OPEN ACCESS
OPUS AUGSBURG

Universität Augsburg
Universitätsbibliothek

A job interview simulation: social cue-based interaction with a virtual character

Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Baur, Tobias, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André. 2013. "A job interview simulation: social cue-based interaction with a virtual character." In *2013 International Conference on Social Computing, 8-14 September 2013, Alexandria, VA, USA*, 220–27. Piscataway, NJ: IEEE. https://doi.org/10.1109/socialcom.2013.39.

Nutzungsbedingungen / Terms of use: **licgercopyright**

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

A Job Interview Simulation: Social Cue-based Interaction with a Virtual Character

Tobias Baur*, Ionut Damian*, Patrick Gebhard†, Kaśka Porayska-Pomsta‡ and Elisabeth André*
*Human Centered Multimedia, Augsburg University
baur@hcm-lab.de, damian@hcm-lab.de, andre@hcm-lab.de
†DFKI GmbH, Saarbrücken, Germany
patrick.gebhard@dfki.de
‡London Knowledge Lab, Institute of Education
K.Porayska-Pomsta@ioe.ac.uk

***Abstract*—This paper presents an approach that makes use of a virtual character and social signal processing techniques to create an immersive job interview simulation environment. In this environment, the virtual character plays the role of a recruiter which reacts and adapts to the user's behavior thanks to a component for the automatic recognition of social cues (conscious or unconscious behavioral patterns). The social cues pertinent to job interviews have been identified using a knowledge elicitation study with real job seekers. Finally, we present two user studies to investigate the feasibility of the proposed approach as well as the impact of such a system on users.**

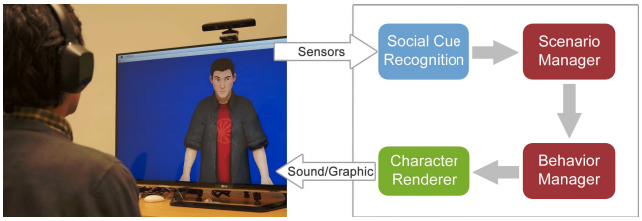

Fig. 1. General setup and main software modules of the system.

I. Introduction

One large issue Europe faces is the rising number of young people Not in Employment, Education or Training (NEETs). NEETs often have underdeveloped socio-emotional and interaction skills [1], [2], such as a lack of self-confidence and sense of their own strengths. This effects their performance in various critical situations, such as job interviews, where they need to convince the recruiter of their fit in a company. To address this issue, many European countries have specialized inclusion centers meant to aid young people secure employment through coaching by professional practitioners. One problem of this approach is that it is very expensive and time-consuming. Considering this, technology-enhanced solutions present themselves as viable and advantageous alternatives to the existing human-to-human coaching practices.

Job interviews are used by the potential future employer as a means to determine whether the interviewee is suited for the company's needs. To make an assessment, interviewers heavily rely on social cues, i.e. actions, conscious or unconscious, of the interviewee that have a specific meaning in a social context, such as a job interview.

In this paper we present an approach to a job interview simulation environment which uses a social virtual character as a recruiter and signal processing techniques to enable the virtual character to react and adapt to the user's behavior and emotions. The purpose of this simulation is to help youngsters improve social skills which are pertinent to job interviews. The system we propose features a real-time social cue recognition system, a dialog/scenario manager, a behavior manager and a 3D rendering environment (Fig. 1).

The next section offers a brief review of the interdisciplinary literature. In Section III, we present the one-on-one mock interviews we conducted with NEETs and practitioners in order to identify the social cues relevant to the job interview scenario. Section IV then introduces the job interview simulation system. Our approach is evaluated in Section V. We conclude the paper in Section VI and take a look at future work in Section VII.

II. Related Work

A growing amount of literature demonstrates the power of social cues that are consciously or unconsciously shown by people in various situations, such as negotiations, group meetings or job interviews. According to a survey by Arvey and Campion [3], nonverbal behaviors, such as eye gaze contact, body movement and voice tone, significantly bias the assessment of the job interviewers. Hence, the use of non-verbal behaviors and their impact on the success of a job interview has become a major focus of research. Curhan and Pentland [4] observed that speech activity, conversational engagement, prosodic emphasis, and vocal mirroring were highly predictive of the outcome of simulated job interviews.

In order to help people train social skills, a variety of techniques have been developed, such as role playing, group discussions or specific exercises [6]. The need for effective social training has also inspired a number of proposals for computer-based simulation environments as additional platforms for delivering such training for a variety of applications including job interviews [7], inter-cultural communication [8], negotiation scenarios [9] or psychotherapy [10].

Recent progress in the area of social signal processing (for an overview see [11]) offers great promise to computer-enhanced environments for social training since they allow users to explore the impact of the social signals they convey in

a safe environment. Most research in the area of social signal processing has focused on the recognition of emotions from speech [12] and facial expressions [13]. Compared to vocal emotions and facial expressions, relatively little attention has been paid to the analysis of gestures [5], [14] and postures [15], [16]. Furthermore, attempts have been made to fuse multiple modalities leading in most cases to an improvement of recognition rates while the gain seems to be significantly higher for acted than spontaneous emotions [17]. Nevertheless, techniques from the area of social signal processing have hardly been explored in interactive scenarios. Noteworthy exceptions include the health care agent by Scherer [18], the Semaine Artificial Listener [19], EmoEmma [20] or the laughter aware agent [21].

