Published in final edited form as:
Proc IEEE Inst Electr Electron Eng. 2013 February 7; 101(5): 1203–1233. doi:10.1109/JPROC.2012.2236291.

Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language:

Computational techniques are presented to analyze and model expressed and perceived human behavior–variedly characterized as typical, atypical, distressed, and disordered–from speech and language cues and their applications in health, commerce, education, and beyond

Shrikanth Narayanan [Fellow IEEE] and **Panayiotis G. Georgiou [Senior Member IEEE]**
The authors are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA

Shrikanth Narayanan: shri@sipi.usc.edu; Panayiotis G. Georgiou: georgiou@sipi.usc.edu

Abstract

The expression and experience of human behavior are complex and multimodal and characterized by individual and contextual heterogeneity and variability. Speech and spoken language communication cues offer an important means for measuring and modeling human behavior. Observational research and practice across a variety of domains from commerce to healthcare rely on speech- and language-based informatics for crucial assessment and diagnostic information and for planning and tracking response to an intervention. In this paper, we describe some of the opportunities as well as emerging methodologies and applications of human behavioral signal processing (BSP) technology and algorithms for quantitatively understanding and modeling typical, atypical, and distressed human behavior with a specific focus on speech- and language-based communicative, affective, and social behavior. We describe the three important BSP components of acquiring behavioral data in an ecologically valid manner across laboratory to real-world settings, extracting and analyzing behavioral cues from measured data, and developing models offering predictive and decision-making support. We highlight both the foundational speech and language processing building blocks as well as the novel processing and modeling opportunities. Using examples drawn from specific real-world applications ranging from literacy assessment and autism diagnostics to psychotherapy for addiction and marital well being, we illustrate behavioral informatics applications of these signal processing techniques that contribute to quantifying higher level, often subjectively described, human behavior in a domain-sensitive fashion.

Keywords

Affective computing; behavior; computational psychology; computational social sciences; emotions; health applications; multimodal signal processing; natural language processing; speech understanding

I. Introduction

A. Human Behavior

Human behavior is complex and multifaceted. It manifests an intricate interplay among the human mind, brain, and the body. Importantly, not only does it represent the natural dynamics of an individual's internal neurological, cognitive, and physiological state, but also it reflects the influence of other agents and the environment. Behavioral expressions of an individual's actions and/or interactions hence can be widely varied depending on the individual's state and the nature of the task engaged in, as well as the external influences and context. More critically, behavioral expressions are in general heterogeneous across individuals. This heterogeneity can arise from, and be attributed to, a wide range of factors ranging from age, gender, sociocultural background, to physical and mental health status and abilities, including possible differences due to illness, disease, or disability. Additional variability in human behavior displays arises from the complex interplay between these and the infinite sources of variability in the cognitive demands of tasks and activities undertaken and variability in contextual factors and the environment, including the behavior of other individuals. All these factors make the understanding and automatic decoding of human behavior cues a challenging engineering problem.

1) Importance of Human Behavior Modeling and Prediction—Understanding, describing, and influencing human behavior is central to many domains of human endeavor. They offer a window into decoding how one is thinking and feeling. This could relate to understanding normative (“typical”) behavior patterns of an individual engaged in a task or activity. More often, it relates to detecting, analyzing, and modeling behavior deviation from what is deemed typical and in factoring out the source attributable to this variability. Consider, for example, a child engaged with a teacher in a learning activity such as reading or a problem solving exercise. In making her formative assessment, and deciding on the next course of action, the teacher may be interested in gauging not only whether the child is getting the correct answer to a question but also behavioral cues of the underlying cognitive state such as the child's confidence or certainty [1] and socio-emotional state such as frustration, engagement, and joy in the activity [2]–[5]. Computing high level states such as uncertainty and engagement from behavioral cues can be used within a spoken dialog system such as for intelligent tutoring [6], [7]. We can draw similar examples for a variety of realms. In a customer care scenario, behavioral analysis may focus on patterns reflecting likes and dislikes of a client toward a product or service or indicators of satisfaction or lack thereof. For example, in a call center interaction, the tone of the spoken dialog can flag an irate customer, a useful element to detect for quality control [8]–[10]. Finally, no other domain exemplifies the centrality of behavioral analysis and modeling more than that related to human health and well being. In particular, research and practice in psychology and psychiatry focus on diagnosing, managing, and treating atypical and distressed behavior by eliciting and/or observing behavioral cues and patterns. For example, in diagnosing whether a child has attention deficit or is on the autism spectrum, an expert often would engage the child in a series of interactive activities, targeting relevant cognitive and socio-emotional aspects, and observe the resulting behavior cues and codify specific patterns of interest (e.g., typicality of prosody, joint attention behavior) [11]. Such direct observations from an expert often are accompanied, or even replaced, by reports of behavior by self or others (e.g., from a parent of a child being screened for autism). In addition to these assessment and diagnostic scenarios, behavior analysis is central to implementing interventions and measuring treatment efficacy. In psychotherapy, the specific behavior patterns of the therapist define the quality and success of an intervention; hence characterizing those patterns is also an important goal of behavior analysis, e.g., [12].

2) Multifaceted Aspects of Human Behavior—The internal human state is expressed and revealed in behavior cues that are multimodal. Likewise, the perceived higher level human state processes are often based on observable multimodal cues of, or about, behavior. Many aspects of human behavior are expressed, displayed, and observable through overt cues. These include verbal and nonverbal expressions of communicative intent and emotions such as through vocalization, intonation, language, facial expressions, and body language. Other relevant behavioral cues are covert. These include physiological signals such as respiratory activity, cardiac activity, or electrodermal activity (EDA) response that can indicate a person's emotional arousal state and neural signals of brain activity obtainable through a variety of measurement techniques such as electroencephalography (EEG) and functional magnetic resonance imaging that can inform us about cognitive and affective states. Access to these overt and covert multimodal signals has significant implications for behavior analysis and modeling in two important ways. First, access to the covert signals can allow *direct modeling of the generative processes* representing the underlying behavior of interest. In other words, they allow the capture of “mind-body” relations by allowing the inference of the oft-latent behavioral state by observing or measuring the overt and covert multimodal cues. We can view this as modeling the behavior production or expression. Second, the availability of the observable overt cues allows us to model how people process and perceive others' behavior. This notion of capturing the observed “felt-sense” of higher level human behavior, i.e., understanding how people assess others' behavior can be viewed as *modeling the behavior perception* or experience. Measuring and quantitatively modeling both these aspects of human behavior is, however, a vastly challenging problem. Developing engineering techniques and technologies to contribute toward solving this problem is precisely the goal of behavioral signal processing (BSP).

Several advances in human centered signal processing and computing provide supporting foundation for behavioral signal processing. For instance, advances in enriched speech processing from novel models of prosody to deriving affective information from speech have contributed to key BSP building blocks (see Section II-B3). Likewise advances in modeling human conversation and turn taking offer a dynamic view of looking at behavioral patterns.Notably social signal processing has emerged out of the interest of engineers in modeling complex human social behaviors in general human–human interaction often with potential applications geared toward various engineering designs of natural human–machine interfaces and social analytics [13]–[15].

3) Challenge and Opportunity—There is a great deal of variety in the needs and goals of human behavioral analysis and modeling. This, coupled with the vast heterogeneity and variability in the possible manifestations of human behavioral patterns, makes the problem of deriving universally useful and valid behavior constructs and methodologies immensely challenging, if not impossible. The typical approach has hence been to develop approaches in a domain or application specific manner both to specify the behavior constructs of interest and to dictate the means for deriving them. The practice in the state of the art to accomplish this has been, however, largely manual, relying often on trained experts. This paper considers the promise, and challenges, offered by engineering techniques in facilitating human behavior analysis and modeling. Specifically, it focuses on deriving behavioral informatics from human speech and language.

B. Behavioral Signal Processing Preliminaries

We first offer an operational definition for BSP, both to contextualize its goals and distinguish it from the significant human-centered signal processing foundations it builds upon.

1) Definition—BSP refers to techniques and computational methods that support the measurement, analysis, and modeling of human behavior signals that are 1) manifested in both overt and covert multimodal cues ("expression"); and 2) processed and used by humans explicitly or implicitly ("experience" and "judgment"). The central goal of BSP is to inform human assessment and decision making; hence the outcome of BSP is referred to as behavioral informatics. Relatedly, BSP can also serve as an enabler of autonomous and hybrid decision-making systems including empowering advanced human–machine interfaces.

2) Facets of BSP—The three facets or ingredients of BSP are as follows.

- **a.** *Acquisition* of rich and ecologically valid data. This includes behavior data sensing, as well as the measurement of the context/environment using multimodal techniques captured with audio, video, physiological, and other sensor measurements in both controlled and natural free-living environments.
- **b.** *Analysis* focused on deriving signal descriptors informing or indicating aspects of "who, what, when, how, where, why" from the multimodal measurements.
- **c.** *Modeling* which involves mapping behavioral constructs from the derived signal descriptors. These behavioral constructs include both high-level descriptions specified and desired by domain experts as well as quantitative descriptive and predictive models that can serve as tools of scientific discovery and enable novel assessment and intervention possibilities from a variety of perspectives.

3) Technical Challenges—Given the wide variety of behavioral analysis goals and requirements across domains, BSP needs to handle varying types of abstraction in data and descriptions. This process faces a number of challenges, notably in the following terms.

- **a.** Heterogeneity and variability in how data are generated and used. There are inherent challenges in the nature of how human behavioral data are generated and expressed across individuals and context.
- **b.** Uncertainty and incomplete nature of observations. Due to the latent nature of the behavioral state, any specific measurement typically affords only partial observability of the underlying state. Importantly, many of the cues that are relied upon for behavioral state estimation and tracking often are also secondary to some other function— physiological, cognitive, or socio-emotional. For instance, vocal cues of verbal communication also offer a window into a person's physiological state (e.g., if intoxicated), cognitive (e.g., if confused), or emotional (e.g., if aroused). Moreover, the observations represent complex nonlinear effects of more than one underlying factor and, together with imperfections in sensing the overtly or covertly available cues, lead to uncertainty and noise in the data available for behavioral modeling.
- **c.** Subjectivity in human descriptions of human behavior. There is inherent subjectivity and variability in how humans observe, process, interpret, and respond to human behavioral cues. This process can be explicit (such as in behavioral annotation and assessment where experts map behavioral data into behavioral representations of interest) or implicit (such as is prevalent in human actions and interactions of daily life). Capturing this inherent subjectivity, reflecting the diversity in the human perception and processing of behavioral cues, within the computational representations and models of behavior, is a significant engineering challenge.

The emerging approaches to handle these challenges are a combination of both human-centered methods, which focus solutions directly on domain-specific needs, and technology-centered methods, which aim to design techniques that are broadly applicable across applications. One of the key hallmarks of the former is the centrality of the human (from experts to naive observers and crowds) in the processing loop in BSP, wherein the behavioral representations, models, and outcomes are directly informed, and used, by humans. The latter represents and builds upon the tremendous technological advances being made in sensing, signal processing, and machine learning, especially in acquiring and analyzing vast amounts of human behavioral data. These are further elaborated below.

C. BSP and Human in the Loop

A central aspect of behavioral analysis is the key role played by the human expert. Notably in observational approaches, guided by expert-defined behavioral representations, humans are relied upon to derive behavioral constructs using observed data. The key point here is that the analysis and modeling are codified by the human "annotator." For instance, consider a scenario in which a teacher is formally assessing a child's abilities in learning to read (formative assessment). This involves audiovisual observation of the child to measure accuracy and fluency and also to gauge how certain and confident the child is in the task. Likewise, research and practice in psychology and psychiatry that is focused on diagnosing, managing, and treating atypical and distressed behavior often relies on expert observations of behavior interactions (Fig. 1). To capture this centrality of humans in behavioral modeling, BSP follows a two-pronged approach. On the one hand, it attempts to emulate the decision making of humans to learn signal features and machine learning techniques relevant to human processing of behavioral information. This often entails manually mapping audiovisual observations of verbal and nonverbal cues of social and affective communication, critically manifested in speech, spoken language, and gestures, into behavioral descriptions that the expert defines and desires, a process referred to as behavior coding. On the other hand, the outcomes of machine processing and learning of behavioral cues are fed back to the human expert both to refine the derived representations and to augment their analytical capabilities. This human-in-the-loop notion of BSP is illustrated in Fig. 2. In sum, this exemplifies a key characteristic of BSP to support, rather than supplant, human analysis and decision making.

Technology and computing advances can offer tremendous benefits to the human expert related to observing, analyzing, and modeling human behavior. Integral to interpretation of such information in cognitively diverse and emotionally rich interactions is the development of behavior-centric computational models that encompass the cognitive, social, affective, and communicative state of the interlocutors reflected in their speech and spoken language. Such capability can augment the relevant information made available to the experts, strengthening their ability not only to take appropriate action and to intervene appropriately but also offering tools of scientific discovery.

BSP builds upon tremendous advances in many realms of signal processing that offer foundational capability to measure, analyze, track, and model human behavior. These include the ability to acquire and diarize audio and video streams to automatically segment aspects of who spoke when and is doing what, performing speech recognition for determining what was spoken, visual activity recognition for estimating head pose, face/hand gestures and body posture, and physiological signal processing. In the next sections, we will describe the technology and challenges associated with the acquisition, analysis, and modeling aspects of BSP with a focus on speech and language cues.

II. Aspects of Behavioral Signal Processing

In this section, we review the principles and practice of how to acquire, process, and model behavioral data. First, we briefly review human-expert-based practice and highlight some of the associated challenges, especially as they relate to technology development. Following that, we describe some of the engineering building blocks of BSP with speech and language.

A. Human-Centered Behavioral Analysis and Modeling

The state of the art in behavioral sciences is to rely on measurements either elicited through an appropriately designed instrument of self-report or through observation. The resulting data are codified and analyzed to test hypotheses, devise diagnostics, or plan and manage treatment. Key elements that need to be underscored in any technology design to facilitate these processes are discussed below.

1) Ecologically Valid Measurements in the Laboratory Versus the Real World —The notions of *ecological validity* and *representative design* have been of great interest to the field of psychology since the mid-twentieth century. Initially, the term representative design was used to denote that the conditions of an experiment were representative of the real world [16]. For instance, the behavior of the interlocutors during a clinical observation represented (statistically) the way that they would mostly behave.

> "Generalizability of results concerning… the variables involved [in the experiment] must remain limited unless the range, but better also the distribution… of each variable, has been made representative of a carefully defined set of conditions" [17].

Ecological validity, on the other hand, was used "to indicate the degree of correlation between a proximal (e.g., retinal) cue and the distal (e.g., object) variable to which it is related" [16], [17]. This is a very important aspect to consider in BSP: Are explicitly or implicitly expressed cues (proximal) actually correlated with the subject's intent (distal), be it conscious or subconscious? Given that the experts are observing, and the signal processing algorithms are sensing, only the expressed cues, it is vital that our scenarios of study have very high "ecological validity" in the original sense of the word. Fig. 3 shows that the *expressed* cues (be it in the audiovisual or physiological streams) are the only observable variables in the BSP domain by both the human experts and machine systems.

Over the years, the term ecological validity has been adapted to infer that the experiment approximates the real-life situation that the experts are trying to study, and it is in such a way that we will be using this term. This in a sense assumes both "representative design" and "ecological validity" in the original sense of the words. Efforts are made in ensuring representative conditions and the correlation of proximal and distal cues in psychological study designs and, additionally, such a correlation can be further established in the analysis stages. The validity of any study's conclusions is highly correlated with the ecological validity of the experiment (despite the independence of the two).

2) Subjectivity, Annotation, and "Baselines"—Observational practice is based on subjective assessments of events, expressions, and behaviors. The subjectivity of quantitatively measuring these behaviors is a major practical issue in content analysis. As Neuendorf [18] states "without the establishment of reliability, content analysis measures are useless." Establishing reliability of the measures also allows for distribution of the coding task in a more efficient way.

Ickes [19] also identifies distinct categories of subjective judgments, ranging from personality traits, interlocutors' judgments of each other, perceived affect, and empathic

accuracy, i.e., the ability to gauge the specific content of another person's thoughts and feelings. As the complexity of subjective judgments increases, the quality and confidence in subjective judgments tend to decrease. Expert or trained annotators strive to achieve consistency. The process, however, is challenged by cost and scalability issues [20].

Emerging human computation (coding) approaches such as those based on crowd sourcing [21], [22] alongside novel modeling techniques to accurately measure coder-specific, or even event-specific, reliability metrics can be instrumental toward better, cost-effective, and most importantly rapid coding. Furthermore, studies have shown that the promise of using "naive [coders] has both theoretical and practical advantages for researchers studying emotional expression" in the field of psychology [23]. Baucom *et al.* [24] in addition point to cases where the nonexpert behavior annotations can even lead to annotations superior to those of experts.

Interannotator reliability can be measured in many different ways, e.g., using percentage agreement, Holsti's method [25], Scott's π [26], Cohen's κ [27], Fleiss' κ [28], and Krippendorff's α [29]. The kappa score, for example, is a chance-corrected percentage agreement score, defined as

$$\kappa = \frac{\text{Prob. of Agreement} - \text{Prob. of Chance Agreement}}{1 - \text{Prob. of Chance Agreement}} \quad (1)$$

In our experience, an agreement of $\alpha = 0.7$ is considered acceptable in psychology studies, given the subjectivity of human events. For instance, $\alpha = 0.7$ is the upper bound among the expert generated rigorous codings in our data sets in Section III-A and D.

In BSP, we often find ourselves confronted with the paradox that BSP algorithms must imitate a reference transcription (the human consensus coding), but must also perform consistently, which quite frequently means that the BSP algorithms will agree with the consensus transcription more reliably than human annotators agree with one another. This can be addressed in some cases by improving interannotator agreement through discussion sessions where coders will discuss all disagreeing ratings until they reach consensus, although this is unusual because it is too cost inefficient. Snyder *et al.* [30] provide in-depth discussion on evidence-based approaches, especially relevant to distressed couples but which generalizes at a certain extent to other BSP domains. Considering the inherent subjective nature of behavioral judgments, simple plurality-based techniques such as majority voting may not adequately capture the diversity in the natural characterization of certain behavioral constructs requiring more sophisticated evaluator-dependent modeling techniques [31], [32].

