HONEST SIGNALS IN VIDEO CONFERENCING

Byungki Byun[1,2], Anurag Awasthi[1,3], Philip A. Chou[1], Ashish Kapoor[1], Bongshin Lee[1], Mary Czerwinski[1]

[1]Microsoft Research
Redmond, WA, 98052, USA
{pachou, akapoor, bongshin, marycz}
@microsoft.com

[2]School of ECE
Georgia Institute of Technology
Atlanta, GA, 30332, USA
yorke3@ece.gatech.edu

[3]Department of CSE
Indian Institute of Technology Kanpur
208016, India
anuraga@cse.iitk.ac.in

ABSTRACT

We propose a novel system to analyze gestural and non-verbal cues of participants in video conferencing. These cues have previously been referred to as "honest signals" and are usually associated with the underlying cognitive state of the participants. The presented system analyzes a set of audio-visual, non-linguistic features in real time from the audio and video streams of two participants in a video conference. We show how these features can be used to compute indicators of the overall quality and type of conversation being held. The system also provides visual feedback to the participants, who then have the choice of modifying their conversational style in order to achieve the desired outcome of the video conference. Experiments on real-life data show that the system can predict the type of conversation with high accuracy using the non-linguistic signals only. Qualitative user studies highlight the positive effects of increased awareness amongst the participants about their own gestural and non-verbal cues.

Index Terms— Non-verbal behavior, gesture analysis, video conference, honest signals.

1. INTRODUCTION

All animals have evolved to communicate with each other using non-linguistic signals. For example, they have been observed to use physical appearance, movement, and/or vocalization to convey dominance, cooperation, warning, trust, and so forth.

In recent years, the rapid increase of the capacity of digital communication has enabled humans to invent many ways of communicating with each other. For humans, linguistic means have always been dominant even in these new communication methods. However, as animals, humans have retained the ability to convey and to utilize non-linguistic signals during communication, including during video conferences.

Non-linguistic signals used for communication are termed *honest signals* if, in addition to communicating information intended by the sending animal, they also carry cues about the sender's underlying state (e.g., emotion, fatigue, or confusion) that are difficult for the sender to hide. Since these cues are detectable by the receiving animal(s), there is a good chance that they should be detectable by sensing devices coupled to computers.

One approach to detecting and analyzing such non-linguistic signals is to have the group of people, whose honest signals are to be sensed, wear electronic badges draped around their necks [1][2]. Data collected in such a way has been the subject of numerous experiments and has been shown to be predictive of the outcomes of a variety of social interactions, including speed dating, elevator pitches, sales, salary negotiation, and card games. However, wearing a badge or any other type of device is intrusive and is unlikely to be adopted by real users.

Other approaches have focused on the analysis of facial expressions, utterances, postures and physiology for cues on underlying internal state [3-8]. There is also a body of related work on human interaction analysis. For example, [9][10] show that interactions between people in a meeting can be analyzed by tracking non-verbal cues, and [11][12] demonstrate visualization of the amount of time attendees are talking during a meeting influences people's behavior.

In this work, we aim to sense and analyze honest signals in the context of 1-1 video conferencing. From audiovisual signals recorded by microphones and video cameras, we *non-intrusively* extract low-level audio and video features to characterize honest signals and social roles, and to present diagnostics of the signals back to the users through a graphical interface, all *in real time*. Our hypothesis is that such visualization of users' nonverbal behavior can help the users modify their actions and make conversations more successful.

There are three main contributions of this work:

- We explore and analyze a spectrum of gestural behaviors and non-verbal cues that are important in 1-1 video teleconferencing.
- We show how to build modules that can predict key aspects of the ongoing conference, such as behaviors indicative of quality and type of conversations, from the low-level features extracted from non-verbal behaviors and gestures.
- We present a detailed user study and experimentally show that such a predictive framework works well and has the capability to positively influence the conversations in the video conference.

