CoCo: Collaboration Coach for Understanding Team Dynamics during Video Conferencing

SAMIHA SAMROSE, RU ZHAO*, JEFFERY WHITE, VIVIAN LI, LUIS NOVA, YICHEN LU, MOHAMMAD RAFAYET ALI, and MOHAMMED (EHSAN) HOQUE, University of Rochester, USA

We present and discuss a fully-automated collaboration system, CoCo, that allows multiple participants to video chat and receive feedback through custom video conferencing software. After a conferencing session, a virtual feedback assistant provides insights on the conversation to participants. CoCo automatically pulls audial and visual data during conversations and analyzes the extracted streams for affective features, including smiles, engagement, attention, as well as speech overlap and turn-taking. We validated CoCo with 39 participants split into 10 groups. Participants played two back-to-back team-building games, *Lost at Sea* and *Survival on the Moon*, with the system providing feedback between the two. With feedback, we found a statistically significant change in balanced participation—that is, everyone spoke for an equal amount of time. There was also statistically significant improvement in participants' self-evaluations of conversational skills awareness, including how often they let others speak, as well as of teammates' conversational skills. The entire framework is available at https://github.com/ROC-HCI/CollaborationCoach_PostFeedback.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing design and evaluation methods*;

Additional Key Words and Phrases: Virtual feedback system, Team dynamics, Group discussion, Video conferencing

ACM Reference Format:
Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed (Ehsan) Hoque. 2017. CoCo: Collaboration Coach for Understanding Team Dynamics during Video Conferencing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 160 (December 2017), 24 pages. https://doi.org/10.1145/3161186

1 INTRODUCTION

Consider the following scenario: *"Beth, the only woman on a team of developers, always looks for an appropriate moment to interject during a staff meeting. But whenever an opening arises, louder voices quickly seize the opportunity to fill the silence. Beth's voice is drowned away, and she is rarely, if ever, heard. Even though Beth is by nature a confident and assertive person, her group members label her as shy and unassuming."*

The above scenario reveals how slight differences in preference for conversational mechanics affect which ideas are heard and how people are judged [50]. These behavioral dynamics are common, and if not addressed properly, can impede team productivity. Thus, analyzing and reflecting on these dynamics may positively impact productivity and foster positive and respectful relationships toward other group members. In this paper, we explore that possibility using a video conferencing platform. Despite the importance of understanding group

* Equal contribution as the first author

Authors' address: Samiha Samrose, ssamrose@cs.rochester.edu; Ru Zhao, rzhao2@u.rochester.edu; Jeffery White, jwhite37@ur.rochester.edu; Vivian Li, vli4@u.rochester.edu; Luis Nova, lnova@u.rochester.edu; Yichen Lu, ylu21@u.rochester.edu; Mohammad Rafayet Ali, mali7@cs.rochester.edu; Mohammed (Ehsan) Hoque, mehoque@gmail.com, University of Rochester, Computer Science, Rochester, NY, 14620, USA.

2474-9567/2017/12-ART160 $15.00
https://doi.org/10.1145/3161186

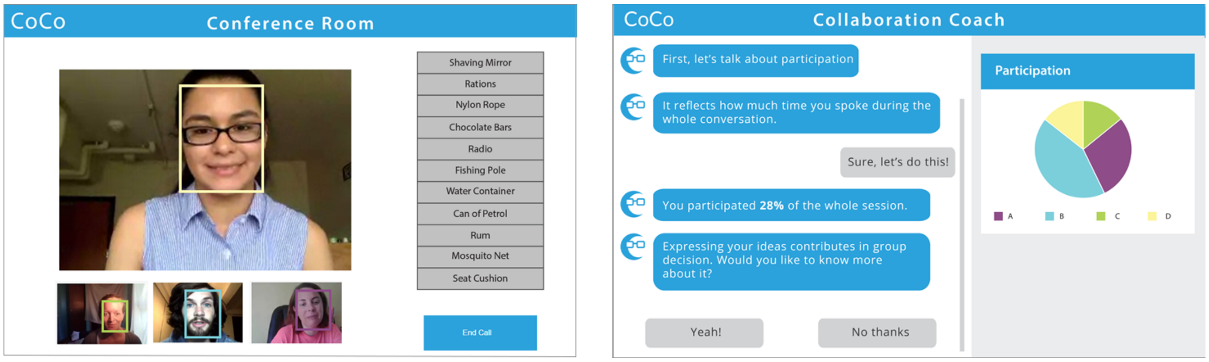


Fig. 1. (Left) Video Conferencing System, (Right) Feedback Interface

dynamics in a video-conferencing environment, it has not been studied extensively in the past. No existing video-conferencing platforms—including Skype, Google Hangouts, GoToMeeting, WebEx, ooVoo, and others—allow easy access to an API to automatically capture the audio and video feeds.

What if, with participants' consent, we could automatically analyze their audial and visual data and offer intelligent analytics on conversation dynamics? Imagine the questions we might answer. Did everyone participate equally, or was it a one-sided conversation? What was the overall emotional valence? Did people talk over one another? Did they share smiles? How did participants act toward one another?

Due to the proliferation of smart phones and tablets, video conferencing is gradually becoming a popular alternative to traditional face-to-face meetings, which require travel, space, and other advance planning. One survey reports that 76 percent of decision-makers use and recommend video conferencing without hesitation [3]. Ninety-six percent of participants surveyed agree that video conferencing increases average attention span from 23 minutes (on regular calls) to 35 minutes. It can also, however, reduce attention devoted to nonverbal cues [57]. Virtual meetings have introduced new challenges. But they have also opened up opportunities for computer-mediated feedback to address their shortcomings.

Another challenge is translating the raw, sensed data into insights that are helpful to participants. The effectiveness of a group discussion depends on many parameters, such as those related to group type (one-time vs. ongoing groups), those specific to individual participants (personality, background, etc.), those related to the topic(s) of discussion (open- vs. set-agenda plans), and more. Some groups may consider the outcome of the discussion the most important factor—in a business meeting, for instance—while other groups may focus on strengthening cohesion among team members. Optimal participant behavior may be difficult to establish, as it hinges on both group dynamics and the desired result. In this paper, we present a fully-automated, online collaboration platform: Collaboration Coach (CoCo). CoCo allows multiple users to video chat and receive feedback on group dynamics as well as individual communication skills. The CoCo platform includes both a video conferencing system and a virtual feedback assistant. The conferencing system allows us to capture participants' audio and video feeds. Before a video conferencing session, participants agree to enable their webcam and to have the conversation's audio and video data transmitted to our server. Using the video feed, we analyze facial features to determine the aggregate emotional valence of participants, participants' attitudes toward one another, and how often users shared a smile with other members. Using the audio feed, we analyze how much participants talk (participation), how much each participant interrupts and gets interrupted (overlap), and who spoke after whom (turn-taking). The virtual feedback assistant, a chat-based interface, relays complex performance metrics in an easy-to-understand format and provides feedback and conversational insights to participants in a positive and constructive way. Fig 1 offers a glimpse of the overall system.

To validate our framework, we conducted a pilot study with 39 participants. We randomly split the participants into groups of four—and one group of three—resulting in 10 groups. Each group participated in two discussion sessions with two popular team-building tasks called *Lost at Sea* [1] and *Survival on the Moon* [24]. The ordering of the tasks are counterbalanced across all groups. In *Lost at Sea*, participants imagine themselves escaping a sinking ship, ranking in order of most useful for survival a pre-defined set of items. (The actual best ranking of the items is annotated by the U.S. Coastguard.) The *Survival on the Moon* task follows a similar pattern in which participants, as crew members of a malfunctioning space ship, have to again rank the available survival items. After receiving their scenario, each participant orders their own list. Then the participants uses the CoCo system to devise a unified, final ranking of the top five items, together representing the team's decision. CoCo records audio and video of the discussion using participants' webcams and microphones. From the audial and visual data, we extracted the facial and prosodic features of the participants. These features include: *participation, interruption, turn-taking, emotional valence, attitude (engagement, attention, surprise, anger),* and *shared smiles*. After each team made its final decision, our chat-based virtual feedback assistant debriefed each member with quantitative metrics about group and individual conversational nuances. The feedback assistant does not advise participants to change their behavior—rather, it informs them of the group and individual effects of the performance features in a positive, constructive way. We opted for this design to allow humans to control their own behavior modifications instead of influencing them with predefined suggestions. After the feedback session, the teams engaged in another round of group discussion through the video conferencing system, focusing on the other task.

We noticed a statistically-significant change in balanced participation—that is, everyone spoke more equally—during the second round. Participants' self-evaluations of their awareness of their own conversational skills, their teammates' conversational skills, and how often they allowed others to speak also showed statistically-significant improvement. We believe the proposed CoCo system improves the efficacy of video conferencing environments with the integration of analysis and feedback tools. Because our framework is publicly available at https://github.com/ROC-HCI/CollaborationCoach_PostFeedback, it could be used by other researchers to collect more virtual meeting data and expand our understanding of group dynamics.

In summary, this paper makes the following contributions:

- The development of a fully-automated video conferencing system that captures nonverbal features of both individuals and teams from participants' audio and video feeds;
- The design of six crucial group interaction features (*participation, speech overlap, turn-taking, emotional valence, attitude, and shared smile*) and four sub-features (*attention, anger, surprise, engagement*, under *attitude*) to better describe the complexities of group discussion;
- The implementation of a chat-based system that converses with participants and constructively reveals nonverbal metrics after the video conferencing session ends;
- The design of a two-stage, team-building experiment with 39 participants (spread over 10 groups) to validate the developed framework and its potential effects;
- The analysis of group performance and behavior metrics from both qualitative and quantitative perspectives to better understand team dynamics, in general.

2 RELATED WORK

An automated system that allows individuals to conduct a video conferencing session and receive automated feedback on team dynamics culls knowledge from affective computing, collaboration, teamwork, and rapport. In this paper, we reviewed two primary sources of information: systems that provide feedback on team dynamics and literature on teamwork and rapport. We first discuss existing systems intended to better understand team dynamics that also provide a feedback mechanism. We highlight the fundamental features that underpin effective group collaboration. We discuss the particular contributions to teamwork of features like turn-taking, overlap, participation, and sentiment.

