Daniel Bone, Chi-Chun Lee, Theodora Chaspari,
James Gibson, and Shrikanth Narayanan

Signal Processing and Machine Learning for Mental Health Research and Clinical Applications

Human behavior offers a window into the mind. When we observe someone's actions, we are constantly inferring his or her mental states—their beliefs, intents, and knowledge—a concept known as *theory of mind*. For example:

- "Is a person growing impatient?"
- "Are they feeling down?"
- "Are they being truthful?"
- "Will they buy my product?"

We can perceive changes in emotion and even if someone is intoxicated. The basis for our "mind reading" abilities is rooted in the influence (conscious and subconscious) that our mental state has on the cognitive and motor processes that control the coordinated production of rich, complex behaviors. Our behaviors are multimodal and encoded at different temporal scales, ranging from rapid facial microexpressions or a quick rise of vocal pitch to indicate a question, to slower body gestures like waving hello.

Behavioral differences across individuals and changes within a person are also critical indicators of behavioral and mental health. For instance, softer, slower, and less articulate speech is a progressive symptom of Parkinson's disease (which can improve with treatment) [1], and three primary indicators of autism spectrum disorder are poor eye contact, lack of conversational skills, and an atypical speech prosody (the rate, rhythm, and intonation of speech) [2]. Behavioral markers are also relevant to relationship quality; for example, extended use of second-person singular pronouns (e.g. you) during conversational speech has been linked to increased levels of blame in romantic relationships [3].

Table 1. The prevalence of selected health conditions in the United States [6]–[9].

Condition	Ages	Prevalence
Autism spectrum disorder	Children*	1.5% (lifetime)
Posttraumatic stress disorder	Adults	3.5% (one year)
Mood disorders (e.g., depression)	Adults	9.5% (one year)
Alcohol addiction/abuse	All	6.6% (one year)
Illicit drug use (nonmarijuana)	All	2.5% (one year)
Parkinson's disease	≥80 years old	1.9% (lifetime)
Dementia (e.g., Alzheimer's disease)	≥60 years old	6.5% (lifetime)

*Typically diagnosed in children but symptoms persist over the life span.

The span of mental and behavioral health conditions is vast (Table 1) and so is the accompanying cost to individuals and society at large. The National Institute of Mental Health (NIMH) has estimated that the total costs of serious mental illness, affecting 6% of the population, is in excess of US$300 billion a year [4]; this doesn't even include developmental or substance abuse disorders. Aside from financial costs, these disorders reduce life expectancy; unlike many other health conditions, years of life lost due to neurological, mental, and behavioral disorders has increased recently, representing a rising burden that will impose new challenges on the health system [5]. Thus, translational research that improves aspects of health awareness, access, treatment quality, and cost will have a profound impact.

The core means of behavioral and mental health assessment have changed little over the past decades: human evaluation (by direct observation, interview, or self-report) is the primary tool. Humans are excellent signal processors, having the ability to transfer knowledge from related experiences and being able to account for, and adapt to, context with relative ease. But humans also have limitations in their perceptual abilities. First, to label "large" data sets, an army of human workers may spend hundreds of hours meticulously annotating videos; this solution does not scale with the pace of collected real-world data. Second, human judgment is subjective and idiosyncratic; reaching an agreement on qualitative judgments such as the level of rapport between individuals is challenging, and it is unlikely a person can precisely estimate quantitative

(continued on page 189)

Digital Object Identifier 10.1109/MSP.2017.2718581
Date of publication: 6 September 2017

 (continued from page 196)

measures as basic as relative speaking times. Additionally, people are not consistent in their judgments over time due, in part, to changes in mood, focus, and fatigue as well as learning effects. Finally, humans can only see and hear what is observable and thus cannot directly measure another's physiology (although in many day-to-day interactions, this is often for the best).

Signal processing is primed to have a transformative impact on behavioral science, a data-rich domain comprising noisy signal data that holds information on individuals' hidden mental states and traits. Clinical experts will always be critical to mental health assessment and treatment, but computational methods can now support their efforts with time-continuous, objective measures of a social scene. In particular, since experts are unable to constantly observe their patients and find quantification of behavior challenging, behavioral signal processing (BSP) methods can augment their abilities [10]. BSP seeks to quantify qualitatively characterized behavioral constructs at scale using low-level behavioral measurements.

Formally, the problem that we present is that of identifying the hidden attributes of the system that modulates the body's signals, uncovered through novel signal processing and machine learning on large-scale multimodal data (Figure 1). Signal processing is the keystone that supports this mapping from data to representations of behaviors and mental states. The pipeline first begins with raw signals, such as from visual, auditory, and physiological sensors. Then, we need to localize information coming from corresponding behavioral channels, such as the face, body, and voice. Next, the signals are denoised and modeled to extract meaningful information like the words that are said and patterns of how they are spoken. The coordination of channels can also be assessed via time-series modeling techniques. Moreover, since an individual's behavior is not isolated, but influenced by a communicative partners' actions and the environment (e.g., interview versus casual discussion, home versus clinic), temporal modeling must account for these contextual effects. Finally, having achieved a representation of behavior derived from the signals, machine learning is used to make inferences on mental states to support human or autonomous decision making.