3) Self Versus Observational Assessment—Self-assessment is a useful and widely adopted approach in many fields such as healthcare and education. In many cases, e.g., in education, the self-assessed metrics are quite reliable [33]. In other domains, such as mental health, that is not always the case. For instance, in a study of distressed couples [34] the "objective observers' inferences" of aggressiveness and feelings were significantly higher than both the self-assessment of the husbands and the nonobjective assessment of the wives. This points to the issue that self-assessments are biased, especially in nontypical/problematic cases, which are the ones of greatest interest. Another issue with self-assessment is the inherently summative nature of annotation. Snyder *et al.* [30, p. 299] discusses in depth advantages and disadvantages of selfevaluation. Of course, self-assessment elicited within the context of a dynamically evolving activity may interfere with that very activity although novel approaches such as ecological momentary assessment (EMA) attempt to minimize

recall bias and maximize ecological validity [35]. In summary, self-assessment offers access to reference information from an individual's generative perspective, conceptually distinct from the perspective of an interlocutor and/or an external observer of the behavioral expression. While some modeling schemes can approximate them interchangeably in deriving informatics, BSP methods can, and should, conceptually distinguish in how these different reference perspectives are used in the modeling.

4) Privacy and Ethical Issues—Privacy and ethical issues in behavioral monitoring of human interactions is another aspect that is of central concern and consideration in BSP. For many existing corpora, the ethical imperative may require only that data access be limited to people authorized by an Institutional Review Board and only after appropriate training to handle human subjects data. Regardless, however, the underlying concerns span a range of legal issues at different levels of federal, state, and local laws and professional standards as well as ethical issues. We need to respect the autonomy of all participants in a behavioral study. We also need to respect the autonomy of the group to avoid perturbing the balance of group dynamics, except that in a very small number of cases precisely delineated by federal and local statutes, our Hippocratic reserve may be countermanded by a required intervention, e.g., to stop apparent child abuse or directly observable bullying. Margolin *et al.* [36] present an in-depth analysis of the privacy and ethical issues in observational studies of behavior with a focus on couple and family research, especially as those relate to computer-assisted research. An excellent summary of challenges and possible approaches to them can be found in a recent National Academies report on computer science research in healthcare informatics [37].

B. The Three Basic Technology Steps in BSP: Acquisition, Feature Engineering, and Modeling

Technology promises an effective way for collecting behavioral data and offers techniques and tools for their analysis and modeling in improving process efficiency and economy. More importantly, the translation of the research conclusions to the practitioners, especially challenging in the reliable and valid application of observational methods in everyday practice [30], can be facilitated by BSP. Speech and language are central to measuring, analyzing, tracking, and modeling human behavior.

1) Enriched Speech and Spoken Language Processing Is a Key BSP Ingredient—The human speech stream is a key information source for behavioral modeling since it offers critical information about not only communicative intent but also speaking style, language/dialect, identity, emotions, attitude, age, gender, and personality. Since such information in the speech signal resides at multiple levels, with complex interplay across these levels, speech understanding by the machine requires utilization of many diverse knowledge sources: spectral shape/intensity, type of articulation, pitch accent, duration, phones, syllabic features, words, part of speech, prosodic events, disfluencies, and linguistic constructs such as syntactic categories and discourse states. Recent advances include methods for robustly estimating voice activity detection, various aspects of speech prosody such as utterance boundary, pitch accent, lexical prominence/stress, speech rate, and emotions, all of which not only provide useful metalinguistic building blocks for enriching vocally expressed intent but can potentially directly inform expert-desired higher level behavior constructs.

2) BSP Acquisition Approaches—Multimodal sensing of human behavior has gained significant engineering interest in recent years, raising research challenges in fields such as signal processing [38], computer vision [39], robotics [40], speech recognition [41], and mobile sensing [42]. Since human behavior observations are desired in a variety of settings,

from constrained structured ones to unconstrained unstructured environments, a wide range of acquisition approaches have been proposed to suit the specific application needs. The sensing methodologies include instrumented environments as well as instrumented people that are being observed. On the one extreme, careful constrained laboratory studies of human behavior allow for sophisticated instrumentation of both the environment (e.g., employing arrays of cameras and microphones) and the people (e.g., use of multimodal wired psychophysiological sensors such as electrocardiography and electroencephalography). Examples of these settings include experimental smart rooms that have been used to monitor meetings [39], [43]–[45]. Real-world behavior observation environments, such as homes, schools, clinics, and urban settings, are often more constrained, limiting the nature of sensing possible due to both technical and human factors. For instance, in observing distressed couple interactions, the norm is for the communication to take place while the couple is seated and undisturbed by external influences; thus fairly rich instrumented environments equipped with arrays of microphones (and other sensors) for sensing vocal and gestural behavior can be employed [46]. On the other hand, a clinical environment for observing an interaction between a doctor and a patient tends to be more constrained, allowing only limited access to audio and video observation from far field, often under nonideal recording conditions. For example, in a study of a diagnostic interaction for autism between a child and a psychologist in a clinic environment, only far-field microphones and cameras mounted discreetly in the room could be employed [47]. Other environments such as homes and schools offer different challenges including robustness issues, collecting the right data at the right time, and contending with energy efficiency and data management issues. Many of these technical challenges are yet to be fully tackled.

People can be instrumented to measure certain other aspects of behavior, although these are not possible under all conditions (where their intrusive nature is a deterrent) and with all people (e.g., children, especially those with sensory issues such as in autism). Advances in body sensing allow for measuring movement (using microaccelerometers) and for psychophysiological (electrodermal activity, blood oxygenation, respiratory and cardiac activity) and neural measurements (electroencephalography). While some of these measurements can be deemed intrusive and are possible only in laboratory settings, rapid miniaturization of sensors, cheap storage, and wireless technologies allow some of these measurements to be possible in real-life conditions as they naturally occur. One such system that has been used in a wide variety of behavioral audio recordings in natural real-life conditions, such as parent–child recordings, is through chest-mounted (in clothing) audio recorders, e.g., the LENA audio system [48]. A similar system for video easily usable in the real world for behavioral observations is the first person vision device being developed at Carnegie Mellon University (CMU, Pittsburgh, PA) which allows for gaze estimation [49], an important piece of evidence for determining joint attention behavior. A number of sensing systems have been proposed and developed for measuring physical movement (using accelerometry) and physiological state. For example, the wrist-worn iCalm sensors proposed at the Massachusetts Institute of Technology (MIT, Cambridge) allow for monitoring and communicating autonomic response such as electrodermal activity (EDA) along with heart rate and heart rate variability as well as movement and posture changes through three-axis accelerometry [50]. A recent study [51] has used such sensor information in conjunction with audiovideo recording of autistic children interacting with an embodied conversational agent [52] to demonstrate a complex interplay between the observed cognitive behavior (in terms of verbal response latency) and the arousal state (in terms of EDA measures). Other work has used such body sensing and computing systems with mobile technologies to measure behavior during free living activities. For example, the KNOWME system [53] uses a suite of multimodal sensors [42] to concurrently measure physical activity using accelerometry and physiological activity (electrocardiography, pulse

oximetry) targeting metabolic health behavior for obesity applications [54], enabling appropriate interventions to encourage physical activity such as based on text messages [55]. In sum, the possibility of instrumenting people allows for behavior measurements that can support and complement audiovideo recordings of interactions in real-world conditions.

Finally, advances in embodied conversational agents [56] and virtual human technologies allow for yet another source of eliciting and observing behavior. A wide number of such systems have been used successfully to collect a range of behavioral data including from children expressing a range of natural dialog behaviors while engaged in cognitive (problem solving) activities [1], [5], [57], [58], including children with autism [52], and in public domains such as museums [59]. Virtual environments also allow for investigating affective behavior with psychophysiological evidence [60]. These have allowed for novel training tools such as training a clinician for interacting with a (virtual) distressed patient and a soldier for negotiating with civilians in challenging settings [61]. These systems strive to serve as an ancillary or surrogate to human-based interactions and allow for systematic elicitation and easier logging of behavior details.

3) BSP Feature Engineering and Modeling From Speech and Language—There are several levels of basic audio and speech signal processing that are carried out to extract the rich information from the recorded multimodal streams.

a) Diarization: The first step after multimodal sensing is the ability to diarize the audiovisual streams to automatically detect the participant's regions of activity (determining “who is talking when”). This by itself is a very significant contribution to domain experts allowing them, for instance, quantitative turn-taking analysis and easier transcription. There are several approaches to speaker diarization. They can be broadly separated by their use of the modalities: audio diarization, audiovisual diarization, and diarization through tracking.

In audio diarization, the effort is to distinguish among participants of an interaction by employing only their acoustic cues. A good review of this work can be found in [62]. Early efforts in audio diarization included broadcast news tasks that are characterized by mostly read speech, long speaking turns, and high-quality audio acquisition, most often from close-talking microphones. Subsequent efforts in conversational speech diarization, usually of multiperson meetings [43], [63], faced still further challenges. The spontaneous nature of meeting interactions results in faster turn taking and more overlaps, and is often accompanied by decreased audio quality due to far-field acquisition. BSP interactions exhibit all the complications of spontaneous interactions in meetings with the challenges compounded by the varied expressive and emotional nature of the interaction including behaviorally salient events that are expressed with subtle acoustic cues.

Audiovisual diarization is less common and is based on the principle that the audio and visual channels are correlated [64]. Due to the complexity of the signals we are dealing with in the BSP domain, the potential contribution of this technique is yet to be determined.

Diarization through tracking can employ both the acoustic and visual modalities in localizing the participants (with granularity constraints depending on the physical setup). This can happen through time-difference-of-arrival (TDOA) information from the speakers to the microphone array(s) and correlation of that information with the nearest detected participant according to the visual modality [65]–[69]. Depending on the BSP application domain, a combination of the above techniques can be adopted.

b) Transcription: Having established the speaking turns, automatic speech recognition (ASR) is the next challenge. Speech recognition research stretches back to the 1950s, and

there is a very large body of research and literature in the field (some of the state-of-the-art methods are represented by articles in this issue of the Proceedings of the IEEE). An excellent historical and scientific look at the field can be found in [70]. State-of-the-art speech recognizers work through examples of previously recorded spoken signals. These serve as a basis for training acoustic models and language models and assume a certain feature extraction technique. These three components are also the major challenges of ASR: robust features that can capture the lexical content but are as much as possible independent of acoustic conditions, speaker identity, emotional content, etc.; acoustic models that capture the acoustic variability in representation of lexically equivalent units; and language models that capture the entropy of language. Despite the great advances in ASR, automated transcripts in the challenging conditions of BSP scenarios, even with the best of recognizers, are far from accurate (we typically see error rates with a 30% WER at the lowest bound in tuned state-of-the-art systems). This is due to the formidable challenges brought forth by BSP domains in all three directions: far-field acquisition reduces the discriminative abilities of the acoustic features; speaking styles (emotional, disfluent, and technologically unaware participants) increase acoustic variability for the same content and even challenge the acquisition due to large dynamic ranges; and the high degree of emotion and spontaneity increase the entropy of the human expression. The traditional (one best) transcription task though may not be always necessary in the lexical annotation of these interactions. Probabilistic transcriptions such as those contained in an output lattice or a tophypotheses list can provide useful information despite the increased noise. Examples of using such ASR lattices for behavioral analysis have been shown to be promising even under high word error rates [71]. In sum, data from BSP domains offer a rich set of opportunities to advance robust ASR research and, in turn, offer a means for useful behavioral informatics.

c) Prosody features—Phrasing and prominence: There is a great deal of information in speech, both linguistic and paralinguistic, that is conveyed in terms of rhythm, intonation, and lexical stress, collectively referred to as prosody. These are typically characterized by one or more of the following: 1) intensity, duration, and fundamental frequency to impart emphasis to certain syllables or words; 2) timing cues, which refer to subtle variations in speech rate, length of syllable nuclei, insertion of pauses, and hesitations that serve to identify punctuation, syntactic boundaries, and separate linguistic and psycholinguistic "phrases" within utterances; and 3) modulation of intonation patterns that reflect different types of speech or dialog acts, as well as the speaker's intent and emotional state. For example, duration and intonation offer the listener a cue to the end of the phrase and can also signal effects such as continuation and questioning. Capturing these key details from the spoken interactions can further enrich and stratify the behavior analysis. A number of methods for capturing key, basic ingredients of prosody from speech acoustic features have been proposed, along with methods to enhance the modeling with a variety of other information sources (e.g., lexical, syntactic, and discourse information). Some highlights are provided below.

Robust pitch processing: Characterization of pitch ($f0$) excursions is an integral part of prosody modeling. There are numerous schemes for robustly estimating pitch broadly categorized as event based (that estimate the pitch period by locating the instant "event"—at which the glottis closes, such as using the wavelet transform [72], [73]) or nonevent based (which estimate pitch period by a direct approach such as the autocorrelation or cepstrum method, e.g., [74] and [75]). Given that information is encoded in the $f0$ signal along multiple time scales, the use of multiresolution signal processing schemes such as wavelet methods are attractive for deriving pitch features. Often an intermediate parametric $f0$ trajectory stylization is also done. This includes approaches such as piecewise linear stylization [76], polynomial-based stylization [77], and perception inspired approaches [78].

The pitch features, or their parametric functions, find wide use in behavioral analysis such as in the study of vocal entrainment in couples (Section III-A) and atypical prosody in autism (Section III-C).

Speech rate estimation: Speech rate variability carries information critical for speech understanding. When phonetic transcription is available (through manual or automatic means), speech rate is usually estimated as the number of linguistic units (phones, syllables, words, etc.) per unit time by first aligning the speech to the text (symbol) sequence, typically with an automatic speech recognizer [79]. Speech rate information can also be directly estimated from the speech signal (without requiring any automatic speech transcription). Techniques for such estimation include those based on spectral subband correlation [80] and those enhanced by including temporal correlation and the use of prominent spectral subbands for improving the signal correlation essential for syllable detection [81]. Speech rate modulations between the interlocutors, in addition to serving as useful features for classifying behavioral states, can shed light into dynamic interaction processes such as entrainment (see Section III-A3d).

Word prominence estimation: In natural conversation, speakers make some words and phrases more prominent than others. For instance, pitch accented words are perceptually more salient to the listener and are presumably employed at least in part to draw the listener's attention to informationally salient words; automatic determination of this information is possible [82]. Speech prominence can be detected based on various acoustic features such as spectral intensity, pitch, and speech rate that are directly extracted from speech without requiring explicit linguistic or phonetic knowledge [83]. These automatically derived acoustics-based measures can be especially useful in offering insights about behavioral processes.

Utterance boundary detection: Utterance boundary information is an essential first step in several behavioral analysis setups. For instance, in understanding the mutual influence of a clinician and a patient in a therapeutic interaction (see Section III for case study examples), considering utterance level analysis units is often found to provide meaningful insights. In addition to using lexical information to detect utterance boundaries [84], acoustic prosodic features can be used directly in utterance boundary detection as well [85], [86].

Combining linguistic information with acoustic features for prosody modeling: Speech acoustic features can be advantageously combined with lexical, syntactic, and discourse information for reliably characterizing prosody such as in terms of symbolic and parametric labeling standards like tones and break indices (ToBI): examples include decision tree methods [87], [88], maximum– likelihood (ML) classification [89], maximum *a posteriori* (MAP) classification [90], [91], and maximum entropy method [92]. For example, the maximum entropy discriminative model showed that the coupled model with both acoustic and syntactic information results in accuracies of 86.0% and 93.1% on pitch accent and boundary tone estimation with the Boston University Radio News Corpus (BU-RNC). These results are state of the art and are comparable to human performance on these tasks and can be used as a starting point to analyze the behavioral interactions in the domains illustrated in this paper. A major open computational challenge is achieving robustness of these techniques in handling the natural spontaneous speech of the target interaction settings that may be atypical, distressed, or reflect effects of other behavioral states.

d) Dialog act modeling: Dialog acts [93] are labels that are used to represent surface level communicative acts in a conversation or dialog. Several elements of behavior coding, including diagnostic instruments such as the autism diagnostic observation schedule

(ADOS) for autism (Section III-C) and therapeutic settings such as motivational interviewing (Section III-D), can benefit from quantitative dialog act characterization. State-of-the-art dialog act taggers typically combine the lexical and prosodic features in a hidden Markov model (HMM) framework with a Markovian discourse model [94]–[96] or use a discriminative framework such as the one that exploits lexical, syntactic, and prosodic cues in a maximum entropy framework [97], again with dialog history captured with an Nth-order Markov model. The performance of over 70% to detect a small set of the most frequent dialog acts in conversational discourse (switchboard corpus) offers a good starting point as an intermediate representation for downstream behavior analysis and modeling. For example, in exploring the efficacy and adherence to the motivational-interviewing-based behavioral therapy, dialog-act-based features offer a means to verify desirable therapy characteristics such as reflection by the therapist [98]; more details are offered in Section III-D.

e) Interaction modeling: A key aspect in understanding social and communicative behavior lies in illuminating the details of the interaction between the agents in the scene, for example, child–parent, patient–therapist, husband–wife, customer–provider, etc. Details of the dynamics of the social communicative and affective cue exchange are critical in creating the behavior map, notably in the study of various mental distress and wellbeing conditions. This includes investigation at several levels including turn taking (e.g., cooperative and disruptive interruption) and mutual coordination (aka entrainment, synchrony) in timing, lexical choices, intonation, and affective patterns.

Significant information regarding the interaction comes from tracking and modeling interlocutor dynamics. Information such as speaker activity and interruption patterns, utterance length duration, and pose of the participants can all inform the understanding of the underlying behavioral processes. For example, gaze can have an impact in the way an argument develops toward or away from recovery in a distressed couple interaction. Likewise, patterns of similarity in pitch and energy can point to more positively valenced affective dynamics [99] or improved dialog coordination and task success [100]. To model interlocutor dynamics, a broad range of techniques are needed to capture the diverse aspects of behavior displays such as the various verbal and nonverbal cues of social communication and affect, many of which can be extracted from audio as explained earlier. Modeling such tracked details resides both within and across turns of interaction and across the interlocutors. Computational dialog models, such as using Markov chains or Bayesian nets, have been used to capture this information and used for enhancing speech recognition [96] and dialog act modeling [97]; similar approaches can be useful in understanding behavior state dynamics.

f) Affect/attitude modeling and recognition from vocal behavior: Affective aspects can be detected from expressed vocal cues using a comprehensive multimodal approach. State-of-the-art methods typically combine phonemic, prosodic, lexical, and discourse features in a variety of ways for affect modeling from speech, such as done, for example, in the detection of user frustration in spontaneous call center interactions [10]. A comprehensive overview of the problem and approaches is given in [101] and a more recent overview of the state of the art based on the first Interspeech Emotion Challenge in [102]. Research has also shown how spoken language information can be combined with visual information for enhancing emotion recognition [103]–[106]. These methods together have been useful for detecting and classifying categorical representations of expressed affect and attitudes (e.g., happy, sad, angry, or domain-specific constructs such as frustration, engagement, politeness [5]). Since categorical descriptions of expressed emotions may not be appropriate in all modeling settings, yet another computational avenue has considered noncategorical (dimensional) representations. The popularly used ones are the three continuous-valued

emotion primitives: 1) valence, describing the negative versus positive nature of an emotion; 2) activation, describing the excitation on a scale from calm to excited; and 3) dominance, describing the appearance of the person on a scale from submissive or weak to dominant or strong. A variety of speech features have been mapped into these representations [107] using both generative and discriminative machine learning approaches.

