Detection and computational analysis of psychological signals using a virtual human interviewing agent

A A Rizzo, S Scherer, D DeVault, J Gratch, R Artstein, A Hartholt, G Lucas, S Marsella, F Morbini, A Nazarian, G Stratou, D Traum, R Wood, J Boberg, L-P Morency

Institute for Creative Technologies, University of Southern California,
12015 East Waterfront Dr., Playa Vista, Los Angeles, California, USA

rizzo@ict.usc.edu, scherer@ict.usc.edu, devault@ict.usc.edu, gratch@ict.usc.edu, artstein@ict.usc.edu, lucas@ict.usc.edu, marsella@ict.usc.edu, morbini@ict.usc.edu, angelana@usc.edu, stratou@ict.usc.edu, traum@ict.usc.edu, rwood@ict.usc.edu,boberg@ict.usc.edu, morency@ict.usc.edu

http://ict.usc.edu/

ABSTRACT

It has long been recognized that facial expressions, body posture/gestures and vocal parameters play an important role in human communication and the implicit signalling of emotion. Recent advances in low cost computer vision and behavioral sensing technologies can now be applied to the process of making meaningful inferences as to user state when a person interacts with a computational device. Effective use of this additive information could serve to promote human interaction with virtual human (VH) agents that may enhance diagnostic assessment. The same technology could also be leveraged to improve engagement in teletherapy approaches between remote patients and care providers. This paper will focus on our current research in these areas within the DARPA-funded "Detection and Computational Analysis of Psychological Signals" project, with specific attention to the *SimSensei* application use case. *SimSensei* is a virtual human interaction platform that is able to sense and interpret real-time audiovisual behavioral signals from users interacting with the system. It is specifically designed for health care support and leverages years of virtual human research and development at USC-ICT. The platform enables an engaging face-to-face interaction where the virtual human automatically reacts to the state and inferred intent of the user through analysis of behavioral signals gleaned from facial expressions, body gestures and vocal parameters. Akin to how non-verbal behavioral signals have an impact on human to human interaction and communication, SimSensei aims to capture and infer from user non-verbal communication to improve engagement between a VH and a user. The system can also quantify and interpret sensed behavioral signals longitudinally that can be used to inform diagnostic assessment within a clinical context.

1. INTRODUCTION

It has long been recognized that facial expression and body gestures play an important role in human communicative signalling (Ekman and Rosenberg, 1997; Russell and Fernandez-Dols, 1997). As well, vocal characteristics (e.g., prosody, pitch variation, etc.) have also been reported to provide additive information regarding the "state" of the speaker beyond the actual language content of the speech (Pentland et al, 2009). While some researchers postulate that the universal expression and decoding of face/body gestures and vocal patterns are indicative of genetic "hardwired" mammalian neural circuitry as Darwin proposed over a hundred years ago (Darwin, 2002), others have placed less emphasis on investigating underlying mechanisms and instead have focused on the empirical analysis of such implicit communication signals and what can be meaningfully derived from them. In the latter category, Pentland's MIT research group has characterized these elements of behavioral expression as "Honest Signals" (Pentland, 2008). Based on his research with groups of people interacting, he suggests: *"…this second channel of communication, revolving not around words but around social relations, profoundly influences major decisions in our lives—even though we are largely unaware of it."* Pentland posits that the physical properties of this signalling behavior are constantly activated, not simply as a back channel or complement to our conscious language, but rather as a separate communication network. It is conjectured that these signalling behaviors, perhaps evolved from ancient primate non-verbal communication mechanisms, provide a useful window into our intentions, goals, values and emotional state.

Based on this perspective, an intriguing case can be made for the development of a computer-based sensing system that can capture and quantify such behavior, and from that activity data make inferences as to a user's cognitive and emotional state. Inferences from these sensed signals could be used to supplement information that is garnered exclusively from the literal content of speech for a variety of purposes. This is one of the major premises of the interdisciplinary research area of affective computing that focuses on the study and development of systems and devices that can recognize, interpret, process, and simulate human affective states. This vision had been discussed early on in the human computer interaction (HCI) literature in the context of perceptual user interfaces (PUI) (Turk and Robertson, 2000). PUIs are user interfaces that maximize the bandwidth of communication between a user and a computational device with such sensing technologies, and aims to enable a user experience with computers that is more similar to the way that people interact with each other face to face. The expectation that PUIs could provide incremental value over traditional HCI methods rests on the premise that more sophisticated forms of bi-directional interaction between a computational device and a human user will produce a more naturalistic engagement between these two complex systems. This is not a new idea, and one can find references to these concepts going back to Picard (1995), and the concept was well summarized on the IBM Almaden legacy website in 2000: *"Just as a person normally expects a certain kind of engagement when interacting with another person, so should a person be able to expect similar engagement when interacting with a computational device. Such engagement requires the computer to carefully observe the user, anticipating user actions, needs, and desires. Such engagement enables users to begin to build personal relationships with computers."* (Blue eyes: Suitor [WWW Document]. URL http://www.almaden.ibm.com/cs/blueeyes/suitor.html, visited 2001, February 2).