978-1-61284-350-6/11/$26.00 ©2011 IEEE

2. HONEST SIGNALS AND SOCIAL ROLES

For completeness, we briefly explain four honest signals considered in this work, originally studied in [2]. They are (a) activity, (b) consistency, (c) influence, and (d) mimicry. *Activity* refers to the energetic state of a person, reflecting his or her degree of involvement in a conversation. *Consistency* refers to the degree of regularity or cadence of behavior, primarily during speech, reflecting mental certainty. *Influence* refers to degree of control of one person over the conversation, reflecting interest or desire to dominate. *Mimicry* refers to a behavioral pattern that mimics the behavior of others, reflecting agreement or empathy.

These honest signals are shown to be predictive of the outcomes of a variety of social interactions. Moreover, these signals are also predictive of social roles. Four particular social roles examined in [2] are *exploring*, *active listening*, *teaming*, and *leading*. Relating honest signals with the social roles, *exploring* (e.g., looking for points in common with another person) is indicated by high activity and low consistency. *Active listening* (e.g., listening and reflecting information back to the talker) is indicated by low activity and low consistency, and *teaming* (i.e., working together in a team, showing cooperative behavior) is indicated by high influence, high mimicry, and high consistency. Finally, *leading* (e.g., dominance or control in a group) is indicated by high activity, high influence, and high consistency.

3. PRELIMINARY STUDY

To test the hypothesis that real time visual presentation of honest signals would influence participants' behaviors during a video conference, we first ran a wizard-of-oz user study in the laboratory bringing in pairs of users and giving them two controversial conversational topics to discuss (e.g., do you prefer Apple or generic PC products?). Each user was to argue for a particular side of an argument to try to influence the other participant via a video conferencing tool. For one conversational topic, no visual representation of their behaviors was included, and for the other, we presented to both users a visual user interface showing the levels of their own honest signals as in Fig. 1. In this preliminary study, we showed 'Excitement' as a proxy for activity, 'Openness' as a proxy for the converse of consistency and/or influence, and 'Agreement' as a proxy for mimicry. We also showed a speaking timeline and overall speaking proportion.

In this study, the signal levels were actually controlled by two researchers, who were adjusting the levels in real time as the conversation unfolded. After each condition, we asked the users to rate the system usefulness, features they used, if any, and behavioral changes if any. Users were satisfied overall with the visual feedback of their behavior. They especially appreciated seeing the amount of time they spent talking compared to the other user. If they saw that they were dominating the discussion, they said they tried to give the other user more opportunity to speak. While some

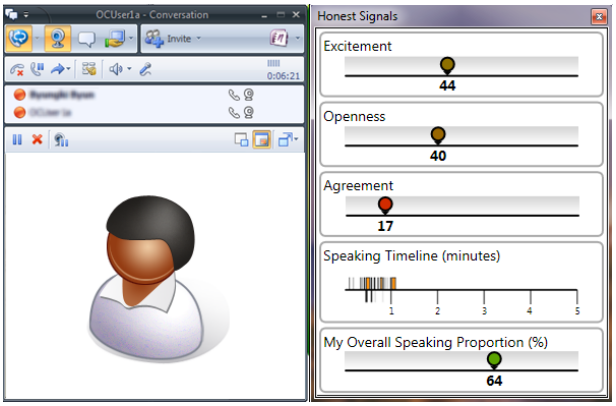

Fig. 1. The graphical user interface showing conversation feedback used in a wizard-of-oz study. The right panel shows the signal levels, speaking timeline and proportion.

of the user interface elements we chose to display were confusing to participants, positive comments indicated the utility of the information. For example:

- *I looked at the time mostly though but it gave me an idea of how our speaking time was shared. I am not sure why but I felt more in control of the video conversation.*
- *Some of it [was useful]. To be honest, I'm not really sure what openness is referring too. Seems vague ... excitement and agreement are good. Using the timeline more than percentage, would be nice to see them together.*
- *It certainly helped moderate the discussion. I also think it might keep meetings from wandering-- given that the time elapsed is shown.*

Thus, we refined the user interface, moving forward to develop a system that could give honest signal-style feedback to users in real time.

