INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do

Vikram Ramanarayanan, David Pautler, Lakshmi Arbatti, Abhishek Hosamath, Michael Neumann, Hardik Kothare, Oliver Roesler, Jackson Liscombe, Andrew Cornish, Doug Habberstad, Vanessa Richter, David Fox, David Suendermann-Oeft and Ira Shoulson

Modality.AI, Inc.
v@modality.ai

Abstract

We present a unified multimodal dialog platform for the remote assessment and monitoring of patients' neurological and mental health. Tina, a virtual agent, guides participants through an immersive interaction wherein objective speech, facial, linguistic and cognitive biomarkers can be automatically computed from participant speech and video in near real time. Furthermore, Tina encourages participants to describe, in their own words, their most bothersome problems and what makes them better or worse, through the Patient Report of Problems (PROP) instrument. The PROP captures unfiltered verbatim replies of patients, in contrast with traditional patient reported outcomes that typically rely on categorical assessments. We argue that combining these patient reports (i.e., what they say) with objective biomarkers (i.e., how they say it and what they do) can greatly enhance the quality of telemedicine and improve the efficacy of siteless trials and digital therapeutic interventions.
Index Terms: multimodal dialog, speech biomarkers, patient self-reports, remote patient monitoring, health assessment.

1. Conversational Agents for Remote Patient Assessment

The World Health Organization (WHO) has highlighted the need to structure health services to expand beyond just a disease-centric focus and organize around individuals seeking care[1]. The use of digital conversational agents in personalized healthcare can potentially fill a gap in both the access to and quality of services and health information [1]. Moreover, because of the subjective nature of clinical assessment and because patients typically present with a complex array of symptoms, continuous objective remote monitoring of symptoms and adverse events has the potential to provide real-time information that can help guide the timing of treatments to improve outcomes and help reduce side effects [2]. Such telehealth solutions are feasible, acceptable, cost effective, and have the potential to improve the efficiency of healthcare while reducing the burden on patients and caregivers, particularly for remote monitoring of motor, pulmonary, cognitive and speech function.

Speech is a rich medium which not only serves as a primary source for communication between individuals, but is also a window into diseases of the heart, lungs, brain, muscles, or vocal folds, which in turn then alter an individual's speech patterns [3, 4]. Indeed, many studies have demonstrated the efficacy of various speech biomarkers that capture how a given disease impacts multiple domains of speech performance – be it motor, anatomical, cognitive, linguistic or affective [5, 6, 7, 8]. Therefore, speech analysis using artificial intelligence opens new opportunities for telehealthcare.

[1]https://www.who.int/teams/primary-health-care/conference/declaration

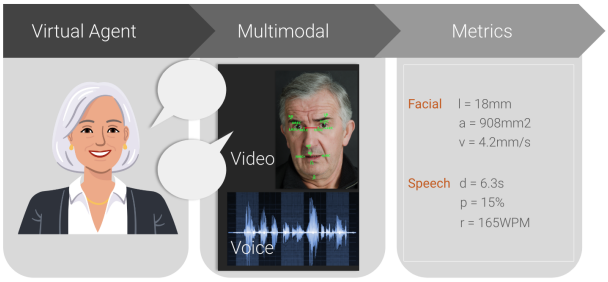

Figure 1: *Multimodal dialogue platform for remote patient assessment driven by a virtual agent, Tina.*

In addition to objective measures of patient behavior captured by such conversational remote patient monitoring technologies, what patients report about their illness is also of critical importance, but has traditionally been captured using categorical scales that are rated by clinicians in research settings. Recent research has demonstrated the efficacy of collecting and analyzing open-ended self-reported responses (called 'verbatims') from patients to questions about what bothers them about their disease and how it affects their daily functioning – also known as the Patient Report of Problems™ (PROP™) [9]. While there are many telehealth solutions for remote patient assessment and monitoring, to our knowledge, there are none that integrate both (i) the measurement of objective biomarkers from multiple modalities and (ii) unfiltered patient verbatim replies about their bothersome problems and functional consequences into one comprehensive solution. This work aims to bridge this gap and demonstrates one such integrative solution.