Recent progress in low cost sensing technologies and computer vision methods have now driven this concept closer to reality and the use cases for such applications can now be extended beyond enhancing basic HCI. Indeed, recent widespread availability of low cost sensors (webcams, Microsoft Kinect, microphones) combined with software advances for facial feature tracking, articulated body tracking, and voice analytics (Baltrusaitis et al, 2012; Morency et al, 2008; Whitehill et al, 2009) has opened the door to new applications for automatic nonverbal behavior analysis. For example, computer vision systems and voice analytic algorithms that are available during a standard clinical session could assist clinicians and health care providers in their daily activities by providing additive predictive information as to patient "state" to supplement the clinician's awareness of subtle behaviors that could enhance clinical decision making. Such automatic behavior descriptors could be unobtrusively captured across the course of a clinical session and this quantitative information on behavior dynamics and intensities could be available to the clinician in real time (via earphones or a personal monitor) as well as providing deeper quantitative analysis for post-session review and longitudinal analysis across multiple sessions. Another promising area is in the enhancement of engagement in Teleheath/Teletherapy approaches between remote patients and care providers. Such new perceptual software could assist clinicians during teletherapy sessions where the capture and delivery technology may provide less than optimal or impoverished audiovisual communication cues relative to those provided in direct face-to-face interactions. In this teletherapy case, sensed behavioral cues could be analyzed and delivered in the form of a real time decision support visualizations to aid the clinicians' awareness of patient state (See Figure 1a). Moreover, short of direct delivery of this information to a clinician, the sensing and quantification of nonverbal behavioral cues can also provide input to an interactive virtual human coach that would be able to offer advice based on perceived indicators of user distress or anxiety during a short interview. This is the primary effort that we will detail in this paper with our presentation of the "SimSensei" interviewing agent (See Figure 1b).

Figure 1. *(a) Telecoach interface concept (on left) and* ***(b)*** *SimSensei virtual health agent (on right).*

2. SIMSENSEI AND MULTISENSE

SimSensei is one application component of our recent research and development within the DARPA-funded "Detection and Computational Analysis of Psychological Signals (DCAPS)" project. This DCAPS application has aimed to explore the feasibility of creating "empathic" virtual human health agents for mental health screening. The

private kiosk-based SimSensei system was envisioned to be capable of conducting interviews with patients who may be initially hesitant or resistant to seeking traditional mental health care with a live provider (See Figure 1b). The system seeks to combine the advantages of traditional web-based self-administered screening (Weisband and Kiesler, 1996), which allows for anonymity, with anthropomorphic interfaces which may foster some of the beneficial social effects of face-to-face interactions (Kang and Gratch, 2012). SimSensei evolves an earlier web-based screening tool, SimCoach (Rizzo et al, 2011), and can engage users in a structured interview using natural language and nonverbal sensing with the aim of identifying behaviors associated with anxiety, depression or PTSD.

The SimSensei capability to accomplish this was supported by the "MultiSense" perception system (Morency et al, http://multicomp.ict.usc.edu/?p=1799; Devault et al, 2014). This is a multimodal system that allows for synchronized capture of different modalities such as audio and video, and provides a flexible platform for real time tracking and multimodal fusion. This is a very important aspect of the system in that it enables fusion of modality "markers" to support the development of more complex multimodal indicators of user state. MultiSense dynamically captures and quantifies behavioral signals such as 3D head position and orientation, type, intensity and frequency of facial expressions of emotion (e.g., fear, anger, disgust and joy), fidgeting, slumped body posture, along with a variety of speech parameters (speaking fraction, speech dynamics, latency to respond, etc.). These informative behavioral signals serve two purposes. First, they produce the capability of analyzing the occurrence and quantity of behaviors to inform assessment. Second, they are broadcast to the other components of SimSensei Kiosk to inform the virtual human interviewer of the state and actions of the participant and assist with turn taking, listening feedback, and building rapport by providing appropriate non-verbal feedback. MultiSense serves to fuse information from web cameras, the Microsoft Kinect and audio capture and processing hardware to identify the presence of any nonverbal indicators of psychological distress and to provide moment-to-moment inferences to the SimSensei virtual agent "who" may act upon that information to provide supportive feedback, deliver acknowledging gestures/facial expressions and drive follow on questions. In depth technical details of the Multisense software as well as the SimSensei dialog management, natural language system, and agent face and body gesture generation methods are beyond the scope of this article and can be found elsewhere (DeVault et al, 2014; Scherer et al, 2013).

3. NON-VERBAL BEHAVIOR AND CLINICAL CONDITIONS

To begin to develop a corpus of automatic nonverbal behavior descriptors that Multisense could track for the SimSensei application, we searched the large body of research that has examined the relationship between nonverbal behavior and clinical conditions. Most of this research resided in the clinical and social psychology literature and until very recently the vast majority relied on manual annotation of gestures and facial expressions. Despite at least forty years of intensive research, there is still surprisingly little progress on identifying clear relationships between patient disorders and expressed behavior. In part, this is due to the difficulty in manually annotating data, inconsistencies in how both clinical states and expressed behaviors are defined across studies, and the wide range of social contexts in which behavior is elicited and observed. However, in spite of these complexities, there is general consensus on the relationship between some clinical conditions (especially depression and social anxiety) and associated nonverbal cues. These general findings informed our initial search for automatic nonverbal behavior descriptors.

