Effects of Damping Head Movement and Facial Expression in Dyadic Conversation Using Real–Time Facial Expression Tracking and Synthesized Avatars

Steven M. Boker

Jeffrey F. Cohn

Barry–John Theobald

Iain Matthews

Timothy R. Brick

Jeffrey R. Spies

Abstract

When people speak with one another they tend to adapt their head movements and facial expressions in response to each others' head movements and facial expressions. We present an experiment in which confederates' head movements and facial expressions were motion tracked during videoconference conversations, an avatar face was reconstructed in real time, and naive participants spoke with the avatar face. No naive participant guessed that the computer generated face was not video. Confederates' facial expressions, vocal inflections, and head movements were attenuated at one minute intervals in a fully crossed experimental design. Attenuated head movements led to increased head nods and lateral head turns, and attenuated facial expressions led to increased head nodding in both naive participants and in confederates. Together these results are consistent with a hypothesis that the dynamics of head movements in dyadic conversation include a shared equilibrium: Although both conversational partners were blind to the manipulation, when apparent head movement of one conversant was attenuated, both partners responded by increasing the velocity of their head movements.

Effects of Damping Head Movement and Facial Expression in Dyadic Conversation Using Real–Time Facial Expression Tracking and Synthesized Avatars

Introduction

When people converse, they adapt their movements, facial expressions, and vocal cadence to one another. This multimodal adaptation allows the communication of information that either reinforces or is in addition to the information that is contained in the semantic verbal stream. For instance, back–channel information such as direction of gaze, head nods, and "uh–huh"s allow the conversants to better segment speaker–listener turn taking. Affective displays such as smiles, frowns, expressions of puzzlement or surprise, shoulder movements, head nods, and gaze shifts are components of the multimodal conversational dialog.

When two people adopt similar poses, this could be considered a form of spatial symmetry (Boker & Rotondo, 2002). Interpersonal symmetry has been reported in many contexts and across sensory modalities: for instance, patterns of speech (Cappella & Panalp, 1981; Neumann & Strack, 2000), facial expression (Hsee, Hatfield, Carlson, & Chemtob, 1990), and laughter (Young & Frye, 1966). Increased symmetry is associated with increased rapport and affinity between conversants (Bernieri, 1988; LaFrance, 1982). Intrapersonal and cross–modal symmetry may also be expressed. Smile intensity is correlated with cheek raising in smiles of enjoyment (Messinger, Chow, & Cohn, 2009) and with head pitch and yaw in embarassment (Ambadar, Cohn, & Reed, 2009; Cohn et al., 2004). The structure of intrapersonal symmetry may be complex: self–affine multifractal dimension in head movements change based on conversational context (Ashenfelter, Boker, Waddell, & Vitanov, in press).

Symmetry in movements implies redundancy in movements, which can be defined as negative Shannon information (Redlich, 1993; Shannon & Weaver, 1949). As symmetry is formed between conversants, the ability to predict the actions of one based on the actions of the other increases. When symmetry is broken by one conversant, the other is likely to be surprised or experience change in attention. The conversant's previously good predictions would now be much less accurate. Breaking symmetry may be a method for increasing the transmission of nonverbal information by reducing the redundancy in a conversation.

This view of an ever–evolving symmetry between two conversants may be conceptualized as a dynamical system with feedback as shown in Figure 1. Motor activity (e.g., gestures, facial expression, or speech) is produced by one conversant and perceived by the other. These perceptions contribute to some system that functions to map the perceived actions of the interlocutor onto potential action: a mirror system. Possible neurological candidates for such a mirror system have been advanced by Rizzolati and colleagues (Rizzolatti & Fadiga, 2007; Iacoboni et al., 1999; Rizzolatti & Craighero, 2004) who argue that such a system is fundamental to communication.

Conversational movements are likely to be nonstationary (Boker, Xu, Rotondo, & King, 2002; Ashenfelter et al., in press) and involve both symmetry formation and symmetry breaking (Boker & Rotondo, 2002). One technique that is used in the study of nonstationary dynamical systems is to induce a known perturbation into a free running system and measure how the system adapts to the perturbation. In the case of facial expressions and head movements, one would need to manipulate conversant A's perceptions of the facial expressions and head movements of conversant B while conversant B remained blind to these manipulations as illustrated in Figure 2.

Recent advances in Active Appearance Models (AAMs) (Cootes, Wheeler, Walker, & Taylor, 2002) have allowed the tracking and resynthesis of faces in real time (Matthews

& Baker, 2004). Placing two conversants into a videoconference setting provides a context in which a real time AAM can be applied, since each conversant is facing a video camera and each conversant only sees a video image of the other person. One conversant could be tracked and the desired manipulations of head movements and facial expressions could be applied prior to resynthesizing an avatar that would be shown to the other conversant. In this way, a perturbation could be introduced as shown in Figure 2.

To test the feasibility of this paradigm and to investigate the dynamics of symmetry formation and breaking, we present the results of an experiment in which we implemented a mechanism for manipulating head movement and facial expression in real–time during a face-to-face conversation using a computer–enhanced videoconference system. The experimental manipulation was not noticed by naive participants, who were informed that they would be in a videoconference and that we had “cut out” the face of the person with whom they were speaking. No participant guessed that he or she was actually speaking with a synthesized avatar. This manipulation revealed the co-regulation of symmetry formation and breaking in two-person conversations.

Methods

Apparatus

Videoconference booths were constructed in two adjacent rooms. Each 1.5m × 1.2m footprint booth consisted of a 1.5m × 1.2m back projection screen, two 1.2m × 2.4m nonferrous side walls covered with white fabric and a white fabric ceiling. Each participant sat on a stool approximately 1.1m from the backprojection screen as shown in Figure 3. Audio was recorded using Earthworks directional microphones through a Yamaha 01V96 multichannel digital audio mixer. NTSC format video was captured using Panasonic IK-M44H “lipstick” color video cameras and recorded to two JVC BR-DV600U digital video decks. SMPTE time stamps generated by an ESE 185-U master clock were used to maintain a synchronized record on the two video recorders and to synchronize the data from a magnetic motion capture device. Head movements were tracked and recorded using an Ascension Technologies MotionStar magnetic motion tracker sampling at 81.6 Hz from a sensor attached to the back of the head using an elastic headband. Each room had an Extended Range Transmitter whose fields overlapped through the nonferrous wall separating the two video booth rooms.

