SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support

David DeVault* , Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo† , Louis-Philippe Morency†
USC Institute for Creative Technologies, 12015 Waterfront Dr., Playa Vista, CA 90094, USA
http://simsensei.ict.usc.edu/

ABSTRACT

We present SimSensei Kiosk, an implemented virtual human interviewer designed to create an engaging face-to-face interaction where the user feels comfortable talking and sharing information. SimSensei Kiosk is also designed to create interactional situations favorable to the automatic assessment of distress indicators, defined as verbal and nonverbal behaviors correlated with depression, anxiety or post-traumatic stress disorder (PTSD). In this paper, we summarize the design methodology, performed over the past two years, which is based on three main development cycles: (1) analysis of face-to-face human interactions to identify potential distress indicators, dialogue policies and virtual human gestures, (2) development and analysis of a Wizard-of-Oz prototype system where two human operators were deciding the spoken and gestural responses, and (3) development of a fully automatic virtual interviewer able to engage users in 15-25 minute interactions. We show the potential of our fully automatic virtual human interviewer in a user study, and situate its performance in relation to the Wizard-of-Oz prototype.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Natural language interfaces*

Keywords

virtual humans; dialogue systems; nonverbal behavior

1. INTRODUCTION

In this paper we present SimSensei Kiosk, an implemented virtual human interview framework designed to create an engaging face-to-face interaction where the user feels comfortable talking and sharing information. The SimSensei Kiosk is embodied in a virtual human named Ellie, pictured in Figure 1. Ellie conducts semi-structured interviews that are intended to create interactional situations favorable to the automatic assessment of distress indicators, defined as verbal and nonverbal behaviors correlated with depression, anxiety or post-traumatic stress disorder (PTSD).

Figure 1: SimSensei Kiosk, our virtual human interviewer for healthcare decision support

The vision behind the development of our SimSensei Kiosk framework is to create clinical decision support tools that complement existing self-assessment questionnaires by giving healthcare providers objective measurements of the user verbal and nonverbal behaviors that are correlated with psychological distress. These distress indicators can allow the clinician or healthcare provider to make a more informed diagnosis. When used over a longer period of time, the distress indicators might also be compared with the previous week or month, allowing health care providers to detect if a change is happening. 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 [13]. Realizing this vision requires a careful and strategic design of the virtual human's behavior.

In this paper, we summarize the design methodology in three main development cycles: (1) analysis of face-to-face

*Corresponding author for this paper: devault@ict.usc.edu
†Principal Investigators (PIs) of the SimSensei project: morency@ict.usc.edu and rizzo@ict.usc.edu

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

human interactions to identify potential distress indicators, dialogue policies and virtual human gestures, (2) development and analysis of a Wizard-of-Oz prototype system utilizing two wizard controllers, one for verbal cues and a second for nonverbal cues, and (3) development of a fully automatic virtual interviewer able to engage users in 15-25 minute interactions. All core functionalities of SimSensei Kiosk (dialogue processing, multimodal perception, and nonverbal behavior generation) followed this design methodology. Another important aspect is the 351 participants who were recorded over the course of two years during one of the three development cycles. In the experiments reported here, we show the potential of our fully automatic virtual human interviewer in a user study, and situate its performance in relation to the Wizard-of-Oz prototype.

We begin in Section 2 by discussing the background for our approach, and then present our design goals in Section 3. In Section 4, we highlight the design methodology we have used. We describe our technical approach to dialogue processing, multimodal perception, and nonverbal behavior generation in Section 5. We present our empirical experiments and results in Section 6. Section 7 concludes.

2. BACKGROUND AND RELATED WORK

Several research efforts have identified relevant advantages for computer mediated communication, compared to face-to-face human interaction, for example in promoting a feeling of anonymity and increased self-disclosure [16]. Virtual human systems in particular have been equipped with skills and strategies to elicit self-disclosure. For example, [11] uses vision and prosodic analysis to implement active listening behaviors such as smiles, head nods, and postural mimicry. [4] utilizes nonverbal skills along with verbal strategies such as expressions of empathy, social dialogue, and reciprocal self-disclosure in health behavior change interventions. Such techniques have been shown to increase self-disclosure and feelings of rapport, self-efficacy and trust [4].

