STATISTICAL & CLINICAL UTILITY OF MULTIMODAL DIALOG BASED SPEECH & FACIAL METRICS FOR PARKINSON'S DISEASE ASSESSMENT

Hardik Kothare[1], Michael Neumann[1], Jackson Liscombe[1], Oliver Roesler[1], William Burke[1], Andrew Exner[2], Sandy Snyder[2], Andrew Cornish[1], Doug Habberstad[1], David Pautler[1], David Suendermann-Oeft[1], Jessica E. Huber[2] & **Vikram Ramanarayanan**[1,3]

[1]Modality.AI, Inc. [2]Purdue University [3]University of California, San Francisco   Contact: v@modality.ai

Motivation & Key Questions

While there is much work on multimodal features and machine learning methods to characterize neurological and mental health, *there remains a gap between these scientific advances and clinical adoption*.

This work aims to bridge this gap by proposing a principled framework for investigating the statistical and clinical utility of various speech/facial metrics.

1. Which metrics show significant differences between people with Parkinson's Disease (pPD) and controls and how reliable are these metrics?

2. For metrics that show differences, what value represents a difference or change above and beyond any measurement errors (*statistical utility*)?

3. For metrics that show differences, what value might represent an actual clinical change tied to physiological manifestations of PD (*clinical utility*)?

Data & Methods

- This study includes data from 60 participants (243 sessions) recruited through the Purdue Motor Speech Lab (Nov '20 - Jan '22). Participants were asked to complete four sessions, a week apart from each other.
- Controls were age- and sex-matched. See Table 1 for demographic info.
- The conversational callflow required participants to do the following speaking exercises: (a) sustained vowel (steady /a/ , up-or-down pitch glide /i/), (b) read speech: speech intelligibility test (SIT) sentences, sentences that elicited variation in intonational prosody, rainbow passage, (c) story retells and (d) spontaneous speech on any topic of their choice.
- Speech acoustic and facial kinematic metrics were automatically extracted (Table 2). Facial metrics were normalised for each participant by the inter-caruncular distance between the eyes. **Non-parametric Kruskal-Wallis tests** were performed to investigate differences between pPD and controls.

Group	Controls	pPD
Sex	18F / 4M	19F / 19M
Age (years)	65; 63.46 (11.08)	71; 67.48 (9.30)
MoCA score	28; 27.55 (1.92)	27; 26.06 (3.63)
Years since diagnosis	n/a	5; 7.89 (6.16)
Region	2 urban, 15 suburban, 5 rural	6 urban, 23 suburban, 9 rural
Session status		
Completed successfully	87	142
User restarted	3	6
User hung up early	0	10
Recoverable system error	0	1

Inclusion Criteria	*Exclusion Criteria*
30 < age < 85; English fluency	non-PD neurological disorder
diagnosis of idiopathic PD	head & neck cancer/surgery
device with mic/camera	pulmonary disease
internet access	MoCA score < 10
no hearing and vision loss	smoking (in the past 5 years)

Table 1. Participant Demographics & pPD inclusion/exclusion criteria

Acoustic measures	• **Fundamental Frequency (F0):** Minimum (Hz) & timepoint (s), Maximum (Hz) & timepoint (s), Mean (Hz), Std Deviation (Hz) • **Formant Frequency Values:** F1, F2, F3 (Hz) and F2 slope (Hz/s) • **Cepstral Peak Prominence** (CPP in dB) • **Harmonics-to-Noise Ratio** (HNR in dB) • **Articulation duration** (in s, excluding pauses) and **speaking duration** (in s, including pauses) • **Articulation rate** and **speaking rate** (words per minute) • **Percent pause duration** (%) • **Signal-to-noise ratio** (SNR in dB) • **Articulation intensity** (dB) • **Jitter** and **shimmer** (%)
Visual measures	velocity, acceleration, and jerk of lower lip and jaw center, lip aperture, lip width, eye opening, vertical eyebrow displacement, eye blinks, area of the mouth, symmetry ratio of the mouth area

Table 2. Automatically extracted acoustic & visual measures.

Measures: Statistical & Clinical Utility

Minimally Detectable Change (MDC) at 95% confidence level is defined as:

$$MDC_{95} = 1.96 \times \sqrt{2} \times SEM \qquad SEM = \sigma \times \sqrt{1-\rho}$$

SEM is the standard error of measurement for a particular metric calculated from all participants across their four sessions.

Minimal Clinically Important Difference (MCID) is defined as the smallest change in a domain that is thought to be clinically relevant or has an impact on patients, clinicians or caregivers . MCID can be considered as a threshold for a change that would be treated as an improvement or deterioration in function.

