Multimedia Systems
DOI 10.1007/s00530-011-0243-z

REGULAR PAPER

Privacy-sensitive recognition of group conversational context with sociometers

Dineshbabu Jayagopi · Taemie Kim · Alex Pentland · Daniel Gatica-Perez

Abstract Recognizing the conversational context in which group interactions unfold has applications in machines that support collaborative work and perform automatic social inference using contextual knowledge. This paper addresses the task of discriminating one conversational context from another, specifically brainstorming from decision-making interactions, using easily computable nonverbal behavioral cues. Privacy-sensitive mobile sociometers are used to record the interaction data. We hypothesize that the difference in the conversational dynamics between brainstorming and decision-making discussions is significant and measurable using speaking activity-based nonverbal cues. We characterize the communication patterns of the entire group by the aggregation (both temporal and person-wise) of their nonverbal behavior. The results on our interaction data set show that the floor-occupation patterns in a brainstorming interaction are different from a decision-making interaction, and our method can obtain a classification accuracy as high as 87.5%.

D. Jayagopi (✉) · D. Gatica-Perez
Idiap Research Institute, Martigny, Switzerland
e-mail: djaya@idiap.ch

D. Gatica-Perez
e-mail: gatica@idiap.ch

T. Kim · A. Pentland
MIT Media Lab, Cambridge, USA
e-mail: taemie@media.mit.edu

A. Pentland
e-mail: sandy@media.mit.edu

D. Gatica-Perez
Ecole Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Keywords Conversational context · Nonverbal behavior · Brainstorming · Decision-making · CSCW · Social inference machines

1 Introduction

Supporting group interactions, both in real time and offline, requires understanding the conversational context reliably and quickly. People routinely perform such social inference in their everyday life, often unconsciously and automatically [20]. With the advent of new sensing and modeling technologies, inferring the conversational context of an interaction—that could potentially include the goal of the group, the type of the interaction (task-oriented vs. casual), and the type of relationship (close friends vs. strangers), etc—could potentially be performed without human intervention.

The automatic recognition of group interaction context is a useful module for Computer-Supported Cooperative Work (CSCW) [19]. With the advent of ubiquitous and mobile sensing platforms, novel ways of collecting and visualizing group interaction behavior have been explored with the primary objective of influencing the group's behavior [8, 26]. Such applications would greatly benefit from the knowledge of the interaction context, i.e., awareness of the interaction type, like a cooperative versus competitive interaction, or a brainstorming versus decision-making phase.

Both supporting interacting groups offline and building social inference machines would also benefit from knowing about the group interaction context [41]. Various social factors related to individual attributes (e.g. personality, hierarchy, roles); relationships among individuals (close friends vs. strangers); and goal at hand (cooperation vs.

Published online: 08 July 2011

competition) have begun to be studied in ubiquitous environments, mostly in indoor environments equipped with microphones, cameras, and other sensors [16]. The joint modeling of nonverbal behavior and social constructs could be facilitated by quickly inferring the conversational context and then employing the models learnt in specific contexts. Apart from infrastructure-based sensors, the availability of privacy-sensitive, mobile platforms to sense conversations is opening the possibility of recording and analyzing behavioral aspects of real-life interactions without breaching the privacy of people, through online extraction of nonverbal cues without recording or storing raw audio [5]. Privacy-sensitivity in our work follows the approach of Wyatt et al. [48] in the sense that the features stored do not reveal the speaker or the words spoken. Such a privacy-sensitive approach for recording interaction data allows studying the behavior of the group members without compromising their freedom of expression.

Within this domain, our work addresses the novel problem of discriminating two types of conversational context categories, namely brainstorming versus decision-making, using computationally simple nonverbal cues extracted from sociometers. Sociometers are wearable, social sensing badges [5]. Laughlin and Ellis [27] postulated that cooperative group tasks may be ordered on a continuum anchored by intellective and judgmental tasks. According to them, intellective tasks are defined as tasks for which a demonstrably correct solution exists, as opposed to decision making or "judgmental" tasks where "correctness" tends to be defined by the group consensus. A different body of research by Mc Grath [30] asserts that group interactions have different dynamics depending on the group's objective. A brainstorming session has a different objective as compared to that of a decision-making session, and therefore demands a different response from the group members as well. Our work investigates whether these differences can be captured through nonverbal behavioral cues automatically extracted from sociometers; and if so whether the interaction type can be inferred. With much of the work in modern societies becoming group-based, such interactions are indeed ubiquitous.