2.1 Existing Systems

Providing feedback on virtual group communication is a powerful way to shape team dynamics. Previous systems have made efforts to leverage the potential of group feedback on a number of group interaction features, including participation, content, and agreement. The thrust of these systems is the idea the more aware group members are of their behaviors, the more they can improve their communication abilities.

One such system is *GroupMeter*, a web-based chat room that lets users instant-message and process feedback visualizations during their conversation [38]. GroupMeter displays feedback on word count, engagement, and agreement between group members. This feedback is shown visually, in real time updates as group members continue to discuss. Results demonstrate that providing visual feedback increased user awareness of language use and agreement, eliciting positive changes in behavior. DiMicco et al. support this finding. In their work, they collect audio from participants in a group discussion and visualize participation rates in a bar chart on a shared screen in the discussion room [22]. Their results show that participants who spoke too often would reduce their speech time as a result of seeing the chart.

Similarly, Tausczik and Pennebaker offer real-time feedback, but unlike the GroupMeter prototype, their system focuses on giving individualized feedback [53]. Instead of showing passive, neutral visuals, Tausczik gives users suggestions, like recommending that users pay more attention or stay on topic. During a conversation, a group member may elicit a comment like "*Your group is working alright, but could be improved. Be sure to pay attention to what others are saying.*" Their user studies show that real-time feedback can be distracting, and that negative feedback can be damaging.

Calacci et al. focused on analyzing speech overlap and turn-taking when they designed their tool, *Breakout*. Breakout is both an interaction platform for online, face-to-face peer learning groups and a research tool to help study real-life team dynamics [13]. The platform collects audial and visual data in real time and then uses metrics like overlap, turn-taking, and response patterns to measure social relationships and engagement to display aggregated feedback. Their modular feedback analysis is based on the Meeting Mediator [33] of Kim et al., which visually maps each teammate as a node. A ball in the center of the visualization moves toward whoever dominates the conversation. However, Calacci et al. have not yet evaluated the effectiveness or influence of this feedback on group interactions. The feedback is provided through real-time visualization, which may affect the natural flow of the group discussion. Although it takes into account various communication features (e.g. turn-taking, speech overlap), it does not include features related to rapport, like smile, valence, etc.

Sellen [49] shows a difference between same-room and video-mediated conversations, especially in terms of floor control and simultaneous talk. It shows no significant effect, however, on turn-taking among speakers when the whole group can be viewed simultaneously. Cohen [16], on the other hand, finds significantly more speaker turnover in face-to-face meetings than video meetings. Notably, in Cohen's study, the video shifts to the current speaker, foreclosing on a view of the whole group. Thus, system settings may make a difference in the behavioral patterns of a group.

Providing feedback for improved social interaction has been studied extensively in assistive technology-based research. This research reveals the complexities of social interactions and the importance of providing feedback. Research has contributed toward improving interactions in different need-based settings [8, 46, 47, 60]. In most cases, the feedback is provided in real time without extensive explanation or analysis. *SayWAT* [10] entails a wearable assistive technology for adults with autism that provides real-time feedback on prosody during face-to-face conversations. Another automated system, *MOSOCO* [54], which was developed on a mobile phone platform, intervenes in real-life situations by providing real-time, step-by-step guidance. Likewise, another mobile-based intervention system called *TalkBetter* [28] provides parents with real-time meta-linguistic analysis during conversations with their children. *Rhema* [52] provides speakers with real-time feedback through Google Glass to improve their public speaking. The *Logue* [19] system explores the instant information flow through Google Glass during social interactions. It shows, notably, that real-time feedback may be distracting during real-life conversations [20, 42, 56]. Reducing the complexity of information flow and the number of interactions can reduce distraction [12, 58]. For group interactions, complexity increases with the number of members and the number of features analyzed. Thus, real-time feedback is liable to be a major distraction in more analytical systems dealing with larger groups.

All these works lay the foundation for navigating feedback design and generation for group communication. However, these works focus on only limited dimensions of team dynamics. For instance, GroupMeter and the work of Tausczik et al. are restricted by the data available through an instant messaging platform, where only text-based information can be extracted. It is subsequently not useful in analyzing team dynamics with respect to nonverbal behavior. Additionally, the mental work required to process real-time feedback often distracts participants from the discussion. The real-time feedback also does not provide participants with explanations for the suggestions. In the CoCo system, we focus on understanding team dynamics by developing a video conferencing platform to extract and display meaningful information from audial and visual data. This enables feature extraction from both verbal and nonverbal behavior, like shared smiles and turn-taking, that users may otherwise have difficulty quantifying and understanding. Our CoCo virtual feedback assistant offers post-discussion feedback, explaining each extracted feature to the participants. Importantly, CoCo makes use of a popular communication mechanism, video conferencing. If successful, the feedback strategies established through CoCo can be readily incorporated into existing video platforms without significantly changing the user experience.

2.2 Teamwork and Rapport

Personalities and behavioral predispositions affect the ways in which people interact in group settings. Social psychology research connects various communicational skill patterns to specific personalities and behaviors. These connections help us better understand group dynamics. Dunbar et al. [23] finds that dominant people tend to be more verbally or physically active during interactions. Evaluating the behavior of less dominant people is difficult because these people demonstrate less activity [31]. Other research [21, 35] indicates that conversational features, like turn-taking, speaking time, prosody, and others, significantly affect the social interaction experience.

Analysis by Hung et al. [27] infers who dominated or led a group discussion based on the turn-taking patterns among the group members. *Meeting Mediator* [32] discusses the effects of continuously updating and showing turn-taking information in group discussions. McLeod et al. [40] show how feedback on speaking time or task focus can influence a group's teamwork tendencies. Feedback on these features can positively influence the group discussion, resulting in better group performance [29, 30]. The AMI meeting corpus [14] develops a meeting browsing technology with hand annotations for transcriptions and extraction of features including emotions, headnods, and more.

There is no standard behavior that defines a "good" team. A number of works have examined teamwork with respect to several social variables, including team size [11], leadership [25, 26], goal-orientation [44], and rapport [34]. Of particular interest to us was rapport, characterized by three related components: mutual attention and involvement, mutual positivity, and coordination [55].

Mutual attention refers to interactions in which group members feel involved and are interested in what others are saying. Mutual positivity is best described as friendliness and caring, and can be quantified through nonverbal behaviors like eye contact and smiling [41, 55]. Coordination requires behavioral cohesion, including attributes like smooth turn-taking, acknowledgments like head-nodding, and back-channeling (e.g. uttering phrases like "uh-huh")[41]. Drawing from these three components of rapport, we look at mutual attention through participation, mutual positivity through smiling, and coordination through overlaps and turn-taking.

In auditory conversations, participation is related to how long a speaker holds the floor. Some argue that equal participation leads to greater group interactivity and an overall better experience because all group members have a chance to share their thoughts [22, 36]. While equal participation is not ideal for all discussions, it is important in collaborative decision-making, where input from all members is needed. Inequality in participation may indicate neglect for some group members' opinions and may subsequently hinder the group's effectiveness [48].

Smiling is conducive to establishing trust and friendliness, and is related to positivity [15]. In both one-on-one interactions and in groups, researchers have found that individual positivity is contagious and helps improve rapport [6, 7]. Barger finds that a service employee's smile can spread to their customers and positively affect the customer's mood and perception of the quality of service [6]. Similarly, in a simulated hiring committee discussion, Barsade shows that a group member's positivity influences not only others' moods but also their judgment when making decisions [7]. In his study, groups with increased positivity saw improved cooperation and a decrease in conflict.

Overlaps are typically considered disruptive or indicative of dominance. But research shows that most overlaps are supportive, rather than obstructive [43, 51]. Peers with strong rapport, for instance, will overlap one another frequently, suggesting that they are socially comfortable with each other [51]. Likewise, one peer may overlap another because they want to be involved in the conversation [43, 45]. Whether an overlap is supportive or obstructive ultimately depends on context, making it difficult to provide feedback.

3 FEATURE DEFINITION

In this section, we describe features that we felt were important in the context of team dynamics based on our examination of relevant literature. We also detail how we extract these features.

Participation is defined by the percentage of time one participant speaks during the conversation. This feature is extracted from audio data using open-source software called *Praat* [9]. *Praat* generates from raw audio data a file containing the start and end times of each "silence" detected in the input audio, correcting for noise. Each participant is given a percentage-based score based on how often and for how long they spoke. Speaking longer and more frequently risks suppression of others' viewpoints, while speaking less squanders chances to express one's own opinion. As a result, equal participation is ideal for small, collaborative discussions [36].

Interruption or *speech overlap* is extracted from the *Praat* analysis and is based on when one person interrupts another. To reduce the effects of background noise, we enforced a threshold of a 0.6-second overlap to qualify as an interruption. We count each interruption as a single unit value. Overlaps that falls under that threshold are considered *back-channeling*, which is supportive of the current speaker. Fig 2 offers examples of interruption and non-interruption. Too much overlap can disrupt the flow of the conversation, creating dissatisfaction.

Turn-taking describes the order in which participants take turns during the conversation. Conventionally, exchange among everyone in the group keeps the group connected. Using the audio features extracted using *Praat*, we are able to merge each individual's recorded audio into a single track with labels at the start and end times of each user. This allows us to identify the sequence of conversation.

Valence expresses emotion in a range from negative to positive (including neutral). Using video analysis through the Affdex SDK [4], we sample a valence score ranging from -100 to +100 every 250 milliseconds during the conversation. Averaging valence scores across the whole conversation yields a valence score for each person.