As in previous work, we make use of signal processing techniques to recognize a user's social cues in order to elicit appropriate backchannel and turn taking behaviors by the virtual character. Our main interest is, however, to analyze social cues as indicators of so-called critical events, i.e. events that have a decisive impact on the outcome of a conversation (see Section III).

III. Knowledge Elicitation

In order to gain a better understanding of what social cues occur in job interview situations and thus to identify the behaviors that need to be recognized automatically, we ran a study with NEETs and practitioners involving one-on-one mock interviews pertaining to real job opportunities. The study was designed to achieve four goals:

1) To gain access to real exemplars of job interview enactments with real practitioners and youngsters
2) To ascertain what different factors, including the specific behaviors of the interviewees, influence the nature of individual interactions
3) To determine the young job-seekers social attitudes, along with the corresponding social cues
4) To test the technicalities involved in and to generate the data needed for real-time detection and interpretation of social cues.

A. Participants

Ten young job seekers (8 females and 2 males, average age 19.2 years) and five practitioners attended the mock interview sessions. The young job seekers were registered with one of the Mission Locale specialized inclusion centers in France. Typically, the young people pre-book their mock interview sessions, but they can also come on a drop-in basis. In either case they are assigned to the next available practitioner on a first-come first-serve basis. Thus, while, the five specific practitioners have agreed to participate in the study a priori, the final sample of NEETs was randomly selected and is also representative (including the evident gender bias) of the young people seeking support in France.

B. Procedure

The procedure for the studies consisted in mock interviews for real jobs advertised in the local area. The interviews involved the interviewee acting as themselves with the practitioner acting as a recruiter for a given job. All of the mock interviews were one-on-one and were conducted in rooms arranged as standard offices (a desk, two chairs on either side, sometimes a desktop computer and bookshelves).

The mock interviews were video recorded using a camcorder and a Kinect camera. Each mock interview was followed by a debriefing of the youngster by the practitioner about how the interview went, what strengths and weaknesses the young person manifested and areas for future improvement. Each session lasted approximately 45 minutes, after which the researchers interviewed the practitioners and youngsters using a semi-structured interview method.

C. Post Hoc Knowledge Elicitation and Data Annotation

We used the answers provided by the practitioners and the young interviewees in the respective post-hoc interviews to design the post-hoc walkthrough with practitioners. The aim of the post-hoc walkthroughs was to elicit detailed knowledge from the practitioners about the individual youngsters' behaviors and about the practitioners' interpretation of the social cues manifested by the youngsters on a moment-by-moment basis. The walkthroughs were facilitated by the use of the Elan [22] annotation tool through which specific enactments were played back to the practitioners. The practitioners were asked to stop the video replays at any point at which they believed a critical incident (or a set thereof) has occurred. A critical incident was defined as a specific behavior on the part of the interviewee, e.g. smile, or a set of behaviors, e.g. persistent smiling and gaze aversion, that the practitioner thought crucial, in a positive or negative way, to the job interview and its outcome.

This procedure and use of tools allowed for the key episodes and behaviors in the given interactions to be identified precisely, i.e. within exact time windows, by the practitioners and to be annotated with their specific comments. Practitioners comments and elaborations, especially the detail related to the particular observable behavioral evidence that the practitioners relied on in making their diagnosis of the youngsters, were explicitly encouraged by the researchers. The researchers also encouraged the practitioners to state whether or not, and if so then what, complex mental states they would/could infer from the specific behaviors. At times, the practitioners found it easier to identify a complex state before trying to identify the specific social cues. Fig. 2 shows a snapshot of the Elan tool used during the post-hoc walkthroughs, with the default tier (top-most line) showing the type of annotations made by the practitioners.

Once annotated with practitioners comments the videos were used as the basis for fine-grained analysis of the discrete social cues perceived by the practitioners. For example, in the same Fig. 2, the episode shown can be further annotated for the finer grained behaviors such as the interviewee looking away, not making direct contact with the interviewer, smiling, etc. and these annotations can be analyzed in relation to a precise time-frame within which each behavior identified took place. The annotations of complex mental states made by the practitioners during the walkthroughs also serve as the baseline reference for further detailed annotations of the videos with respect to the complex mental states which we intend to carry out as part of our future work. These annotations will allow us