Why now?

The data ecosystem continues to grow and become more integrated into our daily lives, enabling multiple opportunities to positively affect human health and well-being. Low-cost physiological wearables that can derive direct measures of a person's internal state are increasingly common as are wearable vision and audio devices. The medical Internet of Things (mIoT) market has been projected to be worth US$117 billion by 2020 [11], representing 40% of the total IoT market, but emerging signal processing research could push that value even higher. Signal processing has become the critical missing link in translating ubiquitous sensors into real impact on mental and behavioral health. There is an immense need for extracting actionable information from multimodal biobehavioral signals. Imagine being able to track someone's complex language use in relation to disease state over years, the effects of an intervention on an autistic child's social functioning, or predicting relationship outcome based on how well people control both their internal physiology (arousal, body temperature) and their expressions toward a spouse during conflicts (e.g., [12]).

Algorithms are now able to handle multiple sources of variability in a wide range of collected data. Applications such as speech recognition and computer vision demonstrate the capabilities of handling this mapping from low-level noisy signals to midlevel behavioral features, and, more frequently, researchers

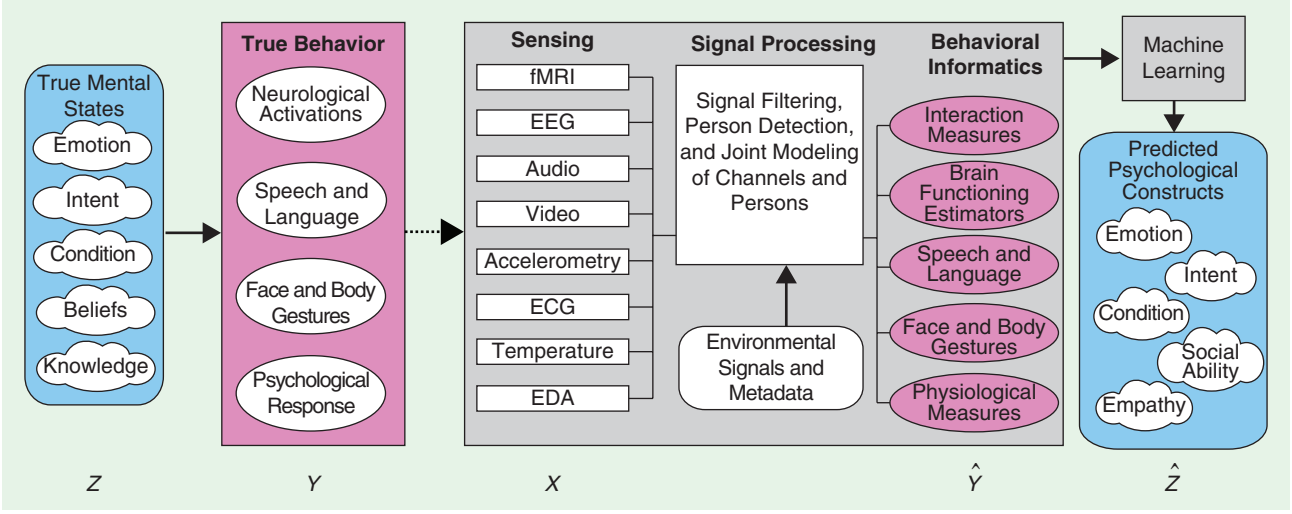

FIGURE 1. A schematic diagram of behavior generation and the mental state prediction process. First, a person's current mental state (Z) affects their behavior (Y). Raw signals are received (X), on which signal processing is performed to produce representations of behavior ($\hat{Y}$). Finally, machine learning is used to predict a person's mental state ($\hat{Z}$).

are extending the mapping to higher level psychological constructs like affect and empathy. Recent technological advances in signal processing and machine learning have enabled new opportunities, while other possibilities will require continued advancement in fundamental mathematical foundations and algorithm development.

Scientific opportunities

Three settings for BSP

In moving from qualitative descriptions to quantitative representations, there are three primary settings in which signal processing and machine learning are used to measure human behavior (as shown in Table 2). The first takes a first-person view of inferring what is happening in one's mind in relation to behavioral (unobtrusively observable) and bodily (measurements often requiring at least some level of intrusion) signals: the mind-body connection. The current standard is to use self-report or human observation of specific behaviors, and then condition on known experimental variables—for example, measuring behavior/physiology during lying and truth-telling in a scientific study of deception. Mathematical formalisms can also determine the possibilities and limitations contained within signals such as the observability, controllability, and stability of an underlying state description.