For example, gaze and mutual attention are critical behaviors for regulating conversations, so it is not surprising that a number of clinical conditions are associated with atypical patterns of gaze. Depressed patients have a tendency to maintain significantly less mutual gaze (Waxer, 1974), show nonspecific gaze, such as staring off into space (Schelde, 1998) and avert their gaze, often together with a downward angling of the head (Perez and Riggio, 2003). The pattern for depression and PTSD is similar, with patients often avoiding direct eye contact with the clinician. Emotional expressivity, such as the frequency or duration of smiles, is also diagnostic of clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity (Perez and Riggio, 2003; Bylsam et al, 2008), fewer mouth movements (Fairbanks et al, 1982; Schelde, 1998), more frowns (Fairbanks et al, 1982; Perez and Riggio, 2003), and fewer gestures (Hall et al, 1995; Perez and Riggio, 2003). Some findings suggest it is not the total quantity of expressions that is important, but their dynamics. For example, depressed patients may frequently smile, but these are perceived as less genuine and often shorter in duration (Kirsch and S. Brunnhuber, 2007) than what is found in non-clinical populations. Social anxiety and PTSD while sharing some of the features of depression, also have a tendency for heightened emotional sensitivity and more energetic responses including hypersensitivity to stimuli: e.g., more startle responses, and greater tendency to display anger (Kirsch and S. Brunnhuber, 2007), or shame (Menke, 2011). Fidgeting is often reported with greater frequency in clinical populations. This includes gestures such as tapping or rhythmically shaking hands or feet and has been reported in both anxiety and depression (Fairbanks et al, 1982). Depressed patients also often engage in "self-adaptors" (Ekman and Friesen, 1969), such as rhythmically touching, hugging or stroking parts of the body or self-grooming, such as repeatedly stroking the

hair (Fairbanks et al, 1982). Examples of observed differences in verbal behavior in depressed individuals include increased speaker-switch durations and diminished variability in vocal fundamental frequency (Cohn et al, 2009), decreased speech output, slow speech, delays in delivery, and long silent pauses (Hall et al, 1995). Differences in certain lexical frequencies have been reported including use of first person pronouns and negatively-valenced words (Rude et al, 2004).

One recent brewing controversy within the clinical literature is whether certain specific categories of mental illness (e.g., depression, PTSD, anxiety, and schizophrenia) reflect discrete and clearly separable conditions or rather, continuous differences along some more general underlying dimensions (Russell and Barrett, 1999). This parallels controversies in emotion research as to whether emotions reflect discrete and neurologically distinct systems in the brain, or if they are simply labels we apply to differences along broad dimensions such as valence and arousal. Indeed, when it comes to emotion recognition, dimensional approaches may lead to better recognition rates than automatic recognition techniques based on discrete labels. The broad dimension receiving the most support in clinical studies is the concept of general psychological distress. For example, (Elhai et al, 2011) examined a large number of clinical diagnostic interviews and found that diagnoses of major depression and PTSD were better characterized by considering only a single dimension of general distress. Several other researchers have statistically re-examined the standard scales and interview protocols used to diagnose depression, anxiety and PTSD and found they highly correlate and are better seen as measuring general distress (Bieling et al, 1998; Marshall et al, 2010; Arbisi, et al, 2012). For this reason, we have investigated if general distress may be a more appropriate concept for recognizing clinical illness in addition to the more conventional discrete categories.

Thus, the key challenge when building such nonverbal perception technology for clinical applications is to develop and validate robust descriptors of human behaviors that are correlated with psychological distress. These descriptors should be designed to probabilistically inform diagnostic assessment or quantify treatment outcomes. However, no descriptor is completely diagnostic by itself, but rather may reveal "tendencies" in user's non-verbal behaviors that are informational to enhance clinical hypothesis testing and/or decision making. As a first step, we relied on three main sources of information to identify such behaviors: a literature review on nonverbal behaviors indicative of psychological conditions as reported by clinical observations and by existing work on automatic analysis (Fairbanks et al, 1982; Hall et al, 1995; Kirsch and Brunnhuber, 2007; Perez and Riggio, 2003), a qualitative analysis based on observations from the videos, and consultation with experts (including trained clinicians) who looked at the data and identified the communicative behaviors that they would use to form a diagnosis. As a next step, selected behaviors were quantified on the face-to-face corpus via manual annotation. The selection criteria for which behaviors to prioritize for annotation was based on diagnostic power and implementability. Initially, face-to-face interview data is utilized as a study ground to identify nonverbal behaviors that are correlated with depression, PTSD, and anxiety. Following the analysis of face-to-face human interactions to identify potential emotional indicators, dialogue policies, and commonality of human gestures, the development and analysis of a Wizard-of-Oz (WoZ) prototype system was required. The WoZ interaction allowed human operators to choose the spoken and gestural responses of a virtual human character (similar to digital puppetry) that interacted with a live research participant. The final step involved the development of a fully automatic virtual interviewer (SimSensei) that is able to engage users in 15-25 minute interactions.

4. DYADIC FACE-TO-FACE INTERACTION DATASET

The fundamental novel research challenge in this project is to endow computers with the ability to recognize clinically-relevant information from the nonverbal behavior of patients. Computer vision and audio signal processing techniques have shown growing success in identifying a number of important nonverbal cues but the limitation of state-of-the-art approaches is that they are data hungry: they require large amounts of annotated data. Thus, our initial milestone was to collect a large dataset of clinical interviews with participants known to have a high likelihood of PTSD, social anxiety and depression and to identify and annotate their nonverbal behaviors relevant to finding indicators of these clinical states.

4.1 Participants

One hundred and seventy seven participants were recruited from two distinct populations. 120 participants (86 male) were recruited from the Los Angeles general population through Craigslist, an online job posting service. 57 participants (49 male) were recruited from *U.S. Vets*, a non-profit organization that helps very troubled military veterans re-integrate into civilian life after leaving the service and has programs tailored for veterans with PTSD and depression. Participants were informed that we are interested in their experience with PTSD and depression.

