ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

Multimodal Technologies for Remote Assessment of Neurological and Mental Health

Vikram Ramanarayanan[a,b]

[a] Modality.AI, Inc., San Francisco, CA [b] Department of Otolaryngology—Head and Neck Surgery, University of California, San Francisco

ARTICLE INFO

Article History:
Received February 28, 2024
Revision received May 6, 2024
Accepted May 7, 2024

Editor-in-Chief: Cara E. Stepp
Editor: Jordan R. Green

https://doi.org/10.1044/2024_JSLHR-24-00142

ABSTRACT

Purpose: Automated remote assessment and monitoring of patients' neurological and mental health is increasingly becoming an essential component of the digital clinic and telehealth ecosystem, especially after the COVID-19 pandemic. This review article reviews various modalities of health information that are useful for developing such remote clinical assessments in the real world at scale.
Approach: We first present an overview of the various modalities of health information—speech acoustics, natural language, conversational dynamics, orofacial or full body movement, eye gaze, respiration, cardiopulmonary, and neural—which can each be extracted from various signal sources—audio, video, text, or sensors. We further motivate their clinical utility with examples of how information from each modality can help us characterize how different disorders affect different aspects of patients' spoken communication. We then elucidate the advantages of combining one or more of these modalities toward a more holistic, informative, and robust assessment.
Findings: We find that combining multiple modalities of health information allows for improved scientific interpretability, improved performance on downstream health applications such as early detection and progress monitoring, improved technological robustness, and improved user experience. We illustrate how these principles can be leveraged for remote clinical assessment at scale using a real-world case study of the Modality assessment platform.
Conclusion: This review article motivates the combination of human-centric information from multiple modalities to measure various aspects of patients' health, arguing that remote clinical assessment that integrates this complementary information can be more effective and lead to better clinical outcomes than using any one data stream in isolation.

Telehealth, or the ability to assess and monitor various aspects of patients' health remotely, and often from the comfort of their own home using their own telecommunication devices, is increasing in demand, particularly in today's post-pandemic world.[1] This includes live health assessments that can be conducted by clinicians over video and audio in patients' homes, as well as automated remote assessments that can be conducted asynchronously using mobile technology and smart devices and do not require manual intervention. The U.S. Department of Health and Human Services has taken a range of administrative and policy steps to expedite the adoption and awareness of telehealth.[2] This is because telehealth, and specifically automated remote patient assessment and monitoring, have the potential to alleviate, if not fully address, several challenges associated with the majority of clinical assessment practice today (see, e.g., Steinhubl et al., 2013). For instance, a typical patient today must (a) come into the clinic to be assessed, which might be a problem for patients with ambulatory difficulties, (b) at specified times, depending on the availability of their chosen medical

Correspondence to Vikram Ramanarayanan: vikram.ramanarayanan@modality.ai. **Publisher Note:** This article is part of the Forum: Research Symposium on Artificial Intelligence. ***Disclosure:*** *Vikram Ramanarayanan is salaried by and receives equity from Modality.AI, Inc.*

[1] https://nibib.nih.gov/science-education/science-topics/telehealth.
[2] https://telehealth.hhs.gov/providers/telehealth-policy.

service provider (which, for certain neurological conditions, can be as far apart as several months, during which time their disease might have progressed significantly). Furthermore, (c) the conduct and interpretation of these assessments can be biased (Gopal et al., 2021) and specific to the clinician involved (particularly for mental health assessments), and therefore not consistently reproducible and scalable. Finally, (d) today's in-clinic assessments can be time-consuming and expensive depending on the specific procedures and personnel involved. Automated remote patient assessments in particular effectively addresses each of these pain points because: (a) as the name suggests, they can be done remotely, from the comfort of people's homes, using widely available consumer-grade electronic devices such as laptops, smart watches, and mobile phones; (b) they can be performed more frequently than in-clinic assessments, with asynchronous and immediate reporting to clinical service providers; (c) they can be performed in an objective, prespecified manner, reducing bias, and enabling scalability; and finally, (d) they can be made available to patients at relatively lower cost as compared to in-clinic assessments (see Dorsey & Topol, 2016). The rest of this review article will therefore focus on automated remote patients telehealth assessments, as opposed to those conducted by a live clinician.

At this point, let us define some important terms. *Artificial intelligence* (AI) refers to the broad field of computer systems capable of performing complex tasks that historically only a human could do, such as analyzing data, reasoning, making decisions, or solving problems. *Machine learning* (ML) is a subset of AI focused on the development of computer systems that learn and adapt automatically from experience and data, without being explicitly programmed.[3]

The recent rapid advances in digital computing, storage and AI technologies have allowed us to measure multiple human-centric signal modalities for the purposes of remote patient assessment (Abernethy et al., 2022). One or more of these modalities—including, but not limited to, speech acoustics, natural language, conversational ability, orofacial movement, limb or body movement, eye gaze, respiration, cardiac signals, or even neural signals—can be impacted in multiple diseases ranging from a simple cold to a neurological disorder (Milling et al., 2022; Ramanarayanan et al., 2022). Furthermore, applying AI techniques in conjunction with multimodal sensing technology has been shown to improve disease detection and diagnosis in mental health conditions (Garcia-Ceja et al., 2018; Shatte et al., 2019; Thieme et al., 2020). This review article argues that integrating multiple modalities into a single platform for the purpose of remote clinical assessment can be more effective and lead to better clinical outcomes than using any of these data streams in isolation. Such multimodal technologies are well-suited to assess communication disorders in particular, because they can capture changes in speech and voice (e.g., slowing of speech rate or decreased loudness), as well as changes in orofacial or full body movement (e.g., reduced lip or jaw speed, facial asymmetry, or impaired gait), which affect communication and movement and may be among the earliest signs of neurologic diseases such as amyotrophic lateral sclerosis (ALS; Neumann et al., 2021), stroke (Liu et al., 2023), Parkinson's disease (PD; Kothare et al., 2022), traumatic brain injury (TBI; Talkar et al., 2020), or mild cognitive impairment (MCI; Roesler et al., in press). In addition, with repeated administrations over time, multimodal analytics can be used to objectively monitor the rate of disease progression (Kothare, Neumann, et al., 2023).

The rest of this review article reviews the current signal sources and health information modalities available and motivates the combination of multiple sources for a more holistic assessment of patient health. We focus on multimodal technologies (and in turn clinical assessments) for remote assessment and monitoring that impact communication and movement, both crucial to patients' daily living and personal well-being.