To track and resynthesize the avatar, video was captured by an AJA Kona card in an Apple 2–core 2.5 GHz G5 PowerMac with 3 Gb of RAM and 160 Gb of storage. The PowerMac ran software described below and output the resulting video frames to an InFocus IN34 DLP Projector. Thus, the total delay time from the camera in booth 1 through the avatar synthesis process and projected to booth 2 was 165ms. The total delay time from the camera in booth 2 to the projector in booth 1 was 66ms, since the video signal was passed directly from booth 2 to booth 1 and did not need to go through a video A/D and avatar synthesis. For the audio manipulations described below, we reduced vocal pitch inflection using a TC–Electronics VoiceOne Pro. Audio–video sync was maintained using digital delay lines built into the Yamaha 01V96 mixer.

Active Appearance Models

Active Appearance Models (AAMs) (Cootes, Edwards, & Taylor, 2001) are generative, parametric models commonly used to track and synthesize faces in video sequences. Recent improvements in both the fitting algorithms and the hardware on which they run allow tracking (Matthews & Baker, 2004) and synthesis (Theobald, Matthews, Cohn, & Boker, 2007) of faces in real-time.

The AAM is formed of two compact models: One describes variation in shape and the other variation in appearance. AAMs are typically constructed by first defining the topological structure of the shape (the number of landmarks and their interconnectivity to form a two-dimensional triangulated mesh), then annotating with this mesh a collection of images that exhibit the characteristic forms of variation of interest. For this experiment, we label a subset of 40 to 50 images (less than 0.2% of the images in a single session) that are representative of the variability in facial expression. An individual shape is formed by concatenating the coordinates of the corresponding mesh vertices, $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$, so the collection of training shapes can be represented in matrix form as $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N]$. Applying principal component analysis (PCA) to these shapes, typically aligned to remove in-plane pose variation, provides a compact model of the form:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{1}$$

where $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the eigenvectors corresponding to the m largest eigenvalues. These eigenvectors are the basis vectors that span the shape-space and describe variation in the shape about the mean. The coefficients p_i are the shape parameters, which define the contribution of each basis in the reconstruction of $\mathbf{s}$. An alternative interpretation is that the shape parameters are the coordinates of $\mathbf{s}$ in shape-space, thus each coefficient is a measure of the distance from $\mathbf{s}_0$ to $\mathbf{s}$ along the corresponding basis vector.

The appearance of the AAM is a description of the variation estimated from a shape-free representation of the training images. Each training image is first warped from the manually annotated mesh location to the base shape, so the appearance is comprised of the pixels that lie inside the base mesh, $\mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$. PCA is applied to these images to provide a compact model of appearance variation of the form:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\ \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients λ_i are the appearance parameters, A_0 is the base appearance, and the appearance images, A_i, are the eigenvectors corresponding to the l largest eigenvalues. As with shape, the eigenvectors are the basis vectors that span appearance-space and describe variation in the appearance about the mean. The coefficients λ_i are the appearance parameters, which define the contribution of each basis in the reconstruction of $A(\mathbf{x})$. Because the model is invertible, it may be used to synthesize new face images (see Figure 4).

Manipulating Facial Displays Using AAMs.

To manipulate the head movement and facial expression of a person during a face-to-face conversation such that they remain blind to the manipulation, an avatar is placed in the feedback loop, as shown in Figure 2. Conversants speak via a videoconference and an AAM is used to track and parameterize the face of one conversant.

As outlined, the parameters of the AAM represent displacements from the origin in the shape and appearance space. Thus scaling the parameters has the effect of either exaggerating or attenuating the overall facial expression encoded as AAM parameters:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i \beta, \tag{3}$$

where β is a scalar, which when greater than unity exaggerates the expression and when less than unity attenuates the expression. An advantage of using an AAM to conduct this manipulation is that a separate scaling can be applied to the shape and appearance to create some desired effect. We stress here that in these experiments we are not interested in manipulating individual actions on the face (e.g., inducing an eye-brow raise), rather we wish to manipulate, in real–time, the overall facial expression produced by one conversant during the conversation.

The second conversant does not see video of the person to whom they are speaking. Rather, they see a re–rendering of the video from the manipulated AAM parameters as shown in Figure 5. To re–render the video using the AAM the shape parameters, $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, are first applied to the model, Equation (3), to generate the shape, $\mathbf{s}$, of the AAM, followed by the appearance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)^{\mathrm{T}}$ to generate the

AAM image, $A(\mathbf{x})$. Finally, a piece–wise affine warp is used to warp $A(\mathbf{x})$ from $\mathbf{s}_0$ to $\mathbf{s}$, and the result is transferred into image coordinates using a similarity transform (i.e., movement in the x–y plane, rotation, and scale). This can be achieved efficiently, at video frame–rate, using standard graphics hardware.

Typical example video frames synthesized using an AAM before and after damping are shown in Figure 6. Note the effect of the damping is to reduced the expressiveness. Our interest here is to estimate the extent to which manipulating expressiveness in this way can affect behavior during conversation.

Participants

Naive participants ($N = 27$, 15 male, 12 female) were recruited from the psychology department participant pool at a midwestern university. Confederates ($N = 6$, 3 male, 3 female) were undergraduate research assistants. AAM models were trained for the confederates so that the confederates could act as one conversant in the dyad. Confederates were informed of the purpose of the experiment and the nature of the manipulations, but were blind to the order and timing of the manipulations. All confederates and naive participants read and signed informed consent forms approved by the Institutional Review Board.