This paper addresses the task of classifying group conversational context, brainstorming versus decision-making. Our recordings included both collocated as well as distributed cases. We measure the performance of single and multiple cues. An off-the-shelf classifier is used to fuse multiple cues. The paper is an extended version of [24], and is organized as follows: Section 2 reviews related works. Section 3 discusses our approach. Section 4 introduces the experimental setup. Section 5 documents the results obtained. Section 6 discusses a few potential applications. Section 7 provides some conclusions.

2 Related work

Our work concerns human behavior analysis in small group interactions using privacy-sensitive sociometers. In this section, we review attempts to recognize and discover patterns of human behavior using mobile sensing devices. We also review works about modeling behavior using infrastructure-based sensors.

2.1 Human behavior analysis using mobile sensors

Mobile sensing devices can integrate many sensors including microphones, accelerometers, or GPS to name a few. A great deal of information about the people wearing them can be derived out of the collected data. This data can be integrated over a large population to understand large-scale human behavior patterns.

The work of Pentland and others has pioneered several wearable platforms to measure different aspects of social context. Choudhury and Pentland [5] created the first sociometer, a wearable sensor package designed to measure face-to-face interactions between people with an infrared (IR) transceiver, a microphone, and two accelerometers. The device was used to learn social interactions from sensor data and to model the structure and dynamics of social networks. The next generation of sociometers, developed by Gips [18], used two types of sensors: proximity sensors and motion sensors. Olguín et al.'s [34] version of the sociometer, used in this work, has a user-friendly design, is light-weight, and incorporates several sensors useful for capturing social signals. Section 3.1 describes the capabilities of this sociometric badge in detail.

Pentland and others have also used mobile phones for understanding human behavior, calling the process as 'reality-mining'. Eagle and Pentland [11] inferred location and activity with both short-range (Bluetooth) and long-range radio network data (GSM) on mobile phones. Madan et al. [29] developed VibeFone, a mobile software application that used location, proximity, and tone of voice to gain understanding of people's social lives by mining their face-to-face and phone interactions. This application augments traditional means of gathering social interaction data (surveys or ethnographic studies). The mobile phone platform is highly conducive to collect long-term continuous data.

This initial work has now expanded to address many interesting research issues about individual, dyadic, and collective behavior. Pentland showed that 'honest signals', which are robustly extractable nonverbal cues often transmitted and received automatically among social beings—predict job performance, negotiation outcomes, dating outcomes, etc. in dyadic relations [38]. These speech and physical activity cues were characterized in terms of

emphasis, activity, influence, and mimicry. While Olguín et al. [33] found these signals derived from sociometers to be correlated with team performance, Waber and Pentland [44] reported the correlation of these features with individual expertise. Recent work by Woolley et al. proposed a general collective intelligence factor in groups that explains a group's performance on a wide variety of tasks [46]. This work shows that this factor is not strongly correlated with the average or maximum individual intelligence of group members, but rather with the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group. In another line of work, individual and collective behavior of individuals in an organization was studied using sociometers, measuring face-to-face interaction time, physical proximity to others, and physical activity levels [34]. Wyatt et al. [48] automatically modeled the creation and evolution of a social network formed through real, spontaneous face-to-face conversations. The work used a Personal Digital Assistant (PDA) with an attached multisensor board containing eight different sensors. Miluzzo et al.'s [31] work explored mobile phone sensing platforms that also perform collaborative sensing and classification to reason about human behavior and context on mobile phones.