Observable differences in the communicative behavior of patients with specific psychological disorders such as depression have previously been investigated in a range of psychological and clinical work. Most work has observed the behavior of patients in human-human interactions, such as clinical interviews and doctor-patient interactions. Some of the differences that have been observed in nonverbal behavior are differences in rates of mutual gaze and other gaze patterns, downward angling of the head, mouth movements, frowns, amount of gesturing, fidgeting, emotional expressivity, and voice quality; see [31] for a recent review.

Examples of observed differences in verbal behavior in depressed individuals include differences in speaker-switch durations and variability in vocal fundamental frequency [6], decreased speech, slow speech, delays in delivery, and long silent pauses [12], as well as differences in certain lexical frequencies including use of first person pronouns and negatively valenced words [28]. Some work has explored automated classification of psychological disorders based on such observed differences in communicative behavior; e.g. [6].

3. DESIGN GOALS

The vision for this effort is to create a fully automated virtual interviewer that creates engaging face-to-face interactions where the user feels comfortable talking and sharing information. At a conceptual level, there are many possible designs for how such a virtual agent could try to interact with users, including factors such as the specific types of questions it asks, what sort of followup dialogue is supported for each question, how it approaches discussion of personal topics related to mental health, its sensitivity to body language as well as spoken responses in the interaction, etc. A particular concern in the design of SimSensei Kiosk was whether users would feel comfortable enough in the interaction to speak openly about their mental health issues to a virtual agent.

The design goals for SimSensei Kiosk's interactions included the following: (1) users should feel comfortable talking and sharing personal information with Ellie; (2) the system should be sensitive to the user's nonverbal behavior; (3) the system should generate appropriate nonverbal behavior itself.

4. DESIGN METHODOLOGY

In this section we summarize the cycles of face-to-face and Wizard-of-Oz development that informed the creation of an engaging virtual human with whom people would feel comfortable talking and sharing information. The first cycle focused on the acquisition and analysis of human-human interactions in the same context of psychological distress assessment. This cycle focused on both sides of the interaction: the interviewee behaviors were analyzed to identify potential indicators of distress while the interviewer was analyzed to identify proper questions and nonverbal behaviors to animate the virtual human. During the second cycle, a Wizard-of-Oz prototype was created allowing two human operators to dictate the virtual human's spoken responses and nonverbal behaviors. Finally, the third cycle focused on the development of our fully automatic system. Section 5 will give more details on how the core modules of SimSensei Kiosk were developed: dialogue processing, multimodal perception and nonverbal behavior generation. The detailed datasets acquired during the three cycles of development are described in [9]; we summarize the datasets from the first two cycles in the remainder of this section.

In the first cycle, a corpus of 120 face-to-face interactions between a confederate interviewer and a paid participant was collected [9, 31, 10]. These interviews began with small talk and a number of neutral questions (for example about the participant's living situation), but as the interview progressed, became more specific about possible symptoms of psychological distress (for example whether the participant has trouble sleeping) and any traumatic events that may have happened in the participant's life.

Analysis of the interaction strategies employed by the face-to-face interviewer highlighted design opportunities for an automated system as well as technical challenges. An important observation was that, apart from simply asking questions, interviewers worked actively to build rapport and to act as good listeners, through factors like providing backchannels and other feedback to indicate surprise, empathy, and engagement during user responses, and also by requesting more information at key moments to encourage participants to keep talking. The open ended range of topics that came up in the dialogues, especially when interviewers would engage in detailed follow up subdialogues (for example discussing at length the reasons why the participant's family moved from one town to another), suggested potential tech-

option type	count	example
nonverbal behaviors	23	head nod to indicate agreement
interview questions	87	*what are you like when you don't get enough sleep?*
neutral backchannels	24	*uh huh*
positive empathy	11	*that's great*
negative empathy	14	*i'm sorry*
surprise responses	5	*wow!*
continuation prompts	26	*could you tell me more about that?*
miscellaneous	24	*i don't know*; *thank you*

Table 1: Wizard-of-Oz option set

nical challenges in automating the interviewer role in SimSensei Kiosk.

The next cycle was a Wizard-of-Oz study. In this setup, a fixed set of 191 utterances and 23 nonverbal behaviors were defined and made available to two Wizards who jointly controlled Ellie's behavior. In addition to asking the top level 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 by providing backchannels, empathy and surprise responses, and continuation prompts. The set of options that was made available to the two Wizards is summarized in Table 1.