Procedure

We attenuated three variables: (1) head pitch and turn: translation and rotation in image coordinates from their canonical values by either 1.0 or 0.5; (2) facial expression: the vector distance of the AAM shape parameters from the canonical expression (by multiplying the AAM shape parameters by either 1.0 or 0.5); and (3) audio: the range of frequency variability in the fundamental frequency of the voice (by using the VoicePro to either restrict or not restrict the range of the fundamental frequency of the voice) in a fully crossed design. Naive participants were given a cover story that video was “cut out” around the face and then participated in two 8 minute conversations, one with a male and one with a female confederate. Prior to debrief, the naive participants were asked if they "noticed anything unusual about the experiment". None mentioned that they thought they were speaking with a computer generated face or noted the experimental manipulations.

Data reduction and analysis

Angles of the Ascension Technologies head sensor in the anterior–posterior (A–P) and lateral directions (i.e. pitch and yaw, respectively) were selected for analysis. These directions correspond to the meaningful motion of a head nod and a head turn, respectively. We focus on angular velocity since this variable can be thought of as how animated a participant was during an interval of time.

To compute angular velocity, we first converted the head angles into angular displacement by subtracting the mean overall head angle across a whole conversation from each head angle sample. We used the overall mean head angle since this provided an estimate of the overall equilibrium head position for each participant independent of the trial conditions. Second, we low–pass filtered the angular displacement time series and calculated angular velocity using a quadratic filtering technique (Generalized Local Linear Approximation; Boker, Deboeck, Edler, & Keel, in press), saving both the estimated displacement and velocity for each sample. The root mean square (RMS) of the lateral and A–P angular velocity was then calculated for each one minute condition of each conversation for each naive participant and confederate.

Because the head movements of each conversant both influence and are influenced by the movements of the other, we seek an analytic strategy that models bidirectional effects (Kenny & Judd, 1986). Specifically, each conversant's head movements are both a predictor variable and outcome variable. Neither can be considered to be an independent variable. In addition, each naive participant was engaged in two conversations, one with each of two confederates. Each of these sources of non–independence in dyadic data needs to be accounted for in a statistical analysis.

To put both conversants in a dyad into the same analysis we used a variant of Actor–Partner analysis (Kashy & Kenny, 2000; Kenny, Kashy, & Cook, 2006). Suppose we are analyzing RMS–V angular velocity. We place both the naive participants' and confederates' RMS–V angular velocity into the same column in the data matrix and use a second column as a dummy code labeled "Confederate" to identify whether the data in the angular velocity column came from a naive participant or a confederate. In a third column, we place the RMS–V angular velocity from the other participant in the conversation. We then use the terminology "Actor" and "Partner" to distinguish which variable is the predictor and which is the outcome for a selected row in the data matrix. If Confederate=1, then the confederate is the "Actor" and the naive participant is the "Partner" in that row of the data matrix. If Confederate=0, then the naive participant is the "Actor" and the confederate is the "Partner." We coded the sex of the "Actor" and the "Partner" as a binary variables (0=female, 1=male). The RMS angular velocity of the "Partner" was used as a continuous predictor variable.

Binary variables were coded for each manipulated condition: attenuated head pitch and turn (0=normal, 1=50% attenuation), and attenuated expression (0=normal, 1=50% attenuation). Since only the naive participant sees the manipulated conditions we also added interaction variables (confederate $\times$ delay condition and confederate $\times$ sex of partner), centering each binary variable prior to multiplying. The manipulated condition may affect the naive participant directly, but also may affect the confederate indirectly through changes in behavior of the naive participant. The interaction variables allow us to account for an overall effect of the manipulation as well as possible differences between the reactions of the naive participant and of the confederate.

We then fit mixed effects models using restricted maximum likelihood. Since there is non–independence of rows in this data matrix, we need to account for this non–independence. An additional column is added to the data matrix that is coded by experimental session and then the mixed effects model of the data is grouped by the experimental session column (both conversations in which the naive participant engaged). Each session was allowed a random intercept to account for individual differences between experimental sessions in the overall RMS velocity. This mixed effects model can be written as

$$\begin{aligned} y_{ij} &= b_{j0} + b_1 A_{ij} + b_2 P_{ij} + b_3 C_{ij} + b_4 H_{ij} + b_5 F_{ij} + b_6 V_{ij} + \\ &= b_7 Z_{ij} + b_8 C_{ij} P_{ij} + b_9 C_{ij} H_{ij} + b_{10} C_{ij} F_{ij} + b_{11} C_{ij} V_{ij} + e_{ij} \qquad (4) \\ b_{j0} &= c_{00} + u_{j0} \qquad (5) \end{aligned}$$

where y_{ij} is the outcome variable (lateral or A–P RMS velocity) for condition i and session j. The other predictor variables are the sex of the Actor A_{ij}, the sex of the Partner P_{ij}, whether the Actor is the confederate C_{ij}, the head pitch and turn attenuation condition H_{ij}, the facial expression attenuation condition F_{ij}, the vocal inflection attenuation condition V_{ij}, and the lateral or A–P RMS velocity of the partner Z_{ij}. Since each session was allowed to have its own intercept, the predictions are relative to the overall angular velocity associated with each naive participant's session.

Results

The results of a mixed effects random intercept model grouped by session predicting A–P RMS angular velocity of the head are displayed in Table 1. As expected from previous reports, males exhibited lower A–P RMS angular velocity than females and when the conversational partner was male there was lower A–P RMS angular velocity than when the conversational partner was female. Confederates exhibited lower A–P RMS velocity than naive participants, although this effect only just reached significance at the $\alpha = 0.05$ level. Both attenuated head pitch and turn, and facial expression were associated with greater A–P angular velocity: Both conversants nodded with greater vigor when either the avatar's rigid head movement or facial expression was attenuated. Thus, the naive participant reacted to the attenuated movement of the avatar by increasing her or his head movements. But also, the confederate (who was blind to the manipulation) reacted to the increased head movements of the naive participant by increasing his or her head movements. When the avatar attenuation was in effect, both conversational partners adapted by increasing the vigor of their head movements. There were no effects of either the attenuated vocal inflection or the A–P RMS velocity of the conversational partner. Only one interaction reached significance — Confederates had a greater reduction in A–P RMS angular velocity when speaking to a male naive participant than the naive participants had when speaking to a male confederate.