Finally, to deal with real-life nonprototypical, often blended and ambiguous, displays of emotions, which are often not well described by a single semantic label [108], a new computational framework has been recently introduced. It quantifies emotional content through "emotion profiles" by providing multiple probabilistic class labels, rather than relying on just raw acoustic features or categorical hard labels [109]–[111]. Such a computational approach, illustrated in Fig. 4, is particularly suited for capturing and further stratifying in detail the heterogeneous, nonprototypical affective patterns expected in these complex behavioral analysis settings particularly when involving distressed or atypical interactions.

Features for emotion detection—Role of F0 and other speech features: The role of F0 in the encoding and perception of emotion information in speech has been widely studied [112]–[114]. The role of F0 in understanding higher level behavior processes in human interaction has also been exemplified in the study of vocal entrainment such as in distressed couple interactions [99]. Many studies have proposed the use of features derived from the fundamental frequency contour as one of the key information sources for automatic emotion recognition [115]. A common approach is to extract as many derived features as possible and then use feature selection techniques to find a reduced subset that maximizes the performance [116].Some of the most common selected F0 features are the mean, range, minimum, maximum, and standard deviation statistics from the *F*0 utterance contour [117].

*F*0 patterns at different segmental and linguistic levels (e.g., phoneme, word, part of speech) offer distinct insights into expressed emotions [10], [118], [119]. For example, it was observed that *F*0 mean significantly differs for angry, happy, sad, and neutral speech and across different vowels. Bänziger and Scherer [120] have suggested that the fundamental frequency is mainly affected by the arousal level of the utterance. They analyzed changes in the *F*0 contour in terms of the degree of activation in the sentences. They also analyzed the change of the *F*0 contour in terms of emotional categories. However, they did not find evidence for the qualitative changes in the *F*0 contour among different emotions. These observations are consistent with the confusion between emotional categories often observed in emotion recognition experiments between certain categories such as between happiness and anger and between sadness and neutral state, which have similar activation levels but differ in the valence domain when only pitch features are used [119]. As suggested by Ladd *et al.* [121], voice quality features may provide information to discriminate in the valence domain.

In addition to *F*0-based features, a comprehensive set of speech-based features [the so-called low-level descriptors (LLDs)] are now commonly used in emotion research, as summarized in [116] and made available in toolkits such as openSMILE [122]. These include prosodic LLDs such as voice activity, speaking rate, and intensity features, spectrum-based LLDs such as Mel-frequency cepstral coefficients (MFCCs) and log Mel-frequency bands (MFBs), and voice quality LLDs such as jitter and shimmer. Modeling and recognition experiments often adopt some form of feature selection and feature reduction to obtain the most useful set of features from this full range of generated features.

Role of spoken language features: The role of emotional information carried by words has been well established and codified in terms of resources such as the dictionary of affect

[123] and the affective norms for English words (ANEW) [124] that map words into affective ratings (activation, valence, dominance). These manually created resources have been at the heart of domain-independent approaches to affective modeling, but they fail to provide good coverage for computing emotions in real-life spoken language data applications. Approaches to address this issue use computational methods to expand an already existing lexicon or create a new (often, domain-adapted) lexicon. Malandrakis *et al.* [125] recently proposed an approach to creation of an affective lexicon where, starting from a small set of manually annotated seed words, continuous valence ratings for new words are estimated using semantic similarity scores and a kernel model. Word level scores are combined to produce sentence-level scores via simple linear and nonlinear fusion. Evaluation on spoken language transcripts in estimating behaviors such as politeness and frustration was found to be promising.

Domain-dependent methods, on the other hand, attempt to learn the affective relevance of words and word sequences in the context of a specific application or interaction setting. Lee *et al.* [126], [127] proposed the notion of emotional salience, i.e., mutual information between a specific word and an emotion class, to identify emotional words in a speech utterance for detecting negative emotion [10]. By adding language information to the acoustic features, they reported a relative improvement of 46% in a valence detection task on call center spoken dialogs. This idea has since been extended to calculate the mutual information between word pairs and emotion classes as well as the use of latent semantic analysis (LSA)-based feature extraction to obtain lexical information for emotion recognition. Encouraging results in detecting politeness and frustration in a conversational dialog task have been reported [3], [5].

Discourse context information can also be used to predict user emotions. Several researchers have attempted to include discourse related information to improve emotion classification. Discourse categories, rejection, repetition, rephrase, ask-start over, and none of the above, were used in [10] to improve their negative/nonnegative emotion classification task. Liscombe *et al.* [128] have considered, in addition to discourse context, features related to changes in prosodic and lexical features between current user turn and previous user turn. Best results in terms of emotion classification accuracy were obtained when they combined prosodic, lexical, and contextual information. In sum, these studies point to the benefit of combining language and speech features within BSP.

Emotion recognition methods: A number of modeling approaches have been proposed depending on the representations being computed. For categorical recognition, typically at the utterance level, both generative methods [such as through HMMs or just Gaussian mixture models (GMMs)] and discriminative approaches [such as support vector machines (SVMs) and logistic regression (LR) schemes] have been found to be useful. Likewise, estimation of continuous-valued dimensional estimates such as activation and valence [107] have employed methods such as rule-based fuzzy logic methods and support vector regression.

Several variants and extensions to these methods exist, both in improving the richness and robustness of representations and their computation. For example, results have consistently shown that combining multiple classifiers can be advantageous [102]. Others have addressed the problem of dealing with real-life expressions of emotions, which can be ambiguous and blended. Lee *et al.* [129], [130] introduced a hierarchical computational structure that maps an input speech utterance into one of the multiple emotion classes through subsequent layers of binary classifications. The key idea is that the levels in the tree are designed to solve the easiest classification tasks first, allowing the mitigation of error propagation. Their results, which yielded the best results in the classification of the 2009 Emotion challenge on

classification, were found to be effective across multiple databases. One of the early steps in such a hierarchy could be the basic discrimination between emotional and nonemotional speech, based on the assumption that emotional speech productions are variants of the nonemotional counterparts in the (measurable) feature space. Busso *et al.* [131] showed that instead of training individual emotional models, building a single, neutral speech model and using it for emotion evaluation either in the categorical approach or in the dimensional approach can be advantageous. This method benefits from the robust modeling of neutral, nonemotional speech due to the availability of large amounts of such speech data.

Another emerging computational trend is a set of advances in modeling the dynamics of emotions [111], [132]– [135]. An individual's affective state variables can be seen as continuously evolving over time during an interaction, manifested, at least partially, in expressive communication cues characterized by the continuous interplay of speech with other multimodal information, such as facial and bodily gestures. Wollmer *et al.* [132], [133] report continuous estimation of valence and activation values from emotional speech using support vector regression and long–short term memory (LSTM) for regression. Metallinou *et al.* [135] report a supervised, GMM-based methodology to continuously track an underlying emotional state using body language (detailed and intuitive descriptions of each participant's body movements, posture, and behavior toward his interlocutor) and speech information. Promising results were reported for tracking trends of participants' activation and dominance values, which outperform other regression-based approaches used in the literature. The method also offers a way to shed light into the way expressive body language is modulated by underlying emotional states in the context of dyadic interactions. While all these efforts are promising, open questions and challenges remain in terms of effectively annotating and evaluating continuous changes in emotions, as well as capturing context effectively, in improving the modeling.

In summary, the advances in modeling and tracking emotions expressed in speech and spoken language (and associated nonverbal cues) are foundational for capturing and further stratifying in detail the heterogeneous, nonprototypical affective patterns expected in the complex behavioral analysis settings, particularly when involving distressed or atypical interactions.

g) Speaker traits: Identity, age, gender, personality: Many aspects of speakers' traits such as their identity, age, gender and personality can be gleaned from their speech and language. The state of the art in this domain is well reviewed in a recent article by Schuller *et al.* [136]. These problems are typically posed as a pattern recognition problem using a range of speech segmental and prosodic features, voice quality features, and language information, much akin to what was detailed in the case of emotion recognition.

III. Bsp from Speech and Language Information: Case Studies

In this section, we describe several case studies, each exemplifying a specific type of spoken interaction and with distinct goals for behavioral analysis. One of the critical aspects of behavioral analysis is the integral role played by human experts in how the behaviors of interest are elicited, analyzed, and used. There is a wide range of interaction types and associated application possibilities along these dimensions. Hence, BSP techniques need to be cognizant of, and incorporate, these diverse types of interactions, each with distinct characteristics, purpose, and types of agents that are involved. To illustrate this diversity in the behavioral analysis scenarios, we describe specific case studies that help to highlight both the commonalities as well as the differences in the computational approaches. In particular, in this paper we focus on speech and spoken language as the key interaction modality and informatics source.

Since in BSP the role of the human expert is critical, we categorize interactions based on whether human experts are actively involved in the behavioral elicitation or just observing either being copresent or not, whether they are performing explicit (online) behavior monitoring and analysis, and whether they are also involved in any active behavior modification (such as in therapeutic settings). In scenarios where experts are not copresent during the interaction, they can perform online (e.g., observing a focus group from behind the scenes) or offline behavior analysis (e.g., analyzing recordings).

The humans being observed can be individuals, dyads, small groups (typically defined by social scientists to be a group large enough to sustain more than one simultaneous conversation), or large groups (too large to be united in a single conversation). The examples chosen for this paper rely on speech and spoken language as the primary behavioral cue and are restricted to case studies with focus on individual level behaviors. Table 1 summarizes the applications considered in this paper in the context of BSP.

A. Example Case Study: Behavior Modeling of Distressed Couples

Herein, we give an overview of a computational study of couples therapy, where distressed couple behavior is analyzed.

Understanding the behavior of the individual spouse and the dynamics of the dyad is a critical step toward aiding distressed couples. The state of the art in couple and family therapy—based on human expert monitoring and intervention—can be expensive and time consuming.Additionally, it suffers from reliability issues and the fundamental limitations of humans in being fully cognizant of multistream, multirate observational information. Finally, it is scalable only to the degree of available experts at the time. BSP tools can support the experts in this process.

The joint human–machine observational and interpretative process in dyadic interactions, as depicted in Fig. 2, promises tremendous benefits to the experts and the patients. Integral to interpretation of such information in emotionally rich, multiperson interactions is the development of behavior-centric computational models that encompass the social, affective, and communicative state of the interlocutors. Such capability can augment the relevant information presented to the experts, strengthening their ability to take appropriate action and to intervene appropriately. BSP's goal is to enable such capability.

1) Behavioral Analysis by Experts in Psychology Research and Practice—

Psychology practitioners and experts observe specific low-level behaviors as indicators of mid- and high-level behaviors such as approach avoidance or mutual blame. For example, experts consider the partners' arousal level, which is reflected in speech (e.g., fundamental frequency and energy [137]), overall body movement [138], and, when available, physiological metrics (e.g., heart rate, galvanic skin response [139]). Likewise, the practitioners carefully analyze the valence level of the partners by observing their facial expressions [140], [141] and the interactional verbal content [142], [143]. Behavioral dominance is also studied, which is frequently broken down into two different aspects: 1) power processes; and 2) power outcomes [144]. Power processes involve what people do to get their way and are usually studied with linguistics, both in terms of verbal content (hedges as low power [145], [146]) and verbal behavior (successful interruptions as high power but not including backchannel interruptions [147]). Power outcomes have to do with attaining what you want and are studied by assessing the degree to which you can get your spouse to become like yourself or to approach a position you are advocating for the spouse, which could either take the form of changing the spouse's mind or getting the spouse to behave as you want. Other important low-level behaviors are approach tendencies (e.g., touch frequency and duration, facial and body orientation, forward lean, frequency and

duration of eye contact, and frequency of nonnervous gestures and backchannel nods) and avoidance tendencies (e.g., changing the topic, disagreement, questioning the other spouse's reasoning) [148]–[150]. These low-level behaviors, as a whole, point to the need for a multimodal approach toward addressing this behavioral monitoring and understanding from a signal processing point of view.

2) A Longitudinal Couples Therapy Case Study—Psychologists depend critically on perceptual judgments made by themselves or other experts to provide appropriate communicative strategy suggestions. To better understand distressed married couples, data were collected as part of a longitudinal study at the University of California, Los Angeles and the University of Washington, Seattle. The resulting corpus consists of audiovideo recordings of couples (wife and husband) during real problem-solving dyadic interactions. For this study [151], 134 seriously and chronically distressed married couples received couples therapy for one year. Participants ranged from 22 to 72 years old, with a median age of 43 years ($SD = 8.8$) for men and a median age of 42 years ($SD = 8.7$) for women. They were, on average, college educated (median level of education for both men and women was 17 years, $SD = 3.2$). The sample was largely Caucasian (77%), with 8% African American, 5% Asian or Pacific Islander, 5% Latino/Latina, 1% Native American, and 4% other. Couples were married an average 10.0 years ($SD = 7.7$). As part of the study, the couples participated in sessions where they discussed a problem in their relationship with no therapist or research staff present. The couple talked for ten minutes about the wife's chosen topic and ten minutes about the husband's chosen topic; these sessions were analyzed separately.

The data have been richly coded by psychology experts with two coding manuals that were designed specifically to capture relevant high-level behaviors during problem-solving couples' interactions: the social support interaction rating system (SSIRS) [152] and the couples interaction rating system 2 (CIRS2) [153]; in total, there are 33 behavioral codes. Both coding manuals were designed to have evaluators watch the entire session, and provide session-level ratings of *each* spouse's overall behavior on an integer scale from 1 to 9; utterance- and turn-level ratings were not obtained. Three or four student evaluators coded each session, producing one set of 33 codes for each spouse. All evaluators underwent a training period to give them a sense for what was “typical” behavior and to help standardize the coding process. The coding is a laborious, expensive, time consuming, and subjective process. Studies such as this are not easily scalable without automated ways of both preparing the data and gauging these metrics. Some of the challenges posed by the manual coding methods include subjective annotation, summative assessment, unknown degree of correlation of the distal and proximal expressions, and what modalities those manifest themselves into.

3) Signal Processing Analysis—For our BSP analysis, we use the couples data from three recordings: before the therapy sessions began, 26 weeks into therapy, and two years after the therapy sessions finished. In total, we have 96 h of data across 574 sessions and we employed 372 of these based on the quality of automated segmentation of the audio.

The audiovideo data consist of a split-screen video (analog NTSC) and a single channel of far-field audio. Since the data were originally only intended for manual coding, the recording conditions were not ideal for automatic analysis; the video angles, microphone placement, and background noise varied across couples and across time periods. We also have access to word transcriptions, in which the speaker was labeled as well (husband/wife). The transcripts lack detailed annotations such as timing and speech overlap indications.

We illustrate three simple speech and language processing applications of behavioral coding using this corpus. For that purpose, we focus mostly on classifying a single behavioral code, namely the level of "blame"[1] expressed from one spouse to the other (the distribution of this and a few other codes of interest for this domain and data are in Fig. 5). Based on this goal, we partitioned the data into two classes: *high* blame and low blame. The high-blame partition consisted of 70 sessions (approximately 20% of the 372 sessions) with the highest average blame score for the wife and 70 sessions with the highest average blame score for the husband. The low-blame partitions consisted of 140 sessions with the lowest average blame score: 70 for the wife and 70 for the husband. The blame scores for the two classes ranged from 1.0 to 1.5 for low blame and from 5.0 to 9.0 for high blame, so they were separable to the human evaluators, as shown at the top row of Fig. 5. The experiments are based on a leave-one-out cross validation.

a) Acoustic classification: Vocal cues have been shown to be informationally relevant in the context of marital interactions [154], [155]. To capture the vocal cues we use a range of acoustic low-level descriptors (LLDs) extracted across each session, some of which were described in the earlier section on feature engineering (Section II-B3). For further details of the LLDs, we refer the readers to [156], [157] and related work in emotion recognition [10], [107], [116], as well as our past voice activity detection (VAD) work [158]. Based on the LLDs, we employed an overgenerative approach to produce session-wide acoustic features based on a range of functionals of the LLDs (Table 2) across five different signal scenarios and at six different temporal granularities. The five different signal scenarios are: whole session with rated partner (two cases), other partner (two cases), and both partners (one case).

The six temporal granularities included one set of global features, in which functionals were computed across the entire session (for each LLD and speaker region), and five sets of hierarchical features based on [159]. The five hierarchical feature sets were computed by first splitting the LLD/speaker region into disjoint windows of durations: 0.1, 0.5, 1, 5, and 10 s. We then computed the 14 functionals listed in Table 2 for each of these windows, producing 14 vectors of functional values for the entire session. Finally, we generated the hierarchical features by computing the six "basic" functionals (Table 2) across each of these vectors. Because of the windowing technique used, these hierarchical features should capture some of the moment-to-moment changes that occur within the interaction.

After removing static features with zero standard deviation, there were about 53 100 features at each cross-validation fold. For the straightforward acoustic classification task, we employed an SVM using LIBSVM [160]. Since there were orders of magnitude more features (50 000+) than instances (280), we used a linear kernel. All features were z-normalized by subtracting the mean value in the training data and dividing by the standard deviation. Results (Table 3) show that critical information about higher level behavior can be discerned from acoustic vocal features.

b) Lexical classification: Lexical content of the interaction is yet another source of information about the couples' behavior [162], [163] and codified in the SSIRS and CIRS2 rating systems. From a natural language processing view, one of the simplest lexical classifiers is that derived from the ML criterion. However, given the noisy audio of spontaneous speech, the automatic transcription can be extremely challenging (with unadapted models, typical WER > 80%). In addition, the training set available in this case study for an ML classifier is small and hence significant data smoothing is needed. This can

[1] A fictitious and exaggerated example of blame would be "it's your fault I cannot sleep at night."

be achieved through the use of a universal background model (UBM), here denoted by B, where the probability of a transcript T given a class Ci

$$p(T|C_i)=\prod_{\text{all } j}[(1-\lambda)p(w_j|C_i)+\lambda p(w_j|B)] \quad (2)$$

where w_i is the ith word in the transcript $T=,w_1\ldots w_l$ and λ is a smoothing weight. The equations above are given in their unigram format for simplicity. Results in Tables 4 and 5 show that this simple interpolation method aids in achieving significant accuracy.