Emotion, in this case, consists of four separate features described by Affdex. *Attention* measures focus. *Anger* and *surprise* are self-explanatory, while *engagement* describes a particpant's expressiveness. These features were

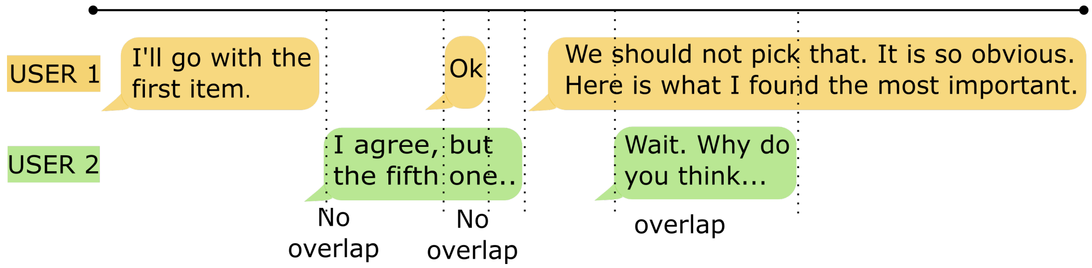


Fig. 2. Example of speech overlap in the context of two users with a threshold of one second.

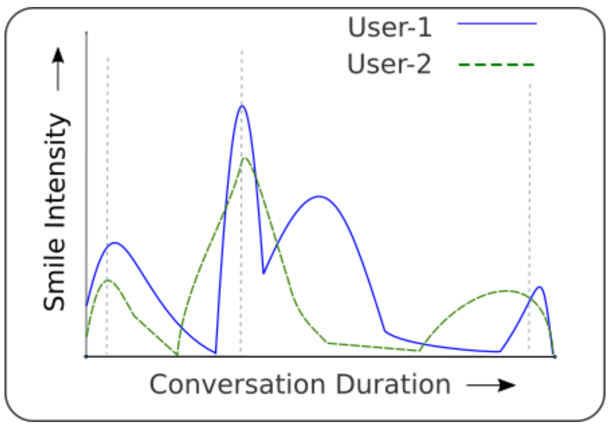


Fig. 3. Calculation of shared smile count.

also sampled every 250 milliseconds during the conversation, with an average score for each shown to the participant at the end.

Shared smile denotes the number of times one person smiles at the same time as another during the conversation. Affdex supports a smile score ranging from 0 to 100—75 beind the threshold at which an expression is considered a smile—which, like other features, was sampled every 250 milliseconds. Shared smiles are, naturally, identified in pairs. When two participants, A and B, cross the score threshold and maintain a smile for at least one second within the same time period, we identified a shared smile between A and B. Fig 3 depicts this scenario.

4 COCO: COLLABORATION COACH

We developed a fully-automated, end-to-end system that allows individuals to video conference, captures their audial and visual data, analyzes them on a remote server, and then deploys a virtual feedback assistant to debrief participants about the team dynamics. The entire system functions in sequence without any significant delays or interruptions. This particular implementation was necessary to support the two rounds of back-to-back collaboration experiments. Below, we provide the details of our design and implementation choices. We used only open-source and freely-available software packages so that others may easily replicate or reuse our framework.

4.1 CoCo Video Conferencing System

In order to provide feedback, CoCo needs to have full access to the audio and video data streams from the conferencing sessions, to which most commercial software only provided limited access. Therefore, we have implemented the CoCo Video Conferencing System to study team dynamics. The system automatically analyzes data on the remote server once the video session has ended. We decided to offer feedback after the session to encourage reflection. While it is possible to analyze the data in real-time, there are risks related to computational overload, which can negatively impact the video conferencing experience.

4.1.1 Browser Application. The video chat platform is developed using open-source WebRTC and a Node.js-based web socket server to enable information exchange between multiple client applications. To establish the connection without interruptions by security settings, we incorporate a Traversal Using Relays around NAT (TURN) server. This enables the system to connect and exchange messages even over a highly-secured network. Fig 1(a) shows our video conferencing browser with four participants. Although our system is customized to hold four participants at a time, this number can be varied without loss of functionality.

4.1.2 Remote Server. Through this platform, we extract the conference data, which are then sent to a remote server for processing. This remote server is hosted on the cloud computing platform Microsoft Azure. As part of the back-end platform, this server holds the audio and video data, analyzes and stores the extracted information, and manages the connected users via web socket communication. Through an open-source, cross-platform database, MongoDB, we handle requests for information from the Virtual Feedback Assistant.

4.1.3 File Transfer. To transfer audio and video data from the web platform to the processing platform, we chose to send it together, not by frame. There is a tradeoff between processing and upload overheads. Keeping the file together allows us to handle the data without any processing overhead for chunking and combining individual file slices. At the same time, this requires more file transfer time from the conference platforms to the server, increasing the upload time overhead. The conference experience terminates before the file transfer begins, however, so this upload overhead does not affect the user experience.

4.1.4 Data Processing. The data processing takes place on the Azure server, where the data is stored using a pipeline of custom Python scripts. Fig 4 explains the processing and feature-extraction, phase by phase. The particular feature criteria are defined in Section 3. We process the audial and visual data to extract the following features:

- Praat-analyzed features: participation, overlap, turn-taking
- Affdex-analyzed features: valence, attitude, shared smile

4.2 CoCo Virtual Feedback Assistant

Our feedback system is a chat-based virtual feedback assistant, designed to hold a conversation with participants in a simple and unbiased manner. Designing this interface was challenging due to the number of features analyzed and the necessity of being unbiased.

4.2.1 Performance Graphs. The performance graphs are shown on the right side of the feedback assistant interface. For each feature, we construct an intuitive visual, all six of which can be seen in Fig 5. The designs for these graphics were determined through several trials and user evaluations.

For shared smiles, we use four-tiered emoticons to depict how often a user shares smiles with each of their group members. If the user shares no smiles, they see an emoticon with a neutral face. If the user shares one to four smiles, the emoticon's smile has slight upturned corners. Five to six smiles results in a steeper smile curve, and seven or more places the participants in the highest tier, with an even steeper smile curve. For participation,

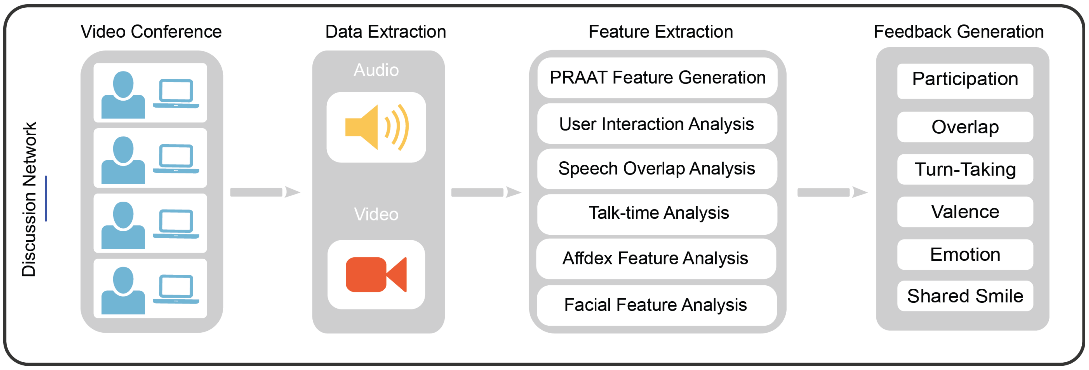


Fig. 4. Our system's data-extraction and -processing procedure.

quantified as a percentage, we used a pie chart to make effective comparisons against a whole. For turn-taking, we sought a visual representation of "You talked after Participant X 30 percent of the time." To achieve this, we have the user's icon positioned on the left, with arrows gesturing toward other users to whom the participant spoke. Users also have the option of clicking on the right-side bubbles to see turn-taking for other users.

For overlaps, we use two images, the top one indicating how often group members interrupt the user and the bottom indicating how often the user interrupts others. The user is represented by the single-person icon on the left, the group members by the the multi-person icon on the right. We chose to combine overlaps from group members into one number to prevent individual identification. For attitude toward group members, we displayed the relevant values on a bar graph. Each value is represented on a scale of 0 to 100. Valence is represented as a graded bar, where the position of the tick mark indicates whether the user's valence was negative, neutral, or positive.

4.2.2 Chatbot. Feedback can be delivered both live and in a post-conversation exchange. The methods of delivery are manifold: for example, animation, graphs, and text (chat-based) are all viable methods of offering feeback. We chose chat-based, post-conversation feedback. Our feedback system concentrates on being informative, not instructive. Instead of telling participant what to do, our system seeks to describe the possible choices the participant can make. This requires paragraphs of text—a significant distraction during a conversation, but suitable for post-conversation feedback. Our feedback system also seeks to be assistive—that is, as helpful as possible. Showing performance graphs, for example, is a useful way of presenting analytics to a user. But to understand the graph, the user must first understand what the graphed scores mean. The chat-based feedback explains the graphs to the user. Finally, our design seeks to be interactive and personable. Explanatory paragraphs meet the aforementioned needs, but may fail to engage users. A chat-based feedback system that employs both visual and textual feedback delivery succeeds here, too.

The virtual feedback assistant "talks" to the user, defining the extracted features, explaining both the performance scores the overall interpretation of the conversation. When it provides a performance score, the feedback assistant first lets the participant know which feature the score describes, then displays the graph and explains the performance metrics.