The above works either correlated human behavior with certain social constructs; or learned supervised models. Other existing works have found typical patterns by clustering human behavior. Indoor daily routines, like commuting and office work, were discovered by Huynh et al. [21] using wearable sensors and accelerometer data with applications in elderly care, or office space management. Outdoor daily routines were discovered by Eagle and Pentland [12] using Principal Component Analysis (PCA) from mobile phone data on subjects' location, proximity, communication, and device usage behavior. Farrahi et al. [13, 14] extended the work using topic models, employing location and proximity data. Candia et al. [3], used phone call data to study mean collective behavior of humans at large-scales. Such works has applications in understanding large-scale human mobility patterns and epidemiology.

2.2 Human behavior analysis using infrastructure-based sensors

Infrastructure-based sensors have been employed in two scenarios—surveillance and social. In surveillance settings, cameras are typically the only source of data recorded, whereas in social settings (e.g. a small group interaction recorded in modern meeting rooms) it is possible to record with multiple microphones and multiple cameras. The research on video-based surveillance has a long history [49] and will not be reviewed here. As our work concerns social behavior, we review in detail works that employ infrastructure-based sensors in social settings.

Regarding individual behavior modeling, attempts have been made to estimate dominant behavior, certain personality traits, and certain roles that individuals are involved in. Jayagopi et al. estimate dominant behavior by computing speaking turns-based features (like speaking time, turns, successful interruptions) along with motion turns and learning supervised models using Support Vector Machines (SVM) on meetings from the AMI (Augmented Multiparty Interaction) corpus [4, 23]. Pianesi et al. [39, 40] estimate personality traits, specifically extraversion (sociable, assertive, playful) versus intraversion (aloof, reserved, shy) using Support Vector Regression and applied to sequences of the MS (Mission Survival) Corpus. Using an influence model, Dong et al. [9] estimated functional roles in meetings related to tasks and socio-emotional roles on the MS Corpus. The work by Lepri et al. [28] estimated individual performance from interaction slices. The above three works employed speaking activity cues, prosodic cues, and visual fidgeting cues. Ad hoc roles were estimated using Dynamic Bayesian Networks (DBN) and turn-taking information in broadcast video by Vinciarelli [43]; and in AMI corpus by [15]. Recently, Sanchez-Cortes et al. [42] modeled emergent leadership using turn-taking patterns and employing score-level fusion techniques. Unlike the above works, Basu et al. [1] using a dynamic Bayesian approach, in an unsupervised approach, estimated pair-wise influence between participants in a group. The observations were speaking activity features, and influence was estimated using a variation of the coupled HMM (Hidden Markov Model) called the influence model.

Regarding group behavior modeling, Zhang et al. [50] and Dielmann et al. [6] characterized group activities employing layered sequential approaches (either HMM or DBN), where the first layer modeled the individuals' behavior, and the second layer the activity (monologue, presentations, or discussions). Otsuka et al. [35] described group activities in terms of conversational regimes (convergence or monologue, dyad-link and divergence). The work in [6, 50] employed speaking-activity and motion-activity in terms of blobs (region of image pixels) as the features, whereas in [35] speaking-activity and visual gaze were used. The latter work was also extended to estimate interpersonal influence, interactivity, and centrality [36]. Gatica-Perez et al. [17] investigated group interest by segmenting meetings temporally into high or neutral interest level segments in a HMM-based supervised framework, fusing audio-visual activity cues. Recently, Dong et al. [10] studied group discussion dynamics with two different corpora (in two languages) and the group performance was estimated using turn-taking patterns and honest signals. The work employed three types of supervised models—SVM,

HMMs, and the influence model. Wrede et al. [47] studied various prosody-related cues correlated with interest 'hot-spots', where the interest level of the meeting participants was perceived to be high. Jayagopi et al. [25] discriminated cooperative and competitive interactions using a small data set of interactions. Unlike the above works, in an unsupervised approach, Jayagopi et al. [22] discovered group interaction patterns resembling prototypical leadership styles in social psychology—autocratic, participative, and free-rein—using probabilistic topic models.