To contend with the noisy automatically derived transcripts, one approach is to operate on the ASR lattices, as illustrated in Fig. 6. The process there is to train the system as previously but to score lattices rather than one-best transcriptions. Practically, the process can be done through two competing two-pass ASR decoders where the first pass (pruning) is done with the same language model and the second pass is done with the two competing language models.

Results in Table 5 show that despite the high WER the system can still result in promising accuracy, while as we can see from Table 4, there are significant gains to be obtained from using an optimized ASR that can result in a more accurate lattice generation.

Lexical significance: Importantly, the language modeling, in collaboration with the domain experts, allows us to examine whether specific words offer insights into specific behavioral codes and whether those are couple specific or generalize broadly. For instance, Table 6 shows that specific words can carry notable insights toward the behavioral codes. For example, the word YOU, which appeared 59 times, had the most contribution toward characterizing “blame” while the filler word UM (23 times) scored as the least blaming unigram. Such quantitative analysis enables us to flag and identify important terms that can be followed up with further detailed experimental and psychological inquiry.

c) Fusion of acoustic and lexical classifiers: The acoustic and lexical information streams, although significantly dependent on each other, also convey complementary information to human observers. This has also been shown in a range of domains such as emotion recognition [10] and with a range of complementary modalities such as in interaction dynamics [45], [69]. In our BSP example, based on the different temporal feature rates of the two classifiers, we decided to fuse the two information streams at the classifier-score level due to the simplicity of such a fusion system.

We again used LIBSVM's SVM for the fusion classifier and z-normalized the fusion features, so they were on a comparable scale. Three pairs of classifier combinations are considered: fusing the static acoustic and ASR-derived lexical classifiers, fusing the static acoustic and oracle lexical classifiers, and fusing the two lexical classifiers; the fusion results confirmed the benefit of combining these information streams. More details can be found in [156].

d) Classification based on user dynamics: In addition to using individual-specific features for quantifying the behavior of each person, features capturing the mutual influence of the interacting dyad can be considered for behavior modeling. This includes measures such as *entrainment* and interaction synchrony [164] between the interlocutors to analyze the quality of their interaction. An excellent summary of entrainment in spoken language communication was presented by Hirshberg [165]. Several previous studies have considered entrainment by using various behavioral cues; for example, McGarva investigated the

mutual entrainment in vocal activity rhymes [166], Nenkova analyzed the high-frequency word usage entrainment [167], Richardson analyzed the entrainment of body movements [168], Pardo showed the phonetic convergence in conversation settings [169], and Edlund *et al.* [170] investigated the correlation of gap and pause durations in interactions. In the context of modeling couple behavioral dynamics [99], we have shown that we can analyze the prosodic entrainment phenomenon and investigate the role of these measures in describing behavioral aspects where they have been implicated in theoretical models in psychology, such as the overall husband's or wife's *positive* or *negative* affect/attitude during interactions.

This work is based on stylizing pitch in a piecewise linear manner ($Y = \Delta X + B$) computed every 100 ms with 50-ms overlap. The slope (Δ) and intercept (B) were calculated using the method of least squares. Hence, we have two parameters for every 100 ms to describe the evolution of prosodic cues instead of 100 raw values. The results illustrated focus only on the slope, which encodes information describing the intonation and the rise or fall of energy values. We hypothesize that changes of pitch slope values across time capture aspects of speaking style, and the covariation of this parameter between speakers can help us identify prosodic entrainment.

Stylizing the pitch and energy contours generated two 1-D feature vectors consisting of several frames of Δs for pitch and energy, respectively, at every automatically aligned speaker turn. Prosodic entrainment was then computed based on three main methods: square of Pearson correlation, mutual information, and mean of spectral coherence across turn on this sequence of Δs to estimate the level of synchronization. Results showed that these measures can explain positive versus negative interaction behavior differences [99]. Using a Markov chain model (similar to an n-gram model) built using quantized values of the Δ coefficients, significant accuracy was demonstrated in the classification task (Table 7). More recent work that has focused on devising direct similarity measures between the vocal feature spaces using principal component analyses has demonstrated further robust results in characterizing the latent vocal behavior similarity and its application in prediction of higher level behavior [171], [172] and in informing outcome-related behaviors [173].

This section highlighted the potential use of speech and language processing techniques in understanding distressed dyadic interactions using acoustic, lexical, and cross-interlocutor dynamics information. Considerable future work is needed in developing new algorithmic techniques for joint modeling of speech with visual and physiological information, understanding salient aspects of feature streams that contribute to specific behavior perception, and developing tools that can jointly leverage human and machine expertise.

B. Example Case Study: Literacy Assessment

Computational speech processing tools offer promising ways for assisting teachers in assessing components of emerging literacy skills and reading skills [174] and in offering innovative automated tutoring support [175], [176]. While much work in this area has targeted specific important (sub)components such as detecting mispronunciations or computing other important cues such as speech rate and emotions, research has shown that these by themselves do not provide a complete picture, i.e., as an expert teacher would desire and implement. To compute such high-level “holistic” assessments that offer interpretive richness, a variety of behavioral and contextual cues used by teachers needs to be gathered and analyzed [177].

What makes the problem interesting and challenging are the multiple sources of variability beyond the cognitive learning differences, such as due to variability in language and socioeconomic backgrounds of the children as well as the expertise levels and backgrounds

of the teachers themselves. BSP offers a way for computing objective features from the observed task performance (e.g., children reading aloud) and creating predictive models that can effectively capture how a set of teachers would evaluate given such data. This example scenario where BSP is used to compute expert judgments of behavior without actual active interaction with the subjects being modeled comes from the Technology-Based Assessment of Language and Literacy (TBALL) project focused on assessing the English literacy skills of young children in early education from multilingual backgrounds [174].

To acquire ecologically valid speech data, assessments were obtained from children from native English- and Spanish-speaking households in actual elementary schools in California using a close-talking microphone [178]. To facilitate consistent and robust speech data acquisition by taking advantage of the structured nature of the activity, behavior elicitation in this work used a child–computer interface to implement age-appropriate reading tasks used to test children in kindergarten to second grade. Since one of the key aspects of BSP is to emulate human expert processing, the study employed diverse human evaluators who rated the children on their overall reading ability based on the audio recordings [179]. Speech signal features inspired by, and correlated with, cues human evaluators stated they used were derived: measures related to pronunciation correctness, speaking rate, and fluency (and disfluency). The pronunciation correctness scores were based on two common pronunciation verification methods: 1) forced alignment with a dictionary of acceptable and foreseeable unacceptable pronunciations of the target word; and 2) goodness of pronunciation (GOP) scoring. The fluency scores were based on constrained ASR using disfluency-specialized grammars, which were designed to detect partial word instantiations of the target word. Finally, the speaking rate scores were based on forced alignment and captured relevant timing information, such as the speech start time (relative to when the word was first displayed on the monitor) and the average speaking rate in units of syllables/s and phones/s.

Simple linear regression techniques with these task-related features automatically predicted individual evaluators' high-level behavior scores with a mean Pearson correlation coefficient of 0.828 and average evaluator's scores with correlation 0.946, both exceeding the mean interevaluator agreement statistics [179]. These results are further improved (correlation coefficient of 0.952) by exploiting actual patterns of evaluation. It was observed that the evaluators' level of agreement significantly varies, depending on the child being judged (Fig. 7): the children for whom the evaluators were more confident in rating are weighted higher in the model [180].

To follow up on the hypothesis that the overall assessment would need to incorporate cues beyond those that are task related, further modeling focused on creating a Bayesian network student model to automate reading assessments of a diverse group of children just as a conscientious teacher would, incorporating cues not just based on expert knowledge of pronunciation variants and their cognitive or phonological sources but also prior knowledge of the student and the test itself. In the spirit of BSP (Fig. 2) human-in-the-loop processing, the model incorporates expert knowledge in terms of a hypothesized structure of conditional dependencies and automation to refine the Bayes net in eliminating unnecessary arcs. This model with a more comprehensive accounting of novel behavioral features was shown not only to outperform a task-only-based (pronunciation scoring) scheme but it also exhibits the same biases along demographic lines as human listeners (e.g., gender effects). Results also showed that the computational algorithm did not choose to exclude very many of the hypothesized (human-knowledge-based) arcs to improve predictive performance. BSP in this case study underscores the promise of emulating expert-like behavior in complex task settings involving subjective judgments and offering computational methods with interpretive capability.

BSP can also enable the design of computer systems that can adapt learning interventions (tutoring) based on a learner's perceived state [181]. To achieve these capabilities, BSP advances for computing learning-specific behaviors is essential [182]. Promising progress has been made recently. Yildirim *et al.* [5] have shown that states such as "frustration" and "politeness" can be automatically discerned from a child's speech cues elicited in spontaneous dialog interactions with computer characters [57]. Their experimental results showed that lexical information has more discriminative power than acoustic and contextual cues for detection of politeness, whereas context and acoustic features perform best for frustration detection. Furthermore, the fusion of acoustic, lexical, and contextual information provided significantly better classification results than using any single information source. In other work, Litman and Forbes-Riley [176] showed that the use of speech and language features for predicting student emotions in human–computer tutoring dialogs improved the accuracy of the system. Likewise, Zhang *et al.* [4] have reported promising results in the combined use of acoustic, spectral, and language information for detecting confidence, puzzlement, and hesitation in their child– machine dialog task. More recent work has shown that such speech and language information can be advantageously combined with visual gestures of interaction such as the movement of the head and facial expressions [1]. Such ability to map higher level behavior from speech and spoken language cues opens up new conversational interface design for educational applications [183].

C. Example Case Study: Behavior Modeling in Autism Diagnosis

Interaction-based behavioral diagnostic settings offer an important venue for BSP. In particular, in these scenarios, the expert performing the diagnosis is often also engaged in eliciting the behavior of interest. BSP hence considers not only just mapping the observed behavior of the target individual into categories desired (and deemed useful) by the expert but also understanding the experts' strategies for the elicitation of desirable patient behavior in the diagnostic interaction process. Here we use the domain of autism to illustrate some of these BSP dimensions and possibilities. Emerging research in this realm includes [47], [184], and [185].

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by a triad of core deficits, including impaired social behaviors, communication, and restricted/ repetitive behaviors [186]. Recent studies indicate that as many as 1 in 110 children are diagnosed with ASD [187]. ASD is considered a "spectrum" disorder because symptom severity in each of the core domains can vary widely. Studies have shown that early diagnosis and intensive early intervention can lead to improved social and communication skills in autistic children [188]. The ADOS is one of the most widely used clinical research instruments for the assessment and diagnosis of ASD and is suitable for individuals of varying ages and verbal abilities [11], [189]. The semistructured 30–60-min interaction provides a trained psychologist with behavioral evidence that can be evaluated along dimensions important in diagnosing autism, e.g., atypical prosody, social interaction, narrative construction. The psychologist is both an active interlocutor who is engaged in eliciting specific responses from the child following the underlying protocol and a judge who simultaneously evaluates the child in accordance with the diagnostic criteria. Not only are BSP techniques useful in stratifying behavioral details of potential use to the experts beyond the categorical ratings that the expert gives, but they can also shed details on the nature of the interaction between the expert examiner and the child, potentially to inform what led to specific diagnoses and specific areas and strategies for intervention.

BSP can assist with this process in a number of ways. On the acquisition side, audiovideo sensors can capture details of the child–clinician interaction in a consistent fashion, while analyses and modeling can quantify behavior patterns therein, including illuminating difficult to observe fine details such as dynamic variability in durations, intonation, voice

quality, instances and types of nonverbal cues, etc. Critically, BSP also allows modeling interaction dynamics and its role in the diagnostic process, as well as in creating analytical capabilities for looking at fine patterning across communication modalities and over time.

The case study considered here illustrating the potential of BSP in autism is based on a corpus of real, spontaneous child–psychologist interactions recorded in a clinical environment in the context of administering ADOS. The collection of the USC CARE ADOS corpus [47] in itself highlights behavioral data gathering in a clinical space, using a portable smart-room solution with multiple far-field audiovideo sensors to unobtrusively record the interaction and to maximize the ecological validity of the experiments. Unlike laboratory environments that can be instrumented more flexibly, sensing behavior in a multiuse clinical space has to consider several factors for maintaining ecological validity by minimizing intrusions and potential distractions to the subjects. For example, only far-field microphones and discreetly mounted wall and ceiling cameras were used. In addition to audio-video recording of the interactions, the data comprise behavioral codes assigned by the administering psychologists as well as the final ADOS diagnosis outcome.

BSP also offers several analytical possibilities. As an example, consider one of the key behavior characterizations in ADOS: atypical prosody, which includes any of the following: slow, rapid, jerky and irregular in rhythm, odd intonation or inappropriate pitch and stress, markedly flat and toneless, or consistently abnormal volume [11]. This is codified on an integer scale from 0 to 2 with 0 designating appropriate prosody, 1 signifying slight deviations from typicality, and 2 used to report “clearly abnormal prosody.” While there is ample documentation of the presence of atypical and impaired prosody in ASD, a precise characterization of the specific details of the nature of the impairment is still lacking [190]. Furthermore, what contextual factors, if any, can best explain the observed patterns is still unclear. A more stratified and objective analysis of the speech properties can help toward a better understanding of the nature of the prosodic deficits, a possibility afforded by BSP: quantitative analysis of the speech signal can shed light on how prosody is expressed and processed by the interlocutors and also model the mutual influence of the interlocutors on prosody.

Speech acoustic features that have been implicated in characterizing atypical prosody include pitch slope, breathiness, and nasality [191], [192], although very few studies have analyzed spontaneous natural speech samples. Using the spontaneous spoken interactions of the USC CARE ADOS corpus, Bone *et al.* [184] have attempted a preliminary computational analysis of prosody in natural interactions. For increasing a perceived typicality, children's prosodic features that suggest “monotonic” speech included variable volume, atypical voice quality, and slower rate of speech, and all correlated with overall diagnostic ratings. More interestingly, results showed that the interacting psychologist's speech features reflect their perception of a child's atypical behavior, e.g., the psychologist's pitch slope and jitter are increasingly variable and their speech rate generally decreases. Importantly, models demonstrated that the psychologist's own speech cues successfully predicted their ratings, suggesting their attunement to the child's behavioral cues. This finding underscores the capability afforded by BSP to offer new insights to the experts.

BSP technologies also offer novel ways for behavior elicitation such as through the use of embodied conversational agents (ECAs) [52]. Mower *et al.* [193] showed that interactions involving children with autism and their parent involving an ECA partner was natural, and elicited communication data similar to that in interactions just between the child and his parent. Such data have enabled further studies relating speech patterns to predicting social cues such as laughter [194]. One of the key aspects of social interaction in autism is the decreased expression of “shared enjoyment.” Analysis of laughter offers an objective insight

into this behavior; it is also known that voiced laughter conveys positive affect while unvoiced laughter functions as a social facilitator. Analysis of speech areas proximal to laughter ("social zones"), 2–10 s preceding and following the laughter event (Fig. 8), showed that acoustic patterns of social zones are indicative of a type of shared enjoyment. Results also showed how BSP afforded a way to quantify the difference in the ways children engaged in social interactions.

D. Example Case Study: Behavior Modeling in Psychotherapy for Addiction

In this section, we describe a case study of psychotherapy to illustrate BSP possibilities. Specifically, we consider motivational interviewing (MI), a goal-oriented psychotherapy, employed in cases such as addiction and overweight issues, which helps patients explore and resolve their ambivalence about the problem at hand in a dialog setting [12]. MI focuses on eliciting and enhancing the intrinsic motivation for change by exploring and resolving client ambivalence through a dyadic spoken dialog between the patient and the therapist. These interactions represent a case where the therapist follows a specific protocol for actively eliciting and influencing specific behavior changes in the patient. Modeling this mutual influence and identifying successful therapy strategies is a challenging BSP opportunity.

The state of the art uses often tedious manual behavioral coding to assess a patient's behavior patterns as well as the counselor's therapy proficiency, as specified by standard methods such as motivational interview skills coding (MISC) [195]. At the most basic level, speech and language processing can facilitate automated turn and utterance segmentation as well as diarization and transcription of the dialog, all key ingredients of the MISC coding since behavioral codes are assigned at the utterance level (and also at the overall session level). The next level of possible BSP is to derive the mapping of desired utterance behavior codes from speech and language cues. For example, on the therapist side, some of the desired codes include marking questions (and whether those are open or closed), affirmation, facilitation, and others, totaling 20 overall. Likewise on the client side, the five desired utterance-level codes include reason (subcodes: desire, ability, need), taking steps, other, commitment, and follow/neutral. This step of mapping is similar to dialog act tagging described earlier (Section II-B3d) and can be accomplished by invoking lexical, acoustic-prosodic, syntactic, and discourse information following the state of the art in dialog act tagging [97]. In addition, these codes are enriched with markers that signal "target behavior change" either toward or away from the therapeutic goals, with a positive (+) or negative (—) valence, respectively. Spoken language representing tendency toward making a change, the so-called "change talk," is considered to lead to positive outcomes. The learning of these specific enrichment labels is guided by the domain expert, and automating the process represents an apt example of the human-in-the-loop learning aspect of BSP.

Beyond assisting in behavioral coding, BSP can also offer novel tools for discovery. For example, one of the desired aspects of any therapy, a human-centered and human-based process is in gauging the quality and efficacy of the therapy. This can illuminate why, when, and how certain strategies of therapy lead to successful outcomes, and how in turn these can be used to inform training and effective implementation [196]. In a recent case study, Can *et al.* [197] have examined a BSP approach to assessing the quality of MI on a particular aspect of the counselor behavior, "reflections," believed to be a critical indicator of therapy quality. Reflections are nontrivial therapist behaviors, and are supposed to "capture and return" to the patient something that they had said. This process is, however, challenging because it requires modeling and using complex contextual knowledge that the expert invokes. Computing reflections is further complicated by the fact that, even though therapists may use prototypical surface linguistic forms (well captured by *N*-gram features [197]), there is a large range of variability in the spoken language forms. BSP can provide insights into how such behavioral modulations unfold, including details of individual specific communication

styles. An example is the idiosyncratic use of discourse markers or phrases such as “yeah” and “it sounds like” or specific intonational patterns. Critically, BSP also offers means for capturing long-term dependencies across various aspects of the dialog. For example, the dynamics of the mutual influence of the interacting agents on the behavior of one another, both causally and noncausally in time, can be modeled using statistical graphical models. Can *et al.* [197] tagged reflection instances in a maximum entropy Markov modeling framework using several linguistic features with rich contextual information obtained from the transcripts of the entire dialog obtaining an F-score of over 80%. For instance, the model showed that the confirmatory discourse markers of the patient's response to a therapist's reflection carries useful information in detecting it, an aspect that was not considered in the MISC system but one that was revealed by BSP. Moreover, in general, the modeling showed that reflections are not mere repetitions or rephrasing of the patient utterances. Superior detection of reflections was obtained with the maximum entropy Markov model that had access to contextual *N*-grams compared to using HMMs that did not have them. Overall, this work illustrates how domain (psychology) inspired knowledge, noted below parenthetically as given in the MI manual, can be translated into computational modeling goals: the speaking style of the counselor (“summarize,” “listen reflectively,” “be collaborative and nonjudgmental”), the response of the client (“confirmation”), the content of the counselor response in relation to prior client talk (“capture and return”), and the dialog flow (“in response to previous client statement”).

