Speech, Facial and Fine Motor Features for Conversation-Based Remote Assessment and Monitoring of Parkinson's Disease

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

***Abstract*— We present a cloud-based multimodal dialogue platform for the remote assessment and monitoring of speech, facial and fine motor function in Parkinson's Disease (PD) at scale, along with a preliminary investigation of the efficacy of the various metrics automatically extracted by the platform. 22 healthy controls and 38 people with Parkinson's Disease (pPD) were instructed to complete four interactive sessions, spaced a week apart, on the platform. Each session involved a battery of tasks designed to elicit speech, facial movements and finger movements. We find that speech, facial kinematic and finger movement dexterity metrics show statistically significant differences between controls and pPD. We further investigate the sensitivity, specificity, reliability and generalisability of these metrics. Our results offer encouraging evidence for the utility of automatically-extracted audiovisual analytics in remote monitoring of PD and other movement disorders.**

I. INTRODUCTION

The need for remote monitoring to support patients, caregivers, and healthcare professionals in their collaborative efforts for better care has never been greater, a situation which has been brought into more acute focus by the SARS-COV-2 pandemic [1]. Indeed, the majority of people with Parkinson's Disease (pPD) have limited access to specialists [2] making outcome improvements through remote patient monitoring (RPM) particularly relevant for PD. Up to 90% of pPD exhibit dysarthria during the course of their disease [3]. Some characteristics of Parkinsonian or hypokinetic dysarthria are monopitch and monoloudness, reduced stress and breathiness [4], thus making acoustic and articulatory parameters related to speech production important indicators of disease progression in PD [5]. Indeed, previous work has demonstrated that speech acoustics [6], [7], articulation [8], [9], orofacial kinematics [10] and motor function [11] can prove to be important biomarkers of PD [12]. Speech, breathing and non-speech oral exercise based therapies have shown encouraging improvements in speech production with a direct impact on intelligibility and an indirect impact on activities of daily living [13]. However, such therapies may be intensive and often involve multiple visits to the clinic, precluding a vast chunk of the population from seeking specialist care. Self-driven conversational RPM that extracts speech acoustic and articulatory features automatically has the potential to revolutionise health care for pPD. Moreover, motor symptoms in PD are traditionally assessed in the clinic through tests like finger tapping [14]. Most existing *contactless* RPM systems that assess dexterity of finger movements are based on smartphone apps in which a variant of the finger tapping task is used by asking the user to alternately tap buttons on the screen, see for instance [15]. While automatic evaluation of finger tapping from videos has shown promising results [16], [17], these studies were so far conducted with on-site video recordings in controlled conditions. Utilising remotely-recorded video data from a participant's webcam combines the benefits of cost-effective, frequent remote monitoring with the ability for clinicians to review the performance (in addition to the scalable automatic processing).

In this paper, we present Tina, a virtual dialogue agent that conducts on-demand automated interviews through a HIPAA-compliant, secure screening portal over an internet browser. During the conversation, Tina engages participants in a mixture of structured speaking exercises and open-ended questions to elicit speech, facial and fine motor behaviour (the latter using novel finger-tapping exercises designed to test limb motor function). We leverage the rich multimodal data collected to answer the following research questions regarding the feasibility of multimodal dialogue technology for remote assessment and monitoring of PD:

1) Which metrics show significant differences between pPD and healthy controls? How reliable are these metrics?
2) What is the unweighted average recall (UAR) for classifying pPD from controls? How generalisable are these findings?
3) What is the relative performance of speech, facial and finger-tapping metrics in distinguishing pPD from healthy controls?

II. SYSTEM

The virtual agent Tina introduced above is powered by NEMSI [18] - the Neurological and Mental health Screening Instrument, which is a cloud-based multimodal dialogue system designed to conduct automated screening interviews that elicit evidence for detection and progress monitoring of neurological and mental health. During each call, analytics modules automatically extract a variety of audio (e.g., speaking rate, duration), facial (e.g., range and speed of movement

[1]Modality.AI, Inc. [2]Purdue University [3]University of California, San Francisco

of lips and jaw) as well as finger-tapping metrics in real time and store them in a database together with meta information of the interaction, like captured participant responses, call duration, or completion status [19]. This information can be accessed by clinicians during and after the interaction through an easy-to-use dashboard, which provides a high-level overview of the interaction and a detailed breakdown of individual interaction turns.