4. SYSTEM OVERVIEW

Here, we provide an overview of the real-time system developed. As illustrated in Fig. 2, a pair of clients establishes an ordinary video conference between them, and starts to stream audio and video data to each other. As soon as the video conference is started, each client automatically establishes a TCP connection to an honest signal server. The server listens for these TCP connections arriving from various clients, matches the connections coming from a communicating pair, and spawns a process to handle each pair of clients in conversation.

The clients intercept their own audio and video signals, extract low-level features (e.g., pitch period, visual motion) from each frame of audio and video, and transmit the low-level features over their connections to the server. The TCP connection carries low bandwidth, constant bit rate information from client to server using about 5 kbps: about 12 bytes for every 30 ms for audio and about 60 bytes for every 67 ms for video. The server process synchronizes the

streams of the low-level features from the two clients, evaluates additional intermediate-level features from these low-level features, estimates the level of the four honest signals, and feeds honest signal estimates back to its respective clients for presentation to the end-users in real time. Transmitting the estimates of the honest signals requires only about 160 bps from the server to each client. The connections are maintained until the end of conference.

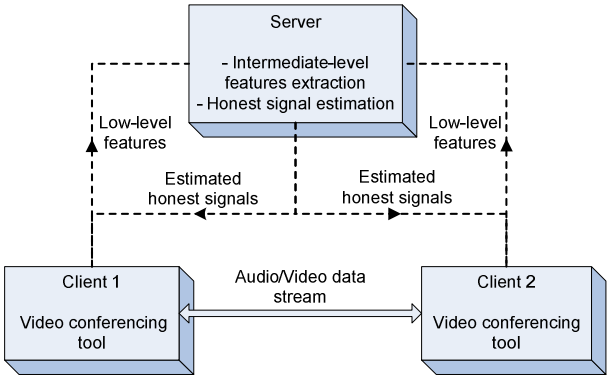

Fig. 2. A block diagram of the real time system.

The key computation at the server is real-time estimation of the users' honest signals from the low-level features of each pair of communicating clients. In particular, at regular intervals (once per second in this work), the low-level features are first processed into a vector of intermediate-level features. Next, these intermediate-level features are fed into four logistic-regression models trained a priori on data collected as described in Section 6. Each of the logistic regression models estimates the probability that the corresponding honest signal is in high state. These probabilities are then sent to the respective client for presentation.

Note that the system can be extended easily to multi-party conversations as computation of the intermediate-level features and estimation of the honest signals are all done at the server.

5. NON-LINGUISTIC FEATURES

In this section, we describe the non-linguistic features used. The low-level features extracted at each client primarily capture individual behavioral patterns. In contrast, the intermediate-level features computed at the server based on the low-level features describe not only individual behavioral patterns but also mutual interactions. We handcrafted the features to be indicative of the four honest signals, which are embedded in participants' behavioral patterns and mutual interactions captured by the audio and video. Later, the logistic regression models that estimate the honest signals from these features are trained from data collected in our user studies.

5.1. Low-level features

The low-level features are captured at 30 ms frame intervals for audio and 67 ms frame intervals (15 fps) for video. For audio, three features are computed per frame: pitch, voice activity, and spectral distance. For pitch, a pitch tracker performs linear predictive coding (LPC) [13] and uses dynamic programming to estimate the pitch (or zero if the frame is unvoiced or silent). For voice activity detection, we use the power spectrum of a frame at time t after identifying voiced frames with the pitch tracker. In particular, voice activity is detected when the total power spectrum of the frame at time t is larger than a threshold and its neighboring frames are determined as voiced by the pitch tracker. Finally, for spectral distance, we compute the Itakura distortion [14] from the LPC coefficients between two consecutive frames.