This case study, even if preliminary, has raised research questions of fundamental interest about the MI therapy process. First, it points to the importance of communication style as much as content in marking a behavior construct such as reflection. Second, it raises questions about how local or globally distributed are specific behavior patterns in defining the overall dynamics of therapy. For example, results of [198] in modeling an observer's perception of an MI therapist's empathy in an interaction (yet another key behavioral marker of the quality of the therapy) were explained by a few isolated salient utterances in the entire session. Potential approaches to addressing some of these challenges are considered in the next section.

IV. Summary of Bsp Challenges and Possible Future Directions

A. Behavioral Data Acquisition

While sensing and signal acquisition advances have opened up new ways of observing human behavior in their natural occurrences and in offering details hitherto not accessible, challenges abound. The foremost of these relates to ensuring representative conditions and the ecological validity of the measurement process and the verity of the observations in relation to the behavior of interest. The former has to be achieved by designing minimally intrusive sensing systems that work within the constraints of the domain. For example, in a clinical observation of a child's interaction with a psychologist engaged in a diagnosis or intervention for attention disorder or autism diagnosis, the placement of microphones and cameras in the environment should be done in a way that offers minimum distraction to the child and the clinical process [47]. In addition to instrumented environments, novel measurement schemes exploiting wireless mobile technologies allow measuring, and communicating in real time, behavior data in real-world settings by directly instrumenting people. For example, using a low-cost mobile-phone-based wireless body area network with multimodal sensors such as tri-axial accelerometers, electrocardiography sensors, a blood oxygen saturation sensor, Global Positioning System (GPS) location sensor, as well as audiovideo data, the USC KNOWME system tracks physical and physiological behavior related to metabolic health allowing for realtime monitoring and intervention [53], [199].

There are a number of technical challenges in implementing multimodal signal capture in unconstrained real-world settings. These include technical constraints such as energy efficiency especially in mobile scenarios, sensor placement and reliability, multimodal data synchrony, noise in the acquisition environment (e.g., home, schools, urban outdoors), and interference in transmission channels. Other challenges include data security, data validity assurance, and "big data" issues such as storage and processing (e.g., continuous collection of multimodal behavior could lead to vast amounts of data to handle). With regard to speech and language data, this signifies developing means for robustly acquiring these data in the various environments (outdoor/indoor, controlled/unstructured, etc.) where the interactions occur with all the appropriate contextual information for allowing meaningful processing, online or offline.

An alternative approach to data acquisition for human communication-centric studies of behavior is the use of simulated conditions and actors to enable systematic exploration of specific characteristics not accessible using only observations of natural behavior. For example, to support studies of affective behavior, researchers have argued that good quality acted databases can be recorded when suitable acting methodologies are used to elicit emotional realizations from experienced actors engaged in dialogs rather than monologues in a goal-oriented approach [200]–[202]. Unlike most naturalistic behavior data from clinical studies or commercial applications, these acted corpora can be freely shared with the research community to help accelerate technology development. Examples include the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which contains approximately 12 h of audiovisual data from five mixed gender pairs of actors interacting with each other in scenarios that encourage emotional expression [203]. Future challenges in behavioral data collection include making available large corpora from across a variety of real domains as well as in supporting specific controlled laboratory studies to afford the scientific community targeted and easily shareable data sets.

B. Behavioral Analysis and Modeling

1) Representations—One of the major challenges in behavioral computing is specifying representations for computing. Given the complexity of human behavior in its expression and its perception, there are many degrees of freedom that vary at different time scales and different interpersonal contexts and across different cultures. Consequently, attempts to describe human behavior are varied, and not clear cut, and tend to be gradual, subjective, and approximate. Deriving suitable descriptions that capture the behavior of interest is a problem with many open challenges. Consider, for example, describing affective behavior (Section II-B3f). Depending on the analysis needs, approaches have focused on computing discrete categorical representations (e.g., happy, sad, engaged) or dimensional representations (e.g., valence or arousal degree). Furthermore, these representations can be based on natural language descriptions or can be numerical (e.g., on a Likert scale). Since language-based descriptions of behavior constitute a key representation approach, the study of natural human behavior warrants closer attention to how everyday people understand and describe emotions with natural language, as exemplified by Kazemzadeh *et al.* [204] in their study of emotions. Typical behavioral analysis schemes tend to adopt one or more of these descriptions; often such descriptions tend to be correlated as well: Methods to understand their interrelations and translating between them are still open research questions both from a scientific and computational perspective despite significant advances. Beyond discrete descriptions that offer a summative view (e.g., Did an utterance convey happiness? Or was that interaction negatively valenced?), current trends attempt to capture continuous variations over time [106], [133]–[135]. Modeling the dynamics of behavior evolution is still a largely poorly understood domain and requires new research.

Given the human-centered nature of behavior modeling, a majority of representations used for computing are either based on a generative perspective on how behaviors are expressed by humans or on a processing perspective based on how behaviors are perceived and described by (other observing) humans. The generative perspective is challenged by the lack of complete knowledge of the underlying behavior production mechanisms and by the inherent challenges posed by the multiple simultaneous roles that speech and language (and other signals) play in encoding linguistic and paralinguistic information that offer critical behavioral cues. The human speech signal carries rich information about a variety of linguistic and paralinguistic phenomena encoding intent and emotions, while manifesting a complex interplay between these phenomena; together with the vast contextual and individual variability, this often makes it a challenging problem for robustly inferring a target behavioral construct from speech observations. The processing perspective is challenged by the inherent subjectivity and diversity in human judgment of human behavior. Computational models of behavior within this framework rely on human-derived annotations, a process that is informed by the domain experts and employs variedly the wisdom of experts, trained annotators, or naive raters, including exploiting various crowd sourcing mechanisms that have become recently popular. The annotation process is wrought with challenges especially in handling abstract behavior. While expert-based approaches strive to achieve consistency, they are often tedious and not scalable; on the other hand, crowd-based approaches suffer from ensuring reliability in deriving the desired representations. Moreover, current approaches largely use simple plurality (such as majority voting) for fusing annotations, but these fail to capture the useful variability inherent in multiple perspectives [31], [32]. Novel computational methods that understand and effectively exploit the diversity of expert ensembles are needed. In particular, these methods also underscore the importance of directly modeling the annotator (the observer) whose judgments often serve as an important ingredient in BSP models.

2) Features—Another key BSP challenge relates to feature engineering in determining the feature-behavior correspondence. Researchers typically take an “exhaustive” approach of generating many possible segmental and suprasegmental speech features, representing both articulation and source properties, along with linguistic features derived from text transcripts, to find the feature subset that best explains a desired behavioral representation. Some of these features are rationalized based on well-understood theoretical and empirical studies in psychology and linguistics, while others are based on a data-driven exploration of the potential explanatory capacity of features. This is particularly the case in working with human-perception-based representations, which are often at a longer temporal scale (utterance or interaction session level). One of the open research challenges herein lies in not just finding the best set of features to predict the human derived representations, through appropriate feature selection methods, but to find those that are interpretable and meaningful to the expert. In particular, it has been shown in perceptual studies that not all portions of the data are equally relevant to the overall behavior judgment [205]. Methods such as those based on multiple instance learning (MIL) offer a promising computational avenue to seek feature representations that are salient with respect to a behavior description [206], [207]. Other knowledge-based approaches, such as those based on neuroscientific evidence of phenomena such as human attentional processes which have been successfully applied to machine audition and vision [208]–[210], offer yet another potential avenue for seeking salient features explaining higher level behavioral constructs. Additionally, combining such knowledge-based approaches with data-driven approaches such as multiple instance learning is also a promising research direction.

Another BSP research direction relates to feature dimensionality. The overgenerative approach for deriving speech and language features often results in thousands of features raising both relevance and computing challenges. The feature descriptors related to a

specific behavior construct are often hypothesized to be in a lower dimensional space, and novel schemes based on sparse representations, low manifold embedding, and other exemplar-based approaches are worth further research in the BSP context. Robustness is yet another BSP challenge, which may arise due to uncertainty in feature computation (e.g., voice quality measures using inverse filtering) or missing features due to inherent signal characteristics (interplay between features) or due to measurement or channel errors. Devising approaches to investigate feature sensitivity to specific behavior characterization is also a potential future direction.

Speech and language provide an important, but still a limited, window into higher level behavioral processes. Visual nonverbal cues offer both complementary and redundant information [211], and their joint modeling with speech derived features has been shown to yield both improved classification accuracy and robustness across a variety of behavioral analysis tasks from emotion recognition [103], [104] to predicting behavioral constructs such as approach avoidance [46], [212]. The knowledge about how multimodal encoding of behavioral information across the verbal and nonverbal channels unfolds dynamically while tracking real-life behavior patterns is limited, especially in the context of distressed and atypical behavior conditions such as those illustrated in Section III-A, pointing to yet another area that needs future research. Important BSP research in this regard should focus on context modeling in capturing the relation across feature (and representation) streams and across time [106].

3) Modeling—There are a number of modeling challenges arising, again, from how behavior cues are expressed by humans and how observers process them. In particular, a major class of behavior studies involves more than one person wherein the interlocutors are both observers and responders to the behaviors of others in the interactions. A key aspect of BSP lies in modeling this mutual influence. For example, prediction of interruption patterns in dialog [213] as well as affective behavior [129] was improved by explicitly modeling the mutual influence. While the mutual influence of the interlocutors has been observed across a variety of settings, computational models of the phenomena need to be further developed, especially in how they help explain the complex behavior patterns in distressed and atypical interaction settings.

One specific phenomenon that has been variedly described in studies of human communication and interaction, and implicated in explaining specific behavior patterning, is entrainment (interaction synchrony). This subtle phenomenon is difficult for direct human computation (and annotation) but is modeled based on features inspired by theoretical considerations (such as, for example, pitch contour features playing an important role in capturing vocal entrainment). Similarly the model's validity is demonstrated indirectly by showing its usefulness in explaining complex behavior phenomena where it is implicated (e.g., vocal entrainment in explaining positive affective dynamics [172], [214]). Advances in computational modeling are, however, sorely needed both to understand the nature of mutual influence across the various verbal and nonverbal channels and how they in turn explain complex higher level behavior patterns. Beyond improving scientific knowledge, such models can inform behavioral assessment, diagnostics, and intervention design. For example, recent modeling results in an autism diagnostic interaction between a child and a psychologist suggest that the behavioral patterns of the interacting psychologist offer predictive power about the ratings they provide about the child's atypical prosody [184]. Similarly, modeling a patient's response patterns offers clues about the therapist's reflective behavior in a motivational interviewing setting [197]. These complex interactions have to be better understood to enhance the explanatory and predictive ability of behavioral modeling.

BSP is inherently an interdisciplinary realm and is human centered. It promises both to improve process efficiency and accuracy in support of human behavior modeling and its applications as well as to create and offer tools for new scientific discovery. Advances in this domain crucially depend on developing productive and collaborative partnerships between domain experts and signal processing and computing researchers.

Acknowledgments

The authors would like to thank the members of the University of Southern California (USC) Signal Analysis and Interpretation Laboratory, and their collaborators, whose contributions were central to this paper.

This work was supported by the National Science Foundation (NSF), the National Institutes of Health (NIH), and the U.S. Department of Defense (DoD). This paper is based on the invited lecture material presented by S. Narayanan at the International Conference on Multimedia and Expo (ICME), Barcelona, Spain, July 2011; the ACM Multimedia International Workshop on Social Signal Processing (SSPW), Scottsdale, AZ, November–December 2011; and the Automatic Speech Recognition and Understanding Workshop (ASRU), Honolulu, HI, December 2011.

References

1. Black M, Chang J, Narayanan S. An empirical analysis of user uncertainty in problem-solving child-machine interactions. Proc Workshop Child Comput Interaction. 2008
2. Arunachalam, S.; Gould, D.; Andersen, E.; Byrd, D.; Narayanan, S. Proc Eurospeech Conf. Aalborg, Denmark: 2001. Politeness and frustration language in child-machine interactions; p. 2675-2678.
3. Yildirim, S.; Lee, C.; Lee, S.; Potamianos, A.; Narayanan, S. Proc Eurospeech Conf. Lisbon, Portugal: 2005. Detecting politeness and frustration state of a child in a conversational computer game; p. 2209-2212.
4. Zhang T, Hasegawa-Johnson M, Levinson S. Cognitive state classification in a spoken tutorial dialogue system. Speech Commun. 2006; 48(6):616–632.
5. Yildirim S, Narayanan S, Potamianos A. Detecting emotional state of a child in a conversational computer game. Comput Speech Lang. 2011; 25(1):29–44.
6. Forbes-Riley K, Litman D. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Commun. 2011; 53(9–10):1115–1136.
7. Pon-Barry H, Shieber SM. Recognizing uncertainty in speech. EURASIP J Adv Signal Process. 2011
8. Petrushin V. Emotion in speech: Recognition and application to call centers. Proc Artif Neural Netw Eng. Nov.1999 :7–10.
9. Lee CM, Narayanan S, Pieraccini R. Recognition of negative emotions from the speech signal. Proc IEEE Workshop Autom Speech Recognit Understand. 2001:240–243.
10. Lee CM, Narayanan S. Towards detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process. Mar; 2005 13(2):293–302.
11. Lord C, Risi S, Lambrecht L, Cook E, Leventhal B, DiLavore P, Pickles A, Rutter M. The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. J Autism Develop Disorders. 2000; 30(3):205–223.
12. Miller, WR.; Rollnick, S. Motivational Interviewing: Preparing People for Change. New York: Guilford; 2002.
13. Vinciarelli A, Pantic M, Bourlard H. Social signal processing: Survey of an emerging domain. Image Vis Comput. 2009; 27:1743–1759.
14. Gravano A, Levitan R, Willson L, Benˇuˇsˇ Sˇ, Hirschberg J, Nenkova A. Acoustic and prosodic correlates of social behavior. Proc 12th Annu Conf Int Speech Commun Assoc. 2011:97–100.
15. Ranganath R, Jurafsky D, McFarland D. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Comput Speech Lang. Jan; 2013 27(1):89–115.
16. Hammond K. Psychology's scientific revolution: Is it in danger? Cntr Res Judgment Policy Tech Rep. 1978

17. Brunswik, E. Perception and the Representative Design of Psychological Experiments. 2nd. Los Angeles, CA: Univ. California Press; 1956.
18. Neuendorf, K. The Content Analysis Guidebook. Thousand Oaks: CA: Sage; 2002.
19. Ickes W. Empathic accuracy. J Personality. 1993; 61(4):587–610.
20. Margolin G, Oliver P, Gordis E, O'Hearn H, Medina A, Ghosh C, Morland L. The nuts and bolts of behavioral observation of marital and family interaction. Clin Child Family Psychol Rev. 1998; 1(4):195–213.
21. Audhkhasi K, Georgiou PG, Narayanan S. Reliability-weighted acoustic model adaptation using crowd sourced transcriptions. Proc InterSpeech Conf. 2011:3045–3048.
22. Audhkhasi K, Georgiou PG, Narayanan S. Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. Proc Int Conf Audio Speech Signal Process. 2011:4980–4983.
23. Waldinger R, Schulz M, Hauser S, Allen J, Crowell J. Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. J Family Psychol. 2004; 18(1):58–71.
24. Baucom KJW, Baucom B, Christensen A. Do the nave know best? The predictive power of nave ratings of couple interactions. Psychol Assessment. 2013
25. Holsti, O. Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley; 1969.
26. Scott W. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quart. 1955; 19(3):321–325.
27. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968; 70(4):213–220. [PubMed: 19673146]
28. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull. 1971; 76(5):378–382.
29. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas. 1970; 30(1):61–70.
30. Snyder D, Heyman R, Haynes S. Evidence-based approaches to assessing couple distress. Psychol Assessment. 2005; 17(3):288–307.
31. Audhkhasi, K.; Narayanan, SS. Proc InterSpeech Conf. Makuhari, Japan: 2010. Data-dependent evaluator modeling and its application to emotional valence classification from speech; p. 2366-2369.
32. Audhkhasi K, Narayanan SS. A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels. IEEE Trans Pattern Anal Mach Intell. 201210.1109/TPAMI.2012.139
33. Ross J. The reliability, validity, and utility of self-assessment. Practical Assessment Res Eval. 2006; 11(10)
34. Clements K, Holtzworth-Munroe A. Aggressive cognitions of violent versus nonviolent spouses. Cogn Therapy Res. 2008; 32(3):351–369.
35. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008; 4:1–32. [PubMed: 18509902]
36. Margolin G, Chien D, Duman S, Fauchier A, Gordis E, Oliver P, Ramos M, Vickerman K. Ethical issues in couple and family research. J Family Psychol. 2005; 19(1):157–167.
37. National Research Council (U.S.) Committee. Computational technology for effective health care: Immediate steps and strategic directions on engaging the computer science research community in health care informatics. National Academies Press (U.S.); 2009.
38. Rozgic V, Busso C, Georgiou P, Narayanan S. Multimodal meeting monitoring: Improvements on speaker tracking and segmentation through a modified mixture particle filter. Proc IEEE 9th Workshop Multimedia Signal Process. 2007:60–65.
39. Cipolla, R.; Pentland, A. Computer Vision for Human-Machine Interaction. Cambridge, U.K: Cambridge Univ. Press; 1998.
40. Chu S, Narayanan S, Kuo CCJ, Mataric MJ. Where am I? Scene recognition for mobile robots using audio features. Proc Int Conf Multimedia Expo. 2006:885–888.