4.2 Procedure

After obtaining informed consent, participants were led to a computer and, in private, completed a series of web-delivered psychometric scales to assess clinically-relevant states and traits. These included the PTSD Checklist-Civilian version (PCL-C) to assess PTSD (Blanchard et al, 1996), Patient Health Questionnaire-Depression 9 (PHQ-9) to assess depression (Kroenke et al, 2001) the State/Trait Anxiety Inventory (STAI) to assess state anxiety (Spielberger et al, 1970), the PANAS to assess current mood (Watson et al, 1988) and the Balanced Inventory of Desirable Responding (BIDR) to assess tendencies to be deceptive in such interview contexts (Paulhus, 1988). The web-based assessment was followed by a 30-minute structured interview that explored if they had been previously diagnosed and are currently experiencing symptoms of PTSD, depression and anxiety and to elicit data relating clinical states with nonverbal behavior. Two possible interviewers conducted the interview which consisted of three phases, as warm-up phase consisting of basic questions designed to establish rapport (e.g., "How is your day going? Where are you from"), an interview phase where participants were asked to elaborate on some of their responses to the scales (e.g., "On the survey you mention you often experience disturbing thoughts; can you tell me a little more about that?"), followed by a wind-down phase designed to return the participant to a more pleasant state of mind (e.g., "If you could travel to any destination, where would you go?"). During the interview phase, both the participant and the interviewer are fitted with a lapel microphone and are recorded with video cameras and the Kinect system to track their body posture. The video cameras and Kinect are placed between the participants. Following the interview, we assessed the quality of the interaction with measures of rapport and social presence.

4.3 Subject Variable Summary Statistics

Overall, about 32% of the subjects were assessed positive for PTSD, 29% for depression, and 62% for trait anxiety. Participants from U.S. Vets were assessed positive more often for each of the disorders, as was expected, and they were demonstrably different from the Craigslist population in several ways. Demographically, U.S. Vets subjects were older, less educated, more likely to be male, and less likely to be employed. They were also much more likely to have been a member of the armed forces, as expected, since we intentionally chose that population for our experiment. Subjects with assessed disorders were significantly different on several measures: They scored significantly higher in neuroticism and were more anxious before the interview. Consistent with the findings on general distress discussed earlier, we observed significant correlations ($p<0.01$) between the disorders (i.e. PTSD, anxiety, and depression). Diagnosis for depression correlated with PTSD ($\rho =0.64$, using Pearson's correlation), depression correlated with anxiety ($\rho = 0.40$) and PTSD correlated with anxiety ($\rho=0.43$). When conserving the scalar severity measure of the three inventories, we found even stronger correlations ($\rho >0.8$). Based on the prior findings on general distress and the comorbidity observed in this dataset, we concluded that at this stage in the research, automatic recognition techniques should focus on recognizing indicators of general distress rather than attempting to distinguish individual conditions. As a result, we used factor analysis to identify a single indicator of distress that is used in subsequent training and analysis.

4.4 Clinical Cue Results

The dataset was annotated with manual and automatic techniques to identify nonverbal (audio and visual) behaviors that might be associated with generalized distress. All manual annotators were trained until they reached high inter-coder agreement. Manual features include hand self-adaptors (i.e., self-touching) and leg fidgeting. Automatic features included head orientation, gaze angle, smile intensity and duration and several features related to vocal quality.

We found several statistically significant differences in the behavior of participants between those that scored positive for general distress and normal controls. (1) There are significant differences in the automatically estimated gaze behavior of subjects with psychological disorders. In particular, an increased overall downwards angle of the gaze could be automatically identified using two separate automatic measurements, for both the face as well as the eye gaze. (2) We could also identify on average significantly less intense smiles for subjects with psychological disorders as well as significantly shorter average durations of smiles (see Figure 2). (3) Based on the manual analysis, subjects with psychological conditions exhibit on average longer self-touches and fidget on average longer with both their hands (e.g. rubbing, stroking) as well as their legs (e.g. tapping, shaking).

We found several significant differences in the vocal patterns of participants with general distress related to the 'coloring' of the voice when compared with normal controls (for this we only analyzed male participants to control for differences in vocal quality that arise from gender). We examined differences in vocal fundamental frequency, speech intensity, measures of monotonicity (i.e. intensity variations and spectral stationarity), and measures of the voice's breathiness (e.g. normalized amplitude quotient (NAQ)). The most promising findings are that the speech intensity variations of distressed subjects are significantly reduced and their voice quality is significantly breathier based on the observed NAQ parameter. These results replicate findings in the

psychological literature and give us confidence that these indicators can be identified automatically in real-time interactions using low-cost sensors.

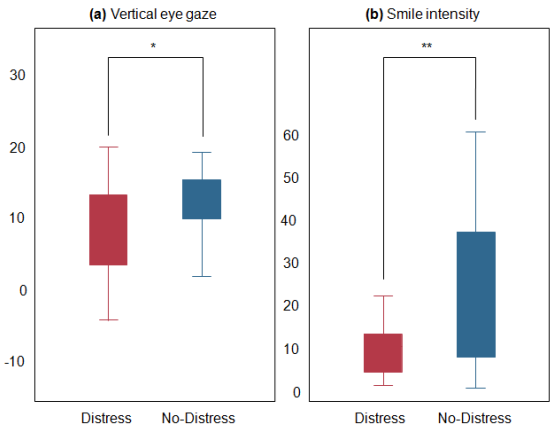

Figure 2. *Example of two automatic behavior descriptors. Boxplots show significantly stronger overall downward angle of the (a) eye gaze ($p<0.05$) and (b) a significantly lowered average smile intensity ($p<0.01$) for "distressed" participants.*