Proposition: *Metric's effect size* $\geq$ *MDC & MCID* $\geq$ *MDC* to have **clinical utility**.

To tie MCID to clinical meaningfulness, we used the Communicative Participation Item Bank (CPIB-S) as an external anchor (clinical gold standard).

pPD were classified into two sub-cohorts based change in CPIB-S T score:

1. *No change*: Change in T score = 0 (n=8)

2. *Decline*: Deterioration in CPIB-S T score < -0.74 or more than the standard error of the mean of the distribution (n=13)

We used ROC curves of a simple binary classifier to determine how well the changes in each metric differentiated between these two sub-cohorts. See Figure 1.

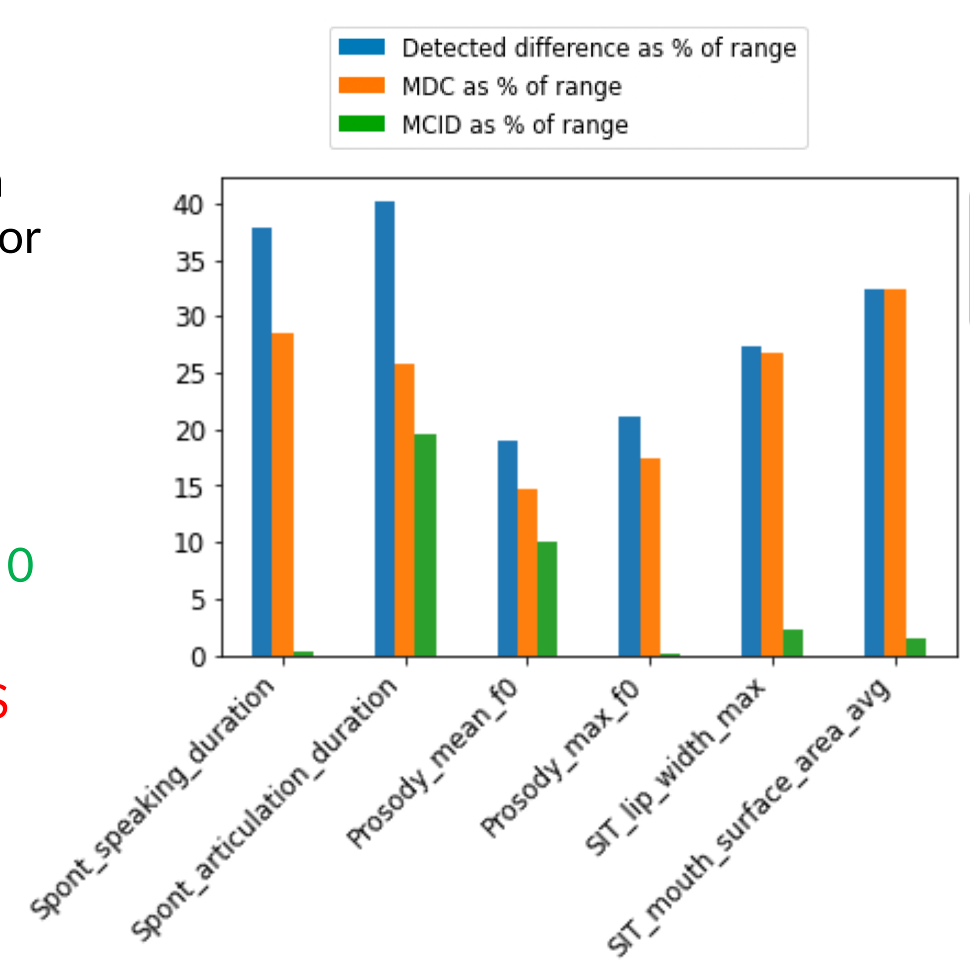

Figure 1. Metrics with detected differences between the median values of the two cohorts (pPD and controls) greater than MDC