41. Potamianos G, Lucey P. Audio-visual ASR from multiple views inside smart rooms. Proc IEEE Int Conf Multisensor Fusion Integration Intell Syst. 2006:35–40.
42. Annavaram M, Medvidovic N, Mitra U, Narayanan S, Sukhatme G, Meng Z, Qiu S, Kumar R, Thatte G, Spruijt-Metz D. Multimodal sensing for pediatric obesity applications. Proc UrbanSense Conf. 2008:21–25.
43. Janin A, Ang J, Bhagat S, Dhillon R, Edwards J, Macias-Guarasa J, Morgan N, Peskin B, Shriberg E, Stolcke A, Wooters C, Wrede B. The ICSI meeting project: Resources and research. Proc Int Conf Acoust Speech Signal Process. 2004 DOI: 10.1.1.63.1514.
44. AMI: Augmented Multi-Party Interaction. online. Available: http://www.amiproject.org/
45. Busso, C.; Hernanz, S.; Chu, C.; Kwon, S.; Lee, S.; Georgiou, P.; Cohen, I.; Narayanan, S. Proc Int Conf Acoust Speech Signal Process. Vol. 2. Philadelphia, PA: Mar. 2005 Smart room: Participant and speaker localization and identification; p. 1117-1120.
46. Rozgi'c V, Xiao B, Katsamanis A, Baucom B, Georgiou PG, Narayanan S. A new multichannel multi modal dyadic interaction database. Proc InterSpeech Conf. 2010:27–30.
47. Black M, Bone D, Williams M, Gorrindo P, Levitt P, Narayanan S. The USC CARE corpus: Child-psychologist interactions of children with autism spectrum disorders. Proc 12th Annu Conf Int Speech Commun Assoc. 2011:1497–1500.
48. Oller DK, Niyogi P, Gray S, Richards JA, Gilkerson J, Xu D, Yapanel U, Warren SF. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proc Nat Acad Sci USA. 2010 DOI: http://10.1073/pnas.1003882107.
49. Tsukada A, Shino M, Devyver M, Kanade T. Illumination-free gaze estimation method for first-person vision wearable device. Proc Int Conf Comput Vis Workshops. 2011:2084–2091.
50. Fletcher R, Dobson K, Goodwin M, Eydgahi H, Wilder-Smith O, Fernholz D, Kuboyama Y, Hedman E, Poh M, Picard R. iCalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. IEEE Trans Inf Technol Biomed. Mar; 2010 14(2):215–223. [PubMed: 20064760]
51. Chaspari T, Lee CC, Narayanan S. Interplay between verbal response latency and physiology of children with autism during ECA interactions. Proc InterSpeech Conf. 2012
52. Mower E, Black M, Flores E, Williams M, Narayanan S. Design of an emotionally targeted interactive agent for children with autism. Proc IEEE Int Conf Multimedia Expo. 201110.1109/ICME.2011.6011990
53. Mitra U, Emken B, Lee S, Li M, Vathsangam H, Zois D, Annavaram M, Narayanan S, Spruijt-Metz D, Sukhatme G. KNOWME: A case study in wireless body area sensor network design. IEEE Commun Mag. May; 2012 50(5):116–125.
54. Li M, Rozgic V, Thatte G, Lee S, Emken A, Annavaram M, Mitra U, Spruijt-Metz D, Narayanan SS. Multimodal physical activity recognition by fusing temporal and cepstral information. IEEE Trans Neural Syst Rehabil Eng. Aug; 2010 18(4):369–380. [PubMed: 20699202]
55. Emken A, Li M, Thatte G, Lee S, Annavaram M, Mitra U, Narayanan S, Spruijt-Metz D. Recognition of physical activities in overweight Hispanic youth using knowme networks. J Phys Activity Health. 2012; 9(3):432–441.
56. Cassell, J.; Bickmore, T.; Billinghurst, M.; Campbell, L.; Chang, K.; Vilhjalmsson, H.; Yan, H. Proc Int Conf Human Factors Comput Syst. Pittsburgh, PA: May. 1999 Embodiment in conversational interfaces: Rea; p. 520-527.
57. Narayanan S, Potamianos A. Creating conversational interfaces for children. IEEE Trans Speech Audio Process. Feb; 2002 10(2):65–78.
58. Cassell J, Ryokai K. Making space for voice: Technologies to support children's fantasy and storytelling. Pers Ubiquitous Comput. 2001; 5(3):169–190.
59. Swartout W, Traum D, Artstein R, Noren D, Paul Debevec KB, Williams J, Leuski A, Shrikanth Narayanan DP, Lane C, Morie J, Aggarwal P, Liewer M, Chiang JY, Gerten J, Chu S, White K. Virtual museum guides demonstration. Proc IEEE Spoken Lang Technol Workshop. 2010:163–164.
60. Wu D, Courtney C, Lance B, Narayanan SS, Dawson M, Oie K, Parsons TD. Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality stroop task. IEEE Trans Affective Comput. Jul-Dec;2010 1(2):109–118.

61. Traum, D.; Roque, A.; Leuski, A.; Georgiou, P.; Gerten, J.; Martinovski, B.; Narayanan, S.; Robinson, S.; Vaswani, A. Hassan: A virtual human for tactical questioning. In: Keizer, S.; Bunt, B.; Paek, T., editors. Proc 8th SIGdial Workshop Discourse Dialogue. Antwerp, Belgium: 2007. p. 75-78.
62. Anguera Miro X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O. Speaker diarization: A review of recent research. IEEE Trans Audio Speech Lang Process. Feb; 2012 20(2):356–370.
63. Wooters C, Huijbregts M. The ICSI RT07s speaker diarization system. Multimodal Technol Percept Humans. 2008:509–519.
64. Friedland G, Hung H, Yeo C. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. Proc IEEE Int Conf Acoust Speech Signal Process. 2009:4069–4072.
65. Rozgic V, Han K, Georgiou P, Narayanan S. Multimodal speaker segmentation and identification in presence of overlapped speech segments. J Multimedia. 2010; 5(4):322–331.
66. Checka N, Wilson K, Siracusa M, Darrell T. Multiple person and speaker activity tracking with a particle filter. Proc Int Conf Acoust Speech Signal Process. 2004; V:881–884.
67. Gatica-Perez D, Lathoud G, McCowan I, Odobez JM, Moore D. Audio-visual speaker tracking with importance particle filters. Proc Int Conf Image Process. 2003; III:25–28.
68. Khan Z, Balch T, Dellaert F. MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans Pattern Anal Mach Intell. Nov; 2005 27(11):1805–1918. [PubMed: 16285378]
69. Busso, C.; Georgiou, P.; Narayanan, S. Proc Int Conf Acoust Speech Signal Process. Honolulu, HI: Apr. 2007 Real-time monitoring of participants' interaction in a meeting using audio-visual sensors; p. II-685-II-688.
70. Furui S. 50 years of progress in speech and speaker recognition. Speech Commun. 2005:1–9.
71. Georgiou, PG.; Black, MP.; Lammert, A.; Baucom, B.; Narayanan, SS. Proc Affective Comput Intell Interaction. Memphis, TN: 2011. ‘That’s aggravating, very aggravating': Is it possible to classify behaviors in couple interactions using automatically derived lexical features?; p. 87-96.
72. Kadambe S, Boudreaux-Bartels G. Application of the wavelet transform for pitch detection of speech signals. IEEE Trans Inf Theory. Mar; 1992 38(2):917–924.
73. Ghosh, P.; Ortega, A.; Narayanan, S. Proc InterSpeech Conf. Antwerp, Belgium: Aug. 2007 Pitch period estimation using multipulse model and wavelet transform; p. 297-300.
74. Talkin D. A robust algorithm for pitch tracking (RAPT). Speech Coding Synthesis. 1995:495–518.
75. Shimamura T, Kobayashi H. Weighted autocorrelation for pitch extraction of noisy speech. IEEE Trans Speech Audio Process. Oct; 2001 9(7):727–730.
76. Wang D, Narayanan S. Piecewise linear stylization of pitch via wavelet analysis. Proc 9th Eur Conf Speech Commun Technol. 2005:3277–3280.
77. Ghosh P, Narayanan S. Pitch contour stylization using an optimal piecewise polynomial approximation. IEEE Signal Process Lett. Sep; 2009 16(9):810–813.
78. d'Alessandro C, Mertens P, et al. Automatic pitch contour stylization using a model of tonal perception. Comput Speech Lang. 1995; 9(3):257.
79. Katsamanis A, Black MP, Georgiou PG, Goldstein L, Narayanan SS. SailAlign: Robust long speech-text alignment. Proc Very-Large-Scale Phonetics Workshop. Jan.2011
80. Morgan N, Fosler-Lussier E. Combining multiple estimators of speaking rate. Proc Int Conf Acoust Speech Signal Process. 1998; 2:729–732.
81. Wang D, Narayanan S. Robust speech rate estimation for spontaneous speech. IEEE Trans Audio Speech Lang Process. Nov; 2007 15(8):2190–2201.
82. Tamburini F, Caini C. An automatic system for detecting prosodic prominence in American English continuous speech. Int J Speech Technol. 2005; 8(1):33–44.
83. Wang D, Narayanan SS. An acoustic measure for word prominence in spontaneous speech. IEEE Trans Audio Speech Lang Process. Feb; 2007 15(2):690–701.
84. Stolcke A, Shriberg E, Bates R, Ostendorf M, Hakkani D, Plauche M, Tur G, Lu Y. Automatic detection of sentence boundaries and disfluencies based on recognized words. Proc 5th Int Conf Spoken Lang Process. 1998 DOI: 10.1.1.53.7127.

85. Shriberg E, Stolcke A, Hakkani-Tu¨r D, Tu¨r G. Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun. 2000; 32(1):127–154.

86. Wang D, Narayanan S. A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues. Proc IEEE Int Conf Acoust Speech Signal Process. 2004; 1:I-525–I-528.

87. Ross K, Ostendorf M. Prediction of abstract prosodic labels for speech synthesis. Comput Speech Lang. 1996; 10(3):155–185.

88. Wightman C, Ostendorf M. Automatic labeling of prosodic patterns. IEEE Trans Speech Audio Process. Oct; 1994 2(4):469–481.

89. Chen K, Hasegawa-Johnson M, Cohen A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. Proc Int Conf Acoust Speech Signal Process. 2004; 1:509–512.

90. Ananthakrishnan S, Narayanan S. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Trans Audio Speech Lang Process. Jan; 2008 16(1):216–228.

91. Ananthakrishnan S, Narayanan S. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. Proc Int Conf Acoust Speech Signal Process. 2005:269–272.

92. Rangarajan Sridhar V, Bangalore S, Narayanan S. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Trans Audio Speech Lang Process. May; 2008 16(4):797–811.

93. Austin, JL. How to do Things With Words. Oxford, U.K.: Clarendon Press; 1962.

94. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema C, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist. 2000; 26(3):339–373.

95. Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; Ess-Dykema, CV. Proc IEEE Workshop Autom Speech Recognit Understand. Santa Barbara, CA: Dec. 1997 Automatic detection of discourse structure for speech recognition and understanding; p. 88-95.

96. Potamianos A, Narayanan S, Riccardi G. Adaptive categorical understanding for spoken dialogue systems. IEEE Trans Speech Audio Process. May; 2005 13(3):321–329.

97. Rangarajan Sridhar V, Bangalore S, Narayanan S. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. Comput Speech Lang. 2009; 23(4):407–422.

98. Amrhein P, Miller W, Yahne C, Palmer M, Fulcher L. Client commitment language during motivational interviewing predicts drug use outcomes. J Consult Clin Psychol. 2003; 71(5):862–878. [PubMed: 14516235]

99. Lee CC, Black MP, Katsamanis A, Lammert A, Baucom BR, Christensen A, Georgiou PG, Narayanan S. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. Proc InterSpeech Conf. 2010:793–796.

100. Levitan R, Gravano A, Hirschberg J. Entrainment in speech preceding backchannels. Proc Annu Meeting Assoc Comput Linguist. 2011; 2:113–117.

101. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. IEEE Signal Process Mag. Jan; 2001 18(1):32–80.

102. Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. 2011; 53:1062–1087.

103. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Proc IEEE Int Conf Multimodal Interfaces. State College, PA: Oct. 2004 Analysis of emotion recognition using facial expressions, speech and multimodal information; p. 205-211.

104. Metallinou, A.; Lee, S.; Narayanan, SS. Proc Int Conf Acoust Speech Signal Process. Dallas, TX: 2010. Decision level combination of multiple modalities for recognition and analysis of emotional expression; p. 2462-2465.

105. Metallinou, A.; Busso, C.; Lee, S.; Narayanan, SS. Proc Int Conf Acoust Speech Signal Process. Dallas, TX: 2010. Visual emotion recognition using compact facial representations and viseme information; p. 2474-2477.

106. Metallinou A, Wollmer M, Katsamanis A, Eyben F, Schuller B, Narayanan S. Context-sensitive learning for enhanced audiovisual emotion classification. IEEE Trans Affective Comput. Apr-Jun;2012 3(2):184–198.
107. Grimm M, Kroschel K, Mower E, Narayanan S. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 2007; 49(10–11):787–800.
108. Mower, E.; Metallinou, A.; Lee, CC.; Kazemzadeh, A.; Busso, C.; Lee, S.; Narayanan, SS. Proc Int Conf Affective Comput Intell Interaction. Amsterdam, The Netherlands: Sep. 2009 Interpreting ambiguous emotional expressions; p. 1-8.
109. Mower, E.; Han, K.; Lee, S.; Narayanan, S. Proc 11th Annu Conf Int Speech Commun Assoc. Makuhari, Japan: 2010. A cluster-profile representation of emotion using agglomerative hierarchical clustering; p. 797-800.
110. Mower E, Mataric M, Narayanan S. A framework for automatic human emotion classification using emotion profiles. IEEE Trans Audio Speech Lang Process. Jul; 2011 19(5):1057–1070.
111. Mower E, Narayanan S. A hierarchical static-dynamic framework for emotion classification. Proc IEEE Int Conf Acoust Speech Signal Process. 2011:2372–2375.
112. Williams C, Stevens K. Emotions and speech: Some acoustical correlates. J Acoust Soc Amer. 1972; 52:1238–1250. [PubMed: 4638039]
113. Scherer K. Vocal communication of emotion: A review of research paradigms. Speech Commun. Apr; 2003 40(1–2):227–256.
114. Bulut, M.; Lee, S.; Narayanan, S. Proc Int Conf Acoust Speech Signal Process. Las Vegas, NV: Apr. 2008 Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech; p. 4629-4632.
115. Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans Audio Speech Lang Process. May; 2009 17(4):582–596.
116. Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L. The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals. Proc InterSpeech Conf. 2007:2253–2256.
117. Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T. Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun. 2008; 50(6):487–503.
118. Bulut, M.; Busso, C.; Yildirim, S.; Kazemzadeh, A.; Lee, CM.; Lee, S.; Narayanan, S. Proc Eurospeech Conf. Lisbon, Portugal: Oct. 2005 Investigating the role of phoneme-level modifications in emotional speech resynthesis; p. 801-804.
119. Yildirim, S.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Busso, C.; Deng, Z.; Lee, S.; Narayanan, S. Proc 8th Int Conf Spoken Lang Process. Jeju Island, Korea: 2004. An acoustic study of emotions expressed in speech; p. 2193-2196.
120. Bänziger T, Scherer K. The role of intonation in emotional expressions. Speech Commun. 2005; 46(3):252–267.
121. Ladd D, Silverman K, Tolkmitt F, Bergmann G, Scherer K. Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. J Acoust Soc Amer. 1985; 78(2):435–444.
122. Eyben F, W¨ollmer M, Schuller B. Opensmile: The Munich versatile and fast open-source audio feature extractor. Proc Int Conf Multimedia. 2010:1459–1462.
123. Whissell C. The dictionary of affect in language. Emotion, Theory Res Exp. 1989; 4:113–131.
124. Bradley, M.; Lang, P. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Miami, FL: Univ Florida; 1999.
125. Malandrakis N, Potamianos A, Iosif E, Narayanan S. Kernel models for affective lexicon creation. Proc 12th Annu Conf Int Speech Commun Assoc. 2011:2977–2980.
126. Lee CM, Narayanan SS, Pieraccini R. Classifying emotions in human-machine spoken dialogs. Proc IEEE Int Conf Multimedia Expo. 2002; 1:737–740.
127. Lee, C.; Narayanan, S.; Pieraccini, R. Proc 7th Int Conf Spoken Lang Process. Denver, CO: Sep. 2002 Combining acoustic and language information for emotion recognition; p. 873-876.
128. Liscombe J, Riccardi G, Hakkani-Tür D. Using context to improve emotion detection in spoken dialog systems. Proc InterSpeech Conf. 2005:1845–1848.

129. Lee CC, Busso C, Lee S, Narayanan S. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. Proc InterSpeech Conf. 2009:1983–1986.
130. Lee C, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. 2011; 53:1162–1171.
131. Busso, C.; Lee, S.; Narayanan, S. Proc InterSpeech Conf. Antwerp, Belgium: Aug. 2007 Using neutral speech models for emotional speech analysis; p. 2225-2228.
132. Wo¨llmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. Proc 9th Annu Conf Int Speech Commun Assoc. 2008:597–600.
133. Wollmer M, Schuller B, Eyben F, Rigoll G. Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. IEEE J Sel Top Signal Process. Oct; 2010 4(5):867–881.
134. Nicolaou M, Gunes H, Pantic M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans Affective Comput. Apr-Jun;2011 2(2):92–105.
135. Metallinou A, Katsamanis A, Wang Y, Narayanan S. Tracking changes in continuous emotion states using body language and prosodic cues. Proc Int Conf Acoust Speech Signal Process. 2011:2288–2291.
136. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Mu¨ller C, Narayanan S. Paralinguistics in speech and language: State-of-the-art and the challenge. Comput Speech Lang. 2012; 27(1):4–39.
137. Juslin, PN.; Scherer, KR. Vocal expression of affect. In: Harrigan, J.; Rosenthal, R.; Scherer, K., editors. The New Handbook of Methods in Nonverbal Behavior Research. Eds Oxford, U.K: Oxford Univ Press; 2005. p. 65-135.
138. Gottman JM, Levenson RW. The social psychophysiology of marriage. Perspect Marital Interaction. 1988:182–200.
139. Cacioppo JT, Berntson GG, Sheridan JF, McClintock MK. Multilevel integrative analyses of human behavior: Social neuroscience and the complementing nature of social and biological approaches. Psychol Bull. 2000; 126(6):829–843. [PubMed: 11107878]
140. Kring AM, Sloan D. The Facial Expression Coding System (FACES): A Users Guide. Tech Manual. 1991
141. Sloan DM, Kring AM. Measuring changes in emotion during psychotherapy: Conceptual and methodological issues. Clin Psychol Sci Practice. 2007; 14(4):307–322.
142. Gonzales AL, Hancock JT, Pennebaker JW. Language style matching as a predictor of social dynamics in small groups. Commun Res. 2010; 37(1):3.
143. Pennebaker JW, Chung CK. Expressive writing, emotional upheavals, and health. Found Health Psychol. 2007:263–284.
144. Cromwell, RE.; Olson, DHL. Power in Families. New York: Sage; 1975.
145. Lakoff R. What you can do with words: Politeness, pragmatics and performatives. Proc Texas Conf Performatives Presuppositions Implicatures. 1977:79–105.
146. Lakoff G. Pragmatics in natural logic. Proc Texas Conf Performatives Presuppositions Implicatures. 1977:107.
147. Sacks H, Schegloff E, Jefferson G. A simplest systematics for the organization of turn-taking for conversation. Language. 1974; 50(4):696–735.
148. Coker DA, Burgoon JK. The nature of conversational involvement and nonverbal encoding patterns. Human Commun Res. 1987; 13(4):463–494.
149. Bousmalis K, Mehu M, Pantic M. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. Proc Affective Comput Intell Interaction. 200910.1109/ACII. 2009.5349477
150. Doohan E. Listening behaviors of married couples: An exploration of nonverbal presentation to a relational outsider. Int J Listening. 2007; 21(1):24–41.

