MACH: My Automated Conversation coacH

Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, Rosalind Picard

MACH: My Automated Conversation coacH

Mohammed (Ehsan) Hoque [1], Matthieu Courgeon[2], Jean-Claude Martin[2], Bilge Mutlu[3], Rosalind W. Picard[1]
MIT Media Lab[1], LIMSI-CNRS[2], University of Wisconsin–Madison[3]
mehoque@media.mit.edu, {courgeon, martin}@limsi.fr, bilge@cs.wisc.edu, picard@media.mit.edu

ABSTRACT
MACH—My Automated Conversation coacH—is a novel system that provides ubiquitous access to social skills training. The system includes a virtual agent that reads facial expressions, speech, and prosody and responds with verbal and nonverbal behaviors in real time. This paper presents an application of MACH in the context of training for job interviews. During the training, MACH asks interview questions, automatically mimics certain behavior issued by the user, and exhibit appropriate nonverbal behaviors. Following the interaction, MACH provides visual feedback on the user's performance. The development of this application draws on data from 28 interview sessions, involving employment-seeking students and career counselors. The effectiveness of MACH was assessed through a weeklong trial with 90 MIT undergraduates. Students who interacted with MACH were rated by human experts to have improved in overall interview performance, while the ratings of students in control groups did not improve. Post-experiment interviews indicate that participants found the interview experience informative about their behaviors and expressed interest in using MACH in the future.

Author Keywords
Social training, automated multimodal affect sensing and behavior recognition, automated feedback, embodied conversational agents.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces – *input devices and strategies, evaluation/ methodology, user-centered design*. H.1.2 Models and Principles: User/Machine Systems – *human factors, software psychology*

INTRODUCTION
Mike, a technically gifted college junior, worries about not getting any internship offers this year and thinks that he needs to improve his interview skills. After a 15-minute session with a career counselor, he receives recommendations to maintain more eye contact with the interviewer, end the interview with a social smile to appear friendly and use intonation and loudness effectively to express enthusiasm. Mike returns to his dorm room with an understanding of several behaviors that he can improve for his upcoming interviews. He wishes to practice with and get feedback from a counselor, but schedule conflicts and limited counselor availability make it difficult. He is also unwilling to ask his peers for help, as he fears social stigma.

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

ACM 978-1-4503-1770-2/13/09…$15.00.

Figure 1. MACH interviews a participant.

Is it possible to help Mike and others like him improve their social skills using an automated system that is available ubiquitously — where they want and when they want?

This paper presents the design, implementation, and evaluation of MACH—My Automated Conversation coacH—a novel technology that automates elements of behavior analysis, embodied conversational analysis, and data visualization to help people improve their conversational skills (Figure 1). MACH automatically processes elements of facial expressions and speech and generates conversational behaviors including speech and nonverbal behaviors, as illustrated in Figure 2. Automated processing and generation allow the system to engage in dialogue with a participant in real-time, creating the illusion that it can "see," "hear," and "respond."

In this paper, we explore the use of MACH in the context of training for job interviews. During an interaction, MACH asks common interview questions used by human interview coaches, and after the interaction, it provides interviewees with personalized feedback to enable self-reflection on the success of the interview.

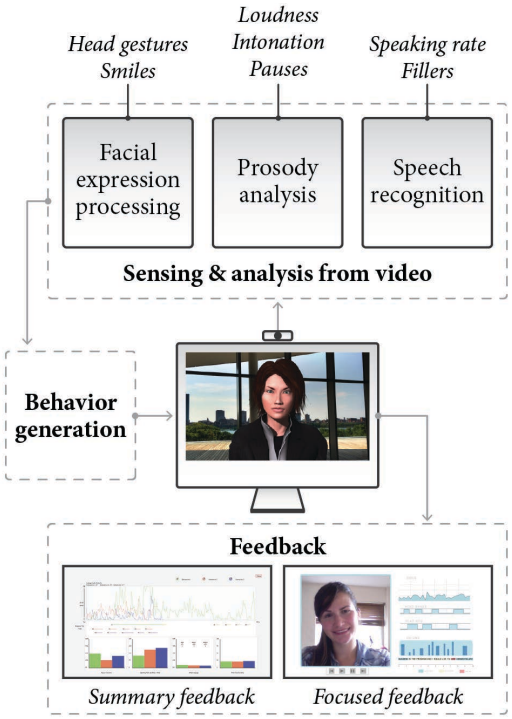

Figure 2. The MACH system works in a regular laptop, which processes the audio and video inputs in real -time. The processed data is used to generate the behaviors of the 3D character that interacts with and provides feedback to participants.

In this context, this paper seeks to address the following research questions:

1) How might computer technology use counseling as a metaphor, offering an individual the ability to practice social interaction and to receive useful feedback?
2) How would a system automatically sense, interpret, and represent conversational multimodal behavioral data into a format that is both intuitive and educational?
3) Will such technology elicit measurable improvements in social skills in participants?

PREVIOUS WORK

The design of an automated coach for social-skills training brings together several disparate bodies of knowledge, including computing research into intelligent virtual agents, affective computing, information visualization and clinical and professional programs for social-skills training. The paragraphs below briefly outline work in these areas.