At the same time, the feedback assistant gives participants the ability to control the flow of the conversation by offering options to move the conversation forward. For example, when the participation graph is displayed, the assistant provides a brief introduction: "Expressing your ideas contributes to group decision-making. Would you like to know more about this?" After describing a concept, the assistant gives participants the option to

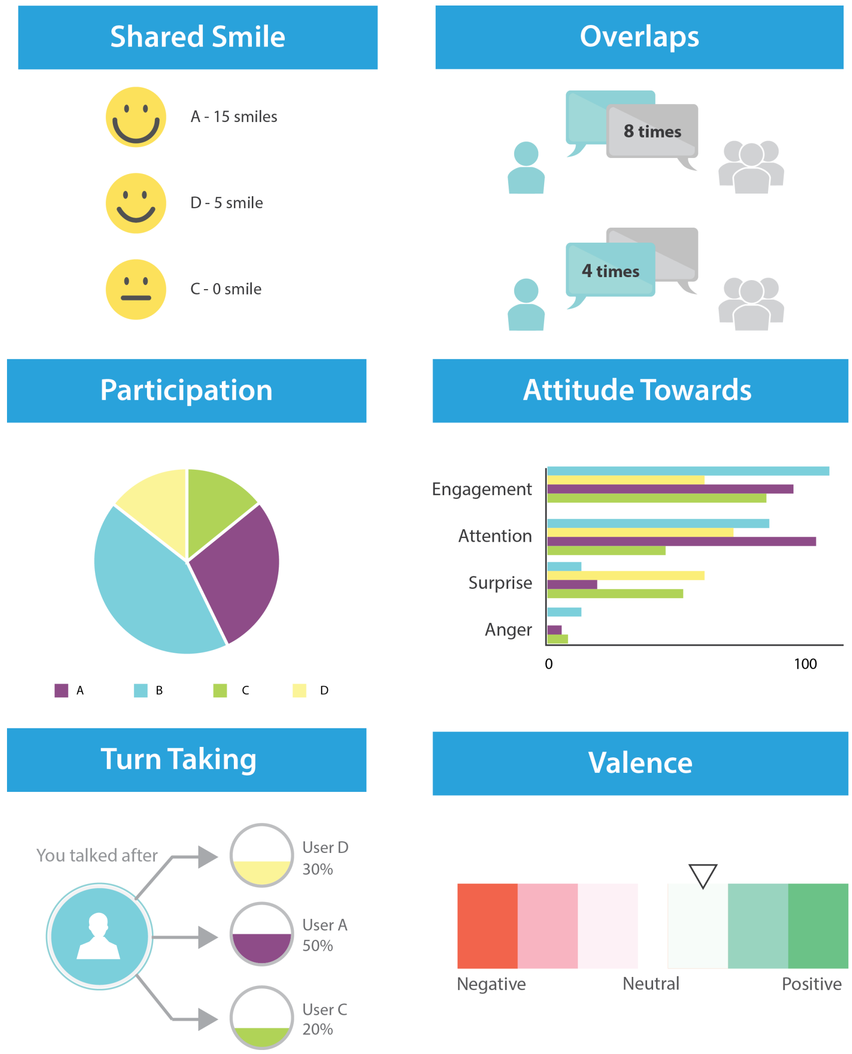


Fig. 5. Performance graph for each feature in the feedback from the perspective of participant B.

learn more about the concept or proceed to another feature. The conversation moves on if the participant selects "Got it." If the user instead chooses to learn more, the assistant provides some information about that feature and describes how it can affect team dynamics. During the conversation, the assistant displays a thinking icon, as if someone is typing a text message before replying. It does this to imitate human conversation and prevent uncanny immediate replies.

When describing a feature, the assistant does not direct the user to take a particular action. Because the interpretation of a feature depends on the conversation topic, the group's environment, the personalities of the participants, and many other factors, suggesting that participants modify their behavior in a certain way is not suitable for all possible scenarios. Subsequently, our feedback assistant explains the performance score in a

Fig. 6. Participant flow for session one

descriptive, not normative, manner. It explains both sides of a score—above and below average, for instance—and explains how the range can affect group discussions.

5 EVALUATION

5.1 Study Design

Study Setup

To better study the effectiveness of our CoCo framework, we conducted an experiment with groups of individuals in our lab. Each group's participants were placed in different rooms stocked with preset laptops and network configurations. This ensured that each participant had a similar experience.

Participants

The study included 39 participants, recruited from the University of Rochester through fliers advertising the experiment. Each of the participants was between 18 and 23 years old, and all were undergraduate students. We kept track of which participants within a group, if any, were acquainted with one another. In two cases out of ten experiments, none of the participants knew one another prior to the study. In one case, two of three participants knew one another, and in another, three of four participants had met before. In the remaining six cases, two of four participants were acquainted. We conducted the study over several weeks.

Procedure

Of the 10 groups, nine had four participants and one had only three, resulting in 39 total participants. The participants were randomly assigned to one of the groups. Each group participated in two conversation sessions, each with a different game. In any one of the sessions, participants received scenario (*Lost at Sea* or else *Survival on the Moon*). Both required that participants individually rank a list of items in order of most helpful to least for survival. The lists were specific to each scenario. For the individual ranking portion, we allowed a maximum of seven minutes. Participants then discussed the scenario over video, deciding as a group on the top five items. For the group discussion portion, we allowed a maximum of 20 minutes. Upon submitting the final group decision, participants were surveyed to evaluate the quality of the discussion and their performances (see Table 1). We also asked them to rank their teammates. After completing the survey, they were redirected to the virtual feedback assistant interface that provided their performance evaluations (Fig 5). Participants were later asked to complete another survey regarding the feedback assistant (see Table 2). For both surveys, the participants were asked questions with answers on a five-point Likert scale, where five suggested strong agreement and one suggested strong disagreement. We term this complete sequence, from the individual item-ranking to the feedback assistant survey, as a single session. The flow of session one can be seen in Fig 6. Participants then moved on to the second session, again using our CoCo system. Unlike the first, there was no feedback after the second session. Accordingly, participants only received the first survey, listed in Table 1. In general, both sessions took 1.5 hours. The only participant incentive we provided was a $20 Amazon gift card.

To address any effects of ordering the tasks, we counterbalanced the order in some groups against the rest. Groups 1, 2, 3, 8, 9, and 10 completed the *Lost in Sea* task in the first session and *Survival on the Moon* in the second. The order was reversed for groups 4, 5, 6, and 7.

Table 1. Discussion survey questions.

1	I was satisfied with the group decision.
2	I was satisfied with my performance in the group decision-making.
3	I was aware of my conversational skills during the discussion.
4	I was aware of my teammates' conversational skills during the discussion.
5	The discussion was productive.
6	The discussion was enjoyable.
7	The group had a positive vibe.
8	I had a positive attitude.
9	I could have spoken less.
10	I could have spoken more.
11	The group often cut off my speech.
12	I would like to have discussions with this group in future.

Table 2. Feedback survey questions.

1	The feedback on participation was easy to understand.
2	The feedback on turn-taking was easy to understand.
3	The feedback on valence was easy to understand.
4	The feedback on attitude was easy to understand.
5	The feedback on shared smiles was easy to understand.
6	The feedback on participation was accurate.
7	The feedback on turn-taking was accurate.
8	The feedback on valence was accurate.
9	The feedback on attitude was accurate.
10	The feedback on shared smiles was accurate.
11	The graphs were easy to understand.
12	The feedback was relevant.
13	The feedback chatbot was engaging.

Discussion Topic

To spontaneously engage participants in discussion, we provided them with survival game scenarios in which they had to individually rank 15 items and then, as a group, rank the top five. This allowed us to compare their selections with standardized lists developed by specialists, which we established as group performance standards. The first of two games we selected for the two sessions was *Lost at Sea* [1], with survival items including a shaving mirror (ranked first), a gasoline mixture (ranked second), a container of water (ranked third), chocolate bars (ranked sixth), nylon rope (ranked eigth), a transistor radio (ranked twelfth), among others. The second game, *Survival on the Moon* [24], included items like tanks of oxygen (ranked first), food concentrate (ranked fourth), parachute silk (ranked eigth), a box of matches (ranked fifteenth), among others. In both cases, the instructions were the same, which requested that participants individually rank 15 items (with one being the highest rank) within seven minutes and video conference with other participants to rank the top five items within 20 minutes.

The scenarios for each game were as follows:

Lost at Sea Situation: *You and your group members have been shipwrecked and are stranded in a life boat. You have a box of matches and a number of items salvaged from the sinking ship. Now, you have to determine which items are the most important for your group's survival.*

Survival on the Moon Situation: *You are a member of a space crew scheduled to rendezvous with the mothership on the lighted surface of the moon. However, due to mechanical difficulties, your own ship was forced to land at a spot 200 miles from the rendezvous point. During re-entry and landing, much of the equipment aboard was damaged and, since survival depends on reaching the mothership, the most critical items available must be chosen for the 200-mile trip.*

5.2 Results

In this section, we compared the performances of participants in the first and second sessions. We looked at the self-reported survey results, analyzed participants' affective features, and presented case studies. In this section, we used t-test to identify any change in behavior, ratings, and affective features. In order to test the normality of data, we have used D'Agostino and Pearson's [17, 18] normality test which tests a null hypothesis that samples are coming from a normal distribution. We performed the t-test on those features where the null hypothesis was accepted ($\alpha = 0.05$). In future, we will investigate features which fails to have a normal distribution using non-parametric tests.

Feedback Survey Results

After each feedback review session, in which participants received feedback from our CoCo system, participants were asked to fill out the feedback survey (shown in Table 2). Fig 7 shows the average ratings of the questions. All the ratings were above three points, on a scale of one to five. A single sample t-test yielded a Bonferroni-corrected p-value of <0.01, which was a good indication of the usability of our system. Among all the ratings, Questions 1 ("The feedback on participation was easy to understand") and 11 ("The graphs were easy to understand") received the highest ratings (mean = 4.6, *SD* = 0.69; and mean = 4.48, *SD* = 0.57, respectively). The feature accuracy–related questions—Q6, Q7, Q8, Q9, and Q10 (concerning whether the feedback on participation, turn-taking, valence, attitude, and shared smiles was accurate, respectively)—received a relatively lower score. Only two questions, however—Q9 ("The feedback on attitude was accurate") and Q10 ("The feedback on shared smiles was accurate")—received average ratings lower than four. The shared smiles and attitude features were the hardest to determine, which may have affected their accuracy.

Discussion Survey Results

After each video conferencing session, participants were asked to fill out the discussion survey, the first of the two surveys (see Table1). We then averaged the ratings for each question within each group for sessions one and two. The average ratings for each question are shown in Fig. 8. To analyze the changes between the sessions, we conducted a paired t-test. We only observed a significant change between the two sessions on Q3 ("I was aware of my conversational skills during the discussion") ($T(40) = -1.88, p < 0.05, [M_{session1} = 4.21, SD_{session1} = 0.83], [M_{session2} = 4.46, SD_{session2} = 0.53]$), Q4 ("I was aware of my teammates' conversational skills during the discussion") ($T(40) = -2.42, p < 0.05, [M_{session1} = 4.04, SD_{session1} = 0.84], [M_{session2} = 4.31, SD_{session2} = 0.8]$), and Q9 ("I could have spoken less") ($T(40) = -4.34, p < 0.05, [M_{session1} = 2.56, SD_{session1} = 0.97], [M_{session2} = 3.36, SD_{session2} = 0.91]$). Because the responses were ordinal, we also performed a Mann-Whitney U test. We found that only responses to Q9 were significantly different ($z - score = -3.24, p = 0.0006$). The responses to Q4 changed only marginally ($s - score = -1.40, p = 0.08$). As we looked for changes in 12 different questions, we used a Bonferroni correction on the p-values, multiplied by 12. This indicated that participants self-reported becoming more aware of their and their groups' conversational skills. This could also mean that there was a priming effect from the initial feedback on their nonverbal cues. The ratings for Q9 indicated that participants realized they should allow others to speak.

