Purpose: Over the past decade, the signal processing and machine learning literature has demonstrated notable advancements in automated speech processing with the use of artificial intelligence for medical assessment and monitoring (e.g., depression, dementia, and Parkinson’s disease, among others). Meanwhile, the clinical speech literature has identified several interpretable, theoretically motivated measures that are sensitive to abnormalities in the cognitive, linguistic, affective, motoric, and anatomical domains. Both fields have, thus, independently demonstrated the potential for speech to serve as an informative biomarker for detecting different psychiatric and physiological conditions. However, despite these parallel advancements, automated speech biomarkers have not been integrated into routine clinical practice to date.

Conclusions: In this article, we present opportunities and challenges for adoption of speech as a biomarker in clinical practice and research. Toward clinical acceptance and adoption of speech-based digital biomarkers, we argue for the importance of several factors such as robustness, specificity, diversity, and physiological interpretability of speech analytics in clinical applications.

Biomarkers are objective indications (i.e., substance, structure, or process) that can be accurately and reproducibly measured from inside or outside the patient (Strimbu & Tavel, 2010). Biomarkers are critical to the rational development of rapid and reliable medical screening, diagnostics, and therapeutics (Califf, 2018). Indeed, the U.S. Food and Drug Administration has deemed the use of novel biomarkers as crucial to improving the success rate and efficiency of drug development for various neurological and mental health conditions (Amur et al., 2015). Speech-based biomarkers are particularly interesting in this regard, as speech requires intricate coordination of multiple cognitive, affective, linguistic, and motoric processes, resulting in a broad range of behaviors that offer rich insights into different aspects of neurological and motor function (Robin et al., 2020).

Recent advances in automated speech processing, machine learning, and computing power have provided the means to automatically extract speech-based biomarkers at scale. Such speech analytics leverage the acoustic, articulatory, and linguistic information that is encoded in speech. Furthermore, the recent surge in remote medicine and mobile health technologies will accelerate the adoption of speech analytics into clinical workflows, especially for the management of patients with mental health and neurological conditions. Early detection or progress monitoring of these conditions is often challenging for patients due to (a) lack of access to neurologists or mental health practitioners; (b) lack of awareness of having a given condition that requires the need to see a specialist; (c) lack of an effective standardized diagnostic tool or end point; and (d) the substantial cost, such as for transportation, involved in conventional or traditional solutions. It is no surprise then that speech analytics and mobile health
technologies are increasingly gaining traction as efficient, effective, widely accessible, and affordable means of monitoring and assessing different medical conditions (Kumar et al., 2012; Steinhubl et al., 2013).

The utility of speech-based digital biomarkers as a window into mental and neurological health has been increasingly recognized in the speech and computer science literature. Recent work has investigated speech as an automated biomarker for psychiatric illness, highlighting features shown to be particularly sensitive to different diseases (Low et al., 2020; Robin et al., 2020). Furthermore, applying machine learning 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; Neumann et al., 2020). Speech also has implications beyond psychiatric disease, which primarily impact the affective domain. Indeed, a growing amount of work has demonstrated the value of speech analytic measures in populations with cognitive, linguistic, motoric, and anatomical deficits (Cordella et al., 2019; Green et al., 2018; Hlavnička et al., 2017; Meilan et al., 2018; Neumann et al., 2021; Perez et al., 2018; Perry et al., 2017; Poellabauer et al., 2015; Vásquez-Correa et al., 2018). Using speech analytics for early detection of disease onset is particularly promising because speech and voice changes (e.g., slowing of speech rate or decreased loudness) may be among the earliest signs of neurologic diseases, such as amyotrophic lateral sclerosis (ALS) and Parkinson’s disease (PD; Hlavnička et al., 2017; Neumann et al., 2021; Rong et al., 2016). In addition, with repeated administrations over time, speech analytics can be used to objectively monitor the rate of disease progression. For example, recent work in ALS used acoustic and kinematic measures of speech abnormalities to identify individuals with a fast-progressing variant of the disease (Rong et al., 2019). There is also a recent surge of interest in using speech biomarkers to track the progression and recovery of respiratory infections. A recent overview paper by Schuller et al. (2020) reviewed the use of standard speech feature-based classifiers in previous studies on cold and influenza detection and speculated that COVID-19 could be a plausible use case. Recent exploratory work in this direction proposed a speech subsystem framework for analyzing both asymptomatic and symptomatic patients with COVID-19 (Quatieri et al., 2020). This approach was motivated by the distinct manifestations of COVID-19, as it results in lower and upper respiratory tract inflammation as well as potential neurological deficits. Yet another promising opportunity for speech analytics is in the pharmaceutical industry, where it could aid in developing surrogate outcome measures for clinical trials of disease-modifying drugs (see, e.g., Green et al., 2018).