One wizard controlled Ellie's nonverbal behavior while the other controlled her verbal 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.

A corpus of 140 Wizard-of-Oz participant interactions was collected using this system [9]. Analysis of these dialogues confirmed the presence of significant differences in the nonverbal behavior of distressed participants when compared to non-distressed participants [33, 32, 31, 30], and also differences in the verbal behavior of distressed participants when compared to non-distressed participants [7]. 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. We report on the performance of the Wizard-controlled Ellie in relation to our design goals in Section 6.

5. DEVELOPMENT

SimSensei Kiosk is based on a general modular virtual human architecture [14], defining at an abstract level the capabilities of a virtual human and how these interact. Capabilities are realized through specific modules and include audio-visual sensing and nonverbal behavior understanding (Section 5.1), natural language understanding and dialogue management (Section 5.2), and nonverbal behavior generation, behavior realization and rendering (Section 5.3). Most modules communicate with each other through a custom messaging system called VHMsg, which is built on top of ActiveMQ.[1] The Unity game engine[2] is used as the renderer for the system.

[1]http://activemq.apache.org
[2]http://unity3d.com

5.1 Perception of nonverbal behavior

The goal is to develop a perception system tailored specifically for this application and based on the design goals for SimSensei Kiosk. Specifically, it should serve a double purpose: i) communicate the necessary nonverbal behavior signals to the other components of the system so that the agent is sensitive to the user's nonverbal behavior, and ii) recognize automatically and quantify the nonverbal behaviors that help indicate the psychological conditions that are being studied (depression and PTSD). As an example, tracking the smile intensity of the participant serves both of these purposes: smile is an important signal that has been tied to investigations of depression (e.g. Reed *et al* [27]) and also plays an important role in a dyadic interaction [5].

As a basis for the perception system, SimSensei Kiosk uses the MultiSense framework, which is a flexible system for multimodal real-time sensing, described further below.

5.1.1 Development methodology

Initially, the face-to-face data is utilized as a study ground to identify nonverbal behaviors that are correlated with depression, PTSD and anxiety. As a first step, three main sources of information were used 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 [8, 12, 19, 26], 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 were based on diagnostic power and implementability. Implementability is important because a fully automatic system is the end goal; for example, hair grooming and physical appearance is one element that a clinician may look at, but it is very difficult to quantify automatically at this point. A similar process is described in Scherer *et al.* [31] where it was shown that such behaviors supported by literature can be identified in face-to-face interactions and are indicative of psychological disorders such as depression, PTSD and anxiety. Moreover, it was shown that some of these signals can be extracted automatically from the videos of the interactions.

With the Wizard-of-Oz configuration, the system is moving towards a standardized interaction (where there is a list of specific stimuli and questions that can be triggered by the wizards) and at this point the main goal from the perception side is to investigate whether the new interaction style allows the participants to express nonverbally in such a way that is still informative about their underlying psychological conditions. Relevant work [30, 33] on audio and video modalities, respectively, showed that a wizard driven agent-human interaction is still rich in nonverbal behaviors that allow for analysis and assessment of depression and PTSD via fully automatic methods. To mention a few examples of indicative nonverbal behaviors in the wizard data: participants scoring positive for depression showcased different measures of voice quality such as normalized amplitude quotient and peak slope than participants that scored negative for depression [30]. Also, participants scoring positive for depression showcased significantly less head motion variation and expressivity on average over the span of the interaction [33],

which aligns with literature reports that motor retardation and emotional flatness are associated with depression.

The next step in the development of the perception system is integration in 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. These behaviors are now being extracted live during the interview and summary statistics can be available automatically at the end of the interview. The list of desired signals, besides behaviors associated with depression and PTSD, includes behavioral signals that can assist the interaction with the virtual human, such as position of the face of the user and smile intensity on a frame level. In this stage the focus was on the implementation of such 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.

5.1.2 Perception system

The MultiSense framework was used as a perception system. 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 because it enables fusion of modalities and development of multimodal indicators.