A closely related experimental setting uses quantified behavior to understand the process by which humans perceive and make decisions. Signal processing can provide measures that scale, are quantitative and objective, and are consistent across time. Moreover, if we can identify the behavioral building blocks for perception, then we can begin to ask how instances are aggregated and weighted toward a global decision: is a single salient behavior most critical to a perceptual construct, or are all relevant events summed to arrive at a decision?

The third setting is that of interaction between multiple humans or humans and machines. Behavioral scientists are interested in systematically modeling the dynamics of multiparticipant processes but often struggle to fully capture the inherent complexities. A rigorous signals and systems approach to human interactions is needed to capture the complexities of interaction dynamics as well as the inherent uncertainty in all facets of the process (perception/cognition/action) and ultimately to build the mathematical foundations for combining human and machine computation such that machines augment human capabilities, not simply replace them.

Interpretable representations

One of the primary reasons that signal processing is critical to behavioral computation is interpretability. It is often not enough to simply make a decision in health care, but the end-to-end system must explain how it arrived at such a decision. Decisions on a person's health are so vital that even if a system is able to diagnose a disorder with 100% accuracy, if that system is a black box such that the decision-making processing is unclear, health-care providers are highly reluctant to trust it. And rightfully so, since there are certainly pitfalls to which machine-learning approaches can succumb.

One way to address this is through a top-down approach, which incorporates human knowledge into feature generation and modeling by taking into account the structure of the data as well as the phenomenon that is being modeled. For instance, consider the case of measuring vocal arousal, or excitement, which is generally expressed vocally by a higher pitch, intensity, and high-frequency energy content. This reliable observation has been incorporated into a rule-based algorithm that has matched state-of-the-art performance in cross-corpus vocal arousal recognition [13]. Similarly, we can quantify affect in any text through semantic similarity metrics by leveraging a small set of seed words that have been carefully annotated by human experts [14]. Another knowledge-inspired approach that will be discussed in later sections is that of quantifying what is atypical about the speech of individuals with autism [15]. Top-down generative approaches dominate the neurosciences because they afford a mechanistic understanding in terms of low-dimensional biophysical constructs, and thus signal processing techniques have potential for strong contributions (for example, in modeling of clinical trajectories in the emerging field of computational psychiatry [16]).

Despite their advantages, top-down approaches might not be able to take into account the entire variability of the data space, which is more likely to be incorporated through data-driven representations. Toward this direction, unsupervised clustering and feature-learning techniques as well as deep and recurrent neural network approaches have been proposed. However, risks with bottom-up approaches include the limited presence of human experts guiding these engineering efforts and the risk of overfitting to inadequate data—either in terms of quantity or quality (e.g., variable and/or bad recording conditions). Semisupervised learning might be an intermediate solution that combines labeled and unlabeled data to build better learners and, at the same time, incorporates directly the knowledge if human experts.

Aiding human decision making

Much of the research community's focus has been on developing the core methods of behavioral sensing; however, computational methods don't need to replace humans but instead can augment their capabilities especially in complex clinical decision making and in understanding its impact. A great challenge going forward continues to be interfacing machine

Table 2. Three settings for BSP.

Setting	Description
Mind-body connection	Studying produced behavior in relation to known internal or contextual variables.
Human perception	Relating the produced behavior to human perception to explain perceptual processes.
Interaction	Quantifying the give-and-take behavioral dynamics that occur in human-human or human-machine interaction.

computation with human computation. Imagine a scenario in which an individual wants to observe the behavior of her employees, patients, or students based on video-recorded data. That person may currently be limited to manually evaluating several random instances. Instead, signal processing could be used to build models for the behaviors of interest and locate the instances in the data that are most relevant, or salient, to an overall judgment. One such recent approach is to identify salient words using attentional recurrent neural networks (RNNs) [17].

Measuring what we can't see or hear

Aside from overt signals that we freely perceive with vision and audition, there are covert signals that can reveal other direct insights into a person's mental state. Consider the case of lying: when we lie, generally our blood pressure rises, we breath faster, our heart rate increases, and we sweat (these are the signals that a polygraph tracks), all while we attempt to express behavioral signals of calmness and honesty. This potential mismatch between internal state and the related behavioral expression has clinical relevance as well. Take, for example, the case of an individual with moderate-to-severe autism spectrum disorder who also displays self-injurious behavior. Such an individual may not give outward indications of building agitation; but it has been shown that, in certain cases, physiological monitoring of arousal can indicate an impending episode, which would be critical knowledge for care providers. While telemetric monitoring of physiological and cognitive behavior has been considered for years, we are (rapidly) improving the sensing and computational tools to make sense of these physiological signals and act upon it in the moment. The great challenge currently is to develop robustness in the sensors and signal processing.