Approach

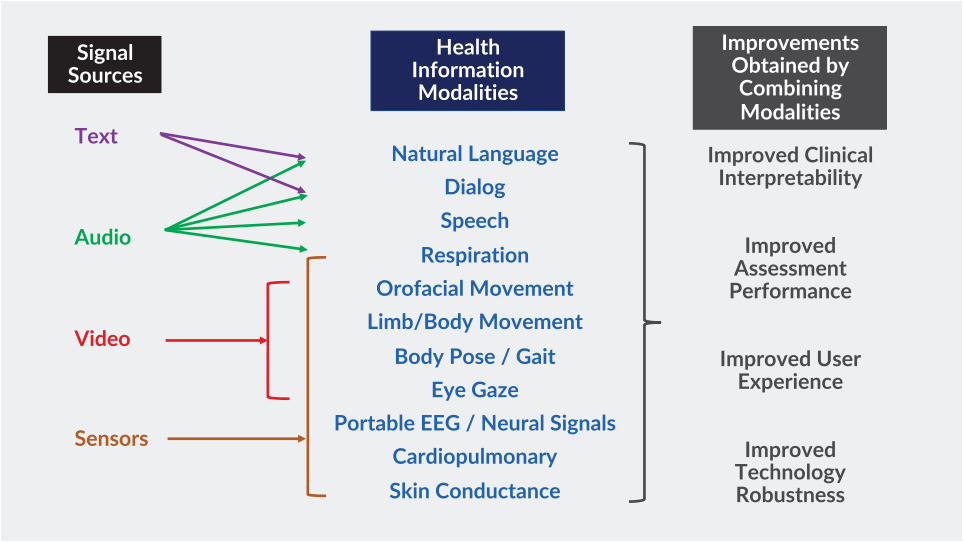
Overview of Health Information Modalities

We draw a distinction at this point between modalities of data collection and modalities of health information. By the former, we refer to signal sources, such as audio, video, text, and sensors.[4] By the latter, we mean information such as speech acoustics, natural language, conversational ability, orofacial movement, limb/body movement, eye gaze, respiration, gait, body posture, cardiac signals, or even neural signals that we can measure from those signal sources. In this review article, we will use the term *modality* to refer to the latter and refer to the former simply as *signal sources*. We further use the term *feature* to denote an individual property or characteristic of a phenomenon, usually numeric or categorical, which is measurable from a signal source (Bishop, 2006). Figure 1 provides a schematic illustration of these ideas, where

[3]https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning.

[4]We use the term *sensor* to broadly refer to any device that detects and responds to an input stimulus, such as heat, light, sound, pressure, magnetism, or a particular motion, from the physical environment and transmits a resulting output measurement. This includes peripheral computing devices like touchscreens or tablets that can be used to record tactile input.

Figure 1. Schematic illustrating the distinction between signal sources and modalities of health information relevant for remote health assessment and monitoring, alongside the benefits obtained by combining information from multiple modalities. All signal sources can be used for both active and passive measurement. Note that while sensors and wearables can theoretically measure a large number of health modalities, different sensor signal sources (or the same sensors positioned differently) are typically required to measure different health information modalities. EEG = electroencephalography.



multiple modalities can be derived from a signal source. For example, we can extract information from multiple modalities—speech, text, conversational, and even respiratory—from an audio signal source depending on the specific recording protocol used. Note that each of these modalities can, in turn, carry information about abnormalities in one or more health domains—motoric, cognitive, linguistic, affective, or anatomical (see Ramanarayanan et al., 2022, for an illustration of this concept in speech). Finally, we use the term *biomarker* to denote objective indications (i.e., substance, structure, or process) that can be accurately and reproducibly measured from inside or outside the patient (Strimbu & Tavel, 2010).

Here, we focus on signal sources relevant to remote health assessment and monitoring. This means other potentially useful multimodal data sources, such as medical imaging scans, blood biomarkers, genomics/proteomics , electronic health records (EHRs), and so forth, which are typically collected in-clinic and relatively infrequently, are out of scope for the purposes of this review article. Additionally, we focus on text, audio, and video signal sources in more depth, given the particular relevance of modalities extracted therefrom to the assessment of communication disorders. For reviews of remote health monitoring based on noninvasive and wearable sensors, see Kim et al. (2019), Majumder et al. (2017), and Vegesna et al. (2017). While text, audio, and video data and some sensor data are easy and affordable to collect, thanks to the ubiquity of cellphones, tablets, laptops, smart watches, and other such devices, modalities collected via other sensors (such as spirometers, ambient monitoring sensors, or portable electroencephalography [EEG] devices) are less accessible and more expensive. All these signal sources—audio, video, text, or sensors—can be used to collect in active mode (where data are recorded at prespecified assessment times) or passive mode (where data recording is always on). While passive monitoring allows one to track patient activity continuously on an intra- and inter-day basis, it also brings with it the challenges of analyzing large amounts of data to extract sparse signals of interest along with significant privacy considerations (patients need to consent to being tracked anywhere anytime depending on the specific signal sources and data collection protocol under consideration).

Speech

The clinical utility of speech-based digital biomarkers computed from multiple modalities as windows into mental and neurological health has been increasingly recognized in the signal processing and computer science literature (Cummins et al., 2018; Low et al., 2020; Ramanarayanan et al., 2023; Robin et al., 2020). An important clinical use case of speech analytics is in the movement disorder space. During the oral motor exam, which is a standard component of the neurological exam, speech is analyzed to help confirm the presence of movement disorders related to regional lesions to one or several components of the motor system. These lesions and their movement disorder consequences have also been causally

linked to different types of dysarthria (i.e., spastic, flaccid, ataxic, hypokinetic, hyperkinetic, and mixed) or apraxia of speech (Ramanarayanan et al., 2022), depending on the specific disorder in question. Therefore, it should come as no surprise that speech biomarkers have demonstrated utility for (a) detection of such disorders (including early detection, during prodromal phases; Hlavnička et al., 2017; Neumann et al., 2021), (b) subgrouping or patient stratification (Daoudi et al., 2022), (c) longitudinal progression tracking (Kothare, Neumann, et al., 2023; Stegmann et al., 2020), and even (d) predicting treatment response in a drug trial (Green et al., 2018; Norel et al., 2020). Indeed, the feasibility of speech-based disease detection or severity prediction has already been demonstrated for a wide spectrum of medical conditions ranging from acute or chronic respiratory diseases, such as cold and flu (Warule et al., 2023), COVID-19 (Deshpande et al., 2022; Quatieri et al., 2020), or asthma (Balamurali et al., 2020); to psychiatric disorders, such as schizophrenia (Rapcan et al., 2010; Richter et al., 2022) or depression (Cohen et al., 2023; Williamson et al., 2019); to developmental disorders, such as autism spectrum disorder (Kothare et al., 2021; Mohanta & Mittal, 2022); and neurodegenerative diseases, such as ALS (Green et al., 2018; Neumann et al., 2021), Alzheimer's disease (AD; König et al., 2015; Meilan et al., 2018), PD (Hlavnička et al., 2017; Kothare et al., 2022; Narendra et al., 2021), Huntington's disease (HD; Chan et al., 2019), or multiple sclerosis (MS; Rusz et al., 2018). Furthermore, recent research has found encouraging evidence for speech signal analysis as a viable alternative to spirometry (or the measurement of lung function, specifically the amount and/or speed of air that can be inhaled and exhaled, using a custom device and sensors) for remote assessment (Vatanparvar et al., 2021) of respiratory function, relevant in conditions such as asthma (Kutor et al., 2019) and ALS (Stegmann et al., 2021). Features of interest computed from the speech signal range from basic, easily interpretable properties of the signal usually captured on a short-term basis through low-level descriptors (LLDs), such as the fundamental frequency (*F*0), formants, jitter, shimmer, speaking rate, and duration to spectral features such as Mel frequency cepstral coefficients, which compute the short-term energy spectrum on a Mel scale to higher level descriptors such as statistical functionals over LLD trajectories (e.g., the extended Geneva minimalistic acoustic parameter set; see Eyben et al., 2016) all the way up to deep learning based features that are not explicitly dependent on expert knowledge (for a comprehensive overview, see Milling et al., 2022).