Examples of real-time systems combining a virtual agent, incremental analysis of behaviors, dialogue management and synthesis of behaviors include *Sensitive Artificial Listener* [1] and *Rapport Agent* [2]. While these systems successfully integrate aspects of affective analysis and interactive characters, they do not include other components that are necessary for an automated coach such as, 1) a realistic task such as training real users, 2) formative affective feedback that provides the user with useful feedback on what behaviors need improvement, and 3) the interpretation or recognition of user utterances to drive the selection of backchannels or formative feedback.

Real-time nonverbal sensing, interpretation, and representation involve significant technical challenges in affective computing. While the research community has made significant progress in nonverbal sensing, abstracting the raw nonverbal data into an interactive, intuitive, and accessible graphical format remains an open problem. Kaliouby et al. [3] developed a real-time system called MindReader to recognize and visualize complex mental states by analyzing facial expressions and displaying the inferred states using radial charts, traffic lights, and lines that change in real-time. However, MindReader focuses on a single modality and displays analysis results only in real-time, lacking support for offline review. CERT [4], another real-time behavior analysis system, was developed to recognize low-level facial features such as eyebrow raise and lip corner pull and graph them as a function of time. However, CERT wasn't designed for people to practice and interpret their behaviors for a given task. While these systems illustrate the promise of real-time, automated behavior sensing, there is a need for combining human-centric designs with multimodal nonverbal sensing that not only pushes the boundary of automated behavior sensing, but also empowers the users to understand and reflect on their own behaviors.

Research in behavioral health has explored interventions for social-skills training to help individuals with social deficits such as individuals on the autism spectrum. For example, to help individuals with autism improve their understanding of social interaction nuances and their responses to social situations, interventions such as s*ocial stories* [5], *social skills group* [6], and *cue cards* [7] have been developed. In most of those cases, the best results were achieved when the interventions targeted specific skills such as modeling and role-play, and when these skills were practiced both in the classroom and at home [8]. Researchers have also explored the use of computerized interventions in treatments for social anxiety. For example, Beard [9] demonstrated that, using a cognitive-bias modification, patients with social anxiety disorder exhibited significantly greater reductions in levels of social anxiety compared to patients in the control group. In addition, during the four-month follow-up, the patients who underwent the intervention continued to maintain their clinical improvement and diagnostic differences. This finding is encouraging, as one of the major challenges of automated behavioral intervention is to ensure that skills generalize beyond the intervention duration.

The importance of social skills and their development is recognized beyond clinical research and practice and includes areas such as professional development and placement. For instance, job seekers seek to assess or improve their interview skills.

Figure 3. Experimental setup of the mock interviews. Camera #1 recorded the video and audio of the interviewee, while Camera #2 recorded the interviewer.

Common assessment strategies include recording one's own behaviors and watching the recordings. Companies such as Walmart, Nike, Starbucks, Dunkin' Donuts, and eBay use automated web-based technologies such as HireVue [10] that require candidates to record answers to interview questions for later assessment. Using these recordings, employers eliminate unfavorable candidates, often using simple behavioral rules. For example, Holiday Inn was reported to eliminate interviewees who smiled less than a given threshold [11]. Such practices highlight the changing nature of and technology use for assessment in professional placement and underline the growing need for technologies that help people improve their communication skills in such contexts.

CONTEXTUAL INQUIRY

To inform the design of MACH as an automated coach for social skills training and contextualize the design in training for job interviews, we sought to understand how expert interviewers carry out mock job interviews. Our first consideration was the appearance of MACH; namely, should it look human-like or cartoon-like? Second, what kind of feedback should MACH provide? Should it provide real-time feedback to indicate how well the interview is going, or should it wait for the interview session to end before debriefing its user, just like in other standard mock interviews? If the latter, how should MACH debrief the user? Should it provide feedback verbally or visually? If the latter, what should the visualizations look like? We approached these questions with a process of iterative design, system development, and a sequence of formative and summative user studies.

Study Setup

To better understand how expert interviewers facilitate mock interviews, we conducted a mock-interview study in a room equipped with a desk, two chairs, and two wall-mounted cameras mounted that captured the interviewer and the interviewee, as shown in Figure 3.

Participants

The study enrolled 28 college juniors (16 females and 12 males), all of whom were native English speakers from the MIT campus, and four professional MIT career counselors (three females and one male) who had an average of over five years of professional experience as career counselors and advanced graduate degrees in professional career counseling.

Procedure

The students were recruited through flyers and emails. They were told that they would have the opportunity to practice interviewing with a professional career counselor and would receive $10 for their participation. They were also informed that their interview would be recorded. Male participants were paired with the male counselor, and female participants were paired with one of the female counselors in order to minimize gender-based variability in behavior, as discussed by Nass et al. [12]. Each participant was debriefed and was asked to give feedback following the interview.

Interaction

With input from the MIT career counselors, we developed 15 likely interview questions that would be applicable to most real-world job positions. The five questions below were the most common questions to all our lists. They were presented in the following order by the counselors to participants:

Q0. How are you doing today? (Or an alternative question that the interviewer uses to initiate the interview.)