The following modules have been integrated in MultiSense: 3D head position-orientation and facial tracking based on GAVAM head tracker [25] and CLM-Z face tracker [3], expression analysis based on SHORE face detector [21] and FACET SDK[3], gaze direction based on OKAO vision[4] and audio analysis based on Cogito software[5]. A subset of these were activated during the automated SimSensei Kiosk study (discussed in Section 6). MultiSense dynamically leverages the above measures into informative signals such as smile intensity, 3D head position and orientation, intensity or lack of facial expressions like anger, disgust and joy, speaking fraction, speech dynamics, gaze direction etc. As mentioned above, these informative signals serve two purposes. First, they contribute to the indicator analysis. Second, they are broadcast to the other components of SimSensei Kiosk using the PML standard [29] to inform the virtual human of the state and actions of the participant and assist with turn taking, listening feedback, and building rapport by providing appropriate non-verbal feedback, as seen in Section 5.3.

5.2 Dialogue processing

The design of SimSensei Kiosk's dialogue processing was guided by a number of technical challenges posed by analysis of the face-to-face and Wizard-of-Oz corpora. Unlike many task-oriented dialogue domains, these interview dialogues are naturally open-ended, as people respond to interview stimuli such as *tell me about the last time you felt really happy* with idiosyncratic stories and events from their lives. Because there is so much variability in the vocabulary and content of participants' answers to such questions, speech recognition is challenging. Further, for language understanding, we cannot simply construct a small semantic ontology and expect to cover the majority of meanings that will be expressed by users. Thus, this is an application in which the dialogue policy needs to be able to create a sense of engagement and empathy despite relatively shallow and limited understanding of user speech.

The finite set of Wizard options summarized in Table 1 was selected to strike a balance between providing enough expressive options for the wizards to sustain an engaging dialogue and limiting the wizard's options to a finite set that could potentially be automated. Development of the automated SimSensei Kiosk's dialogue policy was informed by analysis of the wizards' behavior in the Wizard-of-Oz corpus. Many interview questions had frequent patterns of continuation and empathy responses that could be encoded in rules. However, we found that several simplifications were necessary due to limitations in automated understanding capabilities. For example, wizards had the ability to express surprise using utterances like *wow!*. However, in looking at the situations when these utterances were used, it appeared that quite deep semantic understanding and domain knowledge would be necessary to automatically generate many of these surprise expressions; thus, currently the automated SimSensei Kiosk system does not generate them. Similarly, the wizards seemed to rely on a detailed sensitivity to the implications of participant responses to identify the specific moments in dialogue when continuation prompts would be most likely to elicit additional disclosure. Currently, our automated system uses much shallower heuristics based on the duration of participant speech to generate continuation prompts.

We summarize the implemented dialogue processing in SimSensei Kiosk in the following sections.

5.2.1 Natural Language Understanding

SimSensei Kiosk employs continuous automatic speech recognition (ASR). In the experiments reported in this paper, voice activity detection and speech recognition were performed using PocketSphinx [15]. SimSensei Kiosk currently uses 4 statistically trained utterance classifiers to capture different aspects of user utterance meaning.

The first NLU classifier identifies generic dialogue act types, including statements, yes-no questions, wh-questions, yes and no answers, and several others. This classifier is trained using the Switchboard DAMSL corpus [17] using a maximum entropy model.

The second NLU classifier assigns positive, negative, or neutral valence to utterances. In our current exploratory approach, we use SentiWordNet 3.0 [2], a lexical sentiment dictionary, to assign valence to individual words spoken by users (as recognized by the ASR). The dictionary contains approximately 117,000 entries. In general, each word w may appear in multiple entries, corresponding to different parts of speech and word senses. To assign a single valence score $v(w)$ to each word in the dictionary, we compute the average score across all parts of speech and word senses:

$$v(w) = \frac{\sum_{e \in E(w)} \text{PosScore}_e(w) - \text{NegScore}_e(w)}{|E(w)|}$$

where $E(w)$ is the set of entries for the word w, $\text{PosScore}_e(w)$ is the positive score for w in entry e, and $\text{NegScore}_e(w)$ is the negative score for w in entry e. This is similar to the "averaging across senses" method described in [34]. In [7], it was observed that such lexical valence scores can be used to identify certain significant differences between distressed

[3] *http://www.emotient.com/*
[4] *http://www.omron.com/r_d/coretech/vision/okao.html*
[5] *http://www.cogitocorp.com/*

Speaker	Utterance
Ellie	*How have you been feeling lately ?* (top level question)
Participant	*Not good at all*
Ellie	*Can you tell me about that ?* (continuation prompt)
Participant	*Um... yeah i mean... i recently got released from prison not too long ago... uh i'm older now... and a lot more responsibilities and i feel like my life is not where it's supposed to be... i feel like i could've made a lot more better decisions in my life... you know ... sometimes feeling sorry for myself ... and just things that i've been through and the whole nine yards so yeah*
Ellie	*I'm sorry* (empathy response)

Figure 2: Example of Ellie's subdialogue policies.

and non-distressed individuals in the Wizard-of-Oz corpus.