Natural Language

In recent years, advances in high performance computing and natural language processing (NLP)—a subdiscipline of AI that deals with how computers understand, process, and manipulate human languages—have played an essential role in improving the state of the art in assessing mental and neurological health state information from text data. For example, this problem can be cast as a text classification or sentiment analysis task, where we can apply NLP techniques to data ranging from social media posts to interviews to narrative writing sources and even EHRs across languages to automatically identify early indicators of mental illness and support early detection, prevention, and treatment (for an extensive review of NLP for mental health and neurodegenerative disorders, see Boschi et al., 2017; Zhang et al., 2022). Typical features include linguistic features (such as part-of-speech, bag-of-words, linguistic inquiry and word count, sentiment/emotion scores, semantic similarity features, and topic modeling features), statistical features (such as n-grams, term frequency-inverse document frequency, length of sentences or passages, and word/document embeddings), domain knowledge features (such as unified medical language system labels or linguistic dictionary features), and other auxiliary features such as social connectivity, user profile, or time-series features.

In addition to objective data about patients' language use, we can also capture (from a text- or speech-based signal source) patient-reported outcome measures (PROMs), such as what patients have to say about their disease in their own words. This is important because existing PROMs—which are typically standardized clinical measurement scales such as the Parkinson's Disease Questionnaire–39 (Jenkinson et al., 1997) for PD, or the ALS Functional Rating Scale–Revised (ALSFRS-R; Cedarbaum et al., 1999) for ALS—are often not fine-grained enough to capture disease severity in sufficient granularity (see, e.g., Allison et al., 2017). We can address this granularity problem by allowing patients to describe the problems that bother them the most with respect to their disease, along with how these affect their daily functioning (Shoulson et al., 2022).

Dialog

Interactive conversational systems allow us to estimate deficits in turn-taking and pragmatics aspects of discourse that are affected in neurological disorders such as frontotemporal dementia, AD, and TBI; mental health disorders such as bipolar disorder, clinical depression, and schizophrenia; or neurodevelopmental disorders such as autism. For instance, Nasreen et al. (2021) showed that interactional features such as silent pauses within and between speaker turns, turn switches per minute, standardized phonation time, and turn length are key features of AD conversations, and can classify AD patient conversations versus non–AD patient conversations with 83% accuracy. Rousseaux et al. (2010) demonstrated that TBI affected both patients verbal (fluency, intelligibility, and pragmatics) as well as nonverbal (impaired prosody) communication. Aldeneh et al. (2019) showed that

high-level dialogue features can be used to quantify interaction dynamics in clinical interviews, highlighting how changes in mood episodes can significantly affect the values of the features. Examples of such features include number, duration and frequency of turns, and turn-switches.

Oro-Facial, Limb, and Full Body Movement

Recent developments in computer vision, real-time processing, and feature analysis have allowed image and video to add versatility to the assessment of neurological and mental health states. For example, research has shown that orofacial video analysis during facial gestures and speech provides clinically useful information for assessing neurological conditions such as bulbar ALS (Bandini et al., 2018; Guarin et al., 2022; Neumann et al., 2021) or mental health conditions such as schizophrenia (Richter et al., 2022). Moreover, video allows capture of motor aspects of production (such as finger tapping, see, e.g., Khan et al., 2014; or body pose to quantify gait and balance, see, e.g., Sabo et al., 2020), useful in understanding disorders of neurocognition and movement such as PD and HD. Furthermore, video data allow us to probe specific aspects of micro expression and/or emotion change in patients, both in neurological (Gomez et al., 2023) and mental health conditions (Siam et al., 2022). In addition to video, signals from sensors like accelerometers and gyroscopes can accurately track body movements (or lack thereof, see Kim et al., 2019; Majumder et al., 2017; Vegesna et al., 2017), while passive radio-wave-based sensors can unintrusively monitor gait, home activity, and time in bed for patients with PD and dementia (Kabelac et al., 2019). Still, other novel remote assessments that use signal input from tablets/touchscreens and digital pens such as digital clock drawing tests or handwriting analyses have also shown high sensitivity to screening of neurodegenerative diseases such as AD and PD (Öhman et al., 2021; Vessio, 2019).

Eye Gaze

The advent of wearable eye tracking has demonstrated exciting potential to contribute to pervasive health monitoring and understanding of eye movement pathologies, backed up by a growing body of research in experimental psychology and clinical neuroscience finding strong links between abnormal eye movements and neurological disorders (such as PD, HD, MS, AD, and other dementias; see Anderson & MacAskill, 2013, and Vidal et al., 2012, for more details). Saccadic features and smooth pursuit movements are of particular relevance for mental health monitoring. While the improvement of webcam video-based eye-gaze tracking allows us to bring such technology to everyday devices (e.g., Tisdale et al., 2023), there is a need to develop standards for performance, calibration, and evaluation of gaze tracking systems and overcome limitations arising due to camera quality, random illumination changes, patients wearing glasses, head movement/distance, and display properties across devices (Kar & Corcoran, 2017).

Cardiopulmonary and Other Related Physiological Signals

Recent research has explored multiple methods for remote measurement of cardiovascular and cardiopulmonary signals using sensors (see Al-Naji et al., 2017; Majumder et al., 2017). Recent work has also explored multiple approaches toward spirometry, or the monitoring of respiratory function, via wearable and remote electronics (Vitazkova et al., 2024). While completely remote and unsupervised spirometry would be both impactful and desirable, recent research has indicated that the variability and accuracy of measurement is still not nearly as good as in-clinic supervised spirometry to warrant widespread adoption (Anand et al., 2023). That being said, as mentioned earlier, speech audio analysis can be used as a viable alternative to spirometry for remote assessment (Stegmann et al., 2021; Vatanparvar et al., 2021). Yet, other sensors measure variation in skin conductance via electrodermal activity or galvanic skin response that reflects the activity of the sympathetic nervous system and provides a simple, sensitive, and reliable parameter for assessing the sympathetic nervous activities associated with stress and emotion (see Majumder et al., 2017, for a review).

Neural Signals

While neural signals recorded via EEG, magnetic resonance imaging, x-rays, computational tomography, positron emission tomography, and so forth have long been used to obtain gold standard biomarker measurements for the diagnosis and assessment of neurological disorders, the field of portable neural signal measurement for remote assessment and monitoring is still in its infancy. However, recent advancements in portable EEG can potentially allow us to understand how specific brain areas and neural activity are impacted in neurological and mental health in a manner that is much more accessible than in-clinic versions. For example, Gottlibe et al. (2020) recently showed that 10-min epochs of EEG recordings measured using a portable EEG device can statistically differentiate patients diagnosed with ischemic stroke from nonstroke patients. Therefore, while this field is still developing, it has exciting potential to revolutionize the future of automated remote assessment.

Findings

Advantages of Combining Multiple Modalities

This section motivates the use of multimodal technologies for remote health assessment with a four-pronged

set of reasons: improved scientific interpretability, improved performance on downstream tasks such as early detection and progress monitoring, improved technological robustness, and improved user experience (UX).