Table 1. *Evaluation of Automatic Non-verbal Behavior Analysis. Comparison across clinical conditions (distressed, depression, anxiety, PTSD).*

	Tier	Condition μ (σ)	No-Condition μ (σ)	p	g
Distress	VHead Gaze	0.14 (0.11)	0.19 (0.10)	**0.04**	-0.46
	VEye Gaze	8.93 (7.92)	13.65 (6.30)	**0.01**	-0.64
	Smile Int.	12.31 (10.09)	23.76 (18.30)	**<0.01**	-0.75
	Smile Dur.	2.49 (0.87)	3.43 (1.85)	**0.01**	-0.63
Depression	VHead Gaze	0.15 (0.11)	0.17 (0.09)	0.20	-0.19
	VEye Gaze	9.83 (7.35)	11.56 (6.69)	0.16	-0.25
	Smile Int.	12.81 (11.14)	19.94 (16.85)	**0.04**	-0.45
	Smile Dur.	2.59 (0.87)	3.02 (1.69)	0.15	-0.27
Anxiety	VHead Gaze	0.15 (0.10)	0.19 (0.10)	0.06	-0.36
	VEye Gaze	10.05 (7.87)	12.87 (4.04)	**0.04**	-0.41
	Smile Int.	14.77 (13.33)	23.52 (18.15)	**<0.01**	-0.56
	Smile Dur.	2.66 (1.25)	3.33 (1.87)	**0.03**	-0.44
PTSD	VHead Gaze	0.14 (0.11)	0.18 (0.09)	**0.04**	-0.39
	VEye Gaze	9.37 (8.12)	11.86 (6.11)	0.07	-0.36
	Smile Int.	12.25 (10.78)	20.85 (17.11)	**0.01**	-0.55
	Smile Dur.	2.37 (0.81)	3.17 (1.73)	**0.02**	-0.52

5. WIZARD-OF-OZ AND AUTOMATIC VH AGENT INTERVIEW DATASETS

The next step was to conduct a Wizard-of-Oz (WoZ) study where participants interacted with a female VH character named "Ellie" whose speech and behavior responses were controlled by two "behind the curtain" operators. In this setup, a fixed set of 191 speech utterances and 23 nonverbal behaviors were defined and made available to two Wizards who jointly controlled Ellie's behavior (one controlled speech, the other controlled behavior). This two-wizard arrangement was necessary as the task of controlling both Ellie's verbal and nonverbal behavior proved difficult for a single wizard to coordinate. In addition to asking the relevant interview questions, these options provided the Wizard- controlled Ellie with a finite, circumscribed repertoire of response options to try to act as a good listener. Ellie could also provide backchannel activity, empathy and surprise responses, and continuation prompts. The set of options that was made available to the two Wizards is summarized in Table 2.

Table 2. *Wizard-of-Oz Option Set.*

Option Type	Example
nonverbal behaviors	head nod to indicate agreement
interview questions	*what are you like when you don't get enough sleep?*
neutral backchannels	*uh huh*
positive empathy	*that's great*
negative empathy	*i'm sorry*
surprise responses	*wow!*
continuation prompts	*could you tell me more about that?*
miscellaneous	*i don't know; thank you*

A sample of 140 participant interactions were collected using the WoZ system applying the same methodology, sample sources, and assessment devices used in the previous face-to-face condition. Analysis of these dialogues confirmed the presence of significant differences in the non-verbal behavior of distressed participants when

Proc. 10th Intl Conf. Disability, Virtual Reality & Associated Technologies
Gothenburg, Sweden, 2–4 Sept. 2014

compared to non-distressed participants (Scherer, et al, 2013ab; Stratou et al, 2013) and also differences in the verbal behavior of distressed participants when compared to non-distressed participants (DeVault et al, 2013). These significant differences confirmed that the finite set of wizard utterances and non-verbal behavior options was adequate to conduct interviews that could elicit different responses and behaviors from distressed individuals than from non-distressed individuals. WoZ results in the context of comparison with face-to-face and Automatic VH agent are presented in the next section.

6. COMPARATIVE EVALUATION ACROSS INTERVIEWS: FACE-TO-FACE, WOZ, AND AUTOMATIC INTERACTION (AI) WITH A VH AGENT

The next step in the development of the system was integration into a SimSensei Kiosk. More specifically, the perception system's functionality was tuned to automatically track and recognize nonverbal behaviors that are important for psychological condition assessment, as reported from the previous steps, but in the context of an interview with an autonomous VH agent (still Ellie). The key sensed behaviors associated with depression, anxiety, and PTSD were extracted live during the interview, were used to guide Ellie's interactive behavior and the summary statistics were available automatically at the end of the interview. In this stage the focus was on the capture and analysis of such behavioral signals in the real-time system and the validation of the previous analysis of face-to-face data on the new corpus of fully automated interactions. We compared the three interview datasets: face-to-face, Wizard-of-Oz, and "AI interactions" where the VH was controlled by the automated SimSensei Kiosk system (referred to as AI).

6.1 Participants and Procedures

Across all three studies, 351 participants were recruited through Craigslist and from posted flyers. Of the 120 face-to-face participants, 86 were male and 34 were female. These participants had a mean age of 45.56 (SD = 12.26). Of the 140 WoZ participants, 76 were male, 63 were female, and 1 did not report their gender. The mean age of this group of participants was 39.34 (SD = 12.52). Of the 91 AI participants, 55 were male, 35 were female, and 1 did not report their gender. They had a mean age of 43.07 (SD = 12.84).