For video, 15 features are computed per frame. A face tracker [15] provides one rectangular region where a face is detected. The magnitude of the motion of the center of the rectangle between frames is one feature. Two other features are the average magnitude of the motion vectors inside the facial region and the average magnitude of the motion vectors outside the facial region, where the motion vectors are computed through an optical flow computation between consecutive frames. Twelve outputs of a Gabor filter-bank with two parameters, scale and orientation, to account for abrupt changes of facial expressions [16] are also extracted from the facial region for future work.

5.2. Intermediate-level features

The intermediate-level features are computed at regular intervals–once per second in our experiments–for both audio and video. For *each* participant, a vector of intermediate-level features is computed based on the low-level features of *both* participants since the intermediate-level features may need to capture behavioral interactions between participants.

One of the distinct behavioral interactions in a conversation is turn-taking. At one extreme, a single person can dominate a conversation without giving a chance for anyone else to talk, while at the other extreme, a person can never talk but always listen. We define three unique turn-taking patterns in this work: *barge-in*, *grant-floor*, and *suppression*, as illustrated in Fig. 3.

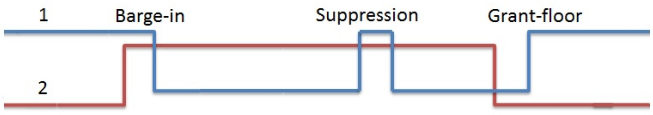

Fig. 3. Example timeline of turn taking between Persons 1 (audience) and 2 (actor). These turn taking patterns can be used to analyze the properties of a conversation.

In Fig. 3, a high value means a person is talking and a low value means the person is silent. Let Person 2 be an actor who wants to control turn-taking and let Person 1 be a member of the audience who gets influenced by the actor. Then, in Fig. 3, from left to right, Person 2 barges in on 1,

suppresses 1, and finally grants the floor to 1. Specifically, by starting to talk, *barge-in* forces the other person to stop talking and by continuing to talk, *suppression* prevents the other person from taking the floor. Finally, converse to barge-in, by stopping to talk, *grant-floor* allows the other person to start talking.

To extract these turn-taking patterns, we use voice activity detection to measure when someone starts and stops talking. Given continuous segments of voice activity along the time axis, let t_1^s and t_2^s be start times of the segments, and let t_1^t and t_2^t be termination times of the segments for Persons 1 and 2, respectively. Then, from the perspective of Person 2, Δt_{barge}, Δt_{grant}, and Δt_{supp} can be defined (where $I(\cdot)$ is an indicator function) as

$$\Delta t_{\text{barge}} = (t_1^t - t_2^s) \cdot I(t_1^t > t_2^s),$$
$$\Delta t_{\text{grant}} = (t_1^s - t_2^t) \cdot I(t_1^s > t_2^t),$$
$$\Delta t_{\text{supp}} = (t_1^t - t_1^s) \cdot I(t_1^s > t_2^s) \cdot I(t_2^t > t_1^t).$$

The number of barge-in, grant-floor, and suppression events within the last T_o (set to 30) seconds, and the average of their durations within the last T_o seconds are the first six intermediate-level features for Person 2. Intermediate-level features for Person 1 are similarly defined.

Four other intermediate-level features are extracted from prosody: (a) average speaking rate, (b) pitch variation, (c) syllabic rate variation, and (d) spectral distance variation. As studied in [17], prosody is closely related to the emotional state of a speaker. For example, if one is overwhelmed by questions, the person will more likely exhibit high variation in prosody than when the person is in a normal emotional state. Among these features, average speaking rate is the simplest to compute: the percentage of active frames over the last T_o seconds.

To compute variations in pitch, syllabic rate and spectral distance, we first compute their averages over the previous T_o seconds, as well as over every t_o (set to 3) second within the T_o seconds. Syllabic rate is determined by counting the number of voiced/unvoiced transitions per second. We then compute the squared variation of these t_o-second averages around the T_o-second average. The variation in syllabic rate indicates changes in the speed of utterances while the variation in spectral distance should capture informative events in utterances such as exclamation.