Improved Interpretability

Speech-based digital biomarkers computed from multiple modalities can serve as windows into mental and neurological health because speech can be viewed as a diagnostic pathway within a biopsychosocial framework (Engel, 1977; Ramanarayanan et al., 2023). The underlying assumption of this viewpoint, supported by multiple studies on AD (cognitive domain); post-stroke aphasia (linguistic domain); depression (affective domain); ALS, PD, and other movement disorders (motoric domain); and face transplants (anatomic domain), is that multiple domains—cognitive, linguistic, affective, motoric, and anatomical—have an influence on speech and orofacial motor output that can in turn be measured by various multimodal speech features. Therefore, combining information from features drawn from multiple modalities can improve clinical interpretability. For example, Cohen et al. (2023) showed that when assessing for clinical depression and suicidality, speech acoustic (such as *F*0 or percent pause time), orofacial (such as velocity and acceleration of the lips), and text (such as relative probabilities of occurrence of specific words or phrases) features were, in turn, the best predictors of depression, anxiety, suicide risk, respectively, and therefore add complementary information toward understanding disease state. Moreover, for neurological disorders such as ALS, which can involve multifocal symptoms such as impaired gross and fine limb motor function (for limb onset of the disease) and impaired speech, swallowing and respiration (bulbar onset), multimodal approaches combining speech, and orofacial information allow us to obtain a more holistic picture of disease state (e.g., Neumann et al., 2021). We could further combine this with sensors that capture limb motor function (e.g., Gupta et al., 2023; Vieira et al., 2022) to improve this picture even further. And this, in turn, could give clinicians, caregivers, and clinical trialists more information for diagnosis or intervention.

Improved Performance

Over and above improved interpretability, a compelling practical reason to use information and features computed from multiple modalities is that the combination thereof often performs better at various ML tasks than the individual features alone. Such tasks could include classification (i.e., discriminating one or more disorder classes from each other or from healthy controls; see, e.g., Neumann et al., 2021, in ALS, Jiang et al., 2024, and Cohen et al., 2023, in depression and anxiety; Richter et al., 2022, in schizophrenia; Escobar-Grisales et al., 2023, in PD; Talkar et al., 2020, in TBI; Roesler et al., in press, in MCI), regression (i.e., predicting a given clinical score of interest; see, e.g., Neumann et al., 2021, in ALS), or clustering (i.e., stratifying health cohorts or disease progression states of interest; see, e.g., Severson et al., 2021, in PD), among others. In all these use cases, combining information from multiple modalities—primarily speech, orofacial or full body movement, and natural language—results in a significantly better performance than either of those modalities in isolation. For example, Talkar et al. (2020) reports a 6% absolute improvement in performance in classifying patients with mild TBI from healthy controls when combining information from speech and gait as compared to the best performing individual modality (speech), while Escobar-Grisales et al. (2023) reports that a multimodal approach combining speech and language features for classifying PD patients from healthy controls outperforms the best unimodal approach (speech) by 6% as well. Neumann et al. (2021) reports a similar pattern of results when applied to a regression task, showing that in a sample of 54 ALS patients, combining features derived from both speech and orofacial movement contributed more predictive power in predicting the ALSFRS-R score than considering the individual modalities alone.

Improved Robustness

Information from multiple modalities allows us to improve technology robustness because they can capture either complementary or redundant information. For example, let us consider the case of voice activity detection (VAD). VAD modules are important components of speech processing and dialog systems that identify the segments of the signal that contain speech from those that contain only noise and interference in offline, and additionally determine when the current speaker has finished speaking in real-time systems, so that downstream processing (speech recognition, spoken language understanding, dialog management, etc.) can commence, in turn, allowing the dialog system to respond appropriately (Liscombe et al., 2021, 2023). Multiple studies have demonstrated the robustness and superiority of VAD performance when multiple modalities (speech audio, orofacial video that captures lip opening) are combined as compared to using a single modality alone, because each modality provides complementary information (Ariav & Cohen, 2019; Tao & Busso, 2019). In this manner, multimodal systems are also more fault-tolerant as compared to their unimodal counterparts; for instance, if the data stream from one modality (say for instance, speech) is missing, corrupted, or otherwise unreliable due to technical reasons, this would cripple a unimodal system based solely on that modality, while a multimodal system could still function based on data streams from other modalities (such as orofacial video).

Improved Engagement and UX

Using multimodal technologies further allows us to collect higher quality data from more engaged and motivated participants as compared to unimodal methods, such as text-based questionnaires. Virtual humans—computer-generated characters that demonstrate many of the same properties as humans in face-to-face conversation including the ability to produce and respond to verbal and nonverbal communication (also referred to as virtual agents, embodied conversational agents, or avatars in the literature)—have been shown to improve retention and UX, especially in young (where they provide a context for the elicitation of social communicative behavior in child–machine interactions, useful, e.g., in autism; see Chaspari et al., 2012; Kothare et al., 2021; Narayanan & Potamianos, 2002) and elderly populations (e.g., people with dementia; see Shaked, 2017; Tanaka et al., 2017). Furthermore, dialog-capable multimodal systems allow for interactive responses to patient queries/responses, which improves the efficacy, utility, and UX of the interaction (see, e.g., Yu et al., 2019).

Case Study of A Real-World Multimodal Implementation: The Modality Platform

The Modality platform, powered by a cloud-based multimodal dialog system, employs a virtual human guide—note there is no live clinician involved—to conduct structured conversational interactions with participants for active monitoring and remote health assessment (Ramanarayanan et al., 2023; Suendermann-Oeft et al., 2019; see Figure 2). Participants can start a conversation with the virtual human guide, Tina, using any internet-connected device equipped with a microphone, speaker, and webcam—be it a phone, tablet, or laptop/desktop computer—via a personalized web link. At the beginning of the conversational assessment, 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 have sufficient quality. Once all device tests pass, Tina guides participants through a customizable sequence of tasks that elicit speech and facial behaviors, such as vowel phonation, counting up of numbers in a single breath, repeating consonant–vowel–consonant words, diadochokinesis, reading sentences and passages, picture description, and production of spontaneous speech on a topic of their choice. Multimodal analytics modules automatically extract features (see Table 1) that capture information from speech acoustic (e.g., energy, timing, voice quality, spectral), textual (e.g., lexico-semantic, sentiment), orofacial (e.g., articulatory kinematics, range of motion, eye and facial movement), limb motor (e.g., finger tapping kinematics), eye-gaze, and body pose (e.g., balance) modalities during these tasks. Tina can also administer tasks that probe cognitive abilities 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, an instrument that allows them to describe their symptoms and severity in their own words, as well as other clinical survey instruments of interest. We then classify these verbatim responses into multiple, clinically relevant symptoms using a multilabel text classification deep neural network model trained on data collected from over 25,000 patients with PD (Shoulson et al., 2022).

The Modality platform exploits the four advantages of multimodality mentioned earlier—improved interpretability, performance, UX, and robustness. For instance, since the Modality platform extracts information from multiple modalities—speech, text, orofacial, limb motor, eye-gaze, and body pose, in addition to patient self-reports of their problems—one can combine this information for

Figure 2. The Modality multimodal dialog platform for remote patient assessment uses a virtual human guide, Tina, to engage participants in an interactive conversation, record their audio and video signals, and extract multiple modalities of health information therefrom. PROP = Patient Report Of Problems.

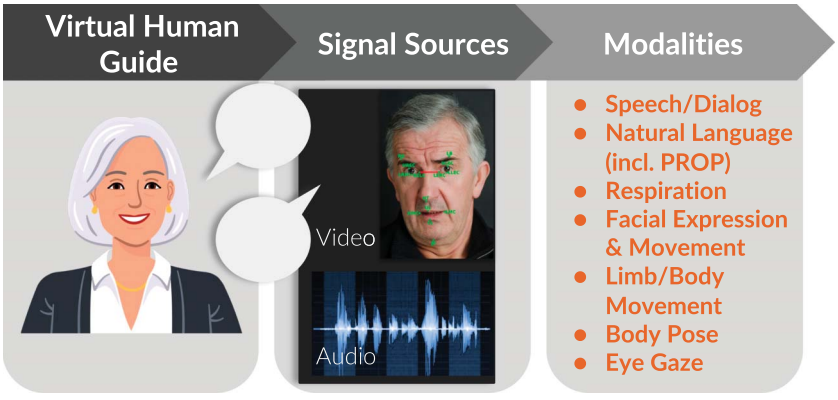


Table 1. Overview of signal sources, modalities of health information, domains captured within each of these modalities and exemplar features extracted by the Modality platform.