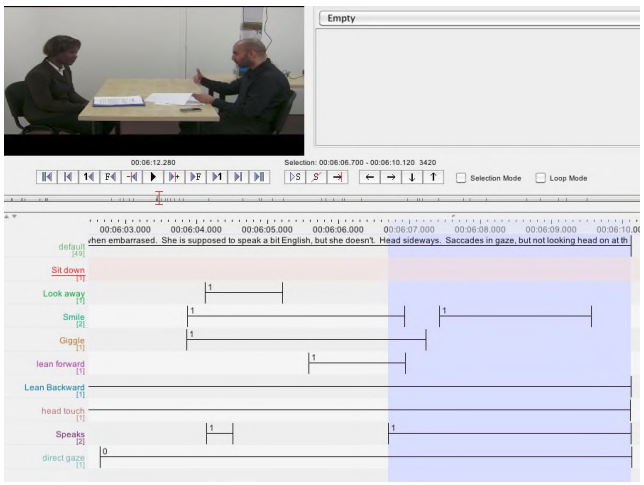

Fig. 2. Elan annotation tool [22] used in the post-hoc walkthroughs. The interviewee is on the left, the interviewer on the right. The practitioners comments to this scene: [She] Smiles when embarrassed. She is supposed to speak a bit of English, but does not. Head sideways. Saccades in gaze, but not looking head on at the interviewer. Head pose shifts/saccades and head shifts in quick succession. Breaths in. She says: I dont speak very much [English], but I understand a little bit. Head moves to the side.

TABLE I. SOCIAL CUES IDENTIFIED BASED ON A COMBINATION OF POST-HOC WALKTHROUGHS WITH PRACTITIONERS, POST-HOC INTERVIEWS WITH PRACTITIONERS AND YOUNG JOB SEEKERS, AND HAND ANNOTATIONS OF VIDEO RECORDED MOCK JOB INTERVIEW INTERACTIONS.

Type	Social Cue
Hands	Hands on table
	Hands under-table
	Restless Hands
	Gesticulating
	Hands to face
Eyes/Head	Look-away
	Saccades
Face	Lip-bite
	Smile

Type	Social Cue
Posture	Lean-forward
	Lean-back
	Rocking
	Sudden movement
Verbal	Interrupting the interviewer
	Laugh
	Low voice
	Clear voice
	Short answers
	Long silence

to create a mapping between the social cues and the complex mental states and thus further inform the design of our system.

D. Preliminary Results

The walkthroughs along with the interviews provided the basis for determining the social cues that the interviewees manifested during job interviews. In total, 19 individual cues have been identified and classified according to the communication means. Table I shows the identified social cues.

IV. THE SYSTEM

For our interactive scenario we rely on a software framework that supports a fine grained multimodal behavior control for virtual characters [23]. It comes with several software modules which are needed for the creation of an interactive character system (e.g. TTS, Character Rendering, Emotion Simulation). And, more important, it allows a standardized integration of additional software components using approved interface standards (e.g. BML, EmotionML) for virtual characters [24]. We extended the framework by two modules: a

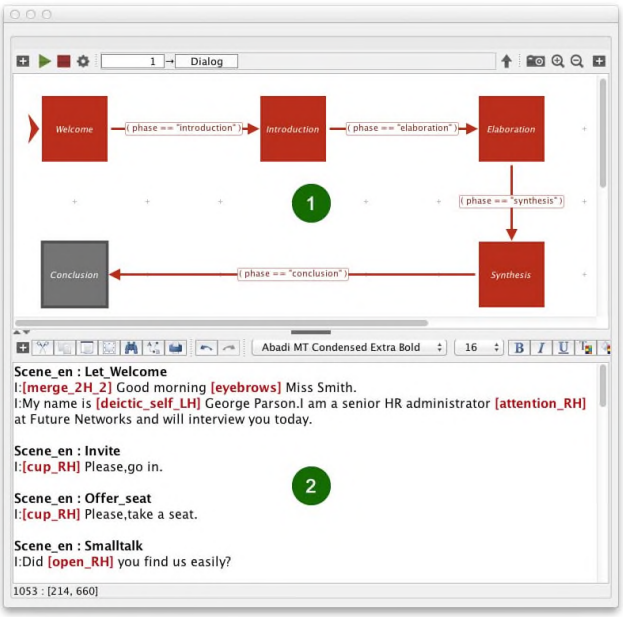

Fig. 3. Main five stages of the recruitment scenario modeled as HSFMs.

Scenario Manager and a Social Cue Recognizer. As shown in Fig. 1, the main user interface is an interactive virtual character that is capable of performing social cue-based interaction with the user. The social cues are recorded by a Microsoft Kinect device and analyzed by the Social Cue Recognizer. Based on the recognized social cues, the Scenario Manager chooses an appropriate reactive behavior model. The behavior manager transforms this model into a sequence of timely aligned multimodal virtual character control commands (e.g. speech, gestures, facial expressions, and head movement) which are then executed by the Character Renderer. In particular, the virtual character is able to react to the user's voice activity, facial expressions and head movements.

A. Scenario Manager - Behavioral Modeling

For the modeling of our interactive virtual recruiter's behavior, we rely on an authoring tool [25] that allows us to model and to execute behavioral aspects at very detailed and abstract level. Central to this tool is the separation of dialog content and interaction structure, see Fig. 3.

The multimodal dialog content is specified with a number of scenes that are organized in a scenescript, see Fig. 3. The scene structure can be compared to those in TV or theater playbooks, which consist of utterances and stage directions for the actors. In our scenes, directions are animation commands for gestures, facial expressions, or postures. The (narrative) structure of our interactive recruitment simulation and the interactive behavior of our virtual recruiter is controlled by parallel hierarchical finite state machines (HFSM) specifying the logic that determines which scenes are played and commands are executed according to user reactions and state of the interactive performance, see Fig. 3.