151. Christensen A, Atkins D, Berns S, Wheeler J, Baucom D, Simpson L. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. J Consult Clin Psychol. 2004; 72(2):176–191. [PubMed: 15065953]
152. Jones, J.; Christensen, A. Univ California. Los Angeles: Tech Manual; 1998. Couples interaction study: Social support interaction rating system.
153. Heavey, C.; Gill, D.; Christensen, A. Couples Interaction Rating System 2 (CIRS2). Los Angeles, CA: Univ. California Press; 2002.
154. Gottman JM, Guralnick MJ, Wilson B, Swanson CC, Murray JD. What should be the focus of emotion regulation in children. Develop Psychopathol. 1997; 9:421–452.
155. Baucom BR, Atkins DC, Simpson LE, Christensen A. Prediction of response to treatment in a randomized clinical trial of couple therapy: A 2-year follow-up. J Consult Clin Psychol. 2009; 77(1):160–173. [PubMed: 19170462]
156. Black, M.; Georgiou, P.; Katsamanis, A.; Baucom, B.; Narayanan, S. Proc InterSpeech Conf. Florence, Italy: Aug. 2011 'You made me do it': Classification of blame in married couples' interaction by fusing automatically derived speech and language information; p. 89-92.
157. Black MP, Katsamanis A, Lee CC, Lammert A, Baucom BR, Christensen A, Georgiou PG, Narayanan SS. Automatic classification of married couples' behavior using audio features. Proc InterSpeech Conf. 2010:2030–2033.
158. Ghosh PK, Tsiartas A, Narayanan SS. Robust voice activity detection using long-term signal variability. IEEE Trans Audio Speech Lang Process. Mar; 2011 19(3):600–613.
159. Schuller B, Wimmer M, M¨osenlechner L, Kern C, Arsic D, Rigoll G. Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? Proc Int Conf Acoust Speech Signal Process. 2008:4501–4504.
160. Chang, CC.; Lin, CJ. LIBSVM: A library for support vector machines. 2001. online. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm
161. Black M, Katsamanis A, Baucom B, Lee C, Lammert A, Christensen A, Georgiou P, Narayanan S. Toward automating a human behavioral coding system for married couples? Interactions Using Speech Acoustic Features. Speech Commun. 2012; 55(1):1–21.
162. Williams-Baucom KJ, Atkins DC, Sevier M, Eldridge KA, Christensen A. 'You' and 'I' need to talk about 'us': Linguistic patterns in marital interactions. Pers Relationships. 2010; 17(1):41–56.
163. Gubbins CA, Perosa LM, Bartle-Haring S. Relationships between married couples' self-differentiation/individuation and Gottman's model of marital interactions. Contemporary Family Therapy. 2010; 32(4):383–395.
164. Kimura M, Daibo I. Interactional synchrony in conversations about emotional episodes: A measurement by 'the between-participants pseudosynchrony experimental paradigm'. J Nonverbal Behav. 2006; 30:115–126.
165. Hirschberg, J. presented at the 12th Annu Conf Int Speech Commun Assoc. Florence, Italy: 2011. Speaking more like you: Entrainment in conversational speech.
166. McGarva AR, Warner RM. Attraction and social coordination: Mutual entrainment of vocal activity rhymes. J Psycholinguist Res. 2003; 32(3):335–354. [PubMed: 12845943]
167. Nenkova, A.; Gravano, A.; Hirschberg, J. Proc Assoc Comput Linguist. Columbus, OH: Jun. 2008 High frequency word entrainment in spoken dialogue; p. 169-172.
168. Richardson MJ, Marsh KL, Schmit RC. Effects of visual and verbal interaction on unintentional interpersonal coordination. J Exp Psychol Human Perception Performance. 2005; 31(1):62–79.
169. Pardo JS. On phonetic convergence during conversational interaction. J Acoust Soc Amer. 2006; 119:2382–2393. [PubMed: 16642851]
170. Edlund J, Heldner M, Hirschberg J. Pause and gap length in face-to-face interaction. Proc 10th Annu Conf Int Speech Commun Assoc. 2009:2779–2782.
171. Lee CC, Katsamanis A, Black MP, Baucom B, Georgiou P, Narayanan S. An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions based vocal entrainment measures in married couples' affective spoken interactions. Proc InterSpeech Conf. Aug.2011 :3101–3104.

172. Lee C, Katsamanis A, Black M, Baucom B, Georgiou P, Narayanan S. Affective state recognition in married couples interactions using PCA-based vocal entrainment measures with multiple instance learning. Proc Int Conf Affective Comput Intell Interaction. 2011:31–41.
173. Lee, CC.; Katsamanis, A.; Baucom, B.; Georgiou, P. S. N. U. of Southern California"Using measures of vocal entrainment to inform outcome-related behaviors in marital conflicts,"; Proc Asia Pacific Signal Inf Process Assoc. 2012. p. 1-5.online. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6412027&isnumber=6411092
174. Price P, Tepperman J, Iseli M, Duong T, Black M, Wang S, Boscardin C, Heritage M, David Pearson P, Narayanan S, Alwan A. Assessment of emerging reading skills in young native speakers and language learners. Speech Commun. 2009; 51(10):968–984.
175. Mostow, J.; Aist, G.; Huang, C.; Junker, B.; Kennedy, R.; Lan, H.; Latimer, DT.; O'Connor, R.; Tassone, R.; Tobin, B.; Wierman, A. 4-month evaluation of a learner-controlled reading tutor that listens. In: Holland, FNFVM., editor. The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice. New York: Routledge; 2008. p. 201-219.
176. Litman D, Forbes-Riley K. Predicting student emotions in computer-human tutoring dialogues. Proc 42nd Annu Meeting Assoc Comput Linguist. 2004:351–358. article 351.
177. Tepperman J, Lee S, Narayanan SS, Alwan A. A generative student model for scoring word reading skills. IEEE Trans Audio Speech Lang Process. Feb; 2011 19(2):348–360.
178. Kazemzadeh A, You H, Iseli M, Jones B, Cui X, Heritage M, Price P, Anderson E, Narayanan S, Alwan A. Tball data collection: The making of a young children's speech corpus. Proc 9th Eur Conf Speech Commun Technol. 2005:1581–1584.
179. Black M, Tepperman J, Narayanan S. Automatic prediction of children's reading ability for high-level literacy assessment. IEEE Trans Audio Speech Lang Process. May; 2011 19(4):348–360.
180. Black M, Narayanan S. Improvements in predicting children's overall reading ability by modeling variability in evaluators' subjective judgments. Proc IEEE Int Conf Acoust Speech Signal Process. 2012:5069–5072.
181. D'Mello S, Graesser A, Picard R. Toward an affect-sensitive autotutor. IEEE Intell Syst. Jul-Aug; 2007 22(4):53–61.
182. Graesser A, D'Mello S. Theoretical perspectives on affect and deep learning. New Perspectives Affect Learn Technol. 2011; 3:11–21.
183. Eskenazi M. An overview of spoken language technology for education. Speech Commun. 2009; 51(10):832–844.
184. Bone D, Black MP, Lee CC, Williams ME, Levitt P, Lee S, Narayanan S. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. Proc InterSpeech Conf. 2012
185. Prud'hommeaux E, Roark B, Black L, van Santen J. Classification of atypical language in autism. Proc Workshop Cogn Model Comput Linguist. 2011:88–96.
186. American Psychiatric Association Task Force. Diagnostic and Statistical Manual of Mental Disorders: Dsm-iv-tr. 2000
187. Centers for Disease Control and Prevention (CDC). National Center on Birth Defects and Developmental Disabilities Autism Spectrum Disorders (ASDs). Mar. 2010 online. Available: http://www.cdc.gov/ncbddd/autism/
188. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, Donaldson A, Varley J. Randomized, controlled trial of an intervention for toddlers with autism: The early start Denver model. Pediatrics. 2010; 125(1):17–23.
189. Gotham K, Risi S, Pickles A, Lord C. The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity. J Autism Develop Disorders. 2007; 37(4):613–627.
190. Pepp'e S, McCann J, Gibbon F, O'Hare A, Rutherford M. Receptive and expressive prosodic ability in children with high-functioning autism. J Speech Lang Hearing Res. 2007; 50(4):1015–1028.
191. Shriberg L, Paul R, McSweeny J, Klin A, Cohen D, Volkmar F. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. J Speech Lang Hearing Res. 2001; 44(5):1097.

192. Wallace MC, Cleary JE, Buder EH, Pettit W, Oller DK. An acoustic inspection of vocalizations in young children with autism spectrum disorders. Proc Int Meeting Autism Res. 2008
193. Mower E, Lee C, Gibson J, Chaspari T, Williams M, Narayanan S. Analyzing the nature of ECA interactions in children with autism. Proc 12th Annu Conf Int Speech Commun Assoc. 2011:2989–2993.
194. Chaspari, T.; Provost, EM.; Katsamanis, A.; Narayanan, S. Proc Int Conf Acoust Speech Signal Process. Kyoto, Japan: 2012. An acoustic analysis of shared enjoyment in ECA interactions of children with autism; p. 4485-4488.
195. Miller, W.; Moyers, T.; Ernst, D.; Amrhein, P. Manual for the Motivational Interviewing Skill Code (MISC). 2003. online. Available: http://casaa.unm.edu/download/misc.pdf
196. Burke B, Dunn C, Atkins D, Phelps J. The emerging evidence base for motivational interviewing: A meta-analytic and qualitative inquiry. J Cogn Psychotherapy. 2004; 18(4):309–322.
197. Can D, Georgiou P, Atkins D, Narayanan S. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. Proc InterSpeech Conf. 2012
198. Xiao, B.; Georgiou, PG.; Narayanan, S. Analyzing the language of therapist empathy in motivational interview based psychotherapy; Proc Asia Pacific Signal Inf Process Assoc. 2012. p. 1-4.online. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6411762url=%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%3D%26arnumber%3D6411762
199. Thatte G, Li M, Lee S, Emken A, Narayanan S, Mitra U, Spruijt-Metz D, Annavaram M. KNOWME an energy-efficient, multimodal body area network for physical activity monitoring. ACM Trans Embedded Comput Syst. 2012; 11(S2) article 48. 10.1145/2331147.2331158
200. Enos F, Hirschberg J. A framework for eliciting emotional speech: Capitalizing on the actors process. Proc 1st Int Workshop Emotion Corpora for Research on Emotion and Affect (Int Conf Lang Resources Eval). 2006:6–10.
201. Bänziger T, Scherer K. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. Affective Comput Intell Interaction. 2007; 4738:476–487.
202. Busso, C.; Narayanan, S. Proc Int Conf Lang Resources Eval/Workshop Emotion Corpora for Research on Emotion and Affect. Marrakech: Morocco; May. 2008 Recording audio-visual emotional databases from actors: A closer look; p. 17-22.
203. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. J Lang Resources Eval. Dec; 2008 42(4):335–359.
204. Kazemzadeh A, Lee S, Georgiou P, Narayanan S. Emotion twenty questions: Toward a crowd-sourced theory of emotions. Affective Comput Intell Interaction. 2011; 6975:1–10.
205. Ambady N, Bernieri F, Richeson J. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. Adv Exp Soc Psychol. 2000; 32:201–271.
206. Katsamanis A, Gibson J, Black M, Narayanan S. Multiple instance learning for classification of human behavior observations. Affective Comput Intell Interaction. 2011; 6974:145–154.
207. Gibson, J.; Katsamanis, A.; Black, MP.; Narayanan, SS. Proc InterSpeech Conf. Florence, Italy: 2011. Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines; p. 1561-1564.
208. Mesgarani N, Slaney M, Shamma S. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans Audio Speech Lang Process. May; 2006 14(3):920–930.
209. Siagian C, Itti L. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans Pattern Anal Mach Intell. Feb; 2007 29(2):300–312. [PubMed: 17170482]
210. Kalinli O, Narayanan S. Prominence detection using auditory attention cues and task-dependent high level information. IEEE Trans Audio Speech Lang Process. Jul; 2009 17(5):1009–1024.
211. Busso C, Narayanan S. Interrelation between speech and facial gestures in emotional utterances: A single subject study. IEEE Trans Audio Speech Lang Process. Nov; 2007 15(8):2331–2347.
212. Xiao B, Rozgic V, Katsamanis A, Baucom B, Georgiou PG, Narayanan S. Overlapped speech detection using long-term spectro-temporal similarity in stereo recording. Proc InterSpeech Conf. 2011

213. Lee CC, Narayanan S. Predicting interruptions in dyadic spoken interactions. Proc Int Conf Acoust Speech Signal Process. 2010:5250–5253.

214. Lee, CC.; Katsamanis, A.; Black, MP.; Baucom, B.; Christensen, A.; Georgiou, PG.; Narayanan, SS. Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. Comput Speech Lang. 2012. online. Available: http://dx.doi.org/10.1016/j.csl.2012.06.006

About The Authors

Shrikanth (Shri) Narayanan (Fellow, IEEE) received the M.S., Engineer, and Ph.D. degrees in electrical engineering from the University of California Los Angeles, Los Angeles, CA, USA, in 1990, 1992, and 1995, respectively. He is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology and as the Founding Director of the Ming Hsieh Institute. Prior to USC, he was with AT&TBell Labs and AT&T Research from 1995to2000.AtUSC,hedirects the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics. He has published over 500 papers and has 14 granted U.S. patents. Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the *Computer Speech and Language Journal* and an Associate Editor for the IEEE Transactions on Affective Computing, *APSIPA Transactions on Signal and Information Processing* and the *Journal of the Acoustical Society of America*. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000– 2004), IEEE Signal Processing Magazine (2005–2008), and the IEEE Transactions on Multimedia. He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011. Papers coauthored with his students have won awards at the 2012 Annual Conference of the International Speech Communication Association (Interspeech) Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2010, InterSpeech 2009/Emotion Challenge, the 2009 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), the 2006 and 2007 IEEE Workshops on Multimedia Signal Processing (MMSP), the 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), and the 2002 International Conference of Spoken Language Processing (ICSLP).

Panayiotis G. Georgiou (Senior Member, IEEE) received the B.A. and M.Eng. degrees (with honors) from Cambridge University (Pembroke College), Cambridge, U.K., in 1996, where he was a Cambridge-Commonwealth Scholar, and the M.Sc. and Ph.D. degrees from the University of Southern California (USC), Los Angeles, in 1998 and 2002, respectively. Since 2003, he has been a member of the Signal Analysis and Interpretation Lab at USC, where he is currently a Research Assistant Professor. His interests span the fields of multimodal and behavioral signal processing. He has worked on and published over 100 papers in the fields of behavioral signal processing, statistical signal processing, alpha stable distributions, speech and multimodal signal processing and interfaces, speech translation, language modeling, immersive sound processing, sound source localization, and speaker identification. He has been a Principal Investigator (PI) and co-PI on federally funded projects notably including the DARPA Transtac “SpeechLinks” and currently the National Science Foundation (NSF) “An Integrated Approach to Creating Enriched Speech Translation Systems” and “Quantitative Observational Practice in Family Studies: The case of reactivity.” His current focus is on behavioral signal processing, multimodal environments, and speech-to-speech translation. Prof. Georgiou is currently serving as a Guest Editor of the *Computer Speech And Language* journal and as a member of the Speech and Language Technical Committee. Papers coauthored with his students have won best paper awards for analyzing the multimodal behaviors of users in speech-to-speech translation at the 2006 International Workshop on Multimedia Signal Processing (MMSP) and for automatic classification of married couples' behavior using audio features at the 2010 Annual Conference of the International Speech Communication Association (Interspeech).

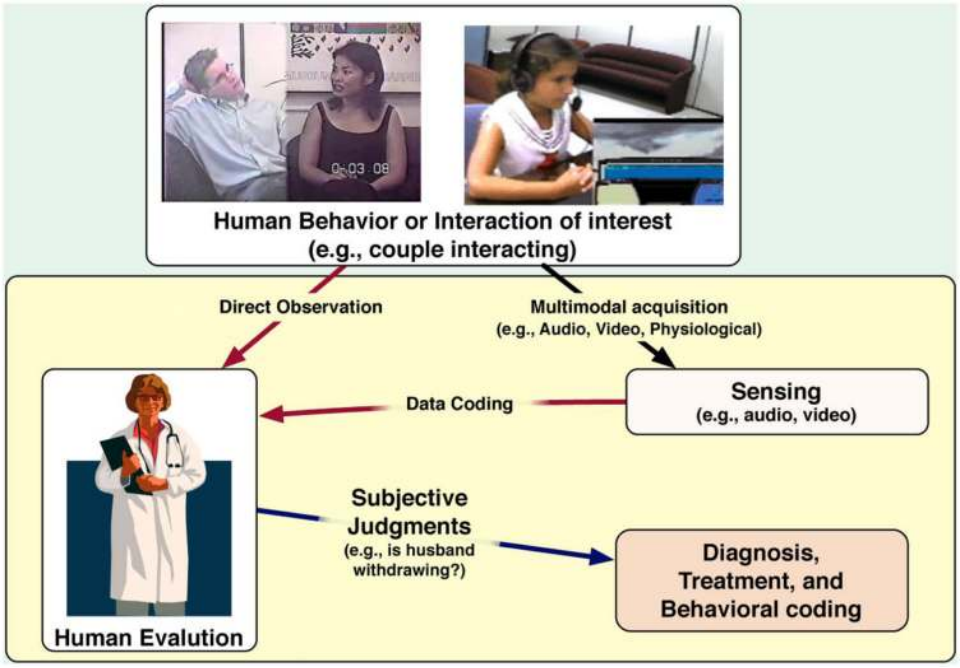

Fig. 1.