Signal source	Modality	Domain	Exemplar features
Audio	Speech	Energy timing	Shimmer (%), intensity (dB), signal-to-noise ratio (dB) speaking and articulation duration (s), articulation and speaking rate, percent pause time (%), canonical timing agreement (%)
		Voice quality frequency	Cepstral peak prominence (dB), harmonics-to-noise ratio (dB) mean, max., min. Fundamental frequency (Hz), first three formants (Hz), slope of 2nd formant (Hz/s), jitter (%)
		Cognition	Reaction times and percentage of correct words (immediate and delayed word recall), digit span forward/backward score (ranges from 0 to 2)
	Respiration	Breathing	Maximum phonation time on a single breath (sec.)
Text	Natural Language	Lexical	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	Positive cosine similarity, negative cosine similarity
		Patient Report Of Problems	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?"
Video	Body	Limb motoric	Finger tapping rate and duration, jitter and shimmer
		Balance & body pose	Time from sit to stand as measured by the Berg Balance Scale
	Orofacial	Mouth (distances)	Lip aperture/opening, lip width, mouth surface area
		Lip/jaw movement	Mean symmetry ratio between left and right half of the mouth
		Oro-motor exam	Velocity, acceleration, jerk, and speed of lower lip and jaw center
		Eyes	Range of motion of lips and jaw, head pose number of eye blinks per second, eye opening, vertical displacement of eyebrows
		Eye gaze	Saccade rates, reaction times and fixation durations for smooth pursuit, saccade, free and directed image exploration, and congruent and incongruent Stroop tasks
		Cognition	Accuracy of three step tasks (touch three points on the face in a specified order)

Note. Observe the one to many mapping between signal sources and modalities. For example, both speech and respiratory information can be derived from audio. The text modality (and corresponding features) is derived from the speech through automatic speech recognition software. For visual features, functionals (minimum, maximum, average) are applied to produce one value across all video frames of an utterance. Unless otherwise noted in parentheses, all features are unitless. ML = machine learning.

enhanced clinical interpretability and a more holistic picture of disease state. Multiple research studies have demonstrated the clinical validity and performance benefits of the Modality platform, combining information from multiple modalities for remote assessments at scale in a variety of domains, including ALS early diagnosis and progression tracking (Kothare, Neumann, et al., 2023; Neumann et al., 2021), PD (Kothare et al., 2022), schizophrenia (Richter et al., 2022), depression (Cohen et al., 2023; Neumann et al., 2023), MCI (Roesler et al., in press), and autism (Kothare et al., 2021). Moreover, the use of the animated virtual human guide, Tina, to guide participants through interactive conversations has been demonstrated to improve UX (Kothare, Habberstad, et al., 2023). These advantages make the platform well-suited for adoption in clinical care and clinical trials because it allows for more frequent, objective remote assessments and finer-grained resolution of features on continuous scales, while mitigating the reliability issues that plague subjective evaluations. These factors, in turn, lead to more effective treatment and improved health and communication outcomes.[5] Importantly, the Modality platform monitors patients actively, only when they consent to undertake the assessment, unlike passive monitoring technologies such as sensors and wearables that can potentially track them all the time, potentially bringing with it privacy concerns.

Discussion and Outlook

Telehealth and remote patient assessments offer an exciting opportunity to improve health care delivery and patient outcomes because they can be performed remotely, as often as required, in an objective manner, and at relatively lower cost as compared to in-clinic assessments. This review article has argued that integrating multiple

[5]https://academy.pubs.asha.org/2020/08/how-will-artificial-intelligence-reshape-speech-language-pathology-services-and-practice-in-the-future/.

modalities into a single platform for the purpose of remote clinical assessment can be more effective and lead to better clinical outcomes than using any of these data streams in isolation, with a four-pronged set of motivating reasons: improved scientific interpretability, improved performance on downstream health applications such as early detection and progress monitoring, improved technological robustness, and improved UX. Given the ease of access to smartphones and other mobile devices that allow remote collection of audio, text, and video signals, such multimodal technologies stand particularly well-equipped to assess disorders of communication associated with neurological and mental disease. Leveraging such multimodal data along with recent advances in AI has the potential to accelerate progress toward personalized precision health, digital clinical trials and digital twin technologies, remote health monitoring via a "hospital-at-home," and virtual AI health assistance (Acosta et al., 2022). It is also important to temper the promise of such multimodal technologies with a realistic appraisal of the open challenges and considerations for real-world deployment of such multimodal technologies. Ramanarayanan et al. (2022) provide a comprehensive set of challenges and requirements for the adoption of digital biomarkers, including robustness in the face of many different conditions that affect signals from different modalities differently, robustness to atypical speech diversity, heterogeneity and comorbidities involved in progression of disease, recording environments and application settings, and generalizability and statistical power of models as promoted by abundant, good-quality training data. There are additional implementation, modeling, and privacy challenges to overcome when AI is added to the mix (Acosta et al., 2022). Over and above these, Dorsey and Topol (2016) point out overarching practical challenges, including, but not limited to, economic issues (reimbursement, insurance coverage), clinical issues (potential for lowering quality of care, potential for abuse, fragmentation of care), legal issues (licensure laws, liability concerns), and social issues (differential access to technologies based on socioeconomic background). Working as a community to address these gaps, while harnessing the power of the rapidly evolving digital revolution—characterized by an increasing number of high-fidelity signal measurement techniques, powerful cloud-based computing and storage, and outage-robust network connectivity—will, in turn, accelerate progress toward the next generation of multimodal digital medicine for precision clinical trials and personalized health care.

Data Availability Statement

This is a review article that references data and results presented in previously published works, which are, in turn, cited in the article. No other data are available.

Acknowledgments

This article stems from the Research Symposium at the 2023 ASHA Convention, which was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under Award R13DC003383. Research reported in this publication was also supported by the NIDCD under Award R42DC019877. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute of Health.