However, despite the potential of speech analytics to inform diagnostics and progress monitoring, speech features have not been widely integrated into clinical practice. We posit that the major barrier to such speech biomarker adoption is the lack of correspondence between the utility of various speech analytics, as suggested by multiple studies in the signal processing and machine learning literature on the one hand and their verification, analytical validity, and clinical validity on the other. To that end, this article (a) proposes a causal framework that links health and disease to speech changes, (b) discusses how one can leverage signal analytics to extract biomarkers that relate these frameworks and bridge this “interpretability gap,” and (c) elaborates on several barriers that must be overcome to strengthen the relationship between frameworks and analytics in establishing clinical validity of speech-based biomarkers.

### Speech as a Diagnostic Pathway: A Biopsychosocial Framework

The central thesis for viewing speech as a diagnostic pathway (and thereby moving toward an understanding of how to better establish clinical validity of speech-based biomarkers) can be understood within a biopsychosocial framework (Engel, 1977). This influential interdisciplinary framework has been used in prior work (George & Engel, 1980) to explain the contributions of biology, psychology, and socio-environmental factors to complex behaviors (see Figure 1). Our adaptation of the framework is focused on how speech can detect abnormalities in multiple domains, in particular, cognitive, linguistic, affective, motoric, and anatomical. This includes studies on Alzheimer’s disease (cognitive domain), poststroke aphasia (linguistic domain), depression (affective domain), ALS (motoric domain), and face transplant (anatomic domain). Table 1 provides several examples from the literature that demonstrate the potential clinical application of speech analytics in a wide range of medical and mental health conditions. The underlying assumption of this rapidly growing body of literature is that each domain has an influence on speech output that can be evidenced by acoustic and kinematic speech features. For example, abnormal formant correlations may reflect deficits in the affective domain for individuals with depression (Williamson et al., 2019), and prosodic features (pitch, energy, chroma [melody]) contribute strongly to change in affect in post-traumatic stress disorder (PTSD; Marmar et al., 2019). Whereas reduced range of motion may reflect deficits in the anatomical domain, for example, for individuals with face transplant (Perry et al., 2017).

The most established clinical use case of speech analytics is in the movement disorder space. During the oral motor exam, which is a standard component of the

---

1For precise definitions of these terms, see Robin et al. (2020).
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 (e.g., upper and lower motor neurons, the basal ganglia, and the cerebellum; Darley et al., 1969). These lesions and their movement disorder consequences have also been causally linked to the different types of dysarthria (i.e., spastic, flaccid, ataxic, hypokinetic, hyperkinetic, and mixed). For example, damage to the basal ganglia due to PD results in hypokinesia (“decreased movement amplitude”), which, when manifested in speech muscles, results in reduced oral volume and imprecise articulation. In addition to dysarthria, the other primary speech motor subtype, apraxia of speech, is characterized by deficits in the left frontal lobe, which often result in affected motor planning (Duffy et al., 2014). For over 4 decades, clinical researchers have been actively searching for speech and articulator movement correlates that best capture the unique signature of each movement disorder. For example, cerebellar damage may lead to impaired movement timing, which can be quantified using voice onset time (Ackermann & Hertrich, 1997). Similarly, a dopamine deficiency in the basal ganglia may lead to reduced range of movement, which can be quantified using the range of the second formant (Volkmann et al., 1992). Importantly, specific lesions can result in clusters of impairments (e.g., impaired range and timing of movement) rather than singular deficits that are independent from one another (Darley et al., 1969). Thus, beyond work on individual speech features, recent studies have explored the validity of acoustic-based impairment profiles or phenotypes (Rowe & Green, 2019; Rusz et al., 2018).

