Automated Coach to Practice Conversations

Mohammed (Ehsan) Hoque
MIT Media Lab
75 Amherst Street,
Cambridge, MA, USA
mehoque@media.mit.edu

Rosalind W. Picard
MIT Media Lab
75 Amherst Street,
Cambridge, MA, USA
picard@media.mit.edu

Abstract— **We present a real-time system including a 3D character that can converse, capture, analyze and interpret subtle and multidimensional human nonverbal behaviors for possible applications such as job interviews, public speaking, or even automated speech therapy. The system works in a personal computer and senses nonverbal data from video (i.e., facial expressions) and audio (i.e., speech recognition and prosody analysis) using a standard webcam. We contextualized the development and evaluation of our system as a training scenario for job interviews. Using user-centered design and iterations, we determine how the nonverbal data could be presented to the user in an intuitive and educational manner. We tested efficacy of the system in the context of job interviews with 90 MIT undergraduate students. Our results suggest that the participants who used our system to improve their interview skills were perceived to be better candidates by human judges. Participants reported that the most useful feature was being given feedback on their speaking rate, and overall they reported strong agreement that they would consider using this system again for self-reflection.**

Virtual agent; nonverbal behaviors; feedback; prosody; facial expressions.

I. Introduction

We describe an automated system — My Automated Conversation coacH (MACH) — that interacts, captures, analyzes, and interprets human nonverbal behaviors in real-time to help people explore and understand their subtle communicative behaviors system (Figure 1). MACH is designed to be standardized, repeatable, and to work autonomously on a standard personal computer. The evaluation of MACH is contextualized in a "job-interview" training scenario. In this context, we followed a user-centered iterative design approach to define the interaction, and determine how to present the nonverbal parameters to the users using information visualization techniques.

II. Technical System

The system consists of the following sensing modules, as shown in Figure 1.

Nonverbal behavior synthesis: MACH has been developed on an existing life-like 3D character platform called Multimodal Affective Reactive Characters (MARC) [1]. It was designed to appear and behave human-like by adjusting its behaviors based on the interaction. This realism was achieved by integrating the following four components into the animation of the virtual coach: arm and posture movements, facial expressions, gaze behavior, and lip synchronization [2].

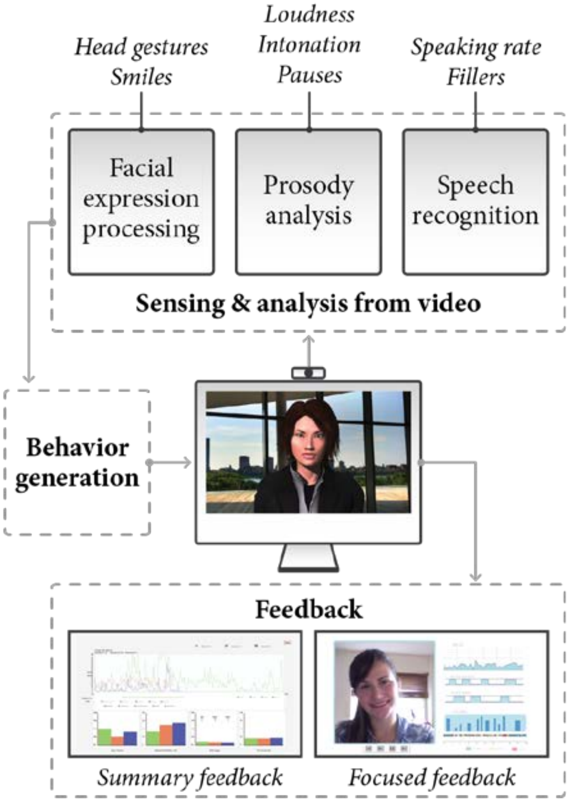

Figure 1. The MACH system works in a regular laptop, which processes the audio and video inputs in real time. The processed data is used to generate the behaviors of the 3D character that interacts with and provides feedback to participants.

Nonverbal Behavior Sensing: The sensing module consists of smiles, head gestures, speech recognition, pitch, loudness and pauses. Our decision of which nonverbal features we sense was determined by our ability to adequately recognize them using real-time automated computing.