Multimodal temporal modeling of individuals and interactions conditioned on context

Most psychological phenomena are complex and multifaceted, therefore requiring a multimodal set of information for detecting and quantifying them. Integrating multiple sources of data can afford us a more complete view of the human state but, at the same time, imposes challenges in terms of information perception and data fusion. Different modalities can be complementary, redundant, or even conflicting. Multimodal patterns and interactions can also be situation dependent and person specific. The conceptualization of the acquired multimodal information (e.g., through human-derived annotations, self-reports, metadata, etc.) can help us identify the most relevant bits and streams of information at a given timepoint and for a given situation.

To elaborate on these points, we will draw an example from the family studies domain, which investigates close relationships between family members and romantic partners. Researchers in this domain are particularly interested in the way partners experience interpersonal conflict since conflict affects quality of life, mental health, and physical well-being. Recent advances in mIoT and wearable technology allow us to capture interpersonal conflict in everyday life via multimodal sensing, enabling answers to a variety of questions related to this phenomenon, i.e., when and why conflict occurs, whether it is efficiently handled, and when it escalates.

Different people experience conflict in different ways and through various signal cues depending on their personality, family history, and previous life experiences. For example, conflict in avoidant people might be expressed through changes in their physiology, while in competing personalities such episodes might be apparent through acoustic and visual cues. This variability is enhanced by other contextual factors, such as the individuals' physical conditions (e.g., medication intake, caffeine/drug/alcohol use) and locations (e.g., home, work, etc.). Taking these into account, reliable conflict detection systems should consider a multimodal stream of information (e.g., audio, language, physiology, etc.) complemented by contextual data streams (e.g., global positioning system location, self-reports of nutrition and substance use, measures of body temperature and acceleration); an example of such ubiquitous sensing is [12]. Since every individual and every couple have a different baseline functionality, attention should be paid to establishing personalized models through knowledge-driven priors (e.g., background of family aggression, previous relationships, etc.) and data-dependent approaches (e.g., adapting system decisions based on data from similar participants).

Beyond the typical multimodal feature fusion, explicitly capturing the interaction of various modalities within person and across people can be very informative. Aggregate statistics can only provide information about the overall relationship between participants but not how it evolved. This can still be useful; for example, previous work has demonstrated that in diagnostic interactions for autism, the psychologist—who is conducting a semistructured interaction—must adapt his or her behavior to the child with whom they are interacting [15]. The results showed that children with higher-severity autism spoke less, while the psychologist spoke more; the child spoke with higher prosodic variability, as did the psychologist; and so on. In fact, the psychologist's prosodic features were at least as predictive of the child's autism severity as those same features from the child.

Forthcoming advances can help in understanding interaction through dynamic modeling (the third BSP setting). It is with such models that capture emotional and behavioral evolution that computational tools will gain widespread validity and utility. General mathematical models that have been considered include dynamical systems equations, hidden Markov models, conditional random fields, and long- and short-term deep neural networks. Knowledge-driven algorithms are also being created. One such example is the quantification of behavioral entrainment, or synchrony, which is the mutual influence displayed between interacting partners; this influence is said to be stronger for more positive interactions, but human observers have great difficulty explaining "how" and "why" some pairs are more entrained than others. Hence, Lee et al. [18] have proposed a signal similarity metric that computes the convergence of two interacting people with respect to their representative signal-spaces, i.e., principal component (PCA) space. Generally, the vocal behavior of partners

should become more similar the more they are entraining. Specifically, data space similarity between two people was quantified through the temporal convergence between their corresponding PCA axes, i.e., through the angles formed by each pair of these principal directions:

$$\text{ssim}(X_1, X_2) = \text{trace}(W_{1L}^T W_{2L} W_{2L}^T W_{1L}) = \sum_{i=1}^{k} \sum_{j=1}^{k} \cos^2(\theta_{ij}), \quad (1)$$

where X_1, X_2 represent expressed behaviors of speakers 1 and 2, and W_{1L} and W_{2L} correspond to the reduced rank projection matrix of each individual toward their signal spaces; θ_{ij} is the angle formed between the ith principal component of the data space from participant 1 and the jth principal component from participant 2. Lee et al. showed that this signal processing-derived synchrony measure related to couples' attachment styles and in extension can be a good indicator of conflict episodes.

Connection between physiology and the brain

The true internal state can be thought of as a particular configuration of brain functioning resulting from complex neural activities within. Proxy measures of neural activities can now be captured and processed via advances in brain activity sensing, e.g., the recently prevalent functional magnetic resonance imaging (fMRI), and widely used electrophysiological signals (electroencephalography, magnetoencephalography, electromyography, etc.) Each of these signals is a loaded signal, differing in spatial-temporal resolution and utility as a proxy measure of human brain activity. Many of these signals require advanced signal processing methodologies to provide a valid objective representation on a particular neural functioning from the raw data. Additionally, these measures need to be further conditioned on appropriate expert-designed stimulation protocols (i.e., controlling factors to address the extreme variability within human brain).