Q1. So, please tell me about yourself.

Q2. Tell me about a time when you demonstrated leadership.

Q3. Tell me about a time when you were working in a team and faced with a challenge. How did you solve that problem?

Q4. What is your weakness, and how do you plan to overcome it?

Q5. Now, why do you think we should hire you?

Data Analysis and Findings

The data was manually analyzed and annotated by two Facial Action Coding System (FACS) [13] trained coders for nonverbal behavior analysis who coded the data for response patterns and average smile duration. The following three behaviors emerged from the analysis of the interview data.

Expressiveness: Our data and observations suggested that counselors maintained a neutral composition during the interviews; however, counselors also matched the expressions of the interviewees by reciprocating smiles and other behaviors.

Listening behavior: In almost all of the interactions, the counselors of both sexes asked a question, carefully listened to the answer, briefly acknowledged the answer, and then moved on to the next question. The listening behavior of the counselor included subtle periodic head nods and occasional crisscrossing of the arms.

Figure 4. The female and male coaches used in the MACH system

Acknowledgements: The counselors used a similar set of acknowledgements at the end of each answer by the interviewee. Examples include "That's very interesting," "Thanks for that answer," "Thank you," and "I can understand that."

Duration: The average duration of the interview sessions between the male counselor and male interviewees were 7 minutes (SD: 2) whereas sessions between female counselors and female interviewees lasted for about 9 minutes in average (*SD*: 2.3). Both male and female counselors spent equal amount of time engaging in feedback with the interviewees.

INTERACTION DESIGN

Our goal in designing MACH was to create an autonomous system that appears responsive, has backchanneling and mirroring abilities, acts aware with real-time processing of facial expressions, recognizes spoken words along with its intonation, seems life-like in size and resolution, provides real-time affective feedback, and positively impacts interview skills. The interaction must be non-intrusive with all the sensing conducted via a standard microphone-enabled webcam (i.e., no wearable headset or microphone or any other physiological sensors should be required).

MACH should enable the following scenario for Mike and others like him.

Mike chooses to use the MACH system after class and over the weekend to improve his interview skills. He chooses one of the two counselors (Figure 4), John, who appears on the screen, greets him, and starts the interview. When Mike speaks, John periodically nods his head to express acknowledgement, shares smiles, and mirrors Mike's head movements. After the interview, John asks Mike to review the feedback on the side of the screen. Mike sees his own smile track for the entire interaction and notices that he never smiled after the opening. He also gets measurements of his speaking rate, intonation, and duration of pauses. Mike also chooses to watch the video of himself during the interview, which helps him identify when his speaking volume was not loud enough for some segment of his answers and when his intonation went flat. He decides to continue practicing at his dorm room, at his parents' home over the weekend, at the cafe where he usually studies while the system keeps tracks of his performance across sessions, allowing for objective comparisons and self-reflection.

Design of the Automated Coach

In order to create a realistic interview experience, we focused on three main components: 1) appearance of the coach—life-like and professional-looking to create a close-to-real-life feeling of being in an interview; 2) interaction between the coach and the user—fluid, interactive, and responsive, including rich nonverbal interaction; and, 3) affective feedback—summaries of the interviewee's behaviors for later review.

Appearance: The use of highly humanlike representations might elicit feelings of eeriness, an outcome often described as the "Uncanny Valley Effect" [14]. However, in this work, we take the position that interviews are often stressful and a visual and behavioral representation that supports the appearance of an intelligent and dominant conversational human partner [15] [16] and elicits stress can help in creating a realistic interview experience.

Interaction: Cassell [17] states that a virtual character's behaviors make up a larger part of its affordances for interaction than does its appearance. Cassell and Tartaro [18] argued that along with embodiment and anthropomorphism, virtual characters should also follow the "behavioral" affordances of human communication. For example, nodding and smiling at appropriate times to acknowledge the interviewee's answers to questions might make MACH more credible than a virtual character that stares at the interviewee the entire interview. Thus, one of the design considerations for MACH was to make it appear responsive to and aware of the interviewee.

Feedback: In our human interview study, career counselors maintained a neutral expression during the entire interview process with occasional nonverbal backchanneling behaviors (nodding of head, sharing smiles) and used more expressive language during the feedback session after the interview. Therefore, we designed MACH to display neutral acknowledgements in response to user behaviors and provide more detailed feedback at the end of the interview. In addition, we decided to design the summative feedback at the end of the interview in the form of interactive visualizations in order to capture the finer aspects of the interviewee's behaviors, which the virtual agent might not be able to effectively communicate using speech.

Because what constitutes "good interview performance" is largely subjective, and the development of an objective metric for interview performance is an open question, we chose to design visualizations that enabled users to engage in a process of guided self-exploration and learning in interaction context.

Figure 5. The two forms of feedback provided by MACH. The summary feedback (left) captures the overall interaction. Participants can practice multiple rounds of interviews and compare their performance across sessions. The focused feedback (right) enables participants to watch their own video. As they watch the video, they also can see how their nonverbal behaviors, such as smiles, head movements, and intonation change over time.