The results for RMS lateral angular velocity of the head are displayed in Table 2. As was true in the A–P direction, males exhibited less lateral RMS angular velocity than females, and conversants exhibited less lateral RMS angular velocity when speaking to a male partner. Confederates again exhibited less velocity than naive participants. Attenuated head pitch and turn was again associated with greater lateral angular velocity: Participants turned away or shook their heads either more often or with greater angular velocity when the avatar's head pitch and turn variation was attenuated. However, in the lateral direction, we found no effect of the facial expression or vocal inflection attenuation. There was an independent effect such that lateral head movements were negatively coupled. That is to say in one minute blocks when one conversant's lateral angular movement was more vigorous, their conversational partner's lateral movement was reduced. Again, only one interaction reached significance — Confederates had a greater reduction in A–P RMS angular velocity when speaking to a male naive participant than the naive participants had when speaking to a male confederate. There are at least three differences between the confederates and the naive participants that might account for this effect: (1) the confederates have more experience in the video booth than the naive participants and may thus be more sensitive to the context provided by the partner since the overall context of the video booth is familiar, (2) the naive participants are seeing an avatar and it may be that there is an additional partner sex effect of seeing a full body video over seeing a "floating head", and (3) the reconstructed avatars have reduced number of eye blinks than the video since some eye blinks are not caught by the motion tracking.

Discussion

Automated facial tracking was successfully applied to create real–time resynthesized avatars that were accepted as being video by naive participants. No participant guessed that we were manipulating the apparent video in their videoconference converstations. This technological advance presents the opportunity for studying adaptive facial behavior in natural conversation while still being able to introduce experimental manipulations of rigid and non–rigid head movements without either participant knowing the extent or timing of these manipulations.

The damping of head movements was associated with increased A–P and lateral angular velocity. The damping of facial expressions was associated with increased A–P angular velocity. There are several possible explanations for these effects. During the head movement attenuation condition, naive participants might perceive the confederate as looking more directly at him or her, prompting more incidents of gaze avoidance. A conversant might not have received the expected feedback from an A–P or lateral angular movement of a small velocity and adapted by increasing her or his head angle relative to the conversational partner in order to elicit the expected response. Naive participants may have perceived the attenuated facial expressions of the confederate as being non–responsive and attempted to increase the velocity of their head nods in order to elicit greater response from their conversational partners.

Since none of the interaction effects for the attenuated conditions were significant, the confederates exhibited the same degree of response to the manipulations as the naive participants. Thus, when the avatar's head pitch and turn variation was attenuated, both the naive participant and the confederate responded with increased velocity head movements. This suggests that there is an expected degree of matching between the head velocities of the two conversational partners. Our findings provide evidence in support of a hypothesis that the dynamics of head movement in dyadic conversation include a shared equilibrium: Both conversational partners were blind to the manipulation and when we perturbed one conversant's perceptions, both conversational partners responded in a way that compensated for the perturbation. It is as if there were an equilibrium energy in the conversation and when we removed energy by attenuation and thus changed the value of the equilibrium, the conversational partners supplied more energy in response and thus returned the equilibrium towards its former value.

These results can also be interpreted in terms of symmetry formation and symmetry breaking. The dyadic nature of the conversants' responses to the asymmetric attenuation conditions are evidence of symmetry formation. But head turns have an independent effect of negative coupling, where greater lateral angular velocity in one conversant was related to reduced angular velocity in the other: evidence of symmetry breaking. Our results are consistent with symmetry formation being exhibited in both head nods and head turns while symmetry breaking being more related to head turns. In other words, head nods may help form symmetry between conversants while head turns contribute to both symmetry formation and to symmetry breaking. One argument for why these relationships would be observed is that head nods may be more related to acknowledgment or attempts to elicit expressivity from the partner whereas head turns may be more related to new semantic information in the conversational stream (e.g., floor changes) or to signals of disagreement or withdrawal.

With the exception of some specific expressions (e.g., Ambadar et al., 2009; Kelner, 1995), previous research has ignored the relationship between head movements and facial expressions. Our findings suggest that facial expression and head movement may be closely related. These results also indicate that the coupling between one conversant's facial expressions and the other conversant's head movements should be taken into account. Future research should inquire into these within–person and between–person cross–modal relationships.

The attenuation of facial expression created an effect that appeared to the research team as being that of someone who was mildly depressed. Decreased movement is a common feature of psychomotor retardation in depression, and depression is associated with decreased reactivity to a wide range of positive and negative stimuli (Rottenberg, 2005). Individuals with depression or dysphoria, in comparison with non–depressed individuals, are less likely to smile in response to pictures or movies of smiling faces and affectively positive social imagery (Gehricke & Shapiro, 2000; Sloan, Bradley, Dimoulas, & Lang, 2002). When they do smile, they are more likely to damp their facial expression (Reed, Sayette, & Cohn, 2007).

Attenuation of facial expression can also be related to cognitive states or social context. For instance, if one's attention is internally focused, attenuation of facial expression may result. Interlocutors might interpret damped facial expression of their conversational partner as reflecting a lack of attention to the conversation.

Naive participants responded to damped facial expression and head turns by increasing their own head nods and head turns, respectively. These effects may have been efforts to elicit more responsive behavior in the partner. In response to simulated maternal depression by their mother, infants attempt to elicit a change in their mother's behavior by smiling, turning away, and then turning again toward her and smiling. When they fail to elicit a change in their mothers' behavior, they become withdrawn and distressed (Cohn & Tronick, 1983). Similarly, adults find exposure to prolonged depressed behavior increasingly aversive and withdraw (Coyne, 1976). Had we attenuated facial expression and head motion for more than a minute at a time, naive participants might have become less active following their failed efforts to elicit a change in the confederate's behavior. This hypothesis remains to be tested.

There are a number of limitations of this methodology that could be improved with further development. For instance, while we can manipulate degree of expressiveness as well as identity of the avatar (Boker, Cohn, et al., in press), we cannot yet manipulate specific facial expressions in real time. Depression not only attenuates expression, but makes some facial actions, such as contempt, more likely (Cohn et al., submitted; Ekman, Matsumoto, & Friesen, 2005). As an analog for depression, it would be important to manipulate specific expressions in real time. In other contexts, cheek raising (AU 6 in the Facial Action Coding System) (Ekman, Friesen, & Hager, 2002) is believed to covary with communicative intent and felt emotion (Coyne, 1976). In the past, it has not been possible to experimentally manipulate discrete facial actions in real–time without the source person's awareness. If this capability could be implemented in the videoconference paradigm, it would make possible a wide–range of experimental tests of emotion signaling.