Capturing relationships between brain functioning, physiological responses, and expressive behaviors is crucial in objectively piecing together the intricate interplay the between mind and body, and it relies on signal processing. Previous studies typically address these three targets separately, but signal processing will become the "glue" between them. For instance, one may simultaneously understand the effects of signal context (e.g., a high-blaming tone of voice) as the trigger of emotional stimuli in changing the brain/physiological responses (measured through another set of quantitative proxy metrics) as well as the resulting reaction (e.g., a stressed, annoyed facial expression). Modeling the timing, coordination, and appropriateness of each in relation to another with foundational mathematics will introduce insights that are not easily obtained as three isolated components.

Clinical opportunities: Enhancing human perception, cognition, and action

The initial step in conducting interdisciplinary research in behavioral science is to consider what the outcome of the research will be. There are two primary computational approaches: predicting a label or generating a rule-based definition of a label (e.g., the entrainment measure discussed previously). What makes this interdisciplinary research even more challenging is the researcher often desires his or her work to be both computationally meaningful and relevant to the application domain; however, even a simple, interpretable computational system may be revolutionary for the behavioral science domain, but might necessitate a reduced computational complexity. Nevertheless, given the promise of signal processing for mental health, we provide a short overview of certain clinical opportunities, with computation ranging from simple feature extraction to end-to-end human-in-the-loop systems.

Screening and diagnosis

For engineers, a straightforward application of technical know-how is to classify data, which in this case means diagnosing or screening for a disorder. In some cases, this may be a viable approach. However, many signal processing techniques are currently limited and trail far behind human perception and judgment, which means humans have a firm hold as the gold standard, and computational techniques ought to support their efforts. Consider image processing: the world's leading experts have only recently been able to robustly identify animals within a photograph; but transfer learning of those models are currently being applied to static medical image-based diagnosis. Yet human behavior is much more complex, nuanced, and dynamic than a static image, so we assert that machines have a long distance to go in mimicking human perception and action, although there are tantalizing possibilities in store for the future.

One viable application of machine learning has been in developing robust diagnostic and screening algorithms. Traditionally this has been done through both hand-chosen features and statistical analysis, which does not optimize the desired objective function directly—the objective function is a combination of sensitivity (recall) and specificity (true negative rate). Machine learning is a perfect fit! Recent work in autism diagnostics has shown that machine-learning based algorithms can effortlessly fuse coded behaviors from multiple diagnostic instruments, are tunable, and can easily reduce the total coded behavior set (feature reduction), effectively shortening the administration time [19].

Taking another angle, consider the case of "atypical" prosody, a prevalent behavioral characteristic marker in neural and motor disorders, which has been said to be the most consistent marker of autism across the life span, although it varies with age and language level and across individuals. Autism researchers are currently constrained to analyze very small amounts of data with meticulous, time-consuming coding; but that coding is often unreliable. One of the primary research thrusts of Bone et al. has been to provide a computational definition of prosody for developmental disorders that could be used in conjunction with human perception of other symptoms [15]. For either purely data-driven or expert systems, one of the greatest advantages may be for monitoring behavior over long periods of time.

Behavior tracking and intervention

Today, metabolic health monitors such as Fitbits are very popular; these monitors are based on extremely simple signals and provide limited metrics, e.g., inferring full body movement and thus exercise from only the motion of the arm. But many people use them as a reminder to better their own health via behavioral regulation. Thus, they are a great example of how one might scale a response to meet the demands of the global community.

Once we quantify behavior objectively using signal processing and machine learning, we can monitor those behavioral constructs over time. This opens up immense opportunities to not only assess the trajectories leading up to and after diagnosis but also to assess the response to a pharmacological or behavioral intervention. Furthermore, we can in turn utilize these relevant behavioral measures to create novel interventions. For example, once we adequately characterize "atypical" speech prosody as it relates to developmental disorders, we can create personalized computerized interventions that would otherwise not be possible. One can imagine a scenario in which behavioral monitoring of self and others is incorporated into ubiquitous computing devices that provide feedback to the wearer; researchers are already considering the use of tech-glasses to assess the emotion of the person they are interacting with, but much more will be possible as this field evolves.

One example utility of monitoring behavior is recent work aiming to monitor a patient's mental health over time based on phone call recordings [20]. In this paradigm, a person who was seeing a psychiatrist for symptoms of a mental health disorder (i.e., major depressive disorder, bipolar disorder, schizophrenia, or schizoaffective disorder) would frequently call into an automated system and leave an update on their status. The patients responded to automated questions on how they were overall, what was going well, and what had been bad. The interdisciplinary team employed vocal analytic and natural language processing techniques to build global and person-specific models of how a person's speech characteristics related the psychiatrist's opinion of their emotional state. Initial findings suggest that speech features are able to quantify, to an extent, how well someone is doing. Relatedly, social network analysis is emerging as a potential means to track a person's state over time at scale.