The design of these visualizations involved an iterative design process comprising several rounds of visual design, implementation, and formative evaluation. The next section describes this process.

Iterative Design of Feedback

Our goal was to design visualizations that were easy for users to understand and interpret and that enabled them to make comparisons across sessions. We performed four iterative design exercises to understand how we might best visualize the transcribed speech, prosodic contour, speaking rate, smiles, and head gestures in an interface that is intuitive, appealing, and insightful. This process and the resulting design are summarized in the paragraphs below.

Our initial prototype used motion, orientation, and other design elements to map the prosodic tone of the speech signal to the words, similar to visualizations developed by Rosenberger [19] in the *ProsodicFont* system. Findings from formative evaluations of this prototype showed that while some users found this form of immediate feedback useful, they also showed interest in seeing how their behaviors changed throughout the interview. Therefore, we chose to incorporate both kinds of feedback in two phases: 1) Summary feedback and 2) Focused feedback. Summary feedback enabled participants to track their progress across many sessions, as shown in Figure 5 (on the left). The top half of the interface contains a visualization of the user's smiles of the entire interaction for each session, displaying data from multiple sessions in different colors. This information provides the user with an idea of when and how much he or she smiled during the interview. The bottom half of the interface contains the following four dimensions of affective cues:

Total pause duration: The percentage of the duration of the pauses in the user's speech.

Speaking Rate: Total number of spoken syllables per minute during the entire interaction.

Weak Language: Filler words such as "like," "basically," "umm," "totally," calculated as the percentage of the interviewee's spoken words. The complete list of words that we considered as weak was obtained from Blue Planet Public Speaking [20].

Pitch Variation: The fourth dimension was the variability in the pitch of the speaker's speech.

Once the user is shown the summary feedback, he or she is given the option to view the focused feedback, as shown in Figure 5 (on the right). The design of the focused feedback was informed by elements of treatment programs developed for social phobia [21]. For example, these treatment programs inform individuals with social phobia of how they might appear to others by asking them to watch videos of their own behaviors. One drawback of this approach is that many such individuals view their video appearance negatively. To resolve this problem and to maximize the discrepancies between the individual's self-image and the video, the individual is asked, 1) to imagine how they will appear before viewing the video, 2) to create a picture of what their negative behaviors will look like, and 3) to ignore their feelings and watch the video as if it is someone else's. Our design seeks to emulate this strategy and elicit those features and feelings by juxtaposing the nonverbal data with their video. The different variables for which data is visualized are described below and illustrated in Figure 5 on the right.

Video: As the participant interacts with MACH, it captures the video of the interaction using the webcam and displays it to the participant, as shown in Figure 5 (right).

Smiles: The system captures intensity of smiles at each frame and displays intensity as a temporal pattern in the upper pane of the interface.

Head Movements: Head nods and shakes are recognized per frame as an output of 0 (not present) or 1 (present). Therefore, they are plotted as binary patterns, as shown in Figure 5 (right).

Spoken Words: The spoken words are plotted at the bottom of the interface with weak and strong language marked in red and blue, respectively.

Loudness: The loudness of each word was plotted as a bar, providing the user with a comparative view of the loudness of all the words in an utterance, shown at the lower right part of Figure 5.

Emphasis and Pauses: The space that each word occupies in the interface corresponds to the amount of time the user took to enunciate it. Thus, the visualization conveys the emphasis that the speaker puts on each word with elongation, which corresponds to enunciation time, and height, which corresponds to loudness. The space between each pair of words represents the length of the pause between them.

In summary, the focused feedback display provides an opportunity for participants to view both their interview video and data on various nonverbal behaviors as a function of time. This allows the users to identify behaviors across multiple modalities that are out of sync, such as using emphasis improperly or smiling inappropriately, with fine resolution and quantifiable patterns.

SYSTEM IMPLEMENTATION

Our implementation sought to realize these design decisions in an autonomous system, combining various sub-systems for simulating the virtual agent, sensing and analysis of user behavior, and visualization of feedback. The implementation of each sub-system is outlined below.

Facial Expression Processing: From the video of the user's face, we tracked smiles and head movements (e.g., nods, shakes, tilts) in every frame. We used the Shore Framework [22] to detect faces and facial features in order to distinguish smiles. The classifier was trained using the Adaboost algorithm with sample images, considering smiling and neutral as the two binary classes. The features from all over the face were used for boosting. The outcome of the classifier was a normalization function that projected the score onto a range, [0,100]. Thus, each face image was scored from 0 to 100 in smile intensity, where 0 and 100 represented no smile and full smile, respectively. We evaluated this smile recognition system using the Cohn-Kanade dataset [23], which includes 287 images of 97 individuals from the United States. Out of these images, 62 were labeled as happy, where happiness is defined as smiling at various levels. Our testing of the smile analysis module for classifying the images in this dataset yielded precision, recall, and F-score values of .90, .97, and .93, respectively.

In addition, we tested the smile module on the JAFEE 178-image dataset [24] of happiness, sadness, surprise, anger, and disgust, including 29 instances of happy faces from 10 Japanese women. The results from the testing with the JAFEE dataset yielded precision, recall, and F-measure values of .69, 1, and 0.81, respectively.