In the experiments reported in this paper, the valence assigned to an utterance is based primarily on the mean valence scores of the individual words in the utterance. Utterances whose mean word valence exceeds a positive threshold are assigned positive valence; utterances whose mean word valence is below a negative threshold are assigned negative valence; other utterances are judged neutral. The positive and negative thresholds are tuned using labeled training data from this domain. Additionally, to reduce certain common types of errors with this approach, utterances whose length is less than three words or which contain explicit negation are treated as neutral.[6]

The third NLU classifier supports domain-specific small talk by recognizing a handful of specific anticipated responses to Ellie's rapport-building questions. For example, when Ellie asks users where they are from, this classifier detects the names of certain commonly mentioned cities and regions. This classifier uses keyword and keyphrase spotting.

The fourth NLU classifier identifies domain-specific dialogue acts, and supports Ellie's follow up responses to specific questions. For example, one of Ellie's questions is "how close are you to your family?". This maximum entropy classifier is trained using face-to-face and Wizard-of-Oz data to detect various forms of positive responses that serve to assert closeness (a domain-specific dialogue act).

5.2.2 Dialogue Management

Ellie currently uses about 100 fixed utterances in total in the automated system. She employs 60 top level interview questions, plus a range of backchannels, empathy responses, and continuation prompts.

The dialogue policy is implemented using the FLoReS dialogue manager [24]. The policy groups interview questions into several phases (rapport-building, diagnostic, and warm-up). Questions are generally asked in a fixed order, with some branching based on responses to specific questions.

Rule-based sub-policies determine what Ellie's follow up responses will be for each of her top-level interview questions. The rules for follow ups are defined in relation to the four NLU classifiers and the duration of user speech (measured in seconds). These rules trigger continuation prompts and empathy responses under specific conditions.

An example of Ellie's subdialogue policies is given in Figure 2. In this example, Ellie selects a continuation prompt based on the short duration of the user's response to her question (using a threshold of less than 4 seconds in this case). In this example, the user provides a much more detailed response following the continuation prompt. Upon detecting negative valence in this response, Ellie responds with an empathy utterance of *I'm sorry.*

5.3 Generation of nonverbal behavior

Beyond the words uttered, nonverbal behavior - including facial expressions, gaze, gestures and postures - powerfully influences face-to-face interaction, impacting a range of relational factors [5]. Given the importance of establishing a relation between Ellie and the participant, and the overall importance of nonverbal behavior in such conversations [12], nonverbal behavior design became a key concern. At a behavioral level, we specifically wanted Ellie's behavior to portray an expressive, but also calming, empathic speaker as well as an attentive listener responsive to the participant's speaking behavior. At a technical level, we wanted this behavior automatically generated, inferred from Ellie's dialog as generated by the dialog manager and the participant's nonverbal behavior as sensed by MultiSense.

To achieve these goals, we used and extended the Cerebella behavior generation system [23, 22] that determines what behaviors a virtual character should exhibit.

5.3.1 Cerebella

Cerebella is a research platform to realize the relation between a character's mental states and processes and its behavior, especially nonverbal behaviors accompanying the virtual human's dialog, responses to perceptual events as well as listening behaviors. In the case of generating nonverbal behavior accompanying dialog, it is designed to be flexible and not make strong assumptions about the inputs it receives. For example, if input containing detailed information about the speaker's mental state, including communicative intent, is provided, a direct mapping to nonverbal behaviors can be made.

However, when only the virtual human's utterance text and audio are given as is the case in SimSensei Kiosk, the system tries to infer mental states through several analyses of the input. In particular, the utterance text is first parsed to derive the syntactic and rhetorical structures (such as contrast). Then, pragmatic, semantic and metaphoric analyses attempt to infer aspects of the utterance's communicative function such as affirmation, inclusivity, and intensification that can have behavioral signatures. Regardless of whether the mental states are provided or inferred, Cerebella uses a model of the relation between mental states and behavior to generate appropriate nonverbal behavior types. A behavior generation phase then maps those behavior types to specific behavior specifications. This mapping can use character-specific mappings designed to support individual differences including personality, culture, gender and body types. The final result is a schedule of behaviors, described in the Behavior Markup Language (BML; [20]), that is passed to the character animation system.