Additionally, to model variation of intonation in speech, two autoregressive models with an order of Q and q are also trained given continuous segments of the estimated pitch frequencies over the last T_o and t_o seconds. If there are multiple such segments, segments that are longer than a certain threshold are selected and concatenated to train R_t and r_t, the models at time t for T_o and t_o, respectively. Given R_t and r_t, we then measure the normalized squared residual errors of the pitch frequency for the interval of *[t, t+1]*, creating our next two features.

Furthermore, we compute the average magnitude and variation in head motion rate from the low-level motion vectors and location of the tracked facial regions as our next two features. Like prosody, physical motion is closely associated with a person's internal state, especially the level of excitement.

Besides 14 intermediate-level features described above, we also compute 12 additional intermediate-level features involving means and variances of pitch, syllabic rate, and spectral distance over both T_o and t_o seconds, which complete 26 features used in our experiments.

For future exploration, we also implement head-nod and shake detectors [18] to identify higher-level cues of the emotional state of a speaker using Hidden Markov Models (HMMs), where there are two hidden states and eight Gaussian mixtures with the motion vectors extracted from the facial region to be observables of the HMMs.

6. EXPERIMENTAL RESULTS

First, we conducted a user study to collect real-world data to train and test our estimators of the honest signals. We then tackled the problem of identifying conversation type to explore predictive power of non-linguistic cues. A second user study with different participants was later conducted using the learned estimators to verify the hypothesis regarding effects of real-time visualization of the honest signals in video conferencing.

6.1. Data collection

In the first user study, we recruited 20 people to carry out 10 pairs of conversations. Each pair was asked to role play in five types of conversations that included salary negotiation, expounding upon or listening to a personal opinion on a controversial topic, expounding upon or listening to a personal dilemma, exploring what is in common, and brainstorming. Two experimenters monitored the conversations and annotated key events on the time line. Such key events consisted of episodes of high/low activity, high/low influence, high/low consistency, and high mimicry. Low mimicry was not explicitly labeled because suitable examples of low mimicry could be chosen arbitrarily. To ensure quality in labeling, each experimenter monitored only one participant.

The length of each conversation was seven to eight minutes resulting in about 350 to 400 minutes of conversation in total (5 conversations from each of 10 pairs of participants). In the end, this exercise resulted in a data set consisting of a collection of instances of varying lengths labeled for high and low levels of each signal during each conversation. The average length of such instances was 11 seconds across all conversations. The episodes for activity and mimicry were the shortest, from three to ten seconds, while episodes indicating influence were the longest and were as much as 30 seconds in duration. These instances were then used to train and test the predictors described in the next sections. The dataset is described in Table 1.

Table 1. Description of the collected dataset. The last row contains the total instances / total durations of time slices.

Activity		Consistency		Influence		Mimicry
High	Low	High	Low	High	Low	High
99 / 1951s	21 / 673s	40 / 574s	60 / 1140s	48 / 391s	41 / 356s	48 / 211s

6.2. Predicting honest signals

Next, we built predictors for the honest signals. For each second of labeled data, we assembled a vector of intermediate-level features as described in the previous section. On these features, we performed binary (i.e., high vs. low) classification using linear logistic regression for all four honest signals. The episodes of low influence were used as a proxy for episodes of low mimicry. Also, we used a leave-one-out strategy, where training was first performed on instances from nine conversation pairs and tested on the instances from tenth. The accuracies were averaged over the complete cycle of leave-one-out and thus reflect predictive power independent of participants in the conversation.

Also note that while activity and consistency are personal traits, influence and mimicry reflect the interaction between people. Consequently, we categorized a subset of the extracted features into two groups as shown in Table 2.

Table 2. 11 features selected to estimate the honest signals.