3 Our approach

We propose the following methodology to classify group conversational context types (Fig. 1). Assume that we have labelled group interaction data, where the interactions differ in their objectives and therefore have a different label. Our approach uses a layered approach for classification. In the first layer, the *individual* nonverbal behavior description is obtained by extracting speaking activity and then computing features which characterize the floor occupation patterns of individuals. In the second layer, *group* nonverbal behavior is inferred by either aggregating these features (e.g. 'how much this group talks per unit time') or by comparing the *individual* nonverbal behavior with others' behavior (e.g. 'does everyone take an equal number of turns or interruptions?'). The group conversational context is classified using a supervised learning approach, with the group behavioral cues as input. We note that while [25] attempted to classify cooperative versus competitive interactions, this work further classifies two common types of cooperative interactions in modern workplaces, i.e. brainstorming and decision-making. Also, our work can be generalized to cues other than audio cues, e.g. motion of the group using accelerometers or gaze and gesture of the group using infrastructure-based sensors.

We discuss the main blocks of our framework in the following subsections. Figure 2 shows explicitly the training and testing phases of our approach.

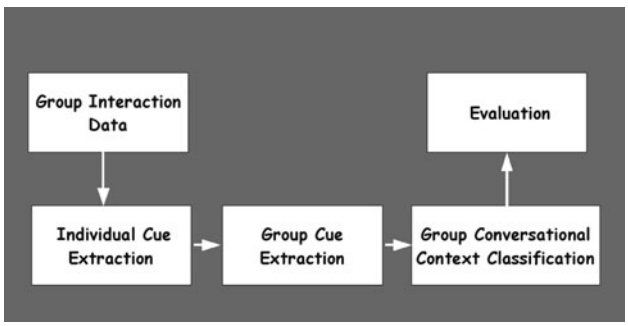

Fig. 1 Overview of our work

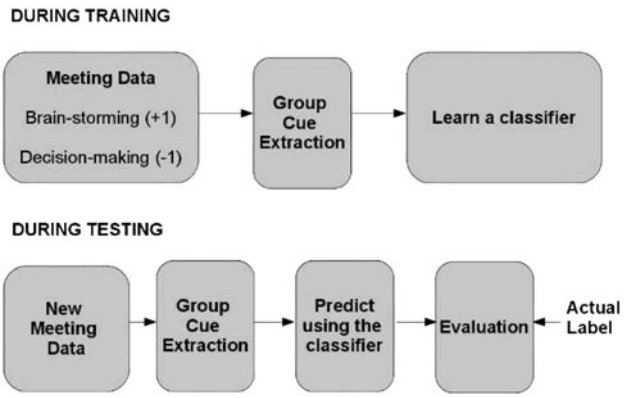

Fig. 2 Our approach. Nonverbal cues are extracted to learn and infer the conversational context

3.1 Meeting data set

The data set was collected from 24 groups of four members each, a total of 96 participants (with almost equal number of male and female participants and mean age of 28). Subjects were recruited on a US university campus and through public internet message boards and were given a standard monetary compensation for their time. Each participant wore a current generation sociometric badge—a wearable electronic badge with multiple sensors collecting interaction data (shown in Fig. 3). By interacting with other badges it can collect proximity data, other badges in direct line of sight, movement data, and speech features. The specific capabilities of this sociometric badge include, as described in [34]:

- Using a three-axis accelerometer combined with a mobile phone containing a second accelerometer, recognize common daily human activities (such as sitting, standing, walking, and running) in real time.

Fig. 3 Sociometer badge

- Using a microphone, extract speech features such as variation, pitch, tone, and volume in real time to capture nonlinguistic social signals such as interest and excitement. The microphone collects speech variation data sampled at 50 Hz, which is immediately processed on the badge so that only the processed data is saved on its SD card.
- Using 2.4 GHz radio transceiver GPS transceivers, communicate with other badges placed at fixed locations or compatible radio base stations, and transfer data.
- Using a Bluetooth module, communicate with Bluetooth-enabled cell phones, personal digital assistants, and other devices to study user behavior, detect people in close proximity.
- Using an IR sensor, capture face-to-face interaction time that can detect when two people wearing badges are facing each other within a 30° cone and 1 m distance.
- Using a microcontroller, perform indoor user localization by measuring received signal strength and using different triangulation algorithms.