Challenges and Recommendations

In this section, we discuss the clinical and technological challenges toward clinical adoption as well as potential paths toward solutions. The first class of challenges lie in the domain of clinical science—although speech analytics has been well explored in the research literature, the lack of rigorous psychometric testing presents a significant barrier to clinical adoption. When biomarkers are used as outcomes in clinical trials, they are considered surrogate end points, which are well-characterized biomarkers with established clinical and neurophysiological relevance and are used as substitutes for clinically meaningful end points. A recent review defined the multistage process required to validate the safety and efficacy of a biomarker (Robin et al., 2020), which included verification (i.e., checking quality of speech recordings), analytical validation (i.e., checking accuracy and reliability), and clinical validation (i.e., checking correspondence to clinical outcomes). In addition to being clinically interpretable and neurophysiologically meaningful, their efficacy must be demonstrated through rigorous testing of diagnostic accuracy (i.e., sensitivity or true-positive rate and specificity or true-negative rate) and responsiveness to change and diversity. These last two considerations also have important implications for operationalizing signal processing and machine learning algorithms and their associated speech analytics for use as biomarkers in clinical and pharmaceutical practice. Indeed, in order to move the field forward particularly from scalability and analytical validity standpoints, powerful machine learning models are crucial. Other relevant considerations for such models include signal attribute
issues, sample size for training and statistical inference, and usability and robustness. We elaborate on these considerations in more detail below.

**Interpretability**

If a measure is uninterpretable, clinicians may be skeptical of its clinical use and added benefit to their perceptual judgment, which often relies on frameworks such as those introduced in Speech as a Diagnostic Pathway: A Biopsychosocial Framework (e.g., the Darley, Aronson, and Brown [DAB] model introduced by Darley et al., 1969, to target regions of the brain associated with a disorder). A primary advantage of models like DAB that has enhanced their adoption among health care professionals is the connection to the brain. In the case of the DAB model, each cluster of perceptual features has a presumed neurological underpinning. However, the biases inherent to perceptual judgment limit the utility of this paradigm as a valid and reliable diagnostic tool. There is, thus, a need to merge the quantitative metrics of speech analytics with domain knowledge from clinicians to achieve the ultimate goal of an interpretable, multidimensional approach.

**Sensitivity and Specificity to Domain**

The approaches used for speech and face analysis often involve data-driven algorithms that automatically extract a large number of features. While these algorithms can obtain high sensitivity in detecting speech abnormalities secondary to dysarthria, it is often difficult to determine the biological or physiological relevance of the extracted features (Tu et al., 2017). In other words, these measures might be sensitive but not specific to a given domain. This challenge can significantly limit the utility of data-driven techniques for identifying proposed biomarkers that are not only sensitive but also specific to diseases with certain impairment profiles. Despite the ongoing effort to phenotype...
speech motor disorders based on the known pathophysiological mechanisms of a disease (Rowe & Green, 2019; Rusz et al., 2018), there is a persistent need for the use of more impairment-specific or diagnosis-specific features in diagnostic models. For example, a prediction model for depression may perform as well on PTSD or traumatic brain injury (TBI), as it may simply be sensitive to deviations from typical function. Related to the problem of specificity is the presence of comorbidities. Comorbidities can introduce significant variation in clinical presentation, which would in turn affect the accuracy of measures that are designed to detect abnormalities in a specific condition. For example, a measure that is developed to detect the cognitive impacts of TBI may be confounded by the affective impacts of comorbid depression. This challenge highlights the need for not only individual features but also a cluster of features that, together, can detect the primary disease and any existing comorbidities.

**Atypical Speech Diversity**

The variety and complexity of abnormal (as well as normal) speech patterns makes it difficult to aggregate data across participants, build models that generalize across speakers, and generate reliable simulations for model training, testing, and validation. Speech attributes can vary significantly across individuals even within the same diagnosis depending on the speech subsystems that are affected and the severity of neurologic involvement. When developing diagnostic models, it is crucial to account for severity to ensure that any between-groups differences found in speech abnormalities are driven by diagnosis and not just by severity level (Weismer et al., 2001). Additionally, speech analytics models must factor in potential differences due to demographic characteristics such as sex (for instance, male and female speakers differ in their pitch ranges), native language/dialect, and accent (which has implications for the use of language-specific lexico-semantic features, such as those proposed for dementia or Alzheimer’s disease, for example; see Boschi et al., 2017). Within-speaker variability is also a common challenge—some motor speech disorders, such as ataxia and apraxia, may be characterized by inconsistent production and prosodic patterns that can fluctuate depending on the patient’s level of fatigue (Duffy, 2013). For progressive conditions, such as ALS or PD, speech typically changes over the course of days or weeks. Speech characteristics can also change naturally from the morning to the evening, further emphasizing the importance of continuous monitoring obtainable with mobile and wearable devices rather than through “snapshot” visits in-clinic or in-lab. This across- and within-speaker variability, as well as natural variability over short time periods, can degrade the performance of many speech analytics and automatic classification and repression approaches (Gupta et al., 2016). It is, therefore, essential to gain a better understanding of how speech features change and cluster under these conditions and to further develop and refine methods to longitudinally track feature variability.