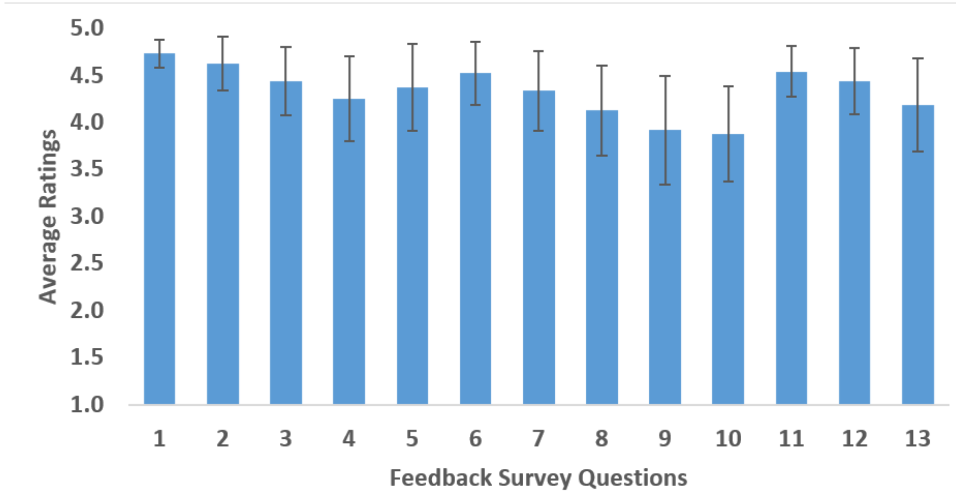


Fig. 7. Average ratings and standard deviation of the feedback survey. The questions are listed in Table 2

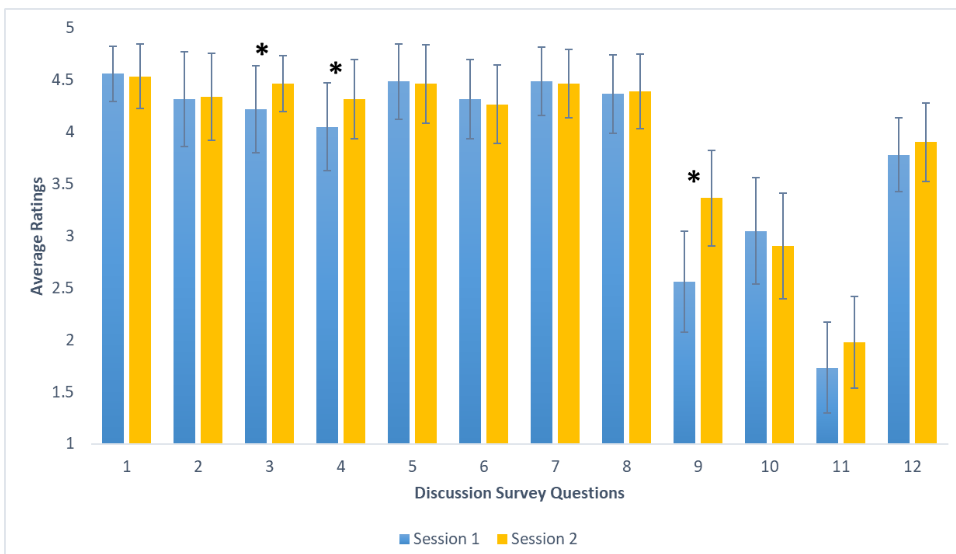


Fig. 8. Average ratings and standard deviation of the discussion survey. The questions are listed in Table 1

Impact on Communication Behavior

We further analyzed the features of the participants and changes therein between the two sessions. We recognized that the duration of any ice-breaking discussion could affect the results of the statistical tests. To avoid this, we considered only the duration of actual discussion of the tasks. We excluded any time participants spent getting acquainted with one another and figuring out how to use the system.

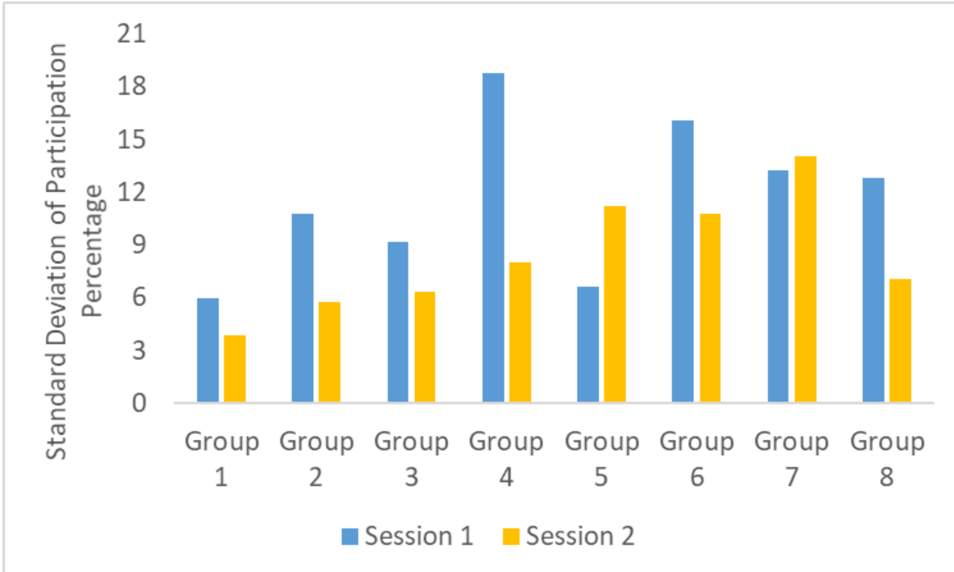
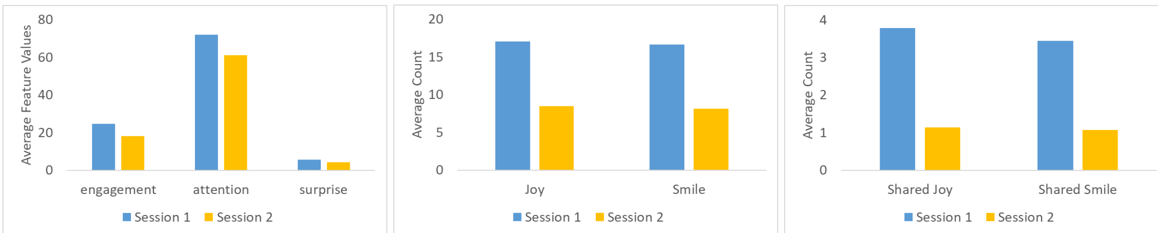


Fig. 9. Standard deviation of participation percentage for eight group discussions.

We found that participants moved toward equal participation in the second discussion, after receiving feedback from the first. By performing a paired t-test, we found this change significant ($T(7) = 2.02, p < 0.05$). Fig 9 shows the standard deviation of the participation in sessions one and two. We had to drop two groups from this analysis due to bad audio quality to ensure that participation was measured correctly. Our virtual feedback system displayed only a participation graph, demonstrating no preference for any particular style. We observed, however, that participants were trying to speak more equally in the second session. In the first four groups, we saw a trend of decreasing standard deviations in participation (that is, an increase in equal participation). We considered the possibility that this might be due to the difficulty of the discussion topics—for example, discussing *Lost at Sea* might be more intense than *Survival on the Moon*. To control against the influence of task difficulty, we reversed the order of the tasks. After reversing, we observed that only two groups did not tend toward equal participation. The overall change was still significant. Also of note is that in the discussion survey, participants rated Q9 ("I could have spoken less") higher in the second session, despite the fact that participation was more balanced in the second session. Next, we looked at the attitude features of the participants in both sessions. These features include engagement, attention, attitude toward others, sadness, disgust, smirk, anger, surprise, fear, valence, and contempt (see Section 3). More specifically, we were interested in the features that changed significantly between the two sessions. Fig. 10(a) shows the average value of these features (uncorrected $p < 0.05$). However, the significance is not visible after a Bonferroni correction for engagement, attention, and surprise. We had no prior hypothesis for this analysis. The uncorrected p-values might suggest a hypothesis for future exploration. Interestingly, the features associated with negative emotions did not show any significant change. For example, there were no noticeable changes in anger, fear, or disgust. One implication is that these expressions are more extreme and therefore less likely to appear in a team task environment. Features like engagement, attention, and surprise decreased in the second session significantly, despite the fact that the system did not recommend the participant behave in a certain way (smile more, for instance). It should be noted that while participants in the



(a) Average of attitude features.

(b) Average count of single features.

(c) Average count of shared features.

Fig. 10. Changes in features.

second session gained awareness of their nonverbal behavior, they did not increase their positive emotions. One possible explanation is that being more aware of nonverbal behavior negatively affected the ways in which they would otherwise express themselves.

We then investigated the single and shared features of the participants. More specifically, we looked at the *smile, joy, shared smile,* and *shared joy* features. We discovered a significant decrease in these features in the second session (T(83)=3.65, p<0.01) after Bonferroni correction. These were consistent with our previous result: positive emotions decreased. Despite these changes, participants showed no difference in the discussion survey's Q1 ("I was satisfied with the group decision") and Q2 ("I was satisfied with my performance in the group decision-making").