III. DATA

All participants were recruited and informed consent was obtained by the Purdue Motor Speech Lab at Purdue University. This study was approved by Purdue's Institutional Review Board. The Montreal Cognitive Assessment (MoCA) [20] was administered to every participant to test for cognitive impairment. Data from 60 participants (see Table 1) collected between November 2020 and January 2022 were used in this study. Although each participant was asked to complete four sessions, spaced a week apart, some participants did more than four sessions. This resulted in a total of 249 sessions. The finger-tapping exercise was introduced in August 2021 and data are available for only 19 participants. A notable highlight of this data collection is that even though we recruited participants from diverse urban, suburban and rural regions, including some with occasional connectivity issues, we observed a $\sim 92\%$ completion rate among our elderly participant pool who reported a successful and favourable interaction with the system.

TABLE I: *Participant Demographics: Age, MoCA scores and years since diagnosis are presented as: median; mean (standard deviation).*

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

The conversational flow elicited speech samples of the following types from participants: (a) sustained vowel phonation (steady / with up-or-down pitch glide), (b) read speech, (c) story retells and (d) spontaneous speech. For (b) read speech, participants were asked to read speech intelligibility test (SIT) sentences, sentences that elicited prosodic variation (Prosody) and a reading passage (Rainbow Passage). For (d), participants were asked to speak about any topic of their choice with a few topics listed on the screen.

For the finger tapping exercises, participants were instructed to hold their hand up to the camera and perform a tapping motion for ten seconds. During the exercise, anatomical landmarks of the participant's hand were derived from the recorded image frames. The coordinates of the tip of the index finger and tip of the thumb were recorded to a database for subsequent calculation of metrics (see Figure 1). The finger tapping assessment comprised three tasks which differed based on the instructed goal of the tap, i.e. participants were instructed to make the tapping movement as (1) wide, (2) fast, or (3) both wide and fast as possible. They were then asked to repeat the same three tasks with the other hand.

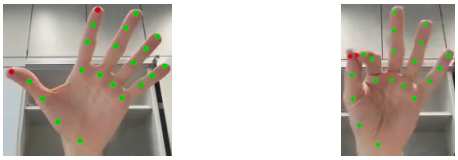


Fig. 1: Illustration of the 21 obtained and 2 used (shown in red) hand landmarks for the two key points of interest, i.e. fingers open and fingers closed.

IV. METHODS

A. Extraction of metrics

All acoustic metrics were automatically extracted using Praat [21] and the Montreal Forced Aligner [22] (to extract timing agreement of read sentences relative to a canonical pronunciation). See Table II for a complete list.

TABLE II: *Extracted acoustic metrics. F0 = fundamental frequency, F1/F2/F3 = first three formant frequencies, PPT = percent pause time, CPP = cepstral peak prominence, HNR = harmonics-to-noise ratio, WPM = words per minute, MFA = Montreal Forced Aligner.*

Speech type	Collected metrics
Vowel	Min, Max, Mean F0 (Hz), jitter (%), shimmer (%), HNR (dB), articulation duration (sec), intensity (dB), F1 (Hz), F2 (Hz), F3(Hz)
Prosody and Story retells	Min, Max, Mean F0 (Hz), jitter (%), shimmer (%), HNR (dB), CPP (dB), intensity (dB), PPT (%), speaking duration (sec)
Read speech	Min, Max, Mean F0 (Hz), jitter (%), shimmer (%), HNR (dB), CPP (dB), intensity (dB), PPT (%), MFA timing agreement (%), speaking duration (sec), speaking rate (WPM), articulation duration (sec), articulation rate (WPM)
Spontaneous Speech	Min, Max, Mean F0 (Hz), jitter (%), shimmer (%), HNR (dB), CPP (dB), intensity (dB), PPT (%), speaking and articulation duration (sec)

Facial metrics were calculated for each utterance in three steps: (i) face detection using the face detector in the dnn module of OpenCV (`https://opencv.org/`) , which uses a Single Shot Detector architecture to determine the (x, y)-coordinates of one or more faces for every input frame, (ii) facial landmark extraction using the Dlib facial landmark detector, which uses an ensemble of regression trees [23] to extract 68 facial landmarks, and (iii) facial metrics calculation, which uses 20 facial landmarks to compute metrics like the speed and acceleration of articulators (jaw, lower lip), surface area of the mouth, etc. See [24] for details.