Our behavior model consists of two parts: 1) facial expressions and head movements, and 2) story structure and reactions

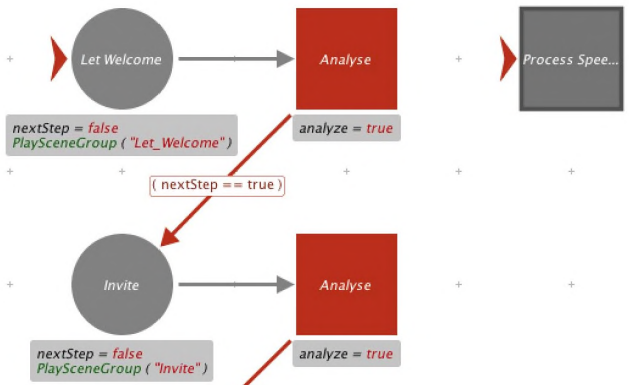

Fig. 4. Segment of the Welcome HFSM with sub HFSMs for the story plot and speech analyze/processing.

to user input.

The virtual character is able to react to the user by mirroring specific behaviors. This is accomplished by two HFSMs. The state machines react on the detected social cues and trigger an overlay behavior. This behavior is blended with any on-going animation the virtual recruiter is performing at any time, e.g. if a user smiles during a question the virtual recruiter asks, the recruiter will return the smile while uttering the question. The same goes with head movements. At this point of time, three head movements (to the left, to the right, and to the front) and two facial expressions (smile, and neutral) are supported.

The story and interaction model is the major part of the virtual recruiter behavior model. Compared to a linear theater scene play, an interactive presentation comes along with another degree of freedom, the reactions of the system on user input. Those have to be covered in the behavior model. In order to achieve this, we have enhanced a linear five stage recruitment story with reactions to user input, in this case speech input. The five content stages of our job recruitment presentation are: 1) Welcome, 2) Introduction, 3) Elaboration, 4) Synthesis, and 5) Conclusion. Users are supposed to answer questions in nearly all phases of the story (except for stage 5 in which the virtual recruiter sums up the interview). The questions are designed to motivate users to give longer and more elaborate answers the further they proceed in the story. Therefore we model the reactive behavior of the virtual recruiter in each phase differently in a separate HFSM, see Fig. 4.

Fig. 4 shows the sub states of the Welcome stage HFSM. It consists of the plot FSM that begins with the state ”Let Welcome” executing a welcome scene, in which the virtual recruiter introduces itself to a user (see Fig. 3). In parallel, the HFSM ”Process Speech” is executed that models how to react on the given answers, and how long to wait, if a user does not answer. The latter depends on the questions of the related plot scene. After the speech is processed, the next plot step and a related scene is activated. Each of the five main stages follows the HSFM pattern presented in Fig. 4.

B. Social Cue Recognition

Based on the data gathered from the user study presented in Section III and literature review, we implemented a module meant to record and analyze social and psychological signals from users and recognize predefined social cues in real time. The module fulfills two important roles in the context of our job interview simulation system. Firstly, it allows the virtual character to react to the user’s behavior generating seamless interaction between the user and the system. Secondly, by recording social cues which occurred during the interaction, we give the user the possibility to review the simulation and see exactly what caused the virtual recruiter to react in a certain way, thus further increasing the learning factor of our system.

The system uses a combination of sensors and software algorithms which offer good results in terms of accuracy and low intrusion. High accuracy ensures that a youngster’s social cues are correctly recognized and allows the game itself to correctly react to them. It is equally important that the approach has a low intrusion factor. For example, biological signal sensors as in [7] are not feasible in this scenario because attaching various sensors to the skin of the users will most likely result in an increase in stress which might have a negative effect on the user’s job interview performance, but may not be actually indicative of the user’s actual abilities. Therefore, in the context studied, remote sensors are preferred.

The use of remote sensors, however, impacts the choice of social cues we can recognize. Physiological cues or cues which are very subtle are difficult to be reliably recognized with remote sensors. Considering this, our efforts concentrated on cues which are both pertinent to the job interview context but are also feasible to be recognized from a technological point of view.

For recording and preprocessing human behavior data, our system relies on the SSI framework which was developed as part of our previous work [26]. It provides a variety of tools for real-time analysis and recognition of such data. The sensor of choice is the Microsoft Kinect as it can be used to achieve recognition of a broad range of social cues and it meets all the requirements of the context. Its low price point and the fact that it is relatively easy to set up means that the system can be used by a wide range of institutions and it would even enable users to use such a system privately in their home. Furthermore, because it is a remote sensor, it has a minimal intrusion level. There are also software development kits for skeleton and face tracking available which provide a good starting point for human behavior analysis.

The system currently allows us to recognize the seven social cues enumerated below. Future developments in sensing technologies and the use of other sensors (e.g. pressure sensors in the seat [15]) may help improve accuracy or enable the recognition of additional social cues. We plan to investigate this as part of our future work.