[6]We are investigating more sophisticated approaches to classifying utterance valence. Our requirements for SimSensei Kiosk include robustness to the potentially high word error rates in recognized speech, and a risk-aversion with respect to false positive and false negative valence labels, which can result in inappropriate expressions of empathy by Ellie.

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 2: Means, standard errors, t-values and effect sizes. * = p <.05

In the case of listening behavior, Cerebella can receive PML messages [29] containing visual and vocal cues of the user to generate behavior such as attending head nods and mimicking smile [36].

5.3.2 SmartBody Character Animation System

To animate Ellie, we use the SmartBody character animation [35]. The system provides many critical capabilities for the representation, interaction and visualization of 3D characters in virtual environments. Using a combination of procedural and keyframe controllers, SmartBody's capabilities include locomotion with collision avoidance, posture and gaze control, facial expressions, blinks, gestures, speech, saccadic eye and head movements, reaching, pointing and grabbing. SmartBody takes BML input and supports a range of rendering and game engines.

5.3.3 Customizing Cerebella for SimSensei Kiosk

Our goal is to design a virtual human that participants feel comfortable sharing personal information with. As noted above, Cerebella is capable of using character-specific mappings designed to support individual differences. To create animation gestures with the desired characteristics, we captured in role play sessions the interaction between a clinician who worked daily with people suffering from depression and PTSD and an actor pretending to suffer from these conditions.[7] We found that almost all of her nonverbal behavior was aimed at making the patient feel comfortable, safe, and listened to. The videos provided an animator with reference material to inform the design of gestures consistent with the communicative functions that Cerebella infers, such as beat gestures used for emphasis/intensification and conveying an empathic personality. Finally, we configured Cerebella's BML generation process to map the inferred function to specific animation gestures.

Since SimSensei Kiosk uses prerecorded audio, Cerebella processed each dialogue line offline, inferring the communicative functions of the sentence and generating an appropriate nonverbal performance.

Listening feedback, however, is tightly coupled to user behavior and cannot be preprocessed. In the Wizard-of-Oz, a set of listening behaviors were manually fired, including different types of head-nods, smiles and facial expressions of concern and surprise. In SimSensei Kiosk, Cerebella receives PML information about the user from MultiSense to determine the timing and type of feedback to express. However the provided visual and acoustic cues are not always sufficient to generate all these behaviors with high certainty. Because it is less risky not to express feedback than to express an inappropriate one, SimSensei Kiosk uses a smaller set of listening feedbacks (head nods and smiles).

[7]Although Ellie is definitely not designed to portray a clinician, the topic of the conversation concerns clinical issues, so we saw a trained clinician as a useful basis for designing the form and dynamics of the gestural animations.

6. EVALUATION

To inform the system design and assess our success in achieving design goals at each stage, we collected three interview datasets: face-to-face interactions with semi-expert human interviewers (referred to as Face-to-Face), Wizard-of-Oz interactions with a virtual human puppet controlled by the same human interviewers (referred to as WoZ), and "AI interactions" where the VH was controlled by the automated SimSensei Kiosk system (referred to as AI).

Pre-experience, all participants were given a series of self-report assessment instruments to index their clinical state. Post-experience, all participants completed a validated measure of rapport [18]. Additionally, participants in WoZ and AI completed nine questions designed to test our success in meeting specific design goals (see Table 2). Examples include questions about disclosure ("I was willing to share information with Ellie"), questions about 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). Finally, participants in WoZ and AI also completed the standard System Usability Scale (SUS; [1]), a measure of a product's perceived system satisfaction and usability.

6.1 Participants

Across all three studies, 351 participants were recruited through Craigslist and posting 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).

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

Table 3: Rapport scores in the three conditions.

This data set includes the face-to-face data, the computer-framed wizard data, and a subset of the computer-framed automated agent data described in [9]. Our evaluation data here include only "computer-framed" sessions, where the AI or WoZ system was presented to the participant as an autonomous agent, and exclude additional sessions where the system was presented as controlled by humans; see [9].