Metric extraction for the finger tapping exercise is performed in three steps: (1) hand detection, (2) hand landmark extraction, and (3) hand metrics calculation. For hand and hand landmarks detection MediaPipe Hands [25] is used, which is a hand tracking pipeline implemented via MediaPipe [26]. MediaPipe Hands first employs a palm detector that outputs a cropped hand bounding box, which is then provided as input to the hand landmarks detection model,

TABLE III: *Extracted finger-tapping metrics.*

Metrics	Description
velocity / acceleration	Maximum (_max) and difference between first half and second half (_diff)
Jitter	Cycle-to-cycle variation of time period
Shimmer	Cycle-to-cycle variation of amplitude

which in turn returns 21 landmarks as illustrated by Figure 1. The positions of the tips of the thumb and index finger are then used to calculate the metrics in Table III.

V. STATISTICAL ANALYSES & RESULTS

All metrics were z-scored by sex to normalize for sex-specific differences. Wherever applicable, metrics were averaged across tasks (except speaking duration and articulation duration). For every acoustic, facial and finger-tapping metric, we performed a non-parametric Kruskal-Wallis test to identify the metrics which showed significant differences between pPD and controls at $\alpha = 0.05$. Figure 2 shows all the metrics that can distinguish between pPD and controls along with effect sizes, measured as Glass' Δ. In terms of acoustic metrics, pPD exhibited a higher articulation rate, greater articulation intensity, shorter duration of speech across various tasks and lesser agreement with the expected duration in the SIT task than controls. pPD also showed lower speed and acceleration of the jaw and lower lip, smaller lip aperture and lesser surface area of the mouth during speaking. Acceleration during left hand finger-tapping tasks was lower in pPD than in controls.

We calculated the test-retest reliability coefficient of these metrics by taking their average absolute Pearson's correlation coefficient between all pairs of sessions (displayed in parentheses in Figure 2). Acoustic metrics showed better test-retest reliability than facial metrics. Finger-tapping metrics that showed statistically significant differences between the two groups displayed large session-to-session variability.

Additionally, to perform a binary classification between the two cohorts, we conducted a 5-fold cross-validation with a random forest classifier. This cross-validation was performed using (a) acoustic metrics alone, (b) facial metrics alone, (c) finger-tapping metrics alone[1]. Receiver operating characteristics (ROC) curves for these classification analyses can be seen in Figure 3. The mean unweighted average recall (UAR) across 5 cross-validation folds when only acoustic metrics were included stood at 0.65 ± 0.16 (Figure 3a), which is above chance. When facial metrics alone were considered (Figure 3b), the average UAR was 0.54 ± 0.07. When finger-tapping metrics alone were included (Figure 3c), average UAR was 0.53 ± 0.15.

VI. DISCUSSION

This work provides preliminary evidence on the feasibility of a multimodal, dialogue-based remote patient monitoring method to track Parkinson's Disease and other movement disorders. The high completion rate of sessions (92%) as seen in Table I is an encouraging indicator of RPM technology adoption by elderly populations. This speaks to the utility of such RPM technology, even in rural locations with occasional internet connectivity issues.

[1]We also investigated fusion of modalities, but this did not improve performance, potentially because of the smaller sample size for sessions that had metrics from all 3 modalities available.