Category I	Category II
Average speaking rate	Average length of barge-in
Variation of pitch	Average length of grant-floor
Variation of syllabic rate	Ave. length of suppression
Variation of spectral dist.	Frequency of barge-in
Average head motion rate	Frequency of grant-floor
	Frequency of suppression

The features in Category I are personal properties, while the features in Category II are properties of how one person interacts with another. Features in both categories can be computed for both sides of a conversation. We hypothesized that determining a person's influence and mimicry depends crucially on joint features from both sides, while activity and consistency depend only on one-sided personal features.

To test our hypothesis, we constructed four different systems. First, we considered two baseline systems that used all 26 features described in Section 5 (including all means and variances, a superset of Table 2) but differed in whether the features were from either individual or both participants. Then, we had another system that included all the features mentioned in Table 2 (Categories I and II) from both participants (11 from each, for a total of 22 dimensions). Finally, we designed a custom system where we included Category I and II features only from one side for activity and consistency, from both sides for mimicry, and just the Category II features along with speaking rate for influence. In sum, we had the following setups:

- System I: All 26 features from both sides (total 52 dim.)

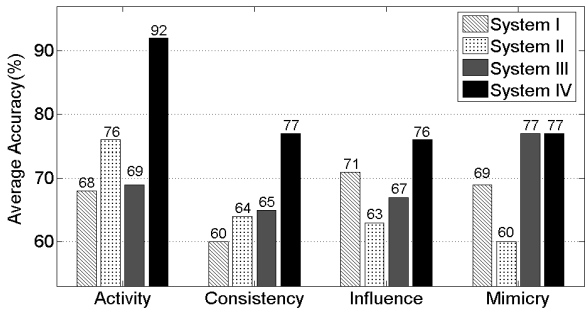

Fig. 4. Average normalized accuracy of predicting the honest signals while varying a set of features used.

- System II: All 26 features from one side (total 26 dim.)
- System III: 11 features from both sides (total 22 dim.)
- System IV: Custom as described above.

Fig. 4 presents the average recognition results of the above four systems. We observe that the accuracies obtained for activity and consistency are higher for System II than I, highlighting that those two signals are more dependent on individual traits as opposed to the joint features. The same tendency is observed when comparing System III with IV. On the other hand, comparing System I vs. II, better accuracies are obtained for influence and mimicry in System I, which verifies our hypothesis about these two honest signals. Furthermore, the custom system, which takes into account the different personal traits and joint properties of the conversation into the classification, performs best showing significant predictive power of the features. In particular, we see a significant improvement in accuracy for activity and consistency from System III to IV (from 69% to 92% and from 65% to 77%, respectively). Moreover, the improvement in accuracy for influence of System IV against System III (67% to 76%) reflects that turn-taking features are fairly important for predicting influence. Overall, the results indicate that the non-linguistic signals have enough discriminative power to predict the four dimensions characterizing the honest signals. These honest signals thus can be estimated from non-verbal behavior and presented back to users while video conferencing.

6.3. Conversation type prediction

We also explored whether non-linguistic cues have enough predictive power to perform conversation type prediction. We first categorized the five conversations that took place during the first user study into four conversation types: (a) negotiation (e.g., salary negotiation), (b) active listening (listening to personal opinions and dilemmas), (c) exploring (exploring what is in common) and (d) brainstorming. We then used a one-vs.-all formulation of logistic regression with the entire set of extracted features (System I). Table 3 summarizes the performances.

We observed an average accuracy of 77% across all the conversation classes. The system performed best in predicting the conversation class of 'Exploring' with 83% accuracy, while 'Listening' achieved the worst performance

with 72% accuracy. Note that our features do not encode any semantics or content of the conversation. This suggests that non-verbal cues can provide important information about the conversation type. Thus, it might eventually be useful in predicting the final success of the video conference.

Table 3. Accuracy for conversation type prediction.