6.2 Results

For all items and scales, participants' total scores were calculated for analysis. Table 2 displays mean total scores and associated standard errors for each of the subsequent analyses. In interpreting these scores, it is important to keep in mind that the AI system does not need to match or replicate the performance of the WoZ system in order for us to achieve our design goals for the automated system.

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 3). 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 ($t(345) = -0.77$, $p = .44$, $d = 0.07$.).

6.3 Discussion

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 (e.g., see [37]), though this will require further study.

As expected, the current version of SimSensei Kiosk 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.

7. CONCLUSION

We have presented SimSensei Kiosk, an implemented virtual human interviewer designed to create an engaging face-to-face interaction where the user feels comfortable talking and sharing information related to psychological distress. We discussed the design process and development of the system, and evaluated several aspects of its performance.

8. 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.

9. REFERENCES

[1] SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, editors, *Usability Evaluation in Industry*. Taylor and Francis.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, 2010.

[3] T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *CVPR, 2012 IEEE Conference*, 2012.

[4] T. Bickmore, A. Gruber, and R. Picard. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59(1):21–30, Oct. 2005.

[5] J. K. Burgoon, L. K. Guerrero, and K. Floyd. *Nonverbal Communication*. Allyn & Bacon, 2009.

[6] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction (ACII)*, September 2009.

[7] D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. S. Rizzo, and L.-P. Morency. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of SIGdial*, 2013.

[8] L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91(2):109–119, 1982.

[9] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency. The Distress Analysis Interview Corpus of human and computer interviews. In *LREC 2014*, to appear.

[10] J. Gratch, L.-P. Morency, S. Scherer, G. Stratou, J. Boberg, S. Koenig, T. Adamson, A. Rizzo, et al. User-state sensing for virtual health agents and telehealth applications. *Studies in health technology and informatics*, 184:151–157, 2012.

[11] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, 2007.

[12] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21 – 37, 1995.

[13] J. Hart, J. Gratch, and S. Marsella. *How Virtual Reality Training Can Win Friends and Influence People*, chapter 21, pages 235–249. Human Factors in Defence. Ashgate, 2013.

[14] A. Hartholt, D. Traum, S. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch. All together now, introducing the virtual human toolkit. In *Intelligent Virtual Agents*, 2013.

[15] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of ICASSP*, 2006.

[16] A. N. Joinson. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192, 2001.

[17] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. 1997.

[18] S.-H. Kang and J. Gratch. Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories. In B. Wiederhold and G. Riva, editors, *Annual Review of Cybertherapy and Telemedicine*, pages 202–207. IOS Press, 2012.

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

[20] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thorisson, and H. Vilhjalmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, 2006.

[21] C. Kublbeck and A. Ernst. Face detection and tracking in video sequences using the modifiedcensus transformation. *Image and Vision Computing*, 24(6):564 – 572, 2006.

[22] M. Lhommet and S. C. Marsella. Gesture with meaning. In *Intelligent Virtual Agents*, 2013.

[23] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the Symposium on Computer Animation*, Anaheim, 2013.

[24] F. Morbini, D. DeVault, K. Sagae, J. Gerten, A. Nazarian, and D. Traum. FLoReS: A forward looking reward seeking dialogue manager. In *Proceedings of IWSDS*, 2012.

[25] L.-P. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *The IEEE Conference on Automatic Face and Gesture Recognition*, 2008.

[26] J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.

[27] L. I. Reed, M. A. Sayette, and J. F. Cohn. Impact of Depression on Response to Comedy: A Dynamic Facial Coding Analysis. *Journal of Abnormal Psychology*, 116:804–809, 2007.

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

[29] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L. P. Morency. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *Intelligent Virtual Agents*, 2012.

[30] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *Proceedings of Interspeech 2013*, pages 847–851. ISCA, 2013.

[31] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[32] S. Scherer, G. Stratou, and L.-P. Morency. Audiovisual behavior descriptors for depression assessment. In *ICMI*, 2013.

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

[34] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), June 2011.

[35] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann. SmartBody: behavior realization for embodied conversational agents. In *AAMAS*, 2008.

[36] Z. Wang, J. Lee, and S. Marsella. Multi-party, multi-role comprehensive listening behavior. *Autonomous Agents and Multi-Agent Systems*, 27(2):218–234, Sept. 2013.

[37] S. Weisband and S. Kiesler. Self-disclosure on computer forms: Meta analysis and implications. In *CHI*, volume 96, pages 3–10, 1996.