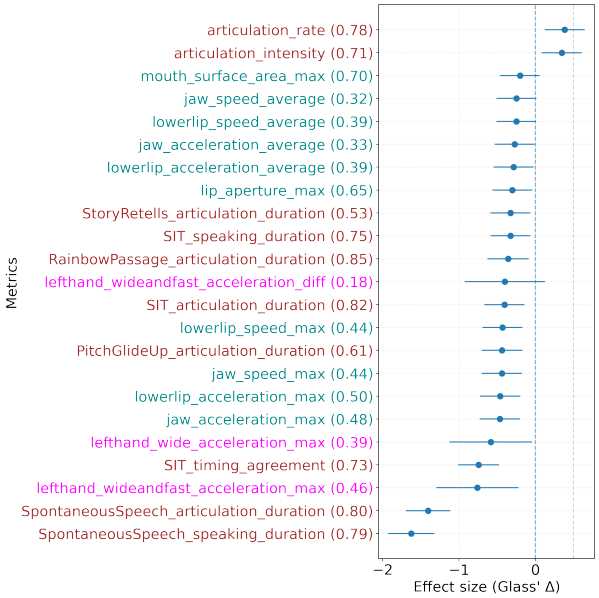
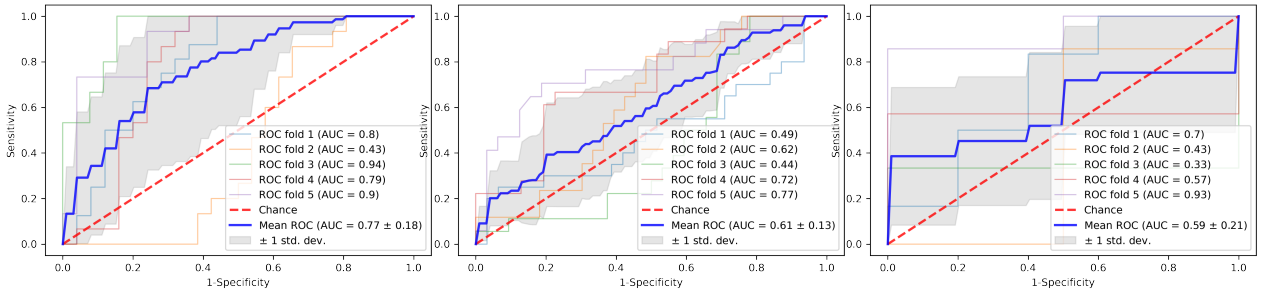


Fig. 2: Effect sizes of **acoustic**, **facial** and **finger-tapping** metrics that show statistically significant differences between controls and pPD at $\alpha = 0.05$. Test-retest reliability measured as the average Pearson's correlation coefficient across all pairs of sessions reported in parentheses.

We found that speech acoustic, facial kinematic and finger tapping metrics can be used to distinguish between pPD and controls. In particular, pPD had a higher articulation rate; abnormalities in articulation and speaking rate in PD are well-documented [27]. pPD also showed greater articulatory intensity, an observation which could be influenced by the distance of the participants from their microphones. Notably, pPD spoke for a shorter duration in most tasks. Facial kinematic metrics like the speed and acceleration of the lower lip and jaw also showed differences between pPD and controls. During finger tapping tasks using the left hand, pPD showed lower acceleration, perhaps indicative of motor rigidity. In general, speech acoustic metrics demonstrated better test-retest reliability from session to session than facial or finger-tapping metrics. Acoustic metrics performed well in the classification experiment in distinguishing between the two groups. Facial metrics performed relatively worse, but this could be due to lower reliability of these features arising from low quality video connections. Finger tapping metrics also did not perform as well as speech metrics, perhaps because of the smaller sample size. It is for this reason that fusion of all modalities was not effective for this limited sample. Additionally, information about participants' handedness was not available for this analysis. Future investigations with larger samples will also take handedness into consideration to better understand finger tapping behaviour in pPD.



(a) **All acoustic metrics (N = 203)** (b) **All facial metrics (N = 249)** (c) **All finger-tapping metrics (N = 46)**

Fig. 3: ROC curves displaying the performance of binary classification with 5-fold cross-validation.