- *Hand to face*. This is a self-manipulation social cue and has been observed multiple times during the mock job interviews conducted as part of our knowledge elicitation study.
- *Looking away*, the most frequently observed social cue in the mock job interviews.

- *Postures (Arms crossed, Arms open, Hands behind head)*. Although not observed in our knowledge elicitation study, these postures have been found to be good indicators of a person's mental state [27].
- *Leaning forward/backward*. Although not as frequently observed as other social cues, leaning back and forward have been identified by the practitioners as very meaningful.
- *Voice activity* (detects whether the user talks or not). This covers three of the social cues presented in Table I: interrupting the interviewer, short answers and long silence.
- *Smile*. An important facial social cue observed during the knowledge elicitation studies.

Voice Activity. In order to detect when the user is talking our system looks at the audio signal provided by a microphone. To ensure accurate results, we decided to use a close-talk microphone instead of the one incorporated in the Microsoft Kinect. The main advantage of the close-talk microphone is that it filters out most of the environmental noise. For the voice activity detection itself we use a binary Signal-To-Noise filter, which uses a threshold-based approach to categorize an audio sample into noise and non-noise, in our case voice activity. The filter also enforces a minimal duration of 0.3s and minimal silence duration of 2.0s. This makes the system more robust towards environmental noise, interjections or short pauses in speech.

Gestures, Postures and Head Gaze. To recognize *Hand to face*, *Looking away*, *Leaning backward/forward*, *Arms crossed*, *Arms open* and *Hands behind head* we use a state machine like approach developed as part of our previous work [28]. It is able to recognize predefined postures and gestures using the skeleton tracking data provided by the Microsoft Kinect SDK[1]. The gestures and postures to be recognized are represented as skeletal configurations using an XML-based specification language. For instance, *hand to face* is defined as the sustained proximity of the left or right hand joint to the head joint for at least 100ms. The gestures and postures we used in our system are exemplified in Figure 5.

In order to recognize the *looking away* social cue, we use the face tracking library within the Microsoft Kinect Developer Toolkit. This library uses both the RGB information and the depth information to track the face of the user and compute several characteristics. Out of these characteristics, the most important one to us is the head pose data. This allows our system to determine the orientation of the user's head, and, using a threshold-based approach, to detect when the user is looking straight ahead, to the left or to the right.

Smiles. To detect smile occurrences we use the SHORE face tracking library[2] [29]. More precisely, we are looking at the facial expression "happy" computed by SHORE. Once this facial expression exceeds a certain intensity threshold, our system reports a smile occurrence. Smiles are important social cues as they are able to convey a broad range of emotions besides happiness, such as friendliness, anxiety and others [30], [31].

[1]http://www.microsoft.com/en-us/kinectforwindows

[2]http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html

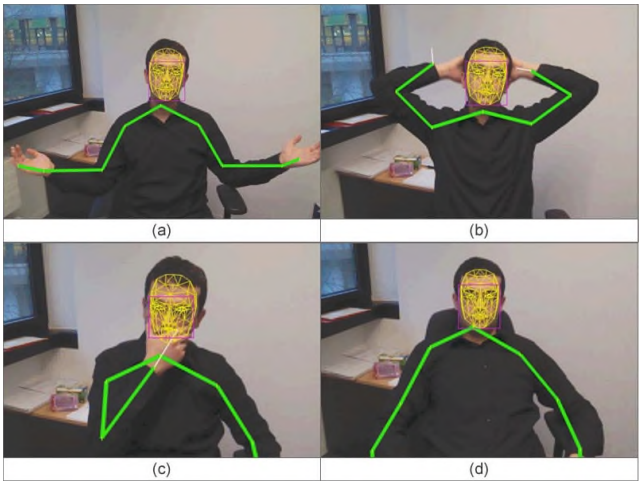

Fig. 5. Examples of the gestures and postures our system can recognize: Arms open (a), Hands behind head (b), Hands to face (c), Leaning backward (d)

Expressivity Features. The system is also able to compute four dimensions of the user's movements' expressivity [32], [33]: energy, overall activation, spatial extent and fluidity. Energy (or Power) gives us the dynamic properties of a movement (weak versus strong) [34]. Caridakis defines overall activation as being the total quantity of movement in a specific time frame [32]. The spatial extent describes the amount of space taken up by the body. Finally, fluidity refers to the continuity of the movement, differentiating smooth movements from jerky ones.

V. Evaluation

In order to evaluate our system, we conducted two user studies. The first study served to acquire a sufficiently comprehensive set of user behaviors for evaluating the robustness of each recognizer. To this end, we set up a scenario in which the participants were instructed to perform specific social cues. This procedure allowed us to obtain accuracy measurements for behaviors we were interested in under controlled conditions. To see how the whole system performs in a more natural environment, we devised a second study in which the participants took part in a prototypical job interview using the system described in Section IV. The objective of the second study was to investigate whether the participants found the system useful as a preparation for job interviews.