Metric Performance: Accuracy & Reliability

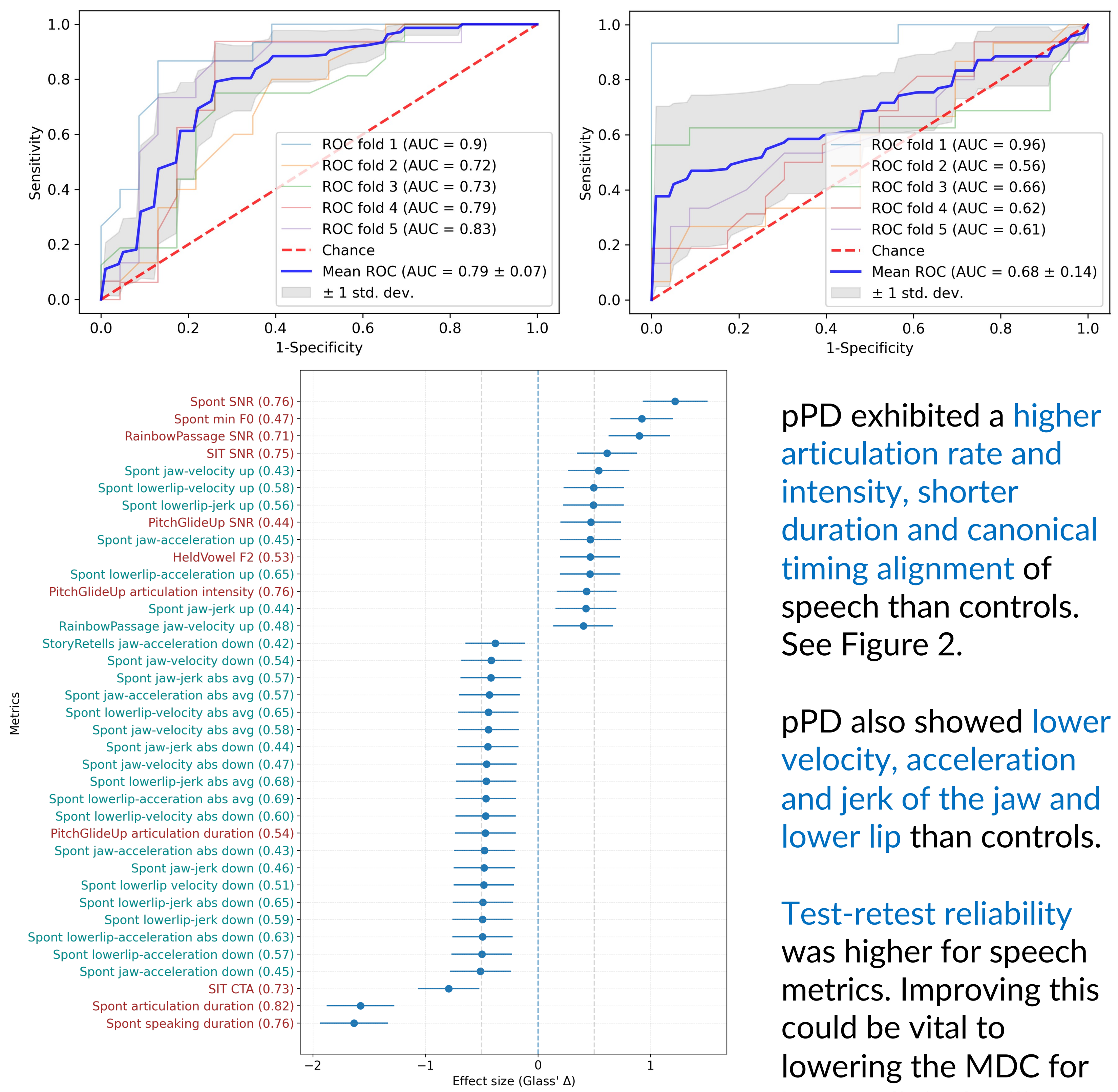

Figure 2. Classification ROC curves (top panel) and effect sizes (bottom panel) of acoustic and facial metrics that show significant differences between pPD and controls (p<0.01).

pPD exhibited a higher articulation rate and intensity, shorter duration and canonical timing alignment of speech than controls. See Figure 2.

pPD also showed lower velocity, acceleration and jerk of the jaw and lower lip than controls.

Test-retest reliability was higher for speech metrics. Improving this could be vital to lowering the MDC for better clinical utility.

Conclusions and Limitations

- We examined a set of measures to characterize the statistical and clinical utility of speech/facial biomarkers of Parkinson's Disease.
- In the case study examined, speaking and articulation duration in particular demonstrated significant effect sizes between pPD and controls greater than the MDC with high reliability.
- The relatively reduced performance of facial metrics could be due to their larger range and lower test-retest reliability; recent experiments show that improving the accuracy of estimation could improve this.
- Future work will examine alternate/better clinical anchors for MCID, and a larger sample size over a longer time period for more improved estimates.

References

1. Stipancic et al. (JSLHR 2018).
2. Ramanarayanan et al. (Interspeech 2020).
3. Kothare et al. (EMBC 2022)
4. Tsanas et al (IEEE Trans Bio Imag 2012)
5. Vasquez-Correa et al. (IEEE Bio Health Inf 2018)
6. Ramanarayanan et al. (ASHA Perspectives 2022)
7. Godino-Llorente et al. (PLOS One 2017)
8. Baylor et al. (2013). CPIB-S.
9. Guarin et al (IEEE FG 2020).