Other limitations include the need for person–specific models, restrictions on head rotation, and limited face views. The current approach requires manual training of face models, which involves hand labeling about 30 to 50 video frames. Because this process requires several hours of preprocessing, avatars could be constructed for confederates but not for unknown persons, such as naive participants. It would be useful to have the capability of generating real–time avatars for both conversation partners. Recent efforts have made progress toward this goal (Lucey, Wang, Cox, Sridharan, & Cohn, in press; Saragih, Lucey, & Cohn, submitted). Another limitation is that if the speaker turns more than about 20 degrees from the camera, parts of the face become obscured and the model no longer can track the remainder of the face. Algorithms have been proposed that address this issue (Gross, Matthews, & Baker, 2004), but it remains a research question. Another limitation is that the current system has modeled the face only from the eyebrows to the chin. A better system would include the forehead, and some model of the head, neck, shoulders and background in order to give a better sense of the placement of the speaker in context. Adding forehead features is relatively straight–forward and has been implemented. Tracking of neck and shoulders is well–advanced (Sheikh, Datta, & Kanade, 2008). The video–conference avatar paradigm has motivated new work in computer vision and graphics and made possible new methodology to experimentally investigate social interaction in a way not before possible. The timing and identity of social behavior in real time can now be rigorously manipulated outside of participants' awareness.

Conclusion

We presented an experiment that used automated facial and head tracking to perturb the bidirectionally coupled dynamical system formed by two individuals speaking with one another over a videoconference link. The automated tracking system allowed us to create resynthesized avatars that were convincing to naive participants and, in real time, to attenuate head movements and facial expressions formed during natural dyadic conversation. The effect of these manipulations exposed some of the complexity of multimodal coupling of movements during face to face interactions. The experimental paradigm presented here has the potential to transform social psychological research in dyadic and small group interactions due to an unprecedented ability to control the real–time appearance of facial structure and expression.

References

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, *33*(1), 17–34.

Ashenfelter, K. T., Boker, S. M., Waddell, J. R., & Vitanov, N. (in press). Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology: Human Perception and Performance.*

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher–student interactions. *Journal of Nonverbal Behavior*, *12*(2), 120–138.

Boker, S. M., & Cohn, J. F. (in press). Real time dissociation of facial appearance and dynamics during natural conversation. In M. Giese, C. Curio, & H. Bültoff (Eds.), *Dynamic faces: Insights from experiments and computation* (pp. ???–???). Cambridge, MA: MIT Press.

Boker, S. M., Cohn, J. F., Theobald, B.-J., Matthews, I., Mangini, M., Spies, J. R., et al. (in press). Something in the way we move: Motion dynamics, not perceived sex, influence head movements in conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *??*(??), ??–??

Boker, S. M., Deboeck, P. R., Edler, C., & Keel, P. K. (in press). Generalized local linear approximation of derivatives from time series. In S.-M. Chow & E. Ferrar (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue.* Boca Raton, FL: Taylor & Francis.

Boker, S. M., & Rotondo, J. L. (2002). Symmetry building and symmetry breaking in synchronized movement. In M. Stamenov & V. Gallese (Eds.), *Mirror neurons and the evolution of brain and language* (pp. 163–171). Amsterdam: John Benjamins.

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed cross–correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(1), 338–355.

Cappella, J. N., & Panalp, S. (1981). Talk and silence sequences in informal conversations: Iii interspeaker influence. *Human Communication Research*, *7*, 117–132.

Cohn, J. F., Kreuze, T. S., Yang, Y., Gnuyen, M. H., Padilla, M. T., & Zhou, F. (submitted). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction (ACII 2009).* Amsterdam: IEEE.

Cohn, J. F., Reed, L. I., Moriyama, T., Xiao, J., Schmidt, K. L., & Ambadar, Z. (2004). Multimodal coordination of facial action, head rotation, and eye motion. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 645–650). Seoul, Korea: IEEE.

Cohn, J. F., & Tronick, E. Z. (1983). Three month old infants' reaction to simulated maternal depression. *Child Development*, *54*(1), 185–193.

Cootes, T. F., Edwards, G., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685.

Cootes, T. F., Wheeler, G. V., Walker, K. N., & Taylor, C. J. (2002). View-based active appearance models. *Image and Vision Computing*, *20*(9–10), 657–664.

Coyne, J. C. (1976). Depression and the response of others. *Journal of Abnormal Psychology*, *85*(2), 186–193.

Ekman, P., Friesen, W., & Hager, J. (2002). *Facial action coding system.* Salt Lake City, UT: Research Nexus.

Ekman, P., Matsumoto, D., & Friesen, W. V. (2005). Facial expression in affective disorders. In P. Ekman & E. Rosenberg (Eds.), *What the face reveals* (pp. 331–341). New York: Oxford University Press.

Gehricke, J.-G., & Shapiro, D. (2000). Reduced facial expression and social context in major depression: Discrepancies between facial muscle activity and self–reported emotion. *Psychiatry Research*, *95*(3), 157–167.

Gross, R., Matthews, I., & Baker, S. (2004). Constructing and fitting active appearance models with occlusion. In *First IEEE Workshop on Face Processing in Video.* Washington, DC: IEEE.

Hsee, C. K., Hatfield, E., Carlson, J. G., & Chemtob, C. (1990). The effect of power on susceptibility to emotional contagion. *Cognition and Emotion*, *4*, 327–340.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–2528.

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (p. 451477). New York: Cambridge University Press.

Kelner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, *68*(3), 441–454.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422–431.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis.* New York: Guilford.

LaFrance, M. (1982). Posture mirroring and rapport. In M. Davis (Ed.), *Interaction rhythms: Periodicity in communicative behavior* (pp. 279–298). New York: Human Sciences Press.