All participants were given a series of self-report assessment instruments to index their clinical state, as described above. Post-experience, all participants completed a validated measure of rapport (Kang and Gratch, 2012). Additionally, participants in WoZ and AI completed nine questions designed to test our success in meeting specific VH design goals (see Table 3). Examples include questions about disclosure ("I was willing to share information with Ellie"), the mechanics of the interaction ("Ellie was sensitive to my body language") and willingness to recommend the system to others. All were rated on a scale from 1 (strongly disagree) to 5 (strongly agree). Note that in the WoZ condition, participants were told that the agent was autonomous and not puppeted by two people. Finally, participants in WoZ and AI also completed the standard System Usability Scale (Brooke, 1996), a measure of a product's perceived system satisfaction and usability.

6.2 Results

For all items and scales, participants' total scores were calculated for analysis. Table 3 displays mean total scores and associated standard errors for each of the subsequent analyses. With regard to the design goals, most participants agreed or strongly agreed they were achieved, whether they interacted with the Wizard-operated or AI system. For example, most people agreed or strongly agreed that they were willing to share information with Ellie (84.2% WoZ; 87.9% AI), were comfortable sharing (80.5% WoZ; 75.8% AI) and did share intimate information (79.3% WoZ; 68.2% AI). Both systems performed less well with regard to their perceived ability to sense and generate appropriate nonverbal behavior. For example, a minority of participants agreed or strongly agreed that Ellie could sense their nonverbal behavior (40.3% WoZ; 27.5% AI). However, this did not seem to seriously detract from the overall experience and majority agreed or strongly agreed they would recommend the system to a friend (69.8% WoZ; 56.1% AI).

We next examined the relative impressions of the AI system when compared with the Wizard-of-Oz. Although the AI is in no way intended to reach human-level performance, this comparison gives insight in areas that need improvement. First, we conducted t-tests to compare Wizard-of-Oz to AI on each of the individual items representing the system's design criteria. Surprisingly, results yielded only one significant difference. WoZ participants reported feeling that the interviewer was a better listener than the AI participants ($t(166) = 3.94$, $p < .001$, $d = 0.61$). Next, we conducted t-tests comparing WoZ to AI on System Usability scores and on ratings of rapport. WoZ participants rated the system as higher in usability than AI participants ($t(229) = 3.24$, $p = .001$, $d = 0.44$) and also felt more rapport ($t(229) = 3.28$, $p = .001$, $d = 0.44$).

Finally, we examined how the WoZ and AI systems compared with the original face-to-face interviews (see Table 4). We conducted an ANOVA to compare ratings of rapport for the three methods. Results revealed a significant effect of method on rapport ($F(2, 345) = 14.16$, $p < .001$, $d = 0.52$). Interestingly, this effect was driven by the WoZ. WoZ participants felt greater rapport than AI participants ($t(345) = 3.87$, $p < .001$, $d = 0.42$ and compared to face-to-face participants ($t(345) = -4.95$, $p < .001$, $d = 0.53$). Surprisingly, AI and face-to-face participants' ratings of rapport did not differ ($t(345) = -0.77$, $p = .44$, $d = 0.07$.).

Table 3. *Means, Standard Errors, t-values and effect sizes on design questions * = <.05.*

Design Goals	Method		t-value	d
	WoZ	AI		
I was willing to share information with Ellie	4.03 (0.83)	4.07 (0.73)	-0.33	0.05
I felt comfortable sharing information with Ellie	3.92 (0.98)	3.80 (1.07)	0.75	0.12
I shared a lot of personal information with Ellie	3.97 (1.04)	3.73 (1.14)	1.47	0.23
It felt good to talk about things with Ellie	3.69 (1.02)	3.60 (0.95)	0.55	0.08
There were important things I chose to not tell Ellie	2.93 (1.19)	2.66 (1.19)	1.48	0.23
Ellie was a good listener	4.10 (0.77)	3.56 (0.98)	3.94*	0.61
Ellie has appropriate body language	3.85 (0.85)	3.84 (.86)	0.05	0.01
Ellie was sensitive to my body language	3.36 (0.72)	3.13 (0.86)	1.87	0.29
I would recommend Ellie to a friend	3.72 (1.10)	3.47 (1.03)	1.52	0.24
System Usability	74.37 (13.63)	68.68 (12.05)	3.24*	0.44
Rapport	80.71 (12.10)	75.43 (11.71)	3.28*	0.44

Table 4. *Rapport scores in the three conditions.*

Face-to-face	WoZ	AI
74.42 (4.89)	80.71 (12.10)	75.43 (11.71)

7. CONCLUSIONS

The results of this first evaluation are promising. In terms of subjective experience, participants reported willingness to disclose, willingness to recommend and general satisfaction with both the WoZ and AI versions of the system. In terms of rapport, participants reported feelings comparable to a face-to-face interview. Unexpectedly, participants felt more rapport when interacting with the WoZ system than they did in face-to-face interviews. One possible explanation for this effect is that people are more comfortable revealing sensitive information to computers than face-to-face interviewers (Weisband and Kiesler, 1996; Lucas et al, 2014), though this will require further study. As expected, the current version of SimSensei does not perform as well as human wizards. This is reflected in significantly lower ratings of rapport and system usability. Participants also felt that the AI-controlled Ellie was less sensitive to their own body language and often produced inappropriate nonverbal behaviors. It should also be noted that our current evaluation focused on subjective ratings and needs to be bolstered by other more objective measures. Such analyses are a central focus of current work. Nonetheless, the overall results are promising and suggest the system is already effective in eliciting positive use-intentions. One key advantage of our SimSensei Kiosk framework over a human interviewer is the implicit replicability and consistency of the spoken questions and accompanying gestures. This standardization of the stimuli allows a more detailed analysis of user responses to precisely delivered interview questions. Another potential advantage is that recent results suggest that virtual humans can reduce stress and fear associated with the perception of being judged and thereby lower emotional barriers to disclosing information (Hart et al, 2013; Lucas et al, 2014). Realizing this vision will require a careful and strategic design of the virtual human's behavior in future efforts.