Providing feedback to the health-care professional

Another avenue to impact mental and behavioral health is through providing objective feedback to the provider about their own actions, serving as an enabler of training. Providing clinicians an in-depth overview of therapy and diagnostic sessions empowers them to both track patient progress as well as alter their own approaches.

A good candidate for such a system is psychotherapy interactions, due to their overall importance in treating mental health as well as their semistructured nature. These interactions typically occur between a limited number of participants (i.e., a therapist-patient dyad) where the primary mode of communication is verbal. The words that are spoken can be analyzed for important counselor behaviors such as reflections (restating what the client says to demonstrate understanding) or the patient's commitments to behavioral change, both of which indicate positive addiction counseling sessions [21]. Beyond the words that are said, the tone of voice, emotion, and politeness can be critical feedback for a counselor. To enable this goal, an automatic system has been developed that segments speaker utterances, assigns them based on automatically determined role, performs transcription, and, finally, predicts behavioral codes. Developing such a tool for clinical use requires incorporating user interface and experience design elements that make the experience both useful and intuitive; a tool presented in [22] allows clinicians and supervisors to select individual utterances and review the automatically transcribed words as well as the behaviors predicted to be present in that particular utterance. It also displays session-level gestalt behaviors and behavioral counts to provide a high-level overview of important measures such as therapist empathy.

Challenges and opportunities in formulation and implementation

While we have established that behavioral modeling from signals is a viable path forward for mental health, as the field grows there are critical questions that must be answered. Some of these challenges are unique to each problem, while others are broadly applicable to this interdisciplinary research.

Data collection and modeling

All stages of the behavioral signal processing pipeline are intertwined; even the problem motivation is not independent of the technical ability and process for achieving a goal. Therefore, it is also imperative that all aspects of BSP problem design are adequately addressed. The first stage is data collection, for which we desire data of high quality and high consistency. Since behavior is inherently multimodal, we must collect multimodal data. This collection should be ecologically valid or natural and not affecting the behavior itself. Another critical factor is time. One of the great contributions of technology will be longitudinal, time-continuous monitoring of behavior. Consider that a psychologist typically has one hour to interact with a child as part of an autism diagnostic evaluation. That child may simply be having a bad day, so assessing progress at the individual level is obstructed. Alternatively, for privacy or data storage reasons, sampling often needs to be done in a time-sensitive (i.e., "sparse") fashion.

The second and third stages are analysis (deriving behavioral features) and modeling (mapping features to behavioral constructs), for which the core challenges stem from the extreme complexity of human behavior and the uncertainty in the measured signals. No two expressions will produce identical signals due to variability within a person and across individuals, as well as external sources of noise. For example, we never actually speak a word the same way twice: intonation will vary with mood, certain syllables will be spoken slightly faster or slower, and the channel conditions can change between recordings. Computational systems must be robust to such heterogeneity and further must translate across data sets; for example, it has

proven surprisingly difficult to translate findings in one emotional database to another—a formidable hurdle in developing a real-world system. One viable approach is to use transfer learning, wherein similar learning tasks are used to bias models to learn more quickly how to perform related tasks. Moreover, analyzing multimodal temporal signals comes with the signal processing challenge of synchronization and addressing diversity in timescales.

Another crucial question from the behavioral science perspective is how to group behaviors and know when behavior shifts—and thus a person's latent state (e.g., mood) may have changed. Given appropriate multimodal, longitudinal data, computational researchers can devise mathematical formulations to differentiate normal data variability and anomalies from medically pertinent behavioral transitions. Relatedly, it is of interest to know what types of behaviors co-occur and whether there are subpopulations within a disorder that exhibit similar tendencies; novel clustering approaches that deal with disparate data types can bring new insights.

Finally, the choice or formulation of a target behavior representation is not always straightforward. Often we rely on human annotation, but this is inherently biased and typically adds another layer of imperfection and variability to the modeling task. Much care must be taken in deciding on a reference behavior of interest. Once a construct is chosen, various mathematical approaches have been proposed to leverage the reality that different raters are more reliable in unique situations (e.g., [23]).

Building community among researchers

The most critical step that computational and behavioral science researchers can take is to develop intimate, sustained, trusting, and productive partnerships from the early stages of a cross-disciplinary research program. Creating technological solutions is a strenuous process that can be wrought with failure points, and these are only multiplied for interdisciplinary projects. Collaborators should communicate often and foster intellectual openness. Furthermore, in this age of interdisciplinary research, there is an opportunity to construct educational programs that jointly train collaborative researchers in both clinical and computational domains. At the intersection of such collaboration will emerge a new team of scientists who have in-depth knowledge in multiple domains and will lead the charge in technology transfer.