5.3 Case Studies

In this section, we discuss the individual study groups and the corresponding results. We present two case studies discussing two categories: Case Study 1, which followed our general findings, and Case Study 2, which did not. Upon observing the qualitative analysis scores (shown in Fig. 9), the groups that underwent the most drastic changes in their behaviors from each category became representative groups. Afterward, the system analyzed quantitatively the audio and video data. For Case Studies 1 and 2, the representative groups are Group 4 and Group 5, respectively. For privacy reasons, we randomly assigned participants a letter ID: A, B, C, and D. Fig 11 presents a comparative analysis of the participation features for these two groups during session-1 and 2.

5.3.1 ***Case Study 1.*** Group 4 is a representative case study for the results of the evaluation shown in Fig 9. In general, the standard deviation of the participation percentage feature decreased after feedback. (Six of eight groups exhibited this.). For Group 4, this value went from SD = 18.80 to SD = 7.97, implying that participants tended toward speaking equally. The average ratings of the discussion survey in Fig 8 further support the idea that participants became more aware of their own and others' conversational skills.

Group 4 had three male and one female members. They discussed the *Survival on the Moon* scenario first, in which participation scores for Participants A, B, C, and D were 55 percent, 12 percent, 26 percent, and 7 percent, respectively (shown in Fig 11). Participants A and C mostly spoke back-to-back, advancing the conversation for the group, even though Participants A and B knew each other beforehand. Participant B mostly used back-channeling responses (e.g. "true" and "that's right") without adding much to the discussion.

Participant D found it tough to get into the conversation and stayed silent for half of the entire first session. They started to get involved afterward.

The first session shows the effects of being unaware of conversational skills. For example, speaking more can provide both useful and useless information. Participants A and C talked more and sometimes deviated from the agenda. Participant B spoke, but contributed little to the final decision. Participant D barely spoke at all. The following conversation snippet shows how the group got sidetracked:

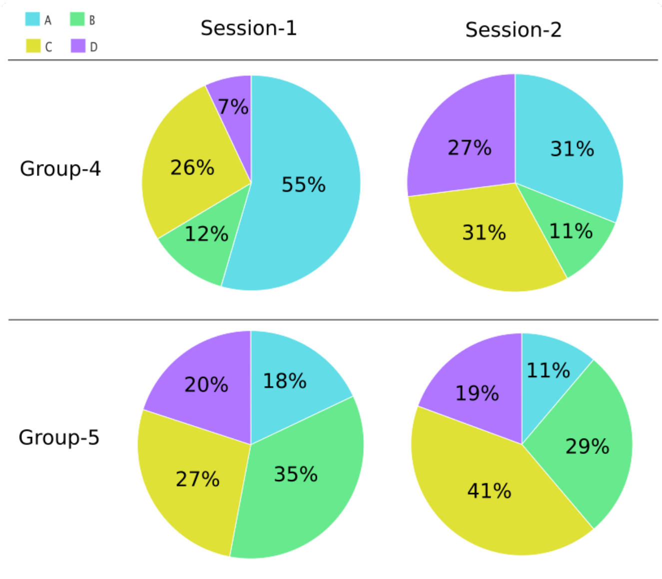


Fig. 11. Change of participation score for Groups 4 and 5 in sessions one and two. In the case of Group 4, an imbalance in participation in session one balanced out in session two, as the participants became more aware and focused after feedback. For Group 5, more balanced participation became imbalanced in the next session because of some out-of-the-box reasoning.

A: Why would you need a gun? (Laughs.)
C: It might be for one of those moon bears. (Laughs.)
B: That's true. (Laughs.)
C: It can be vicious! (Laughing continues.)
A: You realize the only time you will be using it is on blowing up your own allies' oxygen tanks or shooting them, right?
B: That's true. (Laughing continues.)
C: Or you could shoot yourself and wouldn't have to go through the hassle getting to the base.
(The dialogue continues like this for some time. Participant D remained silent.)
A: They (moon bears) don't exist. (Laughing continues.)
C: Okay, maybe let's move on. (Laughs.)

Our survey shows that all participants except one thought they could have spoken more. Notably, our virtual feedback assistant emphasizes both listening and speaking, as well as expressing ideas that contribute to the discussion. After the first session, the participants individually received the following feedback on participation:

Virtual Feedback Assistant [VFA]: First, let's talk about participation. It reflects how much time you spoke during the whole conversation. [Option: Got it.]
Participant: Got it.
VFA: [Shows participation graph.] As you can see from the graph, you participated x% of the whole session. Expressing your ideas contributes to the group decision. Would you like to know more about it? [Options: Sure; No Thanks.]

Participant: Sure.
VFA: Team management experts have found that equal participation is healthy for a group. Speaking about your own opinion can make the discussion stronger, whereas listening to teammates' viewpoints can remove your confusion.

After feedback, in the second session (discussing *Lost at Sea*), the participation scores for Participants A, B, C, and D changed to 31 percent, 11 percent, 31 percent and 27 percent, respectively (Fig 11). From the conversation, it is evident that participants were more focused in the second session. The *joy* feature from Affdex supports this, showing that the score had changed for Participants A (from 4.8 to 4.4), B (from 4.5 to 2.6), C (from 3.2 to 2.5), and D (from 0 to 4.4).

5.3.2 ***Case Study 2.*** The significance of the case study for Group 5 is that the standard deviation of the participation percentage feature increased from $SD = 6.65$ to $SD = 11$ in the second session. This is notable in that in almost all cases, the standard deviation for that feature decreased, meaning that groups generally tended toward equal participation afterward. This group was an exception.

Group 5 consisted of four male participants discussing on *Survival on the Moon* first and *Lost at Sea* second. Participants B and C knew one another beforehand. As shown in Fig 11, in the first session the participation performances for Participants A, B, C, and D were 18 percent, 35 percent, 27 percent, and 20 percent, respectively. Participant D spoke mostly to initiate the discussion on an item or mention his item without much followup. Participant A briefly explained his reasoning if one of his items came up and asked questions in cases of disagreement. Participants B and C generally explained in detail their own reasons for ranking an item in a particular way. All these characteristics built up the participation performances accordingly. The following is a snippet of the conversation:

A: How about signal flares? I think those would be helpful if there are people nearby and they might be looking for us.
B: I'm not sure. It depends on whether or not they would burn without oxygen [C and D agree, saying "Yeah" multiple times]. But I think most flares do glow without oxidizing, though. So that might be useful. Matches are definitely not going to work.
D: Mmm, yeah, nope, matches. I'm not sure about the compass because... [pauses].
B: I'm pretty sure the compass is useless, too. [Laughs.]
C: The moon doesn't have a magnetic field, at least not strong enough to use a compass.

Interestingly, there was even more imbalance in the participation performances in the second session, even though each participant mostly maintained his corresponding characteristics. The participation values for Participants A, B, C, and D were 11 percent, 29 percent, 41 percent, and 19 percent, respectively. This shows an increase of 14 percent in Participant C's performance. Participant C had an out-of-the-box idea to use a combination of three items, supporting it with a detail explanation. This went as follows:

C: So if you have a water container, bottle of rum, and the waterproof sheet, then you can fill the water container with salt water, put the bottle of rum inside at the center of it, and you can cover the water container with the waterproof sheet. And as the water evaporates, it goes up onto the waterproof sheet and then drips down into the bottle of rum.

Impressed, the group decided to put those items on the final top-five list, leaving space for only two items. Participant C proposed more items by supporting his own list, leading the overall discussion. Thus, with the unique idea, Participant C emerged as a natural leader and greatly influenced the discussion. This phenomenon supports our approach of providing feedback on participation by mentioning the importance of both speaking and listening. Our reasoning behind constituting the feedback in this way is that imbalance in participation affects individuals differently. Speaking more can add to the conversation, but others may or may not like what is beind said. This possible variation is supported by the survey, in which the participants were asked which session they preferred:

Comment 1: "Session 1 because there was more general agreement."

Comment 2: "I enjoyed 2 better, because there was more discussion involved."

6 DISCUSSION AND FUTURE WORK

6.1 Observations

We observed some changes in the affective features of the participants across the discussion sessions. Specifically, the *engagement, attention, surprise, joy,* and *smile* features decreased significantly in the second session. Notably, these features are all associated with positive emotions. From observation, we believe that participants became more aware and focused during the second session. It is also worth noting that participants indicated they became more aware of their own nonverbal behavior. This suggests that becoming more aware of one's own conversational performance does not necessarily mean that they will express it more often. This change indicates that as participants were getting acquainted with one another, they did not feel it necessary to communicate as emotionally (for example, fewer polite smiles signaling acknowledgement in the second session). This might also indicate that participants became more goal-oriented and focused on finding the correct items. This could explain the decrease in positive emotion. Indeed, this was echoed by one participant:

"It did not pick up as much on my emotional range, I feel it could be just myself being tired, or maybe the system, I'm not sure."

Additionally, we observed that the features associated with negative emotions (e.g. *anger, disgust*) did not change between the two sessions. These features are more spontaneous and less likely to be displayed in a group discussion with strangers, possibly explaining the lack of change. We would need more groups and participants in order to conclusively say anything about these behavioral changes, or lack thereof.

In the discussion survey, participants agreed more with the statement "I should have spoken less" in the second session. This might be true under the current game settings, which encourage equal participation for a successful outcome. In other scenarios, like learning in a classroom, the opposite might be true. Surprisingly, by just showing the participation percentage, the system was able to make participants consider how often they allowed others to speak. This indicates an effective point at which technology can intervene, allowing individuals to self-correct their behavior. The virtual feedback assistant not only provided the participation graph (in the form of a pie chart), but also talked about balance in speaking and listening. Some aspect of the feedback may have primed participants to change their own behavior.

6.2 Limitations

Our pilot study included only 39 participants spread over 10 groups. More conclusivity about the efficacy of our system in improving team dynamics requires experimentation with many more groups. Separately, all participants were undergrad students, which represent only a small portion of those who have group discussions every day. A larger dataset derived from people of different ages, occupations, etc. may provide more insight into group discussion dynamics. As our framework is rapidly deployable, our immediate future work involves recruiting a gender- and age-diverse set of online workers.

Meetings can have a variety of objectives—for example, problem-solving, information-sharing, or ad hoc meetings are common. Subsequently, the results of the team-building exercise we chose may not fully represent those of different types of meetings. The difficulties of the two team-building exercises likewise may not have been the same. We changed the order of the discussion topics to minimize any effect of the topics' varied difficulties. This needs to be explored in greater detail with other team-building exercises.