**Signal Attribute Issues**

Many speech analytics algorithms rely on spectral speech features (e.g., mel-frequency cepstral coefficients or formants). Yet, speech analytic approaches that rely on spectral features may be invalid or ineffective for some disordered speech types since a disease may affect any one of the core aspects of speech production—articulation, phonation, resonance, respiration, or prosody. For example, severe voice and resonant problems can have a global impact on the speech spectrum; interact with articulatory impairments in unpredictable ways; and, consequently, yield unstable speech analytic estimates as, for example, in measurements of the coordination of speech subsystems (Quatieri et al., 2020). Aperiodicity of vocal fold vibration, in particular, can significantly degrade the many measures that require robust tracking of fundamental frequency (i.e., pitch, jitter, and shimmer), formants, and the interaction of pitch and formants. In addition, hypernasality can have a smearing effect on the speech spectrum and may attenuate or enhance formant frequencies (Eshghi et al., 2019). Therefore, explicitly incorporating such knowledge into the signal processing pipeline could lead to more informed and accurate predictions.

**Usability and Robustness**

A more user-friendly platform such as mobile smartphones and tablets often implies a greater challenge for robustness. Unlike an in-clinic setting, use of mobile devices at home or in the community suffers from environment noise, reverberation, mutlitalker interference, and babble. These issues with audio quality impact the performance of various components of signal analytics pipelines (such as speech activity detection or automatic speech recognition, for instance; see Liscombe et al., 2021), which in turn impacts the detection and assessment of a health condition. Greater clinical acceptance thus requires signal enhancement when needed and signal quality measures, as well as confidence measures of predicted outcome. This issue is also tightly coupled with specificity because one can be fooled in distinguishing across conditions when data from different conditions are recorded on different platforms in distinct environments.

**Sufficiency of Sample Sizes for Training and Statistical Inference**

Many of the challenges above could be addressed with substantial improvements in the amount and quality of training data required on statistical and machine learning-based models—an essential issue we have deliberately not discussed given its established presence in the literature (Gupta et al., 2016). Toward this end, a number of
speech-based challenges have evolved over the last decade, such as those through the annual Interspeech Paralinguistic Challenges and the Audio/Visual Emotion Competitions. These challenges have provided a wealth of open-source data for conditions such as depression, PD, and autism. There is, however, still a paucity of open-source data sets for many other conditions such as ALS, Alzheimer’s disease, TBI, and PTSD. Moreover, when the data sets are available, they often combine multiple diseases with distinct speech motor subtypes or are not well characterized due to privacy issues. The absence of speech labels limits efforts to identify acoustic and kinematic features that are meaningful to a particular diagnosis. Therefore, the signal processing community needs to actively collaborate on curating high-quality training data sets to foster more effective research and development.

**Training of Clinicians or Experts**

The final challenge to clinical adoption is training clinicians to assess and interpret speech analytics and incorporate them into patient assessment, monitoring, and treatment. We envision that a human-in-the-loop strategy might work best here, with online automated speech analytic algorithms crunching through large amounts of speech, facial movement, and other data from patients and producing objective biomarkers and statistics to assist clinicians (and even patients themselves) in disorder severity assessment and progress monitoring. In this paradigm, clinicians and experts would benefit from training to understand and interpret such speech analytics, as well as the aforementioned factors associated with their extraction from patient data.

**Summary**

Speech has received substantial attention from both clinicians and technologists as a potential biomarker of human health, in large part due to its sensitivity to conditions across the biopsychosocial spectrum. Yet, there is a clear gap between the utility of various speech analytics, as suggested by multiple studies in the signal processing and machine learning literature on the one hand and their practical adoption in clinical and pharmaceutical settings on the other. This gap is derived from several criteria of clinical and technological relevance: (a) interpretability (e.g., neurophysiological meaning); (b) specificity of signal attributes in the face of many different conditions that affect speech; (c) robustness of analytics to both atypical speech diversity and application setting; (d) and generalizability and statistical power of models as promoted by abundant, good-quality training data. Indeed, before speech can be implemented as a biomarker for clinical use, its analytical and clinical validity must be demonstrated at scale through rigorous testing of diagnostic accuracy (i.e., sensitivity and specificity) and robustness to within- and across-speaker diversity, disease progression, and delivery platform variability. These criteria are currently perceived as lacking in the development of speech-based biomarkers to varying degrees but, once enhanced, may serve to foster acceptance in such biomarkers from clinicians and technologists alike.

**References**