Head Nod and Shake Detection: Detecting natural head nods and shakes in real-time is challenging, because head movements can be subtle, small, or asymmetric. Our implementation tracked the "between eyes" region, as described by Kawato and Ohya [25]. The head-shaking detection algorithm is described below.

ALGORITHM. HEAD NOD AND SHAKE

If $max\ (X_{i+n}) - min\ (X_{i+n}) \leq 2 \quad (n = -2 \ldots + 2)$

  then *frame i is in stable state*

If $X_i = max\ (X_{i+n}) \quad (n = -2 \ldots + 2)$

  or $X_i = min\ (X_{i+n}) \quad (n = -2 \ldots + 2)$

  then *frame i is in extreme state*

else

  frame i is in transient state

Head nods are also detected using this algorithm, except that only movements in the y-axis are considered. The algorithm assigns one of three states to each frame: 1) stable, 2) extreme, and 3) transient.

At every frame, the system checks whether the state has changed from stable to extreme or transient in order to trigger the head shake evaluation process. Therefore, the system has a two-frame delay. If there are more than two extreme states between the current stable state and the previous stable state, and all the adjacent extreme states differ by two pixels in the *x*-coordinate, then the system records the movement as a head shake. When a user looks to the left or right, the system logs these as stable states instead of extreme states between or after the transient states, and therefore, these head turns do not get mislabeled as head shakes.

Prosody Analysis: For prosody analysis, we automatically recognize pauses, loudness, and pitch variation, which serve as a measure of how well the speaker is modulating his or her speech, as these are useful for assessing expressivity. In order to recognize pauses, we used a minimum pitch of 100 Hz, a silence threshold of −25 dB, minimum silent interval duration of .5 seconds, and minimum sounding interval duration of .5 seconds. Pitch was calculated using acoustic periodicity detection on the basis of an accurate autocorrelation method. To extract prosodic features, we developed an application programming interface (API) using the low level signal processing algorithms included in the *Praat* [26], an open source speech processing toolkit.

Speech Recognition: For real-time speech recognition, we used the Nuance speech recognition software development kit [27]. While the speech recognition system captures the entire transcription of the interaction,

it does not perform any natural language understanding, i.e., there is no assessment of the semantics of speech, as the current application focuses on nonverbal training.

Nonverbal Behavior Synthesis: MACH has been developed on an existing life-like 3D character platform called Multimodal Affective Reactive Characters (MARC) [28]. In order to provide users with a realistic interview experience, MACH must appear and behave humanlike, adapting its behaviors to changes in the interaction. Our implementation sought to achieve this level of realism by integrating the following four components into the animation of the virtual coach: arm and posture movements, facial expressions, gaze behavior, and lip synchronization.

Arm and Posture Animation: We designed a set of arm and postural animations to replicate behaviors that we observed in videos of our human interviewers, such as crossing arms, laying arms on the table, balance shift, and a number of hand gestures that accompanied speech. The animations for some of these movements were created using motion capture, while others were created manually because of occlusions during motion capture.

Lip Synchronization: In MACH, Lip synchronization was achieved using phonemes generated by Cereproc [29] while generating the synthesized voice. Phonemes are converted to visemes, the geometry of the lips, and animated using curved interpolation.

Gaze Behavior: The implementation of the virtual agent's gaze behavior involved directing the agent's eyes and the head toward specific gaze targets such as the user's face, and simulating saccades, rapid movements of the eyes between and around fixation points, and blinks.

Facial Animation: Facial behavior involves several communication channels, including facial expressions of emotions, movements of the eyes, and lip movements. Facial expressions of emotion were created by controlling Facial Action Units (FAU) [13] through spline-based animations using a custom built editor. Common facial expressions that we observed in our contextual inquiry were social signals, such as polite smiles, head nods, and frowning. We designed several variations for each of these behaviors. Using several animations for single expression, such as different ways of nodding, we sought to increase the dynamism and spontaneity of the virtual character's behaviors.

Head orientation and smiles are sensed by MACH and are mirrored. They are dynamically controlled by the behavior sensing and analysis module. Head nods are, however, controlled by a dedicated behavior manager.

Timing and Synchronization: Once designed, we had to ensure that the animations were timed appropriately during the interaction. Because interactions primarily involved MACH asking questions and listening to user responses, the virtual agent employed the majority of the behaviors during listening. In order to coordinate these behaviors in a realistic way, we implemented a *listening behavior module.* An analysis of the videos collected from the mock interviews revealed that in most interviews, counselors nod their heads to signal acknowledgment on average every 4.12 seconds. We utilized this observation by dynamically triggering and combining variations of head nods, arm and postural movements, and real-time mirroring of subtle smiles and head movements in our *listening behavior module*. The listening behaviors were exhibited at a randomly generated rate of frequency. Additionally, an animation was randomly selected from the set of nodding animations that we generated to prevent repetition and achieve spontaneous behavior. The listening module allowed smooth interruptions if the user ended his or her turn in the middle of an animation.

EVALUATION

Our evaluation of MACH sought to answer the following two research questions:

1) How effective is MACH in helping users improve their interview skills?
2) Do users find MACH easy to use and helpful?