References

Abernethy, A., Adams, L., Barrett, M., Bechtel, C., Brennan, P., Butte, A., Faulkner, J., Fontaine, E., Friedhoff, S., Halamka, J., Howell, M., Johnson, K., Long, P., McGraw, D., Miller, R., Lee, P., Perlin, J., Rucker, D., Sandy, L., & Valdes, K. (2022). The promise of digital health: Then, now, and the future. *NAM Perspectives, 6*(22). https://doi.org/10.31478/202206e

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine, 28*(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Aldeneh, Z., Jaiswal, M., Picheny, M., McInnis, M., & Provost, E. M. (2019). *Identifying mood episodes using dialogue features from clinical interviews*. arXiv. https://doi.org/10.48550/arXiv.1910.05115

Allison, K. M., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., & Green, J. R. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 18*(5–6), 358–366. https://doi.org/10.1080/21678421.2017.1303515

Al-Naji, A., Gibson, K., Lee, S. H., & Chahl, J. (2017). Monitoring of cardiorespiratory signal: Principles of remote measurements and review of methods. *IEEE Access, 5,* 15776–15790. https://doi.org/10.1109/ACCESS.2017.2735419

Anand, R., McLeese, R., Busby, J., Stewart, J., Clarke, M., Man, W. D., & Bradley, J. (2023). Unsupervised home spirometry versus supervised clinic spirometry for respiratory disease: A systematic methodology review and meta-analysis. *European Respiratory Review, 32*(169), Article 220248. https://doi.org/10.1183/16000617.0248-2022

Anderson, T. J., & MacAskill, M. R. (2013). Eye movements in patients with neurodegenerative disorders. *Nature Reviews Neurology, 9*(2), 74–85. https://doi.org/10.1038/nrneurol.2012.273

Ariav, I., & Cohen, I. (2019). An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. *IEEE Journal of Selected Topics in Signal Processing, 13*(2), 265–274. https://doi.org/10.1109/JSTSP.2019.2901195

Balamurali, B. T., Hee, H. I., Teoh, O. H., Lee, K. P., Kapoor, S., Herremans, D., & Chen, J. M. (2020). Asthmatic versus healthy child classification based on cough and vocalised /ɑː/ sounds. *The Journal of the Acoustical Society of America, 148*(3), EL253–EL259. https://doi.org/10.1121/10.0001933

Bandini, A., Green, J. R., Taati, B., Orlandi, S., Zinman, L., & Yunusova, Y. (2018, May). Automatic detection of amyotrophic lateral sclerosis (ALS) from video-based analysis of facial movements: speech and non-speech tasks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 150–157). IEEE.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). Springer. https://link.springer.com/in/book/9780387310732

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology, 8,* Article 208495. https://doi.org/10.3389/fpsyg.2017.00269

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., Nakanishi, A., & BDNF Study Group. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences, 169*(1–2), 13–21. https://doi.org/10.1016/S0022-510X(99)00210-5

Chan, J. C., Stout, J. C., & Vogel, A. P. (2019). Speech in prodromal and symptomatic Huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases. *Neuroscience & Biobehavioral Reviews, 107,* 450–460. https://doi.org/10.1016/j.neubiorev.2019.08.009

Chaspari, T., Provost, E. M., Katsamanis, A., & Narayanan, S. (2012, March). An acoustic analysis of shared enjoyment in ECA interactions of children with autism. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4485–4488). IEEE.

Cohen, J., Richter, V., Neumann, M., Black, D., Haq, A., Wright-Berryman, J., & Ramanarayanan, V. (2023). A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. *Frontiers in Psychology, 14,* Article 1135469. https://doi.org/10.3389/fpsyg.2023.1135469

Cummins, N., Baird, A., & Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods, 151,* 41–54. https://doi.org/10.1016/j.ymeth.2018.07.007

Daoudi, K., Das, B., Tykalova, T., Klempir, J., & Rusz, J. (2022). Speech acoustic indices for differential diagnosis between Parkinson's disease, multiple system atrophy and progressive supranuclear palsy. *npj Parkinson's Disease, 8*(1), Article 142. https://doi.org/10.1038/s41531-022-00389-6

Deshpande, G., Batliner, A., & Schuller, B. W. (2022). AI-Based human audio processing for COVID-19: A comprehensive overview. *Pattern Recognition, 122,* Article 108289. https://doi.org/10.1016/j.patcog.2021.108289

Dorsey, E. R., & Topol, E. J. (2016). State of telehealth. *New England Journal of Medicine, 375*(2), 154–161. https://doi.org/10.1056/NEJMra1601705

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science, 196*(4286), 129–136. https://doi.org/10.1126/science.847460

Escobar-Grisales, D., Arias-Vergara, T., Rios-Urrego, C. D., Nöth, E., García, A. M., & Orozco-Arroyave, J. R. (2023). An automatic multimodal approach to analyze linguistic and acoustic cues on Parkinson's disease patients. *Proceedings of Interspeech*, 1703–1707. https://doi.org/10.21437/Interspeech.2023-2287

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing, 51,* 1–26. https://doi.org/10.1016/j.pmcj.2018.09.003

Gomez, L. F., Morales, A., Fierrez, J., & Orozco-Arroyave, J. R. (2023). Exploring facial expressions and action unit domains for Parkinson detection. *PLOS ONE, 18*(2), Article e0281248. https://doi.org/10.1371/journal.pone.0281248

Gopal, D. P., Chetty, U., O'Donnell, P., Gajria, C., & Blackadder-Weinstein, J. (2021). Implicit bias in healthcare: Clinical practice, research and decision making. *Future Healthcare Journal, 8*(1), 40–48. https://doi.org/10.7861/fhj.2020-0233

Gottlibe, M., Rosen, O., Weller, B., Mahagney, A., Omar, N., Khuri, A., Srugo, I., & Genizi, J. (2020). Stroke identification using a portable EEG device—A pilot study. *Neurophysiologie Clinique, 50*(1), 21–25. https://doi.org/10.1016/j.neucli.2019.12.004

Green, J. R., Allison, K. M., Cordella, C., Richburg, B. D., Pattee, G. L., Berry, J. D., Macklin, E. A., Pioro, E. P., & Smith, R. A. (2018). Additional evidence for a therapeutic effect of dextromethorphan/quinidine on bulbar motor function in patients with amyotrophic lateral sclerosis: A quantitative speech analysis. *British Journal of Clinical Pharmacology, 84*(12), 2849–2856. https://doi.org/10.1111/bcp.13745

Guarin, D. L., Taati, B., Abrahao, A., Zinman, L., & Yunusova, Y. (2022). Video-based facial movement analysis in the assessment of bulbar amyotrophic lateral sclerosis: Clinical validation. *Journal of Speech, Language, and Hearing Research, 65*(12), 4667–4678. https://doi.org/10.1044/2022_JSLHR-22-00072

Gupta, A. S., Patel, S., Premasiri, A., & Vieira, F. (2023). At-home wearables and machine learning sensitively capture disease progression in amyotrophic lateral sclerosis. *Nature Communications, 14*(1), Article 5080. https://doi.org/10.1038/s41467-023-40917-3

Hlavnička, J., Čmejla, R., Tykalová, T., Šonka, K., Růžička, E., & Rusz, J. (2017). Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports, 7*(1), Article 12. https://doi.org/10.1038/s41598-017-00047-5

Jenkinson, C., Fitzpatrick, R. A. Y., Peto, V. I. V., Greenhall, R., & Hyman, N. (1997). The Parkinson's Disease Questionnaire (PDQ-39): Development and validation of a Parkinson's disease summary index score. *Age and Ageing, 26*(5), 353–357. https://doi.org/10.1093/ageing/26.5.353

Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Rad, A. B., Kwon, H., Cotes, R. O., & Clifford, G. D. (2024). Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE Journal of Biomedical and Health Informatics, 28*(3), 1680–1691. https://doi.org/10.1109/JBHI.2024.3352075

Kabelac, Z., Tarolli, C. G., Snyder, C., Feldman, B., Glidden, A., Hsu, C.-Y., Hristov, R., Dorsey, E. R., & Katabi, D. (2019). Passive monitoring at home: A pilot study in Parkinson disease. *Digital Biomarkers, 3*(1), 22–30. https://doi.org/10.1159/000498922

Kar, A., & Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access, 5,* 16495–16519. https://doi.org/10.1109/ACCESS.2017.2735633

Khan, T., Nyholm, D., Westin, J., & Dougherty, M. (2014). A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artificial Intelligence in Medicine, 60*(1), 27–40. https://doi.org/10.1016/j.artmed.2013.11.004

Kim, J., Campbell, A. S., de Ávila, B. E. F., & Wang, J. (2019). Wearable biosensors for healthcare monitoring. *Nature Biotechnology, 37*(4), 389–406. https://doi.org/10.1038/s41587-019-0045-y

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment &*

Disease Monitoring, 1(1), 112–124. https://doi.org/10.1016/j.dadm.2014.11.012

Kothare, H., Habberstad, D., Neumann, M., White, S., Pautler, D., & Ramanarayanan, V. (2023, February 21–24). *Impact of synthetic voice and avatar animation on the usability of a dialogue agent for digital health monitoring* [Paper presentation]. International Workshop on Spoken Dialog Systems, Los Angeles, CA.