Behavioral analysis by an expert: in practice, experts observe directly and infer diagnosis and treatment outcomes. In some clinical cases and in observational research the data may be revisited through audiovisual recordings and coded by experts. [Image of couple interacting courtesy of Prof. A. Christensen, Clinical Psychology Department, University of California Los Angeles (UCLA)].

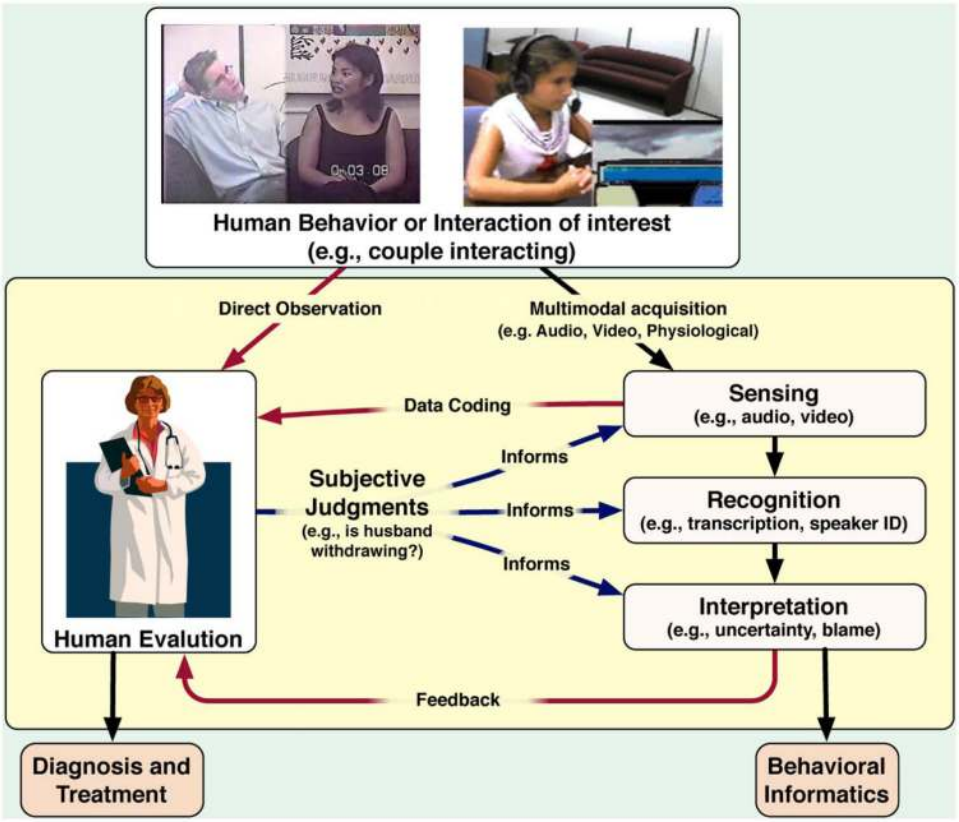

Fig. 2.

Human in the loop: outcomes are both learned from and fed to the human expert to refine machine knowledge and augment the experts' analytical capabilities. Similar behavioral informatics can facilitate decision making across a number of domains such as in education and commerce.

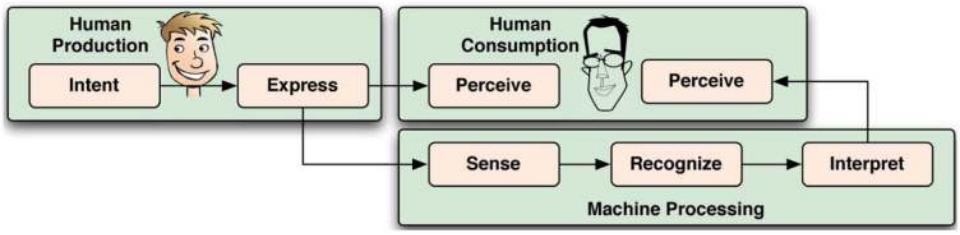

Fig. 3.
Intent (intended, conscious, or subconscious), production (explicit or implicit), observed perception, and machine recognition.

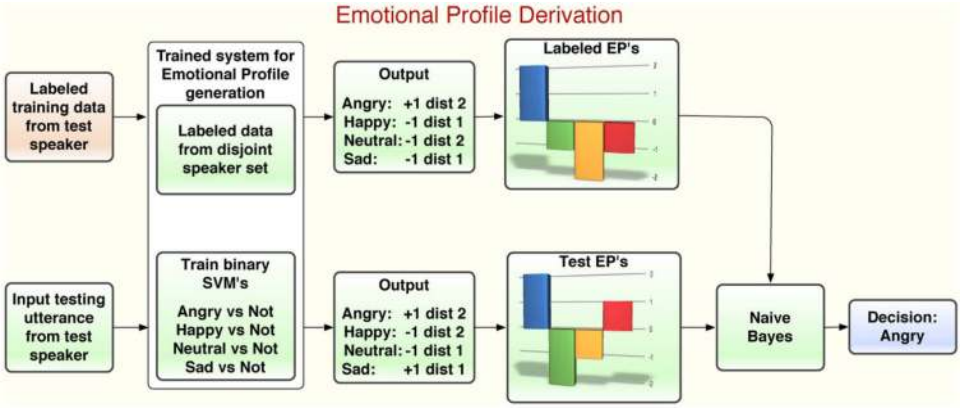

Fig. 4.
Illustration of emotional profile (EP) derivation from signal data for quantifying nonprototypical, ambiguous expressions (after [110]).

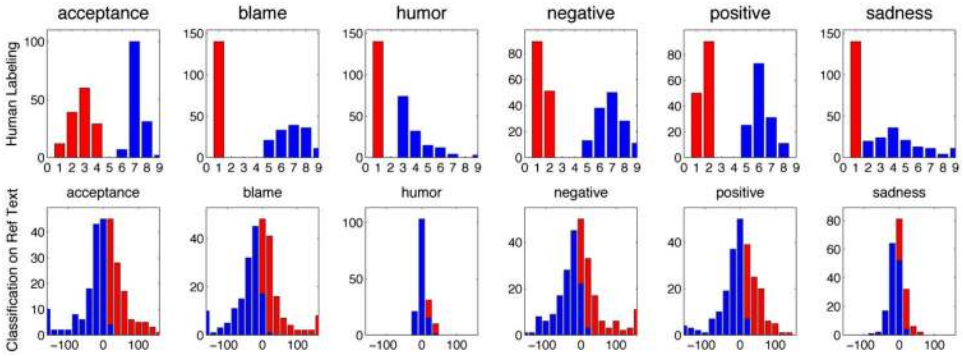

Fig. 5.

Distribution of data based on (top 20% lowest averaged ratings in red and 20% highest averaged ratings in blue) the average ratings provided by multiple human experts and (bottom) the difference in log-likelihoods of the ML model for □= 0.5. Codes where annotators had minimal separation also result in the greatest overlap by the ML model.

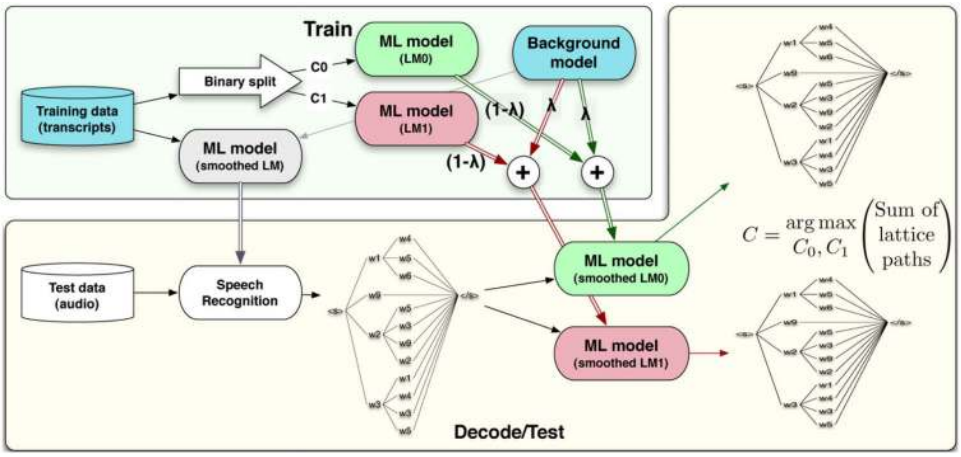

Fig. 6.
Overview of the classification process without human transcripts through the use of ASR lattices.

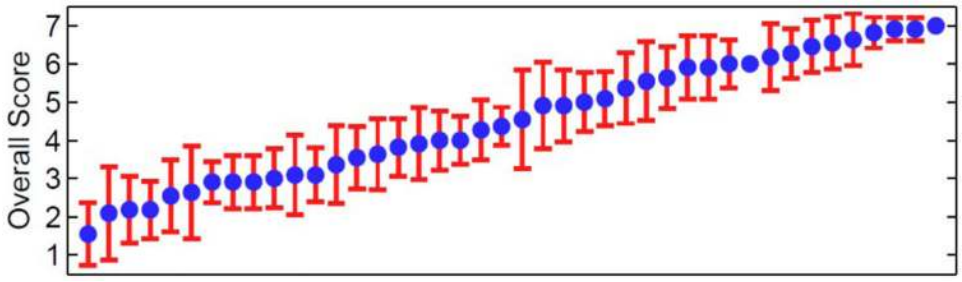

Fig. 7.
Variability in the confidence of evaluator rating across various reading performance levels [180]. Means and standard deviations of ratings obtained from 11 human evaluators across 42 children (arranged in increasing overall score received). Note lower variability, i.e., increased evaluator consistency, for higher performing children.

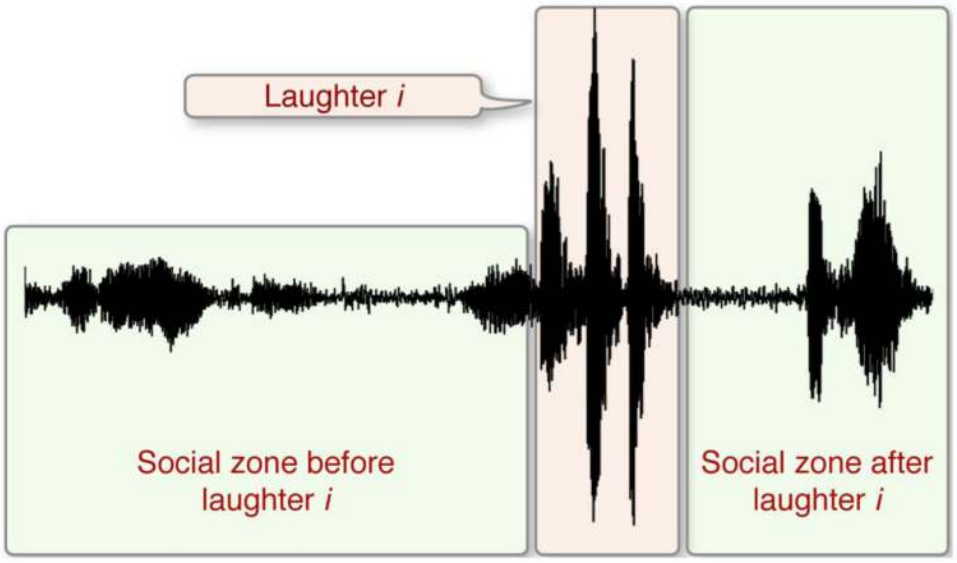

Fig. 8.
Illustration of the speech social zones surrounding the occurrence of a laughter event. BSP models attempt to predict the laughter, a social cue, from the speech in the social zones [194].

Table 1

Illustrative Applications, Some Considered in This Paper, in the Context of BSP

Domain	Activity	Expert Agent Role	Sample BSP Goals
Observational assessment of distressed couples in marital therapy	Problem solving interactions	Observational Screening/ Diagnostics by expert (not engaged in elicitation)	Computational behavior coding for supporting theory and practice
Education/Reading	Cognitive Task Performance	Observational Assessment (minimal/no direct active elicitation)	Quality and efficacy of learning
Assessment in Autism (e.g., Autism Diagnostic Observational Schedule)	Structured dialogic interactions with specific socio affective and cognitive processes	Behavior elicitation and simultaneous assessment by expert	Computational tools for supporting diagnosis for further behavior stratification
Psychotherapy for addiction	Therapeutic intervention using Motivational Interviewing	Dialogic interactions with specific therapeutic goals and structure	Quantifying effective therapy strategies predicting therapy outcomes
Commerce/Customer care/	Information Transactions	Information provisioning, service (direct interaction)	Interaction quality: satisfaction, hotspots
Meeting monitoring, focus groups	Interactions usually focused on specific topic or agenda	Often expert analysis not copresent	Behavior sensing/capture, Quantifying affective dynamics, engagement

Table 2

A List of the Acoustic LLDs and Static Functionals We Used; the Six "Basic" Functionals Are Starred (*)

LLD	speech/non-speech, f_0, intensity, 15 MFCCs, 8 MFBs, jitter, jitter-of-jitter, shimmer
Functional	mean*, median*, standard deviation*, minimum*, maximum*, range*, skewness, kurtosis, min/max positions, lower quartile, upper quartile, interquartile range, linear approximation slope coeff.

Table 3

Interaction Attribute Classification Based on Acoustic LLDs. Classifiers: SVM or LR Using L1 or L2 Error Metric. Details Can Be Found in [161]

Results from Various Acoustic Classifiers					
% correct	**SVM-l^2**	**SVM-l^1**	**LR-l^2**	**LR-l^1**	**Average**
acceptance	71.1	73.9	**77.5**	76.1	**74.6**
blame	79.3	**80.7**	78.9	79.3	**79.6**
humor	64.0	54.3	**66.8**	61.5	**61.6**
negative	82.2	78.9	**82.5**	80.4	**81.0**
positive	73.6	71.1	**73.9**	70.0	**72.2**
sadness	**61.8**	55.4	61.1	59.0	**59.3**

Table 4

Interaction Attribute Classification Using Lexical Analysis on Clean Transcripts, as a Function of UBM Interpolation Weights ( ). Best Score in Each Row Is Boldfaced. Further Details in [71]

	Results on reference transcript (% correct)										
**code vs **	0.01	0.05	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.95	0.99
acceptance	**91.4**	91.0	91.0	90.0	90.3	89.2	88.5	87.5	86.4	75.3	60.5
blame	91.0	91.4	**91.8**	91.0	90.3	89.2	89.2	88.5	88.2	78.1	63.4
humor	71.3	**72.4**	72.0	71.3	69.5	69.9	67.5	67.0	65.2	61.6	57.3
negative	83.8	84.9	**86.7**	**86.7**	86.4	85.7	86.0	86.0	85.3	74.9	60.2
positive	89.6	89.6	89.6	**88.9**	87.5	87.8	87.8	87.5	87.8	76.7	63.8
sadness	59.0	61.6	60.9	**61.3**	60.6	60.2	58.8	59.5	59.1	57.7	58.5

Table 5

Interaction Attribute Classification Using Lexical Analysis of the Noisy Audio Signal Through Lattice Generation, as a Function of UBM Interpolation Weights Best Score in Each Row Is Boldfaced. Further Details in [71]

Results through ASR lattices (% correct)											
**code vs **	**0.01**	**0.05**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.95**	**0.99**
acceptance	71.4	72.9	75.4	73.6	73.6	73.2	71.8	71.1	68.9	64.6	63.6
blame	75.0	76.8	77.9	78.6	78.2	77.9	76.8	76.4	73.9	67.5	63.9
humor	57.9	58.6	58.6	57.5	57.1	56.4	57.9	56.1	55.4	55.0	50.7
negative	64.3	66.1	69.6	71.1	70.4	69.3	69.3	67.9	65.7	60.7	58.9
positive	72.9	73.2	74.6	74.6	72.5	72.9	73.9	73.6	71.4	66.1	64.6
sadness	52.5	55.0	55.7	52.1	50.4	50.7	51.1	51.8	52.1	54.3	52.1

Table 6

The Unigrams With Most Impact Toward the Correct Classification of One Test Sample as a Blaming Session

Most blaming words in terms of discriminative contribution				Least blaming words in terms of discriminative contribution			
Word	**No BI.**	**Blame log prob**	**Δ**	**Word**	**No BI.**	**Blame log prob**	**Δ**
YOU	-95.49	-85.88	-9.61	**EXPECTS**	-16.70	-17.84	1.14
YOUR	-51.24	-47.18	-4.06	**CONSIDERATION**	-16.11	-17.31	1.21
ME	-40.27	-37.74	-2.53	**KNOW**	-35.10	-36.62	1.53
TELL	-33.97	-32.46	-1.51	**INABILITY**	-16.76	-18.32	1.55
ACCEPT	-25.44	-23.99	-1.45	**SESSION**	-20.51	-22.07	1.56
CARING	-27.05	-25.91	-1.14	**OF**	-44.50	-46.26	1.76
KITCHEN	-21.22	-20.21	-1.02	**ANTICIPATION**	-22.22	-24.21	2.00
TOLD	-29.04	-28.19	-0.85	**THINK**	-35.70	-37.77	2.07
NOT	-40.32	-39.59	-0.73	**WE**	-29.39	-31.75	2.36
WHAT	-51.47	-50.77	-0.69	**I**	-99.92	-102.49	2.57
INTIMACY	-43.16	-42.53	-0.63	**THAT**	-91.30	-93.97	2.67
IT	-42.70	-42.18	-0.52	**UM**	-64.75	-70.76	6.01

Proc IEEE Inst Electr Electron Eng. Author manuscript; available in PMC 2013 September 11.

Table 7

Accuracy With Which a Husband's or Wife's Attitude Toward an Interaction Can Be Classified as Positive Versus Negative Based on a Markov Model of Pitch Entrainment [99]

Model	Accuracy (%)
Chance	50%
Full set features	71%
Reduced set features	76%