Experimental Design

To find answers to these questions, we designed a user study with three experimental groups and randomly assigned participants to one of these groups, as shown in Figure 6. In Group 1, the control group, participants watched educational videos on interviewing for jobs that were recommended by the MIT career office. Participants in Group 2 practiced interviews with MACH and watched themselves on video. In Group 3, participants practiced interviews with MACH, watched themselves on video, and received feedback on their behaviors, interacting with all the functionality we incorporated into MACH. This experimental design allowed us to test the effectiveness and usability of our design of MACH against a baseline intervention of watching educational videos and the use of MACH only for practice without feedback.

All participants were brought into the lab for a first interview with a professional career counselor. Participants in the second and third groups were brought back into the lab for an hour-long intervention a few days after the initial interview. All participants were brought back into the lab a week after the initial interview, for the second time for participants in Group 1 and for the third time for those in Groups 2 and 3, to complete an additional interview with the same career counselor but with a different set of questions. The counselor was blind to the study conditions.

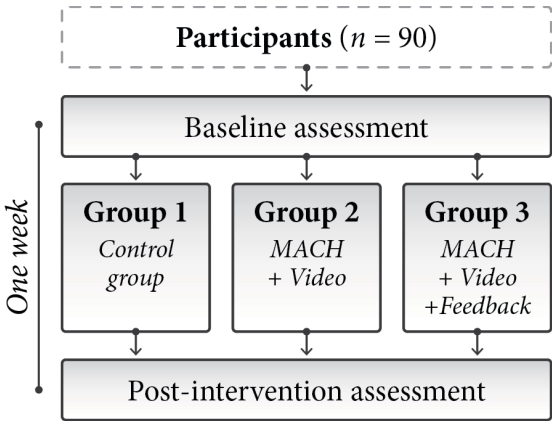

Figure 6. Study design and participant assignment to experimental groups.

Participants

We recruited 90 undergraduate students (53 females, 37 males) from the MIT campus. All of them were native English speakers, were in junior standing, and were likely to be looking for internships. We hired two professional career counselors (one male and one female) with several years of experience in conducting mock interviews and advising students on the interview process.

Procedure

The participants were given a generic job description and were asked to pretend that they were being interviewed for a job position at their favorite company. The study setup of the interview was similar to Figure 3. To minimize gender-interaction variability [12], male students were paired with the male counselor, and female students with the female counselor. After the mock interview, the counselor rated the interviewee's interview performance, and the interviewee rated his or her own performance. The study setup for Groups 2 and 3 was similar to the setup shown in Figure 1, in which MACH was displayed on a 46" Samsung Smart 3D TV. The experimenter left the room during the interview and asked the participant to exit the room once the study was complete. Participants were told that they could practice as many times as they wished, but they had to practice using the system at least once. However, the session automatically terminated after the third practice. During the practice, MACH asked interview questions that we developed based on observations from our contextual inquiry. Students who were in the control group watched 30 minutes of educational videos on tips for successful interviews by professional counselors.

Measures

Following the baseline and post-intervention interviews, the participant and the counselor filled out a questionnaire that evaluated the participant's interview performance (available at http://goo.gl/JeEHR), while the counselor filled out another one (available at http://goo.gl/QKqUm). The questionnaire included items related to the participant's overall interview performance and use of nonverbal cues such as eye contact, body language, and intonation, rated on a scale of 1 to 7. In addition to measuring the interviewer's and the interviewee's evaluations of the interview, we recruited two independent career counselors—one male and one female—from MIT's career services to rate the interview videos. We expect the ratings from these "independent counselors" to be more reliable, because (1) they were blind not only to the study conditions but also to the study phase, i.e., whether an interview was a baseline or post-intervention interview; (2) they did not interact with the participants and thus were less affected by biases that might have been introduced by interpersonal processes such as rapport; and, (3) they could pause and replay the video, which might have enabled them to analyze the interviews more thoroughly.

Following the intervention, participants were asked to fill out a questionnaire (Group2: http://goo.gl/Tfnwo & Group 3: http://goo.gl/dbaUB) to evaluate the quality of the interaction with MACH. In addition, they responded to the System Usability Scale (SUS) [30], a ten-item scale that measures subjective assessments of usability. SUS scores range 0 to 100 and increase as the perceived usability of the system increases. Finally, the participants were asked to provide open-ended verbal feedback after the study debrief. This feedback was recorded for transcription and qualitative analysis.

Results

The reporting of the quantitative data here includes only the ratings of the independent counselors on the item, "What was the overall performance during the interview?" Results on the complete set of measures will be reported in a future publication. Two counselors, blinded to the intervention type, rated the interview videos before and after the intervention. Figure 7 shows the difference in counselor ratings between the ratings after and before intervention, i.e., improvement across the three intervention types. The effect of intervention type was analyzed using one-way analysis of variance (ANOVA), and the effects of intervention type and participant gender were analyzed using two-way ANOVA. Planned comparisons between the "feedback" and "control" and "feedback" and "video" interventions involved Scheffé's method. Statistical details are provided only for significant results.