Although the survey questions are designed conventionally [37], they could have been more extensive. In the future, in extending our work, we intend to adapt the System Usability Scale (SUS) and Networked Minds Questionnaire of Social Presence into a more carefully-designed questionnaire.

Prior acquaintance among participants may have had an effect on the interactions. This may have affected the time needed for ice-breaking discussion during a conversation. We collected information about which participants in a group knew one another beforehand. Most groups had a similar number of acquaintances, so we expect the impact, if any at all, to be similar. However, the impact of prior acquaintance on discussion can neither be nullified nor predicted. During statistical analysis, we excluded the ice-breaking period for each group to obtain the actual discussion period. One participant, hinting at the difference between ice-breaking and task-focused discussion, said,

> *"Kind of awkward at first but then it was productive when we actually started discussing the topic."*

We used the Affdex SDK [4] to extract affective features for its ease of use and extensive API. However, the Affdex SDK is a closed system. We control neither the measure nor the detection of the affective features. In the future, we plan to apply more open-source software (for instance, OpenFace [5]) to detecting and measuring these features, which will allow us to both compare and justify the extracted feature metrics.

6.3 Future Work: Application and Building on the System

Our CoCo system opens up new research opportunities for the ubiquitous computing. The research community can follow-up on our design choices and feedback deliveries in their own experiments. The system code being public can also help the community investigate different types of interactions in different settings. In our current setting, we did not analyze the content of the group discussion. We understand that there can be many types of discussions, some of which may raise privacy concern. The current implementation of CoCo only collects nonverbal features from the audio and video data. It is possible, however, to include a feature that could, with users' consent, analyze the contents of discussions to generate context-relevant feedback. Under such a system, for example, CoCo may analyze how well participants stick to the meeting agenda. We also plan to personalize the feedback, depending on the type of discussion. For example, it is important during a business meeting that a team shares its ideas, keeps track of its progress, and meets its goals. In contrast, discussion in a learning environment does not require equal participation, due largely to the nature of the teacher-pupil relationship. Keeping track of the progress is also highly subjective. These differences should be considered when designing a customizable feedback assistant.

Our system may help in designing new intervention strategies in the context of assistive technology. This system may be modified for individuals with Asperger's Syndrome or certain social phobias who often find it difficult to assert themselves in group settings. If a particular group member does not participate often due to a socially-restrictive condition, other members may be respectfully reminded to be more accommodating and supportive. CoCo can also offer insights into the roles of gender and racial diversity in team environments. In particular, we can examine previously published research where it has been argued that group intelligence has relatively little to do with individual intelligence and teams with more women demonstrate greater collective intelligence [59]. With our online deployable framework, we can potentially identify the composition that makes a group productive.The findings could be applied in the context of online learning, where a computer facilitator could engage every student equally. In addition to standardized test scores, computers can also create objective analytics on students' ability to ask questions, participate effectively and maintain a positive attitude. Recognizing students for their effort may positively impact their scores.

Recently, research on developing cognitive assistants has advanced immensely. *CALO* [39], which eventually evolved into the popular *Siri* [2], is designed with functionalities that include mediating human communications and task management, among other things. Virtual assistants like Google Home, Amazon's Alexa, Jibo Robot, and

Apple's Homepod are gradually evolving to log users' activity and provide useful information. The knowledge generated by this project and the future work to immediately follow from the research community could help existing virtual assistants be more useful and interactive.

For researchers, CoCo can also help in understanding, evaluating, and developing better approaches for delivering feedback. Our feedback assistant provides information about every feature by explaining its importance. It also elaborates on the different possibilities of that feature's use in conversation. Whether this explanatory approach is more useful than providing a user with unexplained and direct feedback remains an open and interesting research question. Future research can build up on the CoCo framework to further investigate.

The appropriateness of highly specific recommendations can depend on the context of a conversation. Interruption, for example, can sometimes be a good thing. In the virtual feedback assistant, we tried to provide unbiased recommendations. This allows users freedom to interpret the feedback as best suits them. Because our system has been developed as an online system, we plan to recruit online workers as participants to more rapidly validate the CoCo framework. Collection of a large dataset can enable the automatic generation of personalized feedback by applying machine-learning techniques. Some participants indicated that the assistant's feedback is inconsistent with what is shown on the graphs. Having more data would help us develop a more personalized and effective virtual feedback assistant.

With respect to the statistical measures that our system displays to its users, we would like to further refine these measures in the future and validate their accuracy and usefulness to participants. For example, the shared smile threshold value of 75, with an overlap duration of one second, was chosen empirically. In order to identify robust parameters, we require human-annotated data, which is currently unavailable. We have collected 39 valid data points from the pilot study, which we can annotate to further refine our system's feedback. This will also provide us with a value that means something to a user.

We have not explored the possibility of using real-time feedback in the CoCo system. In the discussions, we observed that there is one member in each group who takes charge and emerges as the mediator. This is especially true in real-life discussion, where whoever possesses a stronger personality or holds more knowledge about the topic at hand leads the group. It is worthwhile to explore the possibility of a computer playing the role of mediator, steering the conversation by providing feedback in real-time. This might be helpful for specific types of meetings in which participants tend to deviate from the topic often.

7 CONCLUSION

In this paper, we developed a system called CoCo that provides feedback on team dynamics through a web-based video conferencing system. The source code of our system is publicly accessible for other researchers to further expand its functionality and conduct new experiments. We evaluated the usability of the feedback our system provided and its effect on team behavior. We found that the feedback induces significant change in participation rates. We also observed that the feedback affected positive emotional features in the discussion sessions. User survey results revealed that the system made participants more aware of their conversational skills. Additionally, the feedback allowed participants to realize that they could do a better job of letting others speak. This behavioral change indicates that CoCo has the potential to make an impact on team dynamics in a video conferencing setting. Video conferencing is now a widespread utility, popping up in our daily lives for purposes as diverse as business meetings, job candidate interviews, online education, and staying in touch with friends. A system like CoCo can offer valuable insights into how people interact, allowing us to connect more effectively with one another.

8 ACKNOWLEDGMENTS

This work was supported by National Science Foundation Award IIS-1464162, a Google Faculty Research Award, and Microsoft Azure for Research grant. The authors would like to acknowledge the anonymous reviewers of IMWUT-UbiComp for their insights, comments and suggestions.

REFERENCES

[1] 1996. Lost at Sea. (1996). http://insight.typepad.co.uk/lost_at_sea.pdf

[2] 2011. Siri. (2011). https://en.wikipedia.org/wiki/Siri

[3] 2015. *Video Conferencing Trends 2016*. http://resources.idgenterprise.com/original/AST-0162982_Video_Conferencing_Trends_of_2016.pdf

[4] 2016. Affectiva. http://www.affectiva.com/. (2016). Accessed: 2016-10-30.

[5] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical Report. Technical report, CMU-CS-16-118, CMU School of Computer Science.

[6] Patricia B. Barger and Alicia A. Grandey. 2006. Service with a Smile and Encounter Satisfaction: Emotional Contagion and Appraisal Mechanisms. *The Academy of Management Journal* 49, 6 (2006), 1229–1238. http://www.jstor.org/stable/20159829

[7] Sigal G. Barsade. 2002. The Ripple Effect: Emotional Contagion and Its Influence on Group Behavior. *Administrative Science Quarterly* 47, 4 (2002), 644–675. http://www.jstor.org/stable/3094912

[8] JÃĬrÃĬmy Bauchet, HÃĬlÃĬne Pigot, Sylvain Giroux, Dany Lussier-Desrochers, Yves Lachapelle, and Mounir Mokhtari. [n. d.]. Designing judicious interactions for cognitive assistance: the acts of assistance approach. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 11–18. https://doi.org/10.1145/1639642.1639647

[9] Paul Boersma and David Weenink. 2016. Praat: doing phonetics by computer. http://www.praat.org/. (2016). Accessed: 2016-06-01.

[10] LouAnne E. Boyd, Alejandro Rangel, Helen Tomimbang, Andrea Conejo-Toledo, Kanika Patel, Monica Tentori, and Gillian R. Hayes. 2016. SayWAT: Augmenting Face-to-Face Conversations for Adults with Autism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4872–4883. https://doi.org/10.1145/2858036.2858215

[11] Erin Bradner, Gloria Mark, and Tammie D. Hertel. 2005. Team size and technology fit: Participation, awareness, and rapport in distributed teams. *IEEE Transactions on Professional Communication* 48, 1 (2005), 68–77. https://doi.org/10.1109/TPC.2005.843299

[12] T. Prevett. C. Wickens. 1995. Exploring the dimensions of egocentricity in aircraft navigation displays: influences on local guidance and global situation awareness. Journal of Experimental Psychology: Applied, Neuilly-Sur-Seine, France.

[13] Dan Calacci, Oren Lederman, David Shrier, and Alex Sandy Pentland. 2016. Breakout: An Open Measurement and Intervention Tool for Distributed Peer Learning Groups. *CoRR* abs/1607.01443 (2016). http://arxiv.org/abs/1607.01443

[14] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI'05)*. Springer-Verlag, Berlin, Heidelberg, 28–39. https://doi.org/10.1007/11677482_3

[15] Samuele Centorrino, Elodie Djemai, Astrid Hopfensitz, Manfred Milinski, and Paul Seabright. 2015. Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution and Human Behavior* 36, 1 (2015), 8 – 16. https://doi.org/10.1016/j.evolhumbehav.2014.08.001

[16] K. M. Cohen. 1982. Speaker interaction: Video teleconference versus face-to-face meetings *(Teleconferencing and Electronic Communications)*. ACM, University of Wisconsin.

[17] RALPH D'AGOSTINO and Egon S Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b 2 andâĹŽ b. *Biometrika* 60, 3 (1973), 613–622.

[18] Ralph B d'Agostino. 1971. An omnibus test of normality for moderate and large size samples. *Biometrika* 58, 2 (1971), 341–348.