Conversation Type	Accuracy (%)
Negotiation	76
Active listening	72
Exploring	83
Brainstorming	77
Average	77

6.4. User perception and survey

In the second user study, after participating in conversations of the five different types and playing different roles, participants were queried as to how valuable they found the new system, which features they liked or did not find valuable, and whether or not such a tool would influence their conversations. While users varied to the degree that they were able to use the feedback in real time, most agreed that the system could in fact provide value for them during video conferences. Most participants found the speaking time and proportion to be extremely useful, and said that the system did alter their behaviors, whether they used the system as a "coach" to speak less, or as a "prompt" to speak more. Some participants really liked the consistency and influence signals, but many participants wished they could see the signals for the other person as well, to compare how they were doing in reference to their partner. Participants also asked for a cumulative view of how they had been behaving, not just a summary of the last five minutes, for retrospective analysis. A couple of participants liked the activity feedback, with one participant even stating that it would be very useful for her child with attention deficit disorder (ADD) when using Skype. Overall, the system was well received, but it was clear that we needed to iterate on the user interface to make it less distracting, more glanceable, and cumulative for the whole session.

7. CONCLUSION AND FUTURE WORK

We set out to explore whether or not there was viability in building a system that could automatically analyze non-verbal signals for video conference participants, enhance their conversations with appropriate feedback, and also provide predictive information about characteristics of the conversation. An early wizard-of-oz user study provided initial evidence that the concept had merit, but that the feedback user interface needed iteration. We built the system and, using real user data, were able to show that we could not only accurately assess the nonverbal signals we were interested in, but we could also make predictions about meeting types and user roles, and that users found the feedback valuable for modulating their video conference conversation. This is the first known investigation of these kinds of predictions and evaluations using actual audio and visual signals automatically via a videoconferencing system. We believe that there is much more that can be done beyond the work explored in this paper, for example: exploration of alternate UI designs, incorporation of richer modalities such as facial expressions and physiology, and extension to multi-party conversations.

8. REFERENCES

[1] T. Choudhury, "Sensing and modeling human network," *Ph.D thesis*, MIT, 2004.

[2] A. Pentland, *Honest Signals: How they shape our world*, MIT Press, 2008.

[3] S. Basu, "Conversational Scene Analysis," *Ph.D thesis*, MIT, 2002.

[4] R. Picard, *Affective computing*, MIT Press, 1997.

[5] P. Eckman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, 2nd edition, 2004.

[6] A. Kapoor, W. Burleson, and R. Picard, "Automatic prediction of frustration," *Int'l J. Human-Computer Studies*, vol. 65, 724-736, 2007.

[7] J. N. Bailenson *et al.*, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int'l J. Human Machine Studies*, vol. 66, 303-317, 2008.

[8] H. C. van Vugt *et al.*, "Effects of Facial Similarity on User Responses to Embodied Agents," *ACM Trans. Computer-Human Interaction*, vol. 17, 2010.

[9] K. Otsuka *et al.*, "Conversation Scene Analysis with Dynamic Bayesian Network Based on Visual Head Tracking," *Proc. ICME*, 2006.

[10] D. Jayagopi *et al.*, "Characterizing conversational group dynamics using nonverbal behaviour," *Proc. ICME,* 2009.

[11] J. M. Dimicco, "Changing small group interaction through visual reflection of social behavior," *Ph.D thesis*, MIT, 2005.

[12] J. Sturm *et al.*, "Influencing social dynamics in meetings through a peripheral display," *Proc. ICMI*, 2007.

[13] T. F. Quatieri, *Discrete-time speech signal processing: Principles and Practice*, Prentice Hall, 2001.

[14] R. Gray *et al.*, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, 1980.

[15] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," *Microsoft Research Tech. Report*, 2010.

[16] B. Fasei and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, 2003.

[17] R. Barra *et al.*, "Prosodic and segmental rubrics in emotion identification," *Proc. ICASSP*, 2006.

[18] A. Kapoor and R. Picard "A real-time head nod and shake detector," *Proc. PUI*, 2001.