The analysis showed that intervention type significantly affected the change in counselors' ratings ($F[2,83]=4.89$, $p=.010$). Comparisons showed that the change in counselors' ratings of participants in the third group who used MACH with video and feedback was significantly higher than that of the control group ($F[1,83]=7.46$, $p=.008$) and that of participants in the second group who used MACH only with video ($F[1,83]=6.92$, $p=.010$).

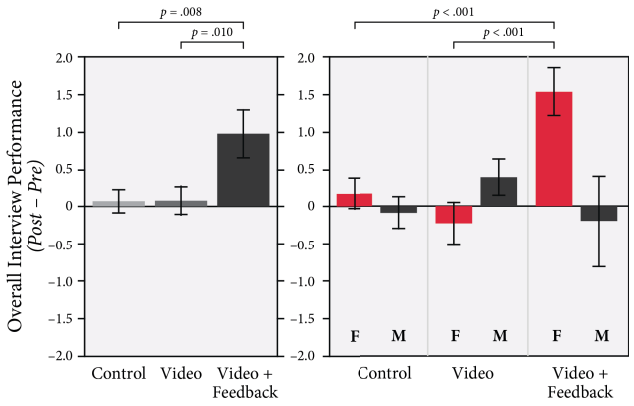

Figure 7. Improvement (post – pre) in independent counselor scores in item, "What was the overall performance during the interview," across conditions (left) and across conditions, broken down to females (F) and males (M) (right).

The analysis that also considered participant gender showed a significant interaction between intervention type and gender (F[2,80]=6.67, p=.002). Comparisons showed that counselors' ratings of females in Group 3 were significantly higher than that of females in Group 2 (F[1,80]=17.71, p<.001) and control group (F[1,80]=12.79, p<.001), while ratings of males did not differ across groups.

Subjective Evaluations: The paragraphs below describe findings from the participants' subjective evaluations of MACH and the open-ended feedback on their experience. The findings are grouped under themes in which qualitative and quantitative results overlap to represent user experience and system usability.

MACH as a Social Facilitator: The responsiveness of MACH's behavior was rated an average of 5.12 (SD = 1.4) by the participants in Groups 2 and 3. Most participants found the character's behavior to be natural.

> *"It has a lot of nonverbal stuff that you would want her to do. Some of the head tilt, acknowledging the speaking, nodding the head, act like it is listening to you and stuff…"*

> *"I was surprised that it knew when to nod its head, especially when it seemed natural to nod."*

> *"Just the way her eyes moved a little a bit… and after you responded… it seemed as if she was listening to you. I thought that made her kinda humanistic as opposed to a system."*

People Accept Humanlike: Overall, participants rated their preference toward a human-like character over a cartoon-like character an average of 4.20 (SD = 0.70), suggesting a preference toward a human-like character. The excerpts below provide further insight into participant preferences:

> *"…being here talking to a machine, I felt quite comfortable, which I didn't think I would feel."*

> *"I think the system is adding more value. I think if you were sitting across the table from me and you were recording, or taking notes, I would feel more intimidated. The fact that nobody is there is really helpful".*

Self-reflective Feedback is Useful: Most of the participants disliked looking at their video during the intervention in Groups 2 and 3. However, all participants overwhelmingly agreed that watching their video was, while discomforting, very useful (average rating of 6.3 out of 7, SD = 0.8). Additionally, participants rated whether they have learned something new about their behaviors at an average of 5.12 (SD = 1.20). This feeling was also reflected in the open-ended feedback from participants:

> *"I didn't like looking at my video, but I appreciated it."*

> *"I think it is really helpful. You don't really know unless you record yourself. This provides more analysis. The pauses may not appear that long, but when you look at it, you see something else."*

Speaking Rate was Most Useful: According to the participants' responses to the questionnaire; speaking rate, weak language (e.g., fillers), smile information, and pauses were the top 4 attributes of the visual feedback.

MACH is Easy to Use: The average of SUS ratings from participants in Groups 2 and 3 was 80 (SD = 11).

CONCLUSIONS

This paper presented a novel ubiquitous computing system called MACH, designed to help individuals improve their social skills by interacting with an automated coach. The system enables social skills training beyond counseling offices and clinical facilities, when and where users want it. The system affords embodied interaction and real-time feedback through a humanlike virtual character and post-interaction feedback on user performance through a visualization interface. The design of the application involved a contextual inquiry of 28 mock interview sessions and an iterative exploration of the design of visual feedback. A study was conducted with 90 college juniors and two professional career counselors to validate the effectiveness and usability of MACH in an interview scenario. Students who interacted with MACH and received automated affective feedback showed statistically significant performance improvement compared to students who were in the control and the video groups. Participants found the automated coach's behavior to be responsive and showed a preference toward interacting with a human-like character over a cartoon-like character. On average, participants reported that MACH enabled them to learn something new about their behaviors and agreed that they would like to use the system again in the future.