[19] Ionut Damian, Chiew Seng (Sean) Tan, Tobias Baur, Johannes SchÃűning, Kris Luyten, and Elisabeth AndrÃĨ. [n. d.]. Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 565–574. https://doi.org/10.1145/2702123.2702314

[20] Richard W. DeVaul, Alex "Sandy" Pentland, and Vicka R. Corey. 2003. The Memory Glasses: Subliminal vs. Overt Memory Support with Imperfect Information. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC '03)*. IEEE Computer Society, Washington, DC, USA, 146–. http://dl.acm.org/citation.cfm?id=946249.946907

[21] J. J. Diehl and R. Paul. 2009. The assessment and treatment of prosodic disorders and neurological theories of prosody. *Int J Speech Lang Pathol* 11, 4 (2009), 287–92. https://doi.org/10.1080/17549500902971887

[22] Joan Morris DiMicco, Anna Pandolfo, and Walter Bender. 2004. Influencing group participation with a shared display. *Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04* (2004), 614–623. https://doi.org/10.1145/1031607.1031713

[23] Norah E. Dunbar and Judee K. Burgoon. 2005. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships* 22, 2 (2005), 207–233. https://doi.org/10.1177/0265407505050944 arXiv:http://dx.doi.org/10.1177/0265407505050944

[24] Jay Hall and Wilfred Harvey Watson. 1970. Survival on the moon. *Human relations* (1970). https://www.psychologicalscience.org/observer/nasa-exercise

[25] Laura A. Hambley, Thomas A. O'Neill, and Theresa J B Kline. 2007. Virtual team leadership: The effects of leadership style and communication medium on team interaction styles and outcomes. *Organizational Behavior and Human Decision Processes* 103, 1 (2007),

1–20. https://doi.org/10.1016/j.obhdp.2006.09.004

[26] Crystal L. Hoyt and Jim Blascovich. 2003. Transformational and Transactional Leadership in Virtual and Physical Environments. *Small Group Research* 34, 6 (2003), 678–715. https://doi.org/10.1177/1046496403257527 arXiv:http://dx.doi.org/10.1177/1046496403257527

[27] H. Hung, Yan Huang, G. Friedland, and D. Gatica-Perez. 2011. Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *Trans. Audio, Speech and Lang. Proc.* 19, 4 (May 2011), 847–860. https://doi.org/10.1109/TASL.2010.2066267

[28] Inseok HWANG, Chungkuk YOO, Chanyou HWANG, Dongsun YIM, Youngki LEE, Singapore Management University, Chulhong MIN, Hjohn KIM, and Junehwa SONG. 2014. TalkBetter: Family-driven Mobile Intervention Care for Children with Language Delay. (2014), 1283. https://doi.org/10.1145/2531602.2531668

[29] Jeroen Janssen, Gijsbert Erkens, and Gellof Kanselaar. 2007. Visualization of Agreement and Discussion Processes During Computer-supported Collaborative Learning. *Comput. Hum. Behav.* 23, 3 (May 2007), 1105–1125. https://doi.org/10.1016/j.chb.2006.10.005

[30] Jeroen Janssen, Gijsbert Erkens, Gellof Kanselaar, and Jos Jaspers. 2007. Visualization of Participation: Does It Contribute to Successful Computer-supported Collaborative Learning? *Comput. Educ.* 49, 4 (Dec. 2007), 1037–1065. https://doi.org/10.1016/j.compedu.2006.01.004

[31] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling Dominance in Group Conversations Using Nonverbal Activity Cues. *Trans. Audio, Speech and Lang. Proc.* 17, 3 (March 2009), 501–513. https://doi.org/10.1109/TASL.2008.2008238

[32] Vaiva Kalnikaitė, Patrick Ehlen, and Steve Whittaker. 2012. Markup As You Talk: Establishing Effective Memory Cues While Still Contributing to a Meeting. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 349–358. https://doi.org/10.1145/2145204.2145260

[33] Taemie Kim, Pamela Hinds, and Alex Pentland. 2012. Awareness As an Antidote to Distance: Making Distributed Groups Cooperative and Consistent. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1237–1246. https://doi.org/10.1145/2145204.2145391

[34] G. Ross. Lawford. 2003. Beyond success: Achieving synergy in teamwork. *The Journal for Quality and Participation* 26, 3 (Fall 2003), 23–27. https://search.proquest.com/docview/219147502?accountid=13567

[35] Yale Child Jane L. Mcsweeny Ami Klin Donald J. Cohen Fred R. Volkmar Lawrence D. Shriberg, Rhea Paul. [n. d.]. Speech and prosody characteristics of adolescents and adults with high functioning autism and Asperger syndrome. In *Journal of Speech, Language, and Hearing Research*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.385.7116

[36] Harold J. Leavitt. 1951. Some effects of certain communication patterns on group performance. *Journal of abnormal psychology* (1951), 38–50.

[37] Gilly Leshed. 2009. *Automated language-based feedback for teamwork behaviors*. Ph.D. Dissertation. https://search.proquest.com/docview/304872230?accountid=13567 Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2016-06-03.

[38] Gilly Leshed, Diego Perez, Jeffrey T. Hancock, Dan Cosley, Jeremy Birnholtz, Soyoung Lee, Poppy L. McLeod, and Geri Gay. 2009. Visualizing Real-time Language-based Feedback on Teamwork Behavior in Computer-mediated Groups. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), 537–546. https://doi.org/10.1145/1518701.1518784

[39] Bill Mark and R Perrault. 2005. Calo: Cognitive assistant that learns and organizes. (2005).

[40] Poppy Lauretta McLeod, Jeffrey K. Liker, and Sharon A. Lobel. 1992. Process Feedback in Task Groups: An Application of Goal Setting. *The Journal of Applied Behavioral Science* 28, 1 (1992), 15–41. https://doi.org/10.1177/0021886392281003 arXiv:http://dx.doi.org/10.1177/0021886392281003

[41] Janice Nadler. 2003. Rapport in negotiation and conflict resolution. *Marq. L. Rev.* 285, 1990 (2003), 875–882. https://doi.org/10.1525/sp.2007.54.1.23.

[42] Eyal Ofek, Shamsi T. Iqbal, and Karin Strauss. 2013. Reducing Disruption from Subtle Information Delivery During a Conversation: Mode and Bandwidth Investigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 3111–3120. https://doi.org/10.1145/2470654.2466425

[43] Jeffery R. Ringer. 1985. Pardon Me, Can I Talk Now?: A Look at the Roles of Interruptions in Conversation. *Annual Meeting of the Organization for the Study of Communication, Language, and Gneder* (1985), 1–29.

[44] Marcello Russo. 2012. Diversity in goal orientation, team performance, and internal team environment. *Equality, Diversity and Inclusion: An International Journal* 31, 2 (2012), 124–143. https://doi.org/10.1108/02610151211202781 arXiv:http://dx.doi.org/10.1108/02610151211202781

[45] Harvey Sacks, Emanuel a Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. (1974), 696–735 pages. https://doi.org/10.2307/412243

[46] M. Satyanarayanan. 2004. From the Editor in Chief: Augmenting Cognition. *IEEE Pervasive Computing* 3, 2 (2004), 4–5. https://doi.org/10.1109/MPRV.2004.1316809

[47] Charles R. Scherl and Jay Haley. 2000. Computer Monitor Supervision: A Clinical Note. *The American Journal of Family Therapy* 28, 3 (2000), 275–282. https://doi.org/10.1080/01926180050081702 arXiv:http://dx.doi.org/10.1080/01926180050081702

[48] Mark Schittekatte and Alain Van Hiel. 1996. Effects of Partially Shared Information and Awareness of Unshared Information on Information Sampling. *Small Group Research* 27, 3 (1996), 431–449. https://doi.org/10.1177/1046496496273006 arXiv:http://dx.doi.org/10.1177/1046496496273006

[49] Abigail J. Sellen. 1992. Speech Patterns in Video-mediated Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. ACM, New York, NY, USA, 49–59. https://doi.org/10.1145/142750.142756

[50] D. Tanen. 1995. The Power of Talk: Who Gets Heard and Why. *Harvard Business Review* (1995).

[51] Deborah Tannen. 1997. Gender and Discourse. *Discourse* (1997), 548–567. https://doi.org/10.2307/2655317

[52] M. Iftekhar Tanveer, Emy Lin, and Mohammed (Ehsan) Hoque. 2015. Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 286–295. https://doi.org/10.1145/2678025.2701386

[53] Yla R. Tausczik and James W. Pennebaker. 2013. Improving teamwork using real-time language feedback. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (2013), 459–468. https://doi.org/10.1145/2470654.2470720

[54] Monica Tentori and Gillian R. Hayes. [n. d.]. Designing for interaction immediacy to enhance social skills of children with autism. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 51–60. https://doi.org/10.1145/1864349.1864359

[55] Linda Tickle-Degnen and Robert Rosenthal. 1990. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry* 1, 4 (1990), 285–293. http://www.jstor.org/stable/1449345

[56] Janet van der Linden, Rose Johnson, Jon Bird, Yvonne Rogers, and Erwin Schoonderwaldt. 2011. Buzzing to Play: Lessons Learned from an in the Wild Study of Real-time Vibrotactile Feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 533–542. https://doi.org/10.1145/1978942.1979017

[57] Joseph B. Walther and Lisa C. Tidwell. 1995. Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *J. Org. Computing* 5, 4 (1995), 355–378. http://dblp.uni-trier.de/db/journals/jocec/jocec5.html

[58] C. Wickens. 1996. Situation awareness: impact of automation and display technology. In NATO AGARD Aerospace Medical Panel Symposium on Situation Awareness.

[59] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (2010), 686–688. https://doi.org/10.1126/science.1193147 arXiv:http://science.sciencemag.org/content/330/6004/686.full.pdf

[60] Qianli Xu, Michal Mukawa, Liyuan Li, Joo Hwee Lim, Cheston Tan, Shue Ching Chia, Tian Gan, and Bappaditya Mandal. [n. d.]. Exploring users' attitudes towards social interaction assistance on Google Glass. In *Proceedings of the 6th Augmented Human International Conference*. ACM, 9–12. https://doi.org/10.1145/2735711.2735831

Received May 2017; revised August 2017; accepted October 2017