Our greatest standing challenge

This brings us to the most imperative challenge that we face in BSP: how to go beyond human abilities, while maintaining human interpretability. Following the standard of practice, we can use BSP to automate behavioral coding (e.g., measurement of empathy [21]). Such systems can bring an expert's opinion (possibly a collection of experts) to a wider audience, automatically, faster, at low cost, and with a certain objectivity/consistency. In specific cases, systems can be independently employed, such as in the example of providing feedback to a therapist on empathic skills. But in many other cases, the human expert will still be next-to-the-loop with their own opinion, which may be equally valid to the automated one (i.e., when the automated system matches inter-human agreement). As we move the field forward, we must create automated systems that profoundly augment human capabilities, effectively integrating into normal workflows.

Fully autonomous perceptual systems would open previously unimagined translational potentials. But how can we possibly go beyond human perception if we don't use human perception for modeling or validation (noting that these systems with independent perceptual outputs would likely still rely on expert utilization)? This is an open challenge for which we are only beginning to find problem-specific solutions. As discussed previously, one approach is to define a computational construct in a top-down manner, and then compare it to peripheral constructs or outcome measures. Recall Lee et al.'s knowledge-based measure of vocal entrainment [18], for which there is no reliable quantitative perceptual measure; validation was made indirectly through (hypothesized) correlation with couple relationship quality. Such top-down approaches require a certain faith in the system's design. In fact, this approach resembles the design of psychometric instruments; thus, similar methods of reliability testing and validation can be employed. Still, autonomous system design is the most challenging and critical problem we should address.

Vision for the future

Computational science undoubtedly has much to contribute to human behavioral study, and this is an exciting time to be a signal processing researcher. It will take a joint, collaborative effort to solve these great challenges posed by mental health research and clinical needs. Primary computational targets include multimodality and interaction modeling as well as behavior (change) prediction. If we can overcome the engineering obstacles, we can provide enduring scientific advances and translational impact in mental health domains. Particularly one special aspect of signal processing—in the service of improving mental health and performance—is the curious fact that the brain may itself be a signal processor. In other words, many of the insights garnered in machine learning and signal processing can be applied to the functioning of the brain in and of itself. Beyond the practical benefits of data assimilation surveyed above, there may be deeper theoretical contributions of signal processing to our understanding of things like theory of mind.

An achievable dream of ours is to see engineering technologies integrated within and supporting all aspects of mental health research and care, helping to fill scientific knowledge gaps, connecting dots, and supporting novel interventions. Signal processing will enable access to a truly dynamic, patient-centric care. With cloud-based architectures and reasonable cost, these technologies can operate on a global scale, overcoming cultural and other boundaries and variables.

Authors

Daniel Bone (dbone@usc.edu) received his B.S. degrees in electrical and computer engineering from University of Missouri in 2009. He received his Ph.D. degree in electrical engineering in 2016 from the University of Southern California (USC), where he is a postdoctoral researcher in

the Signal Analysis and Interpretation Laboratory. His research concerns developing engineering techniques and systems for societal applications in human health and well-being, focusing on affective signal processing, speech and language processing, and machine learning. He received the Alfred E. Mann "Innovation in Engineering" Doctoral Fellowship (2014–2016) and the Achievement Rewards for College Scientists Award (2012–2016), served as a USC Ming Hsieh Institute Scholar (2015–2016), and was on winning teams of two Interspeech Challenges (2012 and 2015). He is a Member of the IEEE.

Chi-Chun Lee (cclee@ee.nthu.edu.tw) received his B.S. and Ph.D. degrees in electrical engineering from the University of Southern California in 2007 and 2012, respectively. He is an assistant professor in the Electrical Engineering Department of the National Tsing Hua University, Taiwan. His research interests are in human-centered behavioral signal processing and multimodal affective computing. He led a team winning the Emotion Challenge in Interspeech 2009. He coauthored a paper that won the Best Paper Award at Interspeech 2010. He is a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is a Member of the IEEE and the International Speech and Communication Association.

Theodora Chaspari (chaspari@usc.edu) received her diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2010 and her master's and Ph.D. degrees from the University of Southern California (USC) in 2012 and 2017, respectively. Since 2010 she has been a member of the Signal Analysis and Interpretation Laboratory at USC. Her research interests lie in the area of biomedical signal processing, speech analysis, and behavioral signal processing. She is a Graduate Student Member of the IEEE.

