Butala, and N. Dehak, "Things to consider when automatically detecting parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues," *Applied Sciences*, vol. 12, no. 3, p. 991, 2022.