The SimSensei system has been further refined via funding from a set of clinical projects. In one ongoing project, U.S. military service members were given a full battery of psychological tests and interviewed by the automatic SimSensei (AI) interviewer prior to a combat deployment in Afghanistan. This unit is still serving on their deployment at the time of this writing and will return in December 2014 for a post deployment round of SimSensei testing and will be studied at 6 months and one year post deployment as well. The primary goal is to determine if both verbal and non-verbal behaviors at pre and post deployment can predict mental health status in an objective fashion. In an upcoming study, the SimSensei clinical interviewer will also be used as part of the assessment package within a clinical trial testing VR Exposure Therapy for the treatment of PTSD due to military sexual trauma. The SimSensei interview will be conducted at pre-, mid- and post-treatment in order to

compare results with a sample whose mental health status is expected to improve over the course of treatment. A video of a user interacting with the AI SimSensei VH agent is available here: http://youtu.be/Yw1c5h_p6Dc

Acknowledgments. The effort described here is supported by DARPA under contract W911NF-04-D-0005 and the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

8. REFERENCES

Arbisi, PA, Kaler, ME, Kehle-Forbes, SM, Erbes, CR, Polusny, MA, and Thuras, P, (2012), The predictive validity of the PTSD checklist in a nonclinical sample of combat-exposed national guard troops, *Psychological Assessment*.

Baltrusaitis, T, Robinson, P, and Morency, L-P, (2012), 3D constrained local model for rigid and non-rigid facial tracking, *Proceedings of The IEEE Computer Vision and Pattern Recognition*, Providence, RI.

Bieling, PJ, Antony, MM, and Swinson, RP, (1998), The state–trait anxiety inventory, trait version: structure and content re-examined, *Behaviour Research and Therapy*, **36**, *7–8*, pp. 777–788.

Blanchard, EB, et al, (1996), Psychometric properties of the PTSD checklist (PCL), *Behaviour Research and Therapy*, **34**, *8*, pp. 669–673.

Brooke, J, (1996), "SUS: a "quick and dirty" usability scale", In PW Jordan, B Thomas, BA Weerdmeester, and AL McClelland, Usability Evaluation in Industry, London: Taylor and Francis.

Bylsam, LM, Morris, BH, and Rottenberg, J, (2008), A meta-analysis of emotional reactivity in major depressive disorder, *Clinical Psychology Review*, *28*, pp. 676–691.

Cohn, JF, Kruez, TS, Matthews, I, Yang, Y, Nguyen, MH, Padilla, MT, Zhou, F, and De la Torre, F, (2009), Detecting depression from facial actions and vocal prosody, In *Affective Computing and Intelligent Interaction (ACII)*, September, 2009.

Darwin, C, (2002), *The Expression of the Emotions in Man and Animals*. (3rd ed.): Oxford University Press, London.

Devault, D, Rizzo, AA, and Morency, L-P, (2014), SimSensei: A Virtual Human Interviewer for Healthcare Decision Support, *In the Proceedings of the Thirteenth International Conference on Autonomous Agents and Multiagent Systems* (AAMAS).

DeVault, D, Georgila, K, Artstein, R, Morbini, F, Traum, D, Scherer, S, Rizzo, AA, and Morency, L-P, (2013), Verbal indicators of psychological distress in interactive dialogue with a virtual human, In *Proceedings of SIGdial2013*.

Ekman, P, and Friesen, WV, (1969), The repertoire of nonverbal behavior: Categories, origins, usage, and coding, *Semiotica*, *1*, pp. 49–98.

Elhai, JD, de Francisco Carvalho, L, Miguel, FK, Palmieri, PA, Primi, R, and Frueh, CB, (2011), Testing whether posttraumatic stress disorder and major depressive disorder are similar or unique constructs, *Journal of Anxiety Disorders*, 25, *3*, pp. 404–410.

Ekman, P, and Rosenberg, EL, (1997), *What the face reveals: Basic and applied studies of spontaneous expressions using the Facial Action Coding System (FACS)*, Oxford University Press, New York.

Fairbanks, LA, McGuire, MT, and Harris, CJ, (1982), Nonverbal interaction of patients and therapists during psychiatric interviews, *Journal of Abnormal Psychology*, **91**,*2*, pp. 109–119.

Gratch, J, Artstein, R, Lucas, G, Stratou, G, Scherer, S, Nazarian, A, Wood, R, Boberg, J, DeVault, D, Marsella, S, Traum, D, Rizzo, AA, and Morency, L,-P, (2014). The Distress Analysis Interview Corpus of human and computer interviews, In *The Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, May, Reykjavik, Iceland, pp. 26-31.

Hall, JA, Harrigan, JA, and Rosenthal, R, (1995), Nonverbal behavior in clinician-patient interaction, *Applied and Preventive Psychology*, **4**,*1*, pp. 21–37.