A. Accuracy Study

For the first study we recruited 11 persons (10 male and 1 female) with an average age of 30.4. Each participant was instructed to perform a series of actions, each representing a social cue: *look away*, *arms open*, *hands behind head*, *arms crossed*, touch face (*hand to face*), *lean back*, *lean forward* and read a text out loud (*voice activity*). The order of these was counterbalanced between the participants to compensate for learning effects or any possible stress users might experience at the start of the study. A Microsoft Kinect was positioned

TABLE II. EVALUATION RESULTS SHOWING THE MEAN RECOGNITION RATES OF EACH SOCIAL CUE.

Social Cue	Recall
Look away	81.8%
Arms open	100.0%
Hands behind head	100.0%
Arms crossed	81.8%
Hand to face	90.9%
Lean backward	72.7%
Lean forward	81.8%
Voice activity	100.0%
Average	88.64%

in front of the participants at a distance of approximately 1.1m from the participants' head and 1.3m of the ground. The participants also wore an AKG C 444 close-talk microphone.

Results and Discussion. The evaluation of the data yielded an overall mean recognition rate of 88%, with three social cues (*arms open*, *hands behind head* and *voice activity*) achieving 100%. The worst recognition rate was obtained for the *lean back* social cue with 72.7%. The results are shown in Table II. The main reason for detection failures was the rather unstable tracking provided by the Microsoft Kinect SDK with the users sitting down. However, if we consider the benefits of the Microsoft Kinect (low cost, minimal intrusion), it becomes immediately clear that it is the best solution currently available for our purposes. Other skeleton tracking sensors, such as motion capturing systems, have a much higher intrusion level and an increased set-up time and complexity. At this point it is also important to note that the aim of the paper is not to provide an overview of the state of the art in terms of recognition algorithms but rather to prove that such techniques are viable in the context of computer-enhanced job interview simulations.

B. Job Interview Study

The purpose of our second study was to get a first impression of the system's impact on real participants. To this end we invited six students (five male, one female) with an average age of 28.83, to participate in a job interview using our system.

Design and Procedure. Each student received the job description one day before the scheduled interview with our system and was asked to prepare for it. In order to keep the study as realistic as possible, the job description was taken from a job exchange website. Each participant was seated at a table in front of a 55 inch display at a distance of 1.5m. Above the display, a Microsoft Kinect was positioned facing the participant. Each participant was asked to wear a headset. The setup is shown in Fig. 6. The first part of the study consisted of the participants taking part in the job interview simulation using our system. The simulated interview was structured in five parts, *Welcome*, *Introduction*, *Elaboration*, *Synthesis* and *Conclusion*, and included a total of 13 questions. The virtual recruiter went through the questions one by one applying turn taking as described in Section IV-A. After the simulation, we performed an informal post-hoc interview with each participant to gather information about the participants impressions and experiences. Once these interviews were done, we showed each participants the data that had been collected

Fig. 6. User study setup showing a user in a job interview with a virtual recruiter.

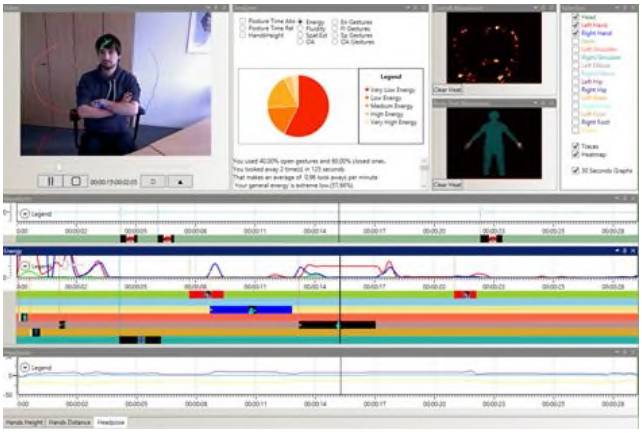

Fig. 7. The NovA (NOnVerbal behavior Analyzer) analysis tool used to debrief the participants.

during the simulation using the NovA[3] visualization tool (see Fig. 7) and asked them again for their feedback.

Results and Discussion. Most of the participants were positively surprised by the system with five out of the six participants feeling that it was helpful for preparing for real job interviews. "Good for practicing behaviors", "I became aware of behaviors I don't normally think about" were some of the comments made. All participants found the visualization very useful saying that it helped them understand what went wrong during the interview and how to improve themselves. Interestingly, the participants quickly forgot that they were participating in an experiment. All participants stated that even though they consciously controlled their movements at the start of the simulation, their non-verbal behavior quickly became unconscious, similar to an interaction with a human. One participant stated that while the simulation did not feel as real as an interview with a person, it was still better than a telephone interview. In general, the attitude towards the virtual recruiter was positive. Five participants thought that the virtual recruiter was sentient and three felt that he was interested in

[3]http://openssi.net/nova

them. When asked to elaborate, the participants indicated as a reason that the recruiter smiled at times.