REFERENCES

[1] F. Motolese, A. Magliozzi, F. Puttini, M. Rossi, F. Capone, K. Karlinski, A. Stark-Inbar, Z. Yekutieli, V. Di Lazzaro, and M. Marano, "Parkinson's disease remote patient monitoring during the covid-19 lockdown," *Frontiers in neurology*, vol. 11, 2020.

[2] M. Achey, J. L. Aldred, N. Aljehani, B. R. Bloem, K. M. Biglan, P. Chan, E. Cubo, E. Ray Dorsey, C. G. Goetz, M. Guttman *et al.*, "The past, present, and future of telemedicine for parkinson's disease," *Movement disorders*, vol. 29, no. 7, pp. 871–883, 2014.

[3] K. Tjaden, "Speech and swallowing in parkinson's disease," *Topics in geriatric rehabilitation*, vol. 24, no. 2, p. 115, 2008.

[4] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.

[5] J. Rusz, T. Tykalová, M. Novotnỳ, E. Ržička, and P. Dušek, "Distinct patterns of speech disorder in early-onset and late-onset de-novo parkinson's disease," *npj Parkinson's Disease*, vol. 7, no. 1, pp. 1–8, 2021.

[6] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, and P. J. Snyder, "Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment," *Journal of Neurolinguistics*, vol. 17, no. 6, pp. 439–453, 2004.

[7] 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.

[8] J. Godino-Llorente, S. Shattuck-Hufnagel, J. Choi, L. Moro-Velázquez, and J. Gómez-García, "Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers," *PloS one*, vol. 12, no. 12, p. e0189583, 2017.

[9] P. Gómez, J. Mekyska, A. Gómez, D. Palacios, V. Rodellar, and A. Álvarez, "Characterization of parkinson's disease dysarthria in terms of speech articulation kinematics," *Biomedical Signal Processing and Control*, vol. 52, pp. 312–320, 2019.

[10] D. L. Guarin, A. Dempster, A. Bandini, Y. Yunusova, and B. Taati, "Estimation of orofacial kinematics in parkinson's disease: Comparison of 2d and 3d markerless systems for motion tracking," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 540–543.

[11] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1618–1630, 2018.

[12] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, pp. 1–8, 2022.

[13] G. M. Schulz and M. K. Grant, "Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease: a review of the literature," *Journal of communication disorders*, vol. 33, no. 1, pp. 59–88, 2000.

[14] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[15] F. Lipsmeier, K. I. Taylor, T. Kilchenmann, D. Wolf, A. Scotland, J. Schjodt-Eriksen, W.-Y. Cheng, I. Fernandez-Garcia, J. Siebourg-Polster, L. Jin *et al.*, "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 parkinson's disease clinical trial," *Movement Disorders*, vol. 33, no. 8, pp. 1287–1297, 2018.

[16] T. Khan, D. Nyholm, J. Westin, and M. Dougherty, "A computer vision framework for finger-tapping evaluation in parkinson's disease," *Artificial intelligence in medicine*, vol. 60, no. 1, pp. 27–40, 2014.

[17] S. Williams, Z. Zhao, A. Hafeez, D. C. Wong, S. D. Relton, H. Fang, and J. E. Alty, "The discerning eye of computer vision: Can it measure parkinson's finger tap bradykinesia?" *Journal of the Neurological Sciences*, vol. 416, p. 117003, 2020.

[18] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill, O. Roesler, and R. Geffarth, "Nemsi: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA)*, Paris, France, July 2019.

[19] V. Ramanarayanan, O. Roesler, M. Neumann, D. Pautler, D. Habberstad, A. Cornish, H. Kothare, V. Murali, J. Liscombe, D. Schnelle-Walka *et al.*, "Toward remote patient monitoring of speech, video, cognitive and respiratory biomarkers using multimodal dialog technology." in *INTERSPEECH*, 2020, pp. 492–493.

[20] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[21] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[23] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, June 2014.

[24] M. Neumann, O. Roessler, D. Suendermann-Oeft, and V. Ramanarayanan, "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, 2020, pp. 47–52.

[25] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, Seattle, WA, USA, 2020.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. http://arxiv.org/abs/1906.08172, 2019.

[27] J. R. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.