Hart, J, Gratch, J, and Marsella, S, (2013). How Virtual Reality Training Can Win Friends and Influence People, chapter 21, Human Factors in Defence, Ashgate, pp. 235–249.

Kang, S-H, and Gratch, J, (2010), Virtual Humans Elicit Socially Anxious Interactants' Verbal Self-Disclosure, *Journal of Visualization and Computer Animation*, **21,** *3-4*, pp. 473-482.

Kang, S-H, and Gratch, J, (2012), Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories, In *The Annual Review of Cybertherapy and Telemedicine* (B Weiderhold and G Riva, Eds.), pp. 202–207. IOS Press.

Kirsch, A, and Brunnhuber, S, (2007), Facial expression and experience of emotions in psychodynamic interviews with patients with PTSD in comparison to healthy subjects, *Psychopathology*, **40**, *5*, pp. 296–302.

Kroenke, K, Spitzer, RL, and Williams, JBW, (2001), The PHQ-9, *Journal of General Internal Medicine*, **16,** *9*, pp. 606-613.

Lucas, GM, Gratch, J, King, A, and Morency, L-P, (2014), It's only a computer: Virtual humans increase willingness to disclose, *Computers in Human Behavior, 37*, pp. 94–100.

Marshall, GN, Schell, TL, and Miles, JNV, (2010), All PTSD symptoms are highly associated with general distress: ramifications for the dysphoria symptom cluster, *Journal of Abnormal Psych.*, **119**, *1*, pp. 126–135.

Menke, R, (2011), Examining nonverbal shame markers among post-pregnancy women with maltreatment histories, PhD thesis, Wayne State University.

Morency, L-P, Whitehill, J, and Movellan, J, (2008), Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation, *Proceedings of The 8th IEEE International Conference on Automatic Face Gesture Recognition* (FG08), pp. 1–8.

Paulhus, D, (1988), Balanced inventory of desirable responding (BIDR). *Acceptance and Commitment Therapy. Measures Package*, pp. 41.

Pentland, A, Lazer, D, Brewer, D, and Heibeck, T, (2009), Using reality mining to improve public health and medicine, *Studies in Health Technology and Informatics*, *149*, pp. 93-102.

Pentland, A, (2008), *Honest signals: How they shape our world*, MIT Press, Cambridge, MA.

Perez, JE, and Riggio, RE, (2003), Nonverbal social skills and psychopathology, In Nonverbal Behavior in Clinical Settings (P Philpot, RS Feldman, EJ Coats, Eds), Oxford University Press, London, pp. 17-44.

Picard, R, (1995), *Affective computing: Technical Report 321*, MIT Media Laboratory: Perceptual Computing, Cambridge, MA.

Rizzo, A, Sagae, K, Forbell, E, Kim, E, Lange, B, Buckwalter, JG, Williams, J, Parsons, TD, Kenny, P, Traum, D, Difede, J, and Rothbaum, BO, (2011), SimCoach: An Intelligent Virtual Human System for Providing Healthcare Information and Support, *In The Proceedings of the Interservice/Industry Training, Simulation and Education Conference*, (I/ITSEC), NTSA.

Rude, S, Gortner, E-M, and Pennebaker, J, (2004), Language use of depressed and depression-vulnerable college students, *Cognition and Emotion*, 18, *8*.

Russell, JA, and Barrett, LF, (1999), Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant, *Journal of Personality and Social Psychology*, **76**, 5, pp. 805–819.

Russell, JA, and Fernandez-Dols, JM, 1997, What does a facial expression mean? In *The Psychology of Facial Expression* (JA Russell and JM Fernandez-Dols, Eds), Cambridge Univ. Press, NY, pp. 3-30.

Schelde, JTM, (1998), Major depression: Behavioral markers of depression and recovery, *The Journal of Nervous and Mental Disease*, 186, *3*, pp. 133–140.

Scherer, S, Stratou, G, Boberg, J, Gratch, J, Rizzo, A, and Morency, L-P, (2013a), Automatic Behavior Descriptors for Psychological Disorder Analysis. *In the Proceedings of the IEEE Conf. on Automatic Face and Gesture Recognition.*

Scherer, S, Stratou, G, Gratch, J, and Morency, L-P, (2013b), Investigating voice quality as a speaker-independent indicator of depression and PTSD, *In Proceedings of Interspeech 2013*, pp. 847–851.

Stratou, G, Scherer, S, Gratch, J, and Morency, L-P, (2013), Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences, In *Affective Computing and Intelligent Interaction*.

Spielberger, CD, et al, (1970), *Manual for the State-Trait Anxiety Inventory*, Consulting Psychologists Press.

Turk, M, Robertson, G, (2000), Perceptual user interfaces, *Communications of the ACM*, **43**, *3*, pp. 32–70.

Watson, DA, Tellegen, A, and Clark, LA, (1988), Development and validation of brief measures of positive and negative affect: The PANAS scale, *Journal of Personality and Social Psychology*, **54,** pp. 1063-1070.

Waxer, P, (1974), Nonverbal cues for depression, *Journal of Abnormal Psychology*, **83**, *3*, pp. 319–322.

Weisband, S, and Kiesler, S, (1996), Self-disclosure on computer forms: Meta-analysis and implications, In *Proceedings of CHI1996*, *96*, pp. 3–10.

Whitehill, J, Littlewort, G, Fasel, I, Bartlett, M, and Movellan, J, (2009), Toward practical smile detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, pp. 2106–2111.