However, the participants also pointed out some problems. The main issue was the timing of the agent's turn taking behavior. In some cases, the system interpreted longer thinking pauses of the participants as ends of a turn and continued with the interview. This behavior gave the participants the feeling that the recruiter was impatient or even interrupting them. The visual appearance of the virtual character was also criticized as being not realistic enough. One participant also stated that he was uncomfortable because the table was too far away and he could not rest his hands. This might explain the low amount of gesticulation observed during the study. However, the rather large distance to the table was necessary to ensure robust tracking with the Microsoft Kinect. We will look into solutions for this issue as part of our future work.

VI. Conclusion

In this paper, we presented an approach that enhances a virtual agent by the ability to interpret and respond to social cues of users participating in a simulated job interview. The ultimate goal of our system is to help young people improve social skills pertinent to job interviews. In order to achieve seamless credible interaction, our system automatically recognizes the user's social cues in real time. Based on these, the virtual recruiter reacts and adapts to the user's behavior. Furthermore, the interaction with the virtual agent can be recorded and presented to the user to enhance the learning effect, for example, by identifying critical incidents during the simulated interview.

As a first step we conducted mock job interviews with actual NEETs (people not in employment, education or training) and trained practitioners from a specialized inclusion center to identify what social cues are actually relevant to this scenario. We then implemented a system which allows user to participate in a job interview simulation. The proposed system is built around a behavior management framework for virtual characters which has been extended by two modules: a scenario manager and a social cue recognizer. The scenario manager was used to model the virtual recruiter's interactive behavior allowing the character to react to various social cues recognized by the social cue recognition module. More precisely, we modeled mirroring and turn taking behavior.

We did not only demonstrate the ability of the recognition component to reliably recognize social cues pertinent to job interviews, but also evaluated the concept by conducting informal interviews with users that had run through simulated job interviews using our system. Despite several reported problems, such as the realism of the character's appearance, all participants' reactions were mainly positive saying they would use such a system to train for real job interviews.

VII. Future Work

As part of our future work we will focus on three areas. First, we want to extend the recognition module to allow it to detect additional social cues, such as voice features, eye gaze or laughter. Voice features, such as pitch functionals, will give us a better understanding of the current affective state of the user enabling more accurate reactions by the virtual recruiter. Extending our current head gaze detection with eye gaze information will drastically increase the precision in detecting when the youngster is looking away. Eye gaze will also allow us to determine more subtle social cues, such as saccades or eye contact. However, such data requires the use of an eye tracker. Therefore, a first step would be to evaluate the impact such a sensor will have on the youngsters and the interview. Another social cue we are currently looking at is laughter. For this we are investigating the use of fusion methods to increase detection rate. Our main goal is to be able to recognize all the social cues observed in the knowledge elicitation study.

Secondly, we are also working on improving the way the system reacts and adapt to the user's actions. For example, we plan to make have the virtual recruiter and the scenario adapt to the user's performance and emotional state. This will allow the virtual recruiter to respond in real-time in a way that is similar to how a real interviewer might do. Enabling a system to emulate socially credible behaviors and thus, to act in socially contingent manner with respect to the user, is aimed to generate user immersion and flow, and ultimately learning especially with respect to their socio-emotional self-regulation.

Finally, additional studies involving actual NEETs are planned to profoundly test the capabilities and the performance of our system in the desired context.

Acknowledgements

We would like to thank Alexis Heloir for being a great help and for providing the Virtual Character Ben within the EMBOTS framework.

This work was partially funded by the European Commission within FP7-ICT-2011-7 (Project TARDIS, grant agreement no. 288578).

References

[1] R. MacDonald, "Disconnected youth? social exclusion, the underclass and economic marginality," *Social Work and Society*, vol. 6, no. 2, pp. 236–247, 2008.

[2] T. Hammer, "Mental health and social exclusion among unemployed youth in scandinavia. a comparative study," *Intl. Journal of Social Welfare*, vol. 9, no. 1, pp. 53–63, 2000. [Online]. Available: http://dx.doi.org/10.1111/1468-2397.00108

[3] R. D. Arvey and J. E. Campion, "The employment interview: A summary and review of recent research," *Personnel Psychology*, vol. 35, no. 2, pp. 281–322, 1982. [Online]. Available: http://dx.doi.org/10.1111/j.1744-6570.1982.tb02197.x

[4] J. Curhan and A. Pentland, "Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes," pp. 802–811, 2007.

[5] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *HumanComputer Interaction*, vol. 28, no. 1, pp. 40–75, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/07370024.2012.688468

[6] J. Greene and B. Burleson, *Handbook of Communication and Social Interaction Skills*, ser. LEA's Communication Series. L. Erlbaum Associates, 2003. [Online]. Available: http://books.google.de/books?id=h3i-refbDKwC

[7] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users' affective states," *Applied Artificial Intelligence*, vol. 19, no. 3-4, pp. 267–285, 2005. [Online]. Available: http://www.ingentaconnect.com/content/tandf/uaai/2005/00000019/F0020003/art00004

[8] B. Endrass, E. André, M. Rehm, and Y. Nakano, "Investigating culture-related aspects of behavior for virtual characters," *Autonomous Agents and Multi-Agent Systems*, 2013.