Our future work includes improving upon the ubiquitous nature of the use of MACH by extending the implementation to mobile platforms and settings as well as providing users with the ability to seamlessly distribute their repeated use of the system to different platforms over time. Additionally, we wish to provide users with the ability to compare their performance to their past

performance through progress charts as well as to that of users with specific characteristics such as educational background, geographic region, and job experience. Another natural extension of our current implementation is the addition of natural language understanding and sentiment analysis to analyze the content of answers. Finally, while we explored the application MACH in a professional development context, we wish to explore other application areas of our system for social skills training such as helping individuals with communicative challenges.

ACKNOWLEDGEMENT

The authors would like to acknowledge Sumit Gogia and Alex Sugurel for their help with UI design. Maarten Bos, Amy Cuddy, Jeffrey Cohn and L.-P. Morency provided valuable input with data analysis and study design. Special thanks to Nuance communications, Samsung, and Media Lab consortium for supporting our research.

REFERENCES

[1] M. Schroder et al., "Building Autonomous Sensitive Artificial Listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 1–20, 2011.

[2] J. Gratch et al., "Virtual Rapport," *Intelligent Virtual Agents*, vol. 4133, pp. 14 – 27, 2006.

[3] R. El Kaliouby and P. Robinson, "Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures," in *In the IEEE International Workshop on Real Time Computer Vision for Human Computer Interaction, CVPR*, 2004.

[4] G. Littlewort et al., "The computer expression recognition toolbox (CERT)," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops FG*, 2011, pp. 298–305.

[5] C. Smith, "Using Social Stories to Enhance Behaviour in Children with Autistic Spectrum Difficulties," *Educational Psychology in Practice*, vol. 17, no. 4, pp. 337–345, 2001.

[6] M. Solomon, B. L. Goodlin-Jones, and T. F. Anders, "A social adjustment enhancement intervention for high functioning autism, Asperger's syndrome, and pervasive developmental disorder NOS," *Journal of autism and developmental disorders*, vol. 34, no. 6, pp. 649–668, 2004.

[7] M. H. Charlop-christy and S. E. Kelso, "Teaching Children With Autism Conversational Speech Using a Cue Card / Written Script Program," *Children*, vol. 26, no. 2, pp. 108–127, 2003.

[8] C. B. Denning, "Social Skills Interventions for Students With Asperger Syndrome and High-Functioning Autism : Research Findings and Implications for Teachers," *Beyond Behavior*, vol. 16, no. 3, pp. 16–24, 2007.

[9] C. Beard, "Cognitive bias modification for anxiety: current evidence and future directions.," *Expert Review of Neurotherapeutics*, vol. 11, no. 2, pp. 299–311, 2011.

[10] "HireVue." [Online]. Available: hirevue.com/. [Accessed: 06-Nov-2013].

[11] M. LaFrance, *Lip Service*. W. W. Norton & Company, 2011.

[12] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices," *Journal of Applied Social Psychology*, vol. 27, no. 10, pp. 864–876, 1997.

[13] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: , 1978.

[14] M. Mori, "The Uncanny Valley," *Energy*, vol. 7, no. 4, pp. 33–35, 1970.

[15] J. Cassell, "Embodied conversational agents: representation and intelligence in user interfaces," *AI Magazine*, vol. 22, no. 4, pp. 67–84, 2001.

[16] B. Mutlu, "Designing Embodied Cues for Dialog with Robots," *AI Magazine*, vol. 32, no. 4, pp. 17–30, 2011.

[17] J. Cassell, "More than just a pretty face: conversational protocols and the affordances of embodiment," *Knowledge-Based Systems*, vol. 14, no. 1–2, pp. 55–64, 2001.

[18] J. Cassell and A. Tartaro, "Intersubjectivity in human–agent interaction," *Interaction Studies*, vol. 3, pp. 391–410, 2007.

[19] T. Rosenberger, "PROSODIC FONT: the Space between the Spoken and the Written," Massachusetts Institute of Technology, 1998.

[20] "Blue Planet Public Speaking." [Online]. Available: http://www.blueplanet.org/. [Accessed: 6-June-2013].

[21] D. M. Clark, "A cognitive perspective on social phobia," in *International Handbook of Social Anxiety Concepts Research and Interventions Relating to the Self and Shyness*, vol. 42, no. 1, W. R. Crozier and L. E. Alden, Eds. John Wiley & Sons Ltd, 2001, pp. 405–430.

[22] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 91–96.

[23] T. Kanade, J. F. Cohn, and Y. T. Y. Tian, "Comprehensive database for facial expression analysis," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[24] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[25] S. Kawato et al.,"Real-time detection of nodding and head-shaking by directly detecting and tracking the 'between-eyes'," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 40–45.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." [Online]. Available: www.praat.org. [Accessed: 21-Jun-2013].

[27] "Nuance Communications." [Online]. Available: http://www.nuance.com/for-developers/dragon/index.htm. [Accessed: 28-Jun-2013].

[28] M. Courgeon et al.,"Impact of Expressive Wrinkles on Perception of a Virtual Character's Facial Expressions of Emotions," in *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, 2009, pp. 201–214.

[29] "CereProc." [Online]. Available: http://cereproc.com. [Accessed: 28-Jun-2013].

[30] J. Brooke, "SUS - A quick and dirty usability scale," *Usability evaluation in industry*, pp. 189–194, 1996.