Lucey, S., Wang, Y., Cox, M., Sridharan, S., & Cohn, J. F. (in press). Efficient constrained local model fitting for non–rigid face alignment. *Image and Vision Computing Journal*, *??*(??), ??–??

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*(2), 135–164.

Messinger, D. S., Chow, S. M., & Cohn, J. F. (2009). Automated measurement of smile dynamics in mother–infant interaction: A pilot study. *Infancy*, *14*(3), 285–305.

Neumann, R., & Strack, F. (2000). "mood contagion": The automatic transfer of mood between persons. *Journal of Personality and Social Psychology*, *79*, 158–163.

Redlich, N. A. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, *5*, 289–304.

Reed, L. I., Sayette, M. A., & Cohn, J. F. (2007). Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of Abnormal Psychology*, *116*(4), 804–809.

Rizzolatti, G., & Craighero, L. (2004). The mirror–neuron system. *Annual Reviews of Neuroscience*, *27*, 169–192.

Rizzolatti, G., & Fadiga, L. (2007). Grasping objects and grasping action meanings the dual role of monkey rostroventral premotor cortex. In P. Ekman & E. Rosenberg (Eds.), *Novartis Foundation Symposium 218 – Sensory Guidance of Movement* (pp. 81–108). New York: Novartis Foundation.

Rottenberg, J. (2005). Mood and emotion in major depression. *Current Directions in Psychological Science*, *14*(3), 167–170.

Saragih, J., Lucey, S., & Cohn, J. F. (submitted). Probabilistic constrained adaptive local displacement experts. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. ??–??). Miami, Florida: IEEE.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: The University of Illinois Press.

Sheikh, Y. A., Datta, A., & Kanade, T. (2008). *On the sustained tracking of human motion.* (Paper presented at the IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam)

Sloan, D. M., Bradley, M. M., Dimoulas, E., & Lang, P. J. (2002). Looking at facial expressions: Dysphoria and facial EMG. *Biologial Psychology*, *60*(2–3), 79–90.

Theobald, B., Matthews, I., Cohn, J. F., & Boker, S. (2007). Real–time expression cloning using appearance models. In *Proceedings of the 9th international conference on multimodal interfaces* (pp. 134–139). New York: Association for Computing Machinery.

Young, R. D., & Frye, M. (1966). Some are laughing; some are not — why? *Pychological Reports*, *18*, 747–752.

Author Note

Steven Boker, Timothy Brick, and Jeffrey Spies are with the University of Viginia, Jeffrey Cohn is with the University of Pittsburgh and Carnegie Mellon University, Barry John Theobald is with the University of East Anglia, and Iain Matthews is with Disney Research and Cargnegie Mellon University. Preparation of this manuscript was supported in part by NSF grant BCS05 27397, EPSRC Grant EP/D049075, and NIMH grant MH 51435. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We gratefully acknowledge the help of Kathy Ashenfelter, Tamara Buretz, Eric Covey, Pascal Deboeck, Katie Jackson, Jen Koltiska, Sean McGowan, Sagar Navare, Stacey Tiberio, Michael Villano, and Chris Wagner. Correspondence may be addressed to Jeffrey F. Cohn, Department of Psychology, 3137 SQ, 210 S. Bouquet Street, Pittsburgh, PA 15260 USA. For electronic mail, jeffcohn@cs.cmu.edu.

Table 1

Head A–P RMS angular velocity predicted using a mixed effects random intercept model grouped by session. "Actor" refers to the member of the dyad whose data is being predicted and "Partner" refers to the other member of the dyad. (AIC=3985.4, BIC=4051.1, Groups=27, Random Effects Intercept SD=1.641)

	Value	SE	DOF	t–value	p
Intercept	10.009	0.5205	780	19.229	$< .0001$
Actor is Male	-3.926	0.2525	780	-15.549	$< .0001$
Partner is Male	-1.773	0.2698	780	-6.572	$< .0001$
Actor is Confederate	-0.364	0.1828	780	-1.991	0.0469
Attenuated Head Pitch and Turn	0.570	0.1857	780	3.070	0.0022
Attenuated Expression	0.451	0.1858	780	2.428	0.0154
Attenuated Inflection	-0.037	0.1848	780	-0.200	0.8414
Partner A–P RMS Velocity	-0.014	0.0356	780	-0.389	0.6971
Confederate × Partner is Male	-2.397	0.5066	780	-4.732	$< .0001$
Confederate × Attenuated Head Pitch and Turn	-0.043	0.3688	780	-0.116	0.9080
Confederate × Attenuated Expression	0.389	0.3701	780	1.051	0.2935
Confederate × Attenuated Inflection	0.346	0.3694	780	0.937	0.3490

Table 2

Head lateral RMS angular velocity predicted using a mixed effects random intercept model grouped by dyad. (AIC=9818.5, BIC=9884.2, Groups=27, Random Effects Intercept SD=103.20)

	Value	SE	DOF	t–value	p
Intercept	176.37	22.946	780	7.686	$< .0001$
Actor is Male	-60.91	9.636	780	-6.321	$< .0001$
Partner is Male	-31.86	9.674	780	-3.293	0.0010
Actor is Confederate	-21.02	6.732	780	-3.122	0.0019
Attenuated Head Pitch and Turn	14.19	6.749	780	2.102	0.0358
Attenuated Expression	8.21	6.760	780	1.215	0.2249
Attenuated Inflection	4.40	6.749	780	0.652	0.5147
Partner A–P RMS Velocity	-0.30	0.034	780	-8.781	$< .0001$
Confederate × Partner is Male	-49.65	18.979	780	-2.616	0.0091
Confederate × Attenuated Head Pitch and Turn	-4.81	13.467	780	-0.357	0.7213
Confederate × Attenuated Expression	6.30	13.504	780	0.467	0.6408
Confederate × Attenuated Inflection	10.89	13.488	780	0.807	0.4197

Figure Captions

Figure 1. Dyadic conversation involves a dynamical system with adaptive feedback control resulting in complex, nonstationary behavior.