Kothare, H., Neumann, M., Liscombe, J., Green, J., & Ramanarayanan, V. (2023). Responsiveness, sensitivity and clinical utility of timing-related speech biomarkers for remote monitoring of ALS disease progression. *Proceedings of Interspeech*, 2323–2327. https://doi.org/10.21437/Interspeech.2023-2002

Kothare, H., Neumann, M., Liscombe, J., Roesler, O., Burke, W., Exner, A., Snyder, S., Cornish, A., Habberstad, D., Pautler, D., Suendermann-Oeft, D., Huber, J. E., & Ramanarayanan, V. (2022). Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for Parkinson's disease assessment. *Proceedings of Interspeech,* 3658–3662. https://doi.org/10.21437/Interspeech.2022-11048

Kothare, H., Ramanarayanan, V., Roesler, O., Neumann, M., Liscombe, J., Burke, W., & Demopoulos, C. (2021). Investigating the interplay between affective, phonatory and motoric subsystems in autism spectrum disorder using a multimodal dialogue agent. *Proceedings of Interspeech*, 1967–1971. https://doi.org/10.21437/Interspeech.2021-1796

Kutor, J., Balapangu, S., Adofo, J. K., Dellor, A. A., Nyakpo, C., & Brown, G. A. (2019). Speech signal analysis as an alternative to spirometry in asthma diagnosis: Investigating the linear and polynomial correlation coefficient. *International Journal of Speech Technology, 22*(3), 611–620. https://doi.org/10.1007/s10772-019-09608-7

Liscombe, J., Kothare, H., Habberstad, D., Cornish, A., Roesler, O., Neumann, M., Pautler, D., Suendermann-Oeft, D., & Ramanarayanan, V. (2021). *Voice activity detection in dialog agents for dysarthric speakers* [Paper presentation]. International Workshop on Spoken Dialog Systems, Singapore.

Liscombe, J., Kothare, H., Neumann, M., Pautler, D., & Ramanarayanan, V. (2023, February 21–24). *Pathology-specific settings for voice activity detection in a multimodal dialog agent for digital health monitoring* [Paper presentation]. International Workshop on Spoken Dialog Systems, Los Angeles, CA.

Liu, J., Du, X., Lu, S., Zhang, Y. M., An-ming, H. U., Ng, M. L., Su, R., Wang, L., & Yan, N. (2023). Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis. *Biomedical Signal Processing and Control, 79*(Pt. 2), Article 104161. https://doi.org/10.1016/j.bspc.2022.104161

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*(1), 96–116. https://doi.org/10.1002/lio2.354

Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. *Sensors, 17*(12), Article 130. https://doi.org/10.3390/s17010130

Meilan, J. J., Martinez-Sanchez, F., Carro, J., Carcavilla, N., & Ivanova, O. (2018). Voice markers of lexical access in mild cognitive impairment and Alzheimer's disease. *Current Alzheimer Research, 15*(2), 111–119. https://doi.org/10.2174/1567205014666170829112439

Milling, M., Pokorny, F. B., Bartl-Pokorny, K. D., & Schuller, B. W. (2022). Is speech the new blood? Recent progress in AI-based disease detection from audio in a nutshell. *Frontiers in Digital Health, 4,* Article 886615. https://doi.org/10.3389/fdgth.2022.886615

Mohanta, A., & Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Computer Speech & Language, 72,* Article 101287. https://doi.org/10.1016/j.csl.2021.101287

Narayanan, S., & Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing, 10*(2), 65–78. https://doi.org/10.1109/89.985544

Narendra, N. P., Schuller, B., & Alku, P. (2021). The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 1925–1936. https://doi.org/10.1109/TASLP.2021.3078364

Nasreen, S., Rohanian, M., Hough, J., & Purver, M. (2021). Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science, 3,* Article 640669. https://doi.org/10.3389/fcomp.2021.640669

Neumann, M., Kothare, H., Habberstad, D., & Ramanarayanan, V. (2023). A multimodal investigation of speech, text, cognitive and facial video features for characterizing depression with and without medication. *Proceedings of Interspeech*, 1219–1223. https://doi.org/10.21437/Interspeech.2023-2194

Neumann, M., Roesler, O., Liscombe, J., Kothare, H., Suendermann-Oeft, D., Pautler, D., Navar, I., Anvar, A., Kumm, J., Norel, R., Fraenkel, E., Sherman, A. V., Berry, J. D., Pattee, G. L., Wang, J., Green, J. R., & Ramanarayanan, V. (2021). Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale. *Proceedings of Interspeech*, 4783–4787. https://doi.org/10.21437/Interspeech.2021-1801

Norel, R., Agurto, C., Heisig, S., Rice, J. J., Zhang, H., Ostrand, R., Wacnik, P. W., Ho, B. K., Ramos, V. L., & Cecchi, G. A. (2020). Speech-based characterization of dopamine replacement therapy in people with Parkinson's disease. *npj Parkinson's Disease, 6*(1), Article 12. https://doi.org/10.1038/s41531-020-0113-5

Öhman, F., Hassenstab, J., Berron, D., Schöll, M., & Papp, K. V. (2021). Current advances in digital cognitive assessment for preclinical Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13*(1), Article e12217. https://doi.org/10.1002/dad2.12217

Quatieri, T. F., Talkar, T., & Palmer, J. S. (2020). A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology, 1,* 203–206. https://doi.org/10.1109/OJEMB.2020.2998051

Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups, 7*(1), 276–283. https://doi.org/10.1044/2021_PERSP-21-00174

Ramanarayanan, V., Pautler, D., Arbatti, L., Hosamath, A., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Cornish, A., Habberstad, D., Richter, V., Fox, D., Suendermann-Oeft, D., & Shoulson, I. (2023). When words speak just as loudly as actions: Virtual agent based remote health assessment integrating what patients say with what they do. *Proceedings of Interspeech*, 678–679.