Facial Expression Processing: From the face, we track smiles and head gestures (e.g., nods, shakes, tilts) in every frame. We used the Shore Framework [3] to detect faces

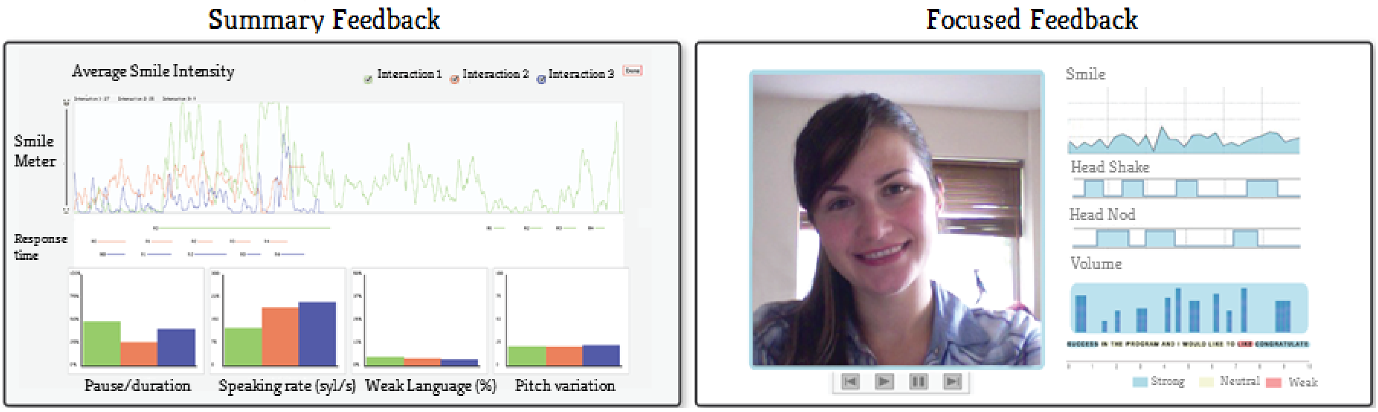

Figure 2. The two forms of feedback provided by MACH. The summary feedback (left) captures the overall interaction. Participants can practice multiple rounds of interviews and compare their performance across sessions. The focused feedback (right) enables participants to watch their own video. As they watch the video, they also can see how their nonverbal behaviors, such as smiles, head movements, and intonation change over time.

and facial features for recognizing smiles.

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 is motivated by tracking the "between eyes" region, as described by Kawato and Ohya [4].

Prosody Analysis: For prosody analysis, we automatically recognize pauses, loudness and pitch variation (e.g., how well one is modulating his/her voice), since these are useful for nonverbal expressivity assessment. We developed an API using low level signal processing algorithms from *Praat* [5], an open source speech processing toolkit, towards extracting prosodic features.

Speech Recognition and forced alignment: In order to perform word level prosody analysis, we use the Nuance SDK to first recognize the text, and then the forced aligner [6] to recognize the beginning and end of each word.

Speaking rate: Speaking rate is calculated by calculating total number of syllables per minute during the entire interaction.

Weak language: We want the interface to give feedback on the number of filler words (e.g., "like", "basically", "umm", "totally" etc.) that the user produces. We worked with the OwnTheRoom Public Speaking Company [7] to define the weak language dictionary.

III. APPLICATION

Understanding, analysis and interpretation of nonverbal behaviors are dependent on conversational context. To validate the system, we defined the interaction context as job interviews. We recruited 90 MIT undergraduate students for our study. Students who used MACH were perceived to be stronger candidates by human judges compared to the students in the control group. The judges scored videos from the interviews, blinded to the condition that the students were in. More details on the experimental set up and results are available in [2]. A video demonstration of the system is available at: http://tinyurl.com/MIT-MACH

ACKNOWLEDGMENT

We would like to acknowledge Jean-Claude Martin and Matthieu Courgeon for allowing us to build our system on their Multimodal Affective Reactive Characters (MARC) [1] Platform. Special acknowledgements to Sumit Gogia and Alex Suguel for helping with the UI design, and Eunice Lin, Nicole Pan, Kristi Tausk, Abra Shen, Tamanna Urmi for helping with the study. Bilge Mutlu provided valuable suggestions on the data analysis and experimental set up of our job interview experiment. We are grateful to Nuance communications, Samsung, and the Media Lab consortium for supporting our research.

REFERENCES

[1] M. Courgeon, S. Buisine, and J. Martin, "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.

[2] M. E. Hoque, M. Courgeon, J. Martin, B. Mutlu, and R. W. Picard, "MACH: My Automated Conversation coacH," in *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013)*, 2013.

[3] 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.

[4] S. Kawato and J. Ohya, "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.

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

[6] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics*, 2008, vol. 123, no. 5, pp. 5687–5690.

[7] "OwnTheRoom." [Online]. Available: owntheroom.com. [Accessed: 21-Jun-2013].