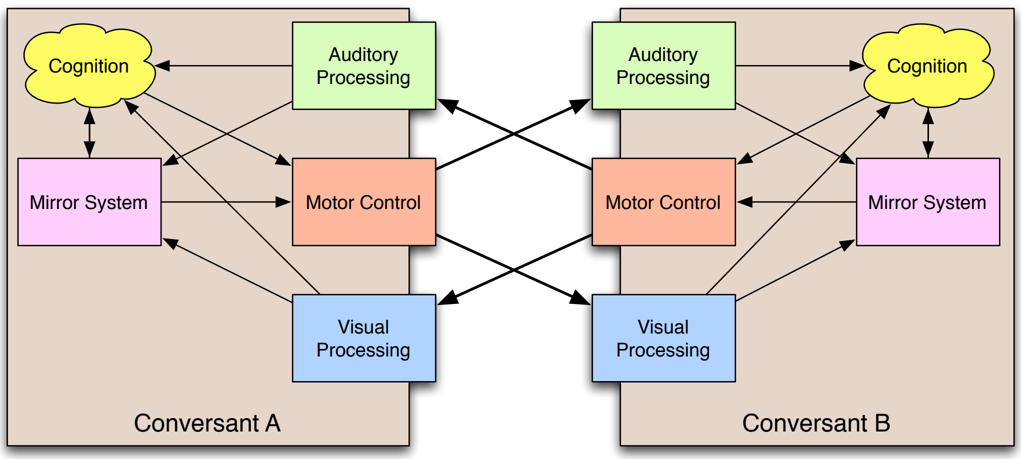
Figure 2. By tracking rigid and nonrigid head movements in real time and resynthesizing an avatar face, controlled perturbations can be introduced into the shared dynamical system between two conversants.

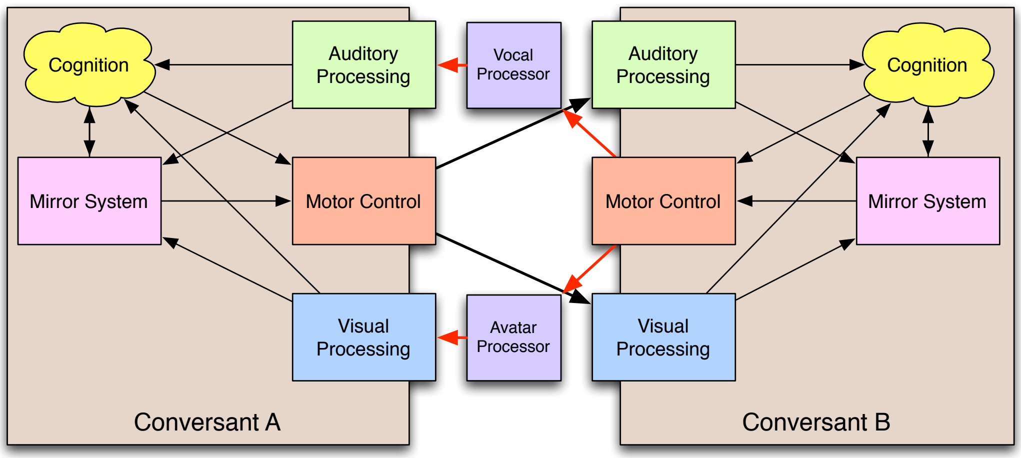
Figure 3. Videoconference booth. (a) Exterior of booth showing backprojection screen, side walls, fabric ceiling, and microphone. (b) Interior of booth from just behind participant's stool showing projected video image and lipstick videocamera.

Figure 4. Illustration of AAM resynthesis. Row (a) shows the mean face shape on the left and first shape modes. Row (b) shows the mean appearance and the first three appearance modes. The AAM is invertible and can synthesize new faces, four of which are shown in row (c). (From Boker & Cohn, 2009)

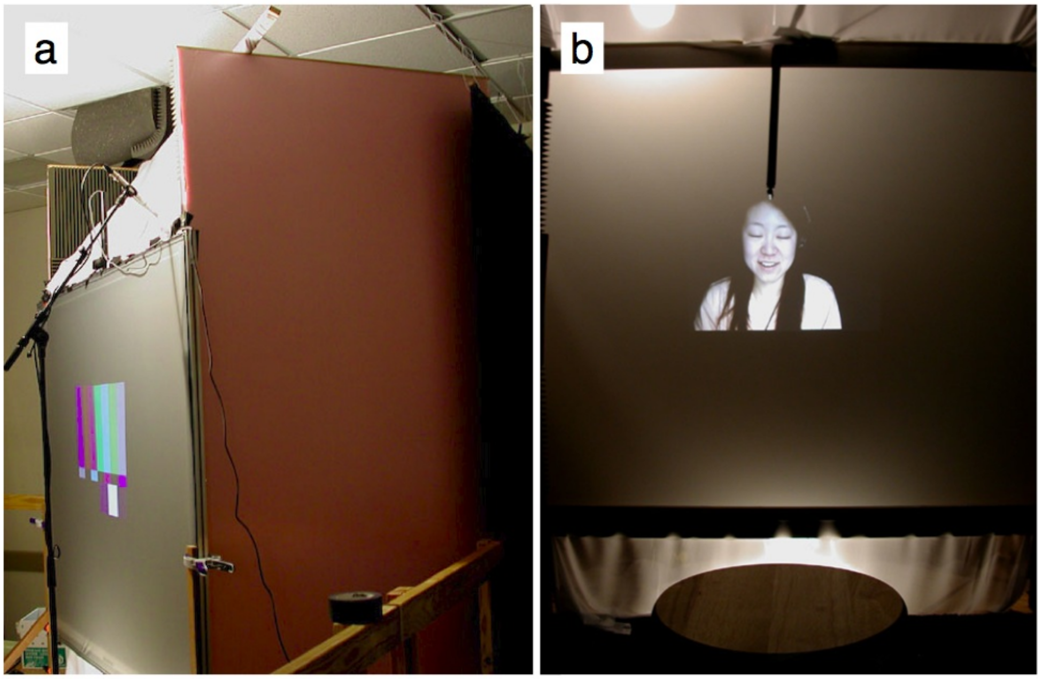
Figure 5. Illustration of the videoconference paradigm. A movie clip can be viewed at `http://people.virginia.edu/~smb3u/Clip1.avi`. (a) Video of the confederate. (b) AAM tracking of confederate's expression. (c) AAM reconstruction that is viewed by the naive participant. (d) Video of the naive participant.