Rapcan, V., D'Arcy, S., Yeap, S., Afzal, N., Thakore, J., & Reilly, R. B. (2010). Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering & Physics, 32*(9), 1074–1079. https://doi.org/10.1016/j.medengphy.2010.07.013

Richter, V., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Suendermann-Oeft, D., Prokop, S., Khan, A., Yavorsky, C., Lindenmayer, J.-P., & Ramanarayanan, V. (2022). Towards multimodal dialog-based speech & facial biomarkers

of schizophrenia. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 171–176. https://doi.org/10.1145/3536220.3558075

Robin, J., Harrison, J. E., Kaufman, L. D., Rudzicz, F., Simpson, W., & Yancheva, M. (2020). Evaluation of speech-based digital biomarkers: Review and recommendations. *Digital Biomarkers, 4*(3), 99–108. https://doi.org/10.1159/000510820

Roesler, O., Liscombe, J., Neumann, M., Kothare, H., Hosamath, A., Arbatti, L., Habberstad, D., & Ramanarayanan, V. (in press). Towards scalable remote assessment of mild cognitive impairment via multimodal dialog. *Proceedings of Interspeech.*

Rousseaux, M., Vérigneaux, C., & Kozlowski, O. (2010). An analysis of communication in conversation after severe traumatic brain injury. *European Journal of Neurology, 17*(7), 922–929. https://doi.org/10.1111/j.1468-1331.2009.02945.x

Rusz, J., Benova, B., Ruzickova, H., Novotny, M., Tykalova, T., Hlavnicka, J., Uher, T., Vaneckova, M., Andelova, M., Novotna, K., Kadrnozkova, L., & Horakova, D. (2018). Characteristics of motor speech phenotypes in multiple sclerosis. *Multiple Sclerosis and Related Disorders, 19,* 62–69. https://doi.org/10.1016/j.msard.2017.11.007

Sabo, A., Mehdizadeh, S., Ng, K. D., Iaboni, A., & Taati, B. (2020). Assessment of Parkinsonian gait in older adults with dementia via human pose tracking in video data. *Journal of Neuroengineering and Rehabilitation, 17*(1), 1–10. https://doi.org/10.1186/s12984-020-00728-9

Severson, K. A., Chahine, L. M., Smolensky, L. A., Dhuliawala, M., Frasier, M., Ng, K., Ghosh, S., & Hu, J. (2021). Discovery of Parkinson's disease states and disease progression modelling: A longitudinal data study using machine learning. *The Lancet Digital Health, 3*(9), e555–e564. https://doi.org/10.1016/S2589-7500(21)00101-1

Shaked, N. A. (2017). Avatars and virtual agents–Relationship interfaces for the elderly. *Healthcare Technology Letters, 4*(3), 83–87. https://doi.org/10.1049/htl.2017.0009

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(09), 1426–1448. https://doi.org/10.1017/S0033291719000151

Shoulson, I., Arbatti, L., Hosamath, A., Eberly, S. W., & Oakes, D. (2022). Longitudinal cohort study of verbatim-reported postural instability symptoms as outcomes for online Parkinson's disease trials. *Journal of Parkinson's Disease, 12*(6), 1969–1978. https://doi.org/10.3233/JPD-223274

Siam, A. I., Soliman, N. F., Algarni, A. D., Abd El-Samie, F. E., & Sedik, A. (2022). Deploying machine learning techniques for human emotion detection. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/8032673

Stegmann, G. M., Hahn, S., Duncan, C. J., Rutkove, S. B., Liss, J., Shefner, J. M., & Berisha, V. (2021). Estimation of forced vital capacity using speech acoustics in patients with ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 22*(Suppl. 1), 14–21. https://doi.org/10.1080/21678421.2020.1866013

Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S., Shelton, K., Suncan, C. J., & Berisha, V. (2020). Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *npj Digital Medicine, 3*(1), Article 132. https://doi.org/10.1038/s41746-020-00335-x

Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2013). Can mobile health technologies transform health care? *JAMA, 310*(22), 2395–2396. https://doi.org/10.1001/jama.2013.281078

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS, 5*(6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

Suendermann-Oeft, D., Robinson, A., Cornish, A., Habberstad, D., Pautler, D., Schnelle-Walka, D., Haller, F., Liscombe, J., Neumann, M., Merrill, M., Roesler, O., & Geffarth, R. (2019). NEMSI: A multimodal dialog system for screening of neurological or mental conditions. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 245–247.

Talkar, T., Yuditskaya, S., Williamson, J. R., Lammert, A. C., Rao, H., Hannon, D. J., O'Brien, A., Vergara-Diaz, G., DeLaura, R., Sturim, D., Ciccarelli, G., Zafonte, R., Palmer, J., Bonato, P., & Quatieri, T. F. (2020). Detection of subclinical mild traumatic brain injury (mTBI) through speech and gait. *Proceedings of Interspeech,* 135–139.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., & Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE Journal of Translational Engineering in Health and Medicine, 5,* 1–11. https://doi.org/10.1109/JTEHM.2017.2752152

Tao, F., & Busso, C. (2019). End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Communication, 113,* 25–35. https://doi.org/10.1016/j.specom.2019.07.003

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer–Human Interaction (TOCHI), 27*(5), 1–53. https://doi.org/10.1145/3398069

Tisdale, D., Liscombe, J., Pautler, D., and Ramanarayanan, V. (2023, February 21–24). *Towards integrating eye gaze tracking into a multimodal dialog agent for remote patient assessment* [Paper presentation]. International Workshop on Spoken Dialog Systems, Los Angeles, CA.

Vatanparvar, K., Nathan, V., Nemati, E., Rahman, M. M., McCaffrey, D., Kuang, J., & Gao, J. A. (2021). Speechspiro: Lung function assessment from speech pattern as an alternative to spirometry for mobile health tracking. *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* 7237–7243. https://doi.org/10.1109/EMBC46164.2021.9630077

Vegesna, A., Tran, M., Angelaccio, M., & Arcona, S. (2017). Remote patient monitoring via non-invasive digital technologies: A systematic review. *Telemedicine and e-Health, 23*(1), 3–17. https://doi.org/10.1089/tmj.2016.0051

Vessio, G. (2019). Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review. *Applied Sciences, 9*(21), Article 4666. https://doi.org/10.3390/app9214666

Vidal, M., Turner, J., Bulling, A., & Gellersen, H. (2012). Wearable eye tracking for mental health monitoring. *Computer Communications, 35*(11), 1306–1311. https://doi.org/10.1016/j.comcom.2011.11.002

Vieira, F. G., Venugopalan, S., Premasiri, A. S., McNally, M., Jansen, A., McCloskey, K., Brenner, M. P., & Perrin, S. (2022). A machine-learning based objective measure for ALS disease severity. *npj Digital Medicine, 5*(1), Article 45. https://doi.org/10.1038/s41746-022-00588-8

Vitazkova, D., Foltan, E., Kosnacova, H., Micjan, M., Donoval, M., Kuzma, A., Kopani, M., & Vavrinsky, E. (2024). Advances in respiratory monitoring: A comprehensive review of wearable and remote technologies. *Biosensors, 14*(2), Article 90. https://doi.org/10.3390/bios14020090

Warule, P., Mishra, S. P., & Deb, S. (2023). Significance of voiced and unvoiced speech segments for the detection of common cold. *Signal, Image and Video Processing, 17*(5), 1785–1792. https://doi.org/10.1007/s11760-022-02389-8

Williamson, J. R., Young, D., Nierenberg, A. A., Niemi, J., Helfer, B. S., & Quatieri, T. F. (2019). Tracking depression

severity from audio and video based on speech articulatory coordination. *Computer Speech & Language, 55,* 40–56. https://doi.org/10.1016/j.csl.2018.08.004

Yu, Z., Ramanarayanan, V., Lange, P., & Suendermann-Oeft, D. (2019). An open-source dialog system with real-time engagement tracking for job interview training applications. In M. Eskenazi, L. Devillers, & J. Mariani (Eds.), *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems* (pp. 199–207). Springer International Publishing. https://doi.org/10.1007/978-3-319-92108-2_21

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine, 5*(1), Article 46. https://doi.org/10.1038/s41746-022-00589-7