James Gibson (jjgibson@usc.edu) received his B.S. degree (magna cum laude) in electrical engineering from the University of Miami, Florida, in 2010, and his M.S. degree in electrical engineering from the University of Southern California (USC), in 2012. He is currently pursuing the Ph.D. degree in the Signal Analysis and Interpretation Laboratory at USC. His research interests focus on human-centered signal processing and machine learning in the context of mental health and well-being domains. He is a member of the IEEE Signal Processing Society and Eta Kappa Nu. He was a recipient of the USC Annenberg Fellowship, 2010–2012. He is a Graduate Student Member of the IEEE.

Shrikanth Narayanan (shri@sipi.usc.edu) received his M.S., engineer, and Ph.D. degrees all in electrical engineering, from the University of California, Los Angeles, in 1990, 1992, and 1995, respectively. He is the Niki and Max Nikias Chair in Engineering at the University of Southern California, where he is a professor of electrical engineering, computer science, linguistics, psychology, neuroscience, and pediatrics and the director of the Ming Hsieh Institute and the Signal Analysis and Interpretation Laboratory. He is a fellow of the National Academy of Inventors, Acoustical Society of America, International Speech Communication Association, and American Association for the Advancement of Science. His research focuses on human-centered signal and information processing and modeling with applications of direct societal relevance in intelligence, defense, health, education and the media arts. He is a Fellow of the IEEE.

References

[1] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," *J. Speech Lang. Hear. Res.*, vol. 50, no. 4, pp. 899–912, 2007.

[2] C. Lord, S. Risi, L. Lambrecht, E. Cook, B. Leventhal, P. DiLavore, A. Pickles, and M. Rutter, "The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Dev. Disord.*, vol. 30, no. 3, pp. 205–223, 2000.

[3] R. A. Simmons, P. C. Gordon, and D. L. Chambless, "Pronouns in marital interaction what do 'You' and 'I' say about marital health?" *Psychol. Sci.*, vol. 16, no. 12, pp. 932–936, 2005.

[4] Annual total direct and indirect costs of serious mental illness. National Institute of Mental Health. (2002). [Online]. Available: https://www.nimh.nih.gov/health/statistics/cost/index.shtml

[5] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, et al. "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010," *Lancet*, vol. 380, no. 9859, pp. 2197–2223, 2013.

[6] NIMH Statistics Home: Prevalence. National Institute of Mental Health. [Online]. Available: https://www.nimh.nih.gov/health/statistics/prevalence/

[7] NIDA nationwide trends. National Institute on Drug Abuse. (2015). [Online]. Available: https://www.drugabuse.gov/publications/drugfacts/nationwide-trends

[8] T. Pringsheim, N. Jette, A. Frolkis, and T. D. Steeves, "The prevalence of Parkinson's disease: A systematic review and meta-analysis," *Mov. Disord.*, vol. 29, no. 13, pp. 1583–1590, 2014.

[9] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: A systematic review and meta-analysis," *Alzheimers Dement.*, vol. 9, no. 1, pp. 63–75, 2013.

[10] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[11] T. J. McCue, $117 Billion market for internet of things in healthcare by 2020. *Forbes*. [Online]. Available: http://www.forbes.com/sites/tjmccue/2015/04/22/117-billion-market-for-internet-of-things-in-healthcare-by-2020/

[12] A. Timmons, T. Chaspari, S. Han, L. Perrone, S. S. Narayanan, and G. Margolin, "Using multimodal wearable technology to detect conflict among couples," *IEEE Computer*, vol. 50, no. 3, pp. 50–59, Mar. 2017.

[13] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," *IEEE Trans. Affect Comput.*, vol. 5, no. 2, pp. 201–213, 2014.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 11, pp. 2379–2392, 2013.

[15] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan. "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech Lang. Hear. Res.*, vol. 57, no. 4, pp. 1162–1177, 2014.

[16] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, "Computational psychiatry," *Trends Cogn. Sci.*, vol. 16, no. 1, pp. 72–80, 2012.

[17] J. Gibson, D. Can, P. Georgiou, D. C. Atkins, and S. S. Narayanan, "Attention networks for modeling behaviors in addiction counseling," in *Proc. Interspeech,* 2017.

[18] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 518–539, 2014.

[19] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *J. Autism Dev. Disord.*, vol. 45, no. 5, pp. 1121–1136, 2015.

[20] A. C. Arevian, D. Bone, N. Malandrakis, V. R. Martinez, K. B. Wells, and S. Narayanan, "Personalized prediction of mental health state through computational analysis of speech," submitted for publication.

[21] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "'Rate My Therapist': Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS ONE*, vol. 10, no. 12, 2015.

[22] J. Gibson, G. Gray, T. Hirsch, Z. E. Imel, S. S. Narayanan, and D. C. Atkins, "Developing an automated report card for addiction counseling: The counselor observer ratings expert for MI (CORE-MI)," in *Proce. CHI Computing and Mental Health Workshop*, 2016.

[23] K. Audhkhasi and S. Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 769–783, 2013.

SP