2. Modality Platform

The Modality platform, powered by a cloud-based multimodal dialog system, leverages a virtual agent user interface to conduct structured conversational interactions with participants for remote health assessment. See Figure 1. Participants can start a conversation with a virtual agent, Tina, through a personalised web URL. At the beginning of the call, tests of the speaker, microphone, and camera need to be passed to ensure that the participants' devices are correctly configured so that the collected data has sufficient quality. Once all device tests pass, Tina guides participants through a battery of tasks that elicit speech and facial behaviours like vowel phonation, counting up of numbers in a single breath, repeating consonant-vowel-consonant (CVC) words, diadochokinesis, reading sentences and passages, picture description and production of spontaneous speech on a topic of their choice. Multimodal analyt-

Table 1: *Overview of extracted metrics. For visual metrics, functionals (minimum, maximum, average) are applied to produce one value across all video frames of an utterance. Visual distance metrics are measured in pixels and are normalized by dividing them by the intercanthal distance (distance between inner corners of the eyes) for each participant.*

	Modality	Domain	Exemplar Metrics
What Patients Do	*Audio*	Energy	shimmer (%), intensity (dB), signal-to-noise ratio (dB)
		Timing	speaking and articulation duration (sec.), articulation and speaking rate (WPM), percent pause time (PPT, %), canonical timing agreement (CTA, %)
		Voice quality	cepstral peak prominence (CPP, dB), harmonics-to-noise ratio (HNR, dB)
		Frequency	mean, max., min. fundamental frequency F0 (Hz), first three formants F1, F2, F3 (Hz), slope of 2nd formant (Hz/sec.), jitter (%)
	Text	Lexico-semantic	word count, percentage of content words, noun rate, verb rate, pronoun rate, noun-to-verb ratio, noun-to-pronoun ratio, closed class word ratio, idea density
		Sentiment	Empath positive cosine similarity, Empath negative cosine similarity [10]
	Video	Mouth (distances)	lip aperture/opening, lip width, mouth surface area, mean symmetry ratio between left and right half of the mouth
		Lip/Jaw Movement	velocity, acceleration, jerk, and speed of lower lip and jaw center
		Eyes	number of eye blinks per sec., eye opening, vertical displacement of eyebrows
		Oro-motor Exam	range of motion of lips and jaw, head pose
		Limb Motoric	finger tapping rate and duration, jitter and shimmer
	Cognitive	Eye gaze	Saccade rates, reaction times and fixation durations for smooth pursuit, saccade, free and directed image exploration, and congruent and incongruent Stroop tasks.
		Cognitive scores	reaction times and percentage of correct words (immediate and delayed word recall), digit span forward/backward score (ranges from 0 to 2)
What Patients Say		*PROP*	Clinical symptom probabilities (predicted by trained ML model) based on responses to: Tell us, in your own words, what bothers you the most about your condition? How does this affect your daily functioning? What makes this better or worse?

ics modules automatically extract features (see Table 1) that capture information from acoustic (energy, timing, voice quality, spectral), facial (articulatory kinematics, range of motion, eye and facial movement), motoric (finger tapping kinematics) and textual (lexico-semantic, sentiment) domains during these tasks. Tina can also administer tasks that probe cognitive abilites of participants – such as working memory, executive function, attention and word fluency – using measures that capture reaction times, recall accuracy, eye gaze saccades and fixations. Finally, participants respond to the Patient Report of Problems™ (PROP™), describing their symptoms and severity, as well as other clinical survey instruments of interest.We then classify these verbatim responses into multiple, clinically-relevant symptoms using a multi-label text classification deep neural network model trained on data collected from over 25,000 patients [9].

3. Summary

We have presented a unified multimodal dialog platform for the remote assessment and monitoring of patients' neurological and mental health. Combining objective audiovisual biomarkers with patient self-report of problems via conversational AI and interpretable machine learning has significant potential to enhance the quality of telemedicine based healthcare and improve the efficacy of siteless trials and digital therapeutic interventions.

4. References

[1] P. Parmar, J. Ryu, S. Pandya, J. Sedoc, and S. Agarwal, "Health-focused conversational agents in person-centered care: a review of apps," *NPJ digital medicine*, vol. 5, no. 1, p. 21, 2022.

[2] T. Sakamaki, Y. Furusawa, A. Hayashi, M. Otsuka, and J. Fernandez, "Remote patient monitoring for neuropsychiatric disorders: a scoping review of current trends and future perspectives from recent publications and upcoming clinical trials," *Telemedicine and e-Health*, vol. 28, no. 9, pp. 1235–1250, 2022.

[3] 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*, vol. 7, no. 1, pp. 276–283, 2022.

[4] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digital biomarkers*, vol. 5, no. 1, pp. 78–88, 2021.

[5] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in psychology*, vol. 8, p. 269, 2017.

[6] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, "Quantifying articulatory impairments in neurodegenerative motor diseases: A scoping review and meta-analysis of interpretable acoustic features," *International Journal of Speech-Language Pathology*, pp. 1–14, 2022.

[7] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, 2022.

[8] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[9] I. Shoulson, L. Arbatti, A. Hosamath, S. W. Eberly, and D. Oakes, "Longitudinal cohort study of verbatim-reported postural instability symptoms as outcomes for online parkinson's disease trials," vol. 12, no. 6, pp. 1969–1978, 2022, publisher: IOS Press. [Online]. Available: https://content.iospress.com/articles/journal-of-parkinsons-disease/jpd223274

[10] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 4647–4657.