[9] D. R. Traum, D. DeVault, J. Lee, Z. Wang, and S. Marsella, "Incremental dialogue understanding and feedback for multiparty, multimodal conversation," in *Intelligent Virtual Agents - 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings*, ser. Lecture Notes in Computer Science, Y. Nakano, M. Neff, A. Paiva, and M. A. Walker, Eds., vol. 7502. Springer, 2012, pp. 275–288.

[10] S.-H. Kang, J. Gratch, C. L. Sidner, R. Artstein, L. Huang, and L.-P. Morency, "Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure," in *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, W. van der Hoek, L. Padgham, V. Conitzer, and M. Winikoff, Eds. IFAAMAS, 2012, pp. 63–70.

[11] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69 –87, jan.-march 2012.

[12] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, july 2005, pp. 474 –477.

[13] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39 –58, jan. 2009.

[14] A. Kleinsmith and N. Bianchi-Berthouze, "Form as a cue in the automatic recognition of non-acted affective body expressions," in *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part I*, ser. ACII'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 155–164. [Online]. Available: http://dl.acm.org/citation.cfm?id=2062780.2062801

[15] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 677–682. [Online]. Available: http://doi.acm.org/10.1145/1101149.1101300

[16] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[17] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 31–38. [Online]. Available: http://doi.acm.org/10.1145/2388676.2388686

[18] 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*, ser. Lecture Notes in Computer Science, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Springer Berlin Heidelberg, 2012, vol. 7502, pp. 455–463. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33197-8_47

[19] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 165 –183, april-june 2012.

[20] M. Cavazza, D. Pizzi, F. Charles, T. Vogt, and E. André, "Emotional input for character-based interactive storytelling," in *Proc. 8th Intl. Conf. on Autonomous Agents and Multiagent Systems Vol1*, ser. AAMAS '09. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 313–320. [Online]. Available: http://dl.acm.org/citation.cfm?id=1558013.1558056

[21] R. Niewiadomski, J. Hofmann, J. Urbain, T. Platt, J. Wagner, B. Piot, H. Cakmak, S. Pammi, T. Baur, S. Dupont, M. Geist, F. Lingenfelser, G. McKeown, O. Pietquin, and W. Ruch, "Laugh-aware virtual agent and its impact on user amusement ," in *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS2013)*, Saint Paul, USA, May 2013.

[22] H. Sloetjes and P. Wittenburg, "Annotation by category: Elan and iso dcr." in *LREC*. European Language Resources Association, 2008. [Online]. Available: http://dblp.uni-trier.de/db/conf/lrec/lrec2008.html#SloetjesW08

[23] A. Heloir and M. Kipp, "Real-time animation of interactive agents: Specification and realization," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 510–529, 2010.

[24] M. Kipp, A. Heloir, M. Schröder, and P. Gebhard, "Realizing multimodal behavior: closing the gap between behavior planning and embodied agent presentation," in *Proceedings of the 10th international conference on Intelligent virtual agents*, ser. IVA'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 57–63. [Online]. Available: http://dl.acm.org/citation.cfm?id=1889075.1889083

[25] P. Gebhard, G. Mehlmann, and M. Kipp, "Visual scenemaker - a tool for authoring interactive virtual characters," *Journal on Multimodal User Interfaces*, vol. 6, no. 1-2, pp. 3–11, 2012. [Online]. Available: http://dx.doi.org/10.1007/s12193-011-0077-1

[26] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework - multimodal signal processing and recognition in real-time," in *Proceedings of ACM MULTIMEDIA 2013*, Barcelona, 2013.

[27] A. Pease and J. Kent, *Body Language: How to Read Others' Thoughts by Their Gestures*. Camel Pub., 1981. [Online]. Available: http://books.google.de/books?id=_EeFQgAACAAJ

[28] F. Kistler, B. Endrass, I. Damian, C. T. Dang, and E. André, "Natural interaction with culturally adaptive virtual characters," *Journal on Multimodal User Interfaces*, vol. 6, pp. 39–47, 2012. [Online]. Available: http://dx.doi.org/10.1007/s12193-011-0087-z

[29] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modifiedcensus transformation," *Image Vision Comput.*, vol. 24, no. 6, pp. 564–572, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2005.08.005

[30] R. E. Kraut and R. E. Johnston, "Social and Emotional Messages of Smiling: An Ethological Approach," *Journal of Personality and Social Psychology*, vol. 37, no. 9, pp. 1539–1553, 1979.

[31] J. Harrigan and K. Taing, "Fooled by a smile: Detecting anxiety in others," *Journal of Nonverbal Behavior*, vol. 21, no. 3, pp. 203–221, 1997. [Online]. Available: http://dx.doi.org/10.1023/A%3A1024921631009

[32] G. Caridakis, A. Raouzaiou, K. Karapouzis, and S. Kollias, "Synthesizing gesture expressivity based on real sequences," *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC Conference Genoa, Italy*, Mai 2006.

[33] H. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, no. 28, pp. 879–896, 1998.

[34] B. Hartman, M. Mancini, and C. Pelachaud, "Implementing expressive gesture synthesis for embodied conversational agents," *Gesture Workshop Vannes*, 2005.