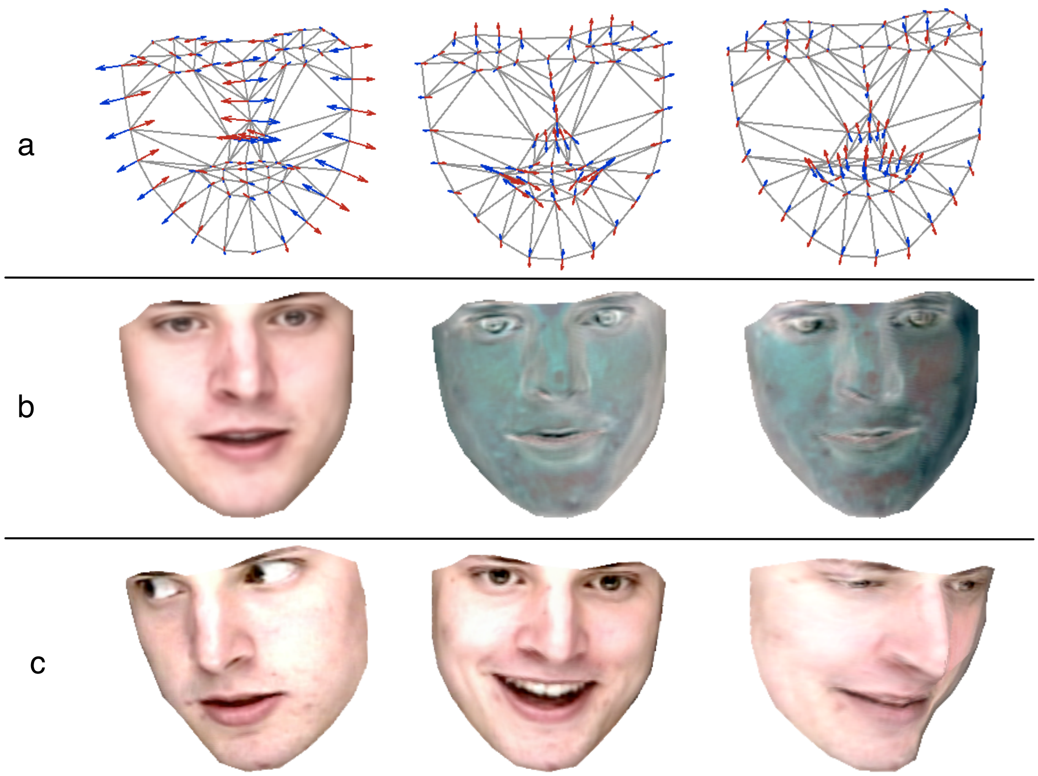
Figure 6. Facial expression attenuation using an AAM. (a) Four faces resynthesized from their respective AAM models showing expressions from tracked video frames. (b) The same video frames displayed at 25% of their AAM parameter difference from each individual's mean facial expression (i.e., $\beta = 0.25$).


Cognition
Auditory Processing
Mirror System
Motor Control
Visual Processing
Conversant A
Auditory Processing
Cognition
Motor Control
Mirror System
Visual Processing
Conversant B

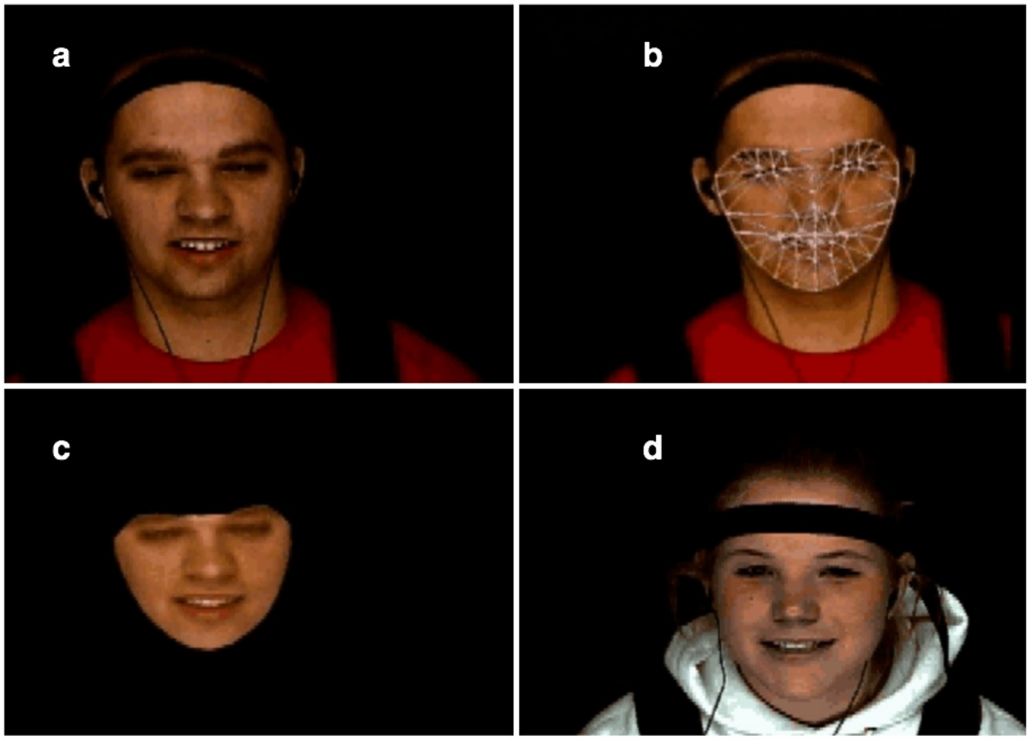


Cognition
Auditory Processing
Vocal Processor
Auditory Processing
Cognition
Mirror System
Motor Control
Motor Control
Mirror System
Visual Processing
Avatar Processor
Visual Processing
Conversant A
Conversant B


a
b



a
b
c


a
b
c
d