Though the sociometer has a number of sensors, for the study of seated face-to-face interactions (as seen in Fig. 4) except the speech features, the data collected by other sensors do not provide information relevant to the task at hand. For example, Bluetooth proximity data and IR sensor data reveal trivial information. Also, as the participants hardly move from their seated location, the accelerometer data is also not useful. The speech features are privacy-sensitive, similar to the work by Wyatt et al. [48], as they capture nonlinguistic information that preserve information about conversation style and dynamics; but do not contain information that may identify the speaker or the spoken words.

The task given to subjects were based on a modification of the game "Twenty-Questions", replicating Wilson's [45] experiments. Each round consisted of two phases. In the first phase, each group was given a set of ten yes/no question-and-answer pairs. The groups were given 8 min to collaboratively brainstorm as many ideas that satisfy the set of question-and-answers. An example of the list of such pairs is shown in Table 1. In this example, the group would have to come up with objects that satisfy all ten question-and-answer pairs. In this example, a towel or a dish sponge would be a possible answer as it satisfies all ten criteria, however, a pot would not be a possible answer as it does not satisfy a few of the question-and-answer pairs such as number 1 and number 7. Groups tried to generate the most number of possible answers during the 8 min provided. We label these interactions as 'brainstorming'.

Fig. 4 Example of an interacting group wearing sociometric badges around the neck

Table 1 Twenty-Questions game

Questions and answers		
1	Is it shiny?	No
2	Do you hold it when you use it?	Yes
3	Can it fit in a shoebox?	Yes
4	Would you use it daily?	Yes
5	Is it flexible?	Yes
6	Is it decorative?	No
7	Does it open?	No
8	Is it found in the home?	Yes
9	Do you clean it regularly?	Yes
10	Is it organic?	No

Ten questions answered yes/no before brainstorming

Then in the second phase, groups were given 10 min to ask the remaining ten questions of the Twenty-Question game to determine the correct solution. For each question asked, the experimenter answered the group and then groups were given time to discuss to come up with their next question. Hence in the example shown in Table 1, groups might ask questions like "is it used in the kitchen?" to narrow down the answer. Based on the answer that they hear from the experimenter, they would generate the next question. As this problem-solving phase mainly involved the group making decisions about the subsequent questions, we regard and label them as 'decision-making' interactions. In the second phase groups were asked to select a leader among them who would be the question-asker who communicates with the experimenter.

Each team began with one practice round and then participated in two rounds where their behavior was measured: one round in collocated settings and the other round separated into pairs into two rooms. When distributed, the group members were not able to see each other but were able to have verbal communication. The sequence of collocated and distribution was counter-balanced to minimize learning effect. The group leader was chosen during the

practice round, and was kept consistent throughout the two measured rounds.

The data set we used for our experiments was 9.8 h of group conversational recordings. As the data set has the same group of people participating in both types of interaction, allowing an important dimension to be controlled in the discrimination study.

3.2 Individual nonverbal cue extraction

Among the speech features, we only use the speech variation data, sampled at 50 Hz for this work. Future work could employ other nonlinguistic features such as pitch, volume. In this work, we hypothesize that turn-taking patterns themselves have sufficient discriminatory power to classify brainstorming and decision-making interactions. For each of the participant the speech variation is thresholded to obtain the speaking status—a binary variable indicating speaking (1) and nonspeaking (0). As the signal-to-noise ratio was high, a fixed threshold was sufficient to detect the voice activity, with higher magnitudes representing speech and lower magnitudes representing non-speech. The speaking status was downsampled to a rate of 10 frames per second (fps). This rate is sufficient to analyze conversational behavior at the level of turns. Short conversational events, e.g. backchannels (sounds such as 'ya', 'aa-ha'), are of the order of 1-s duration.

From the speech segmentation, we compute Total Speaking Length [$TSL(i)$] defined as the total time that participant i speaks, Total Speaking Turns [$TST(i)$], Total Successful interruptions [$TSI(i)$], and Total Unsuccessful interruptions [$TUI(i)$] defined below:

- *Total Speaking Length* (*TSL*) This feature considers the total time that a person speaks according to their binary speaking status.
- *Total Speaking Turns* (*TST*) A speaking turn is the time interval for which a person's speaking status is active. The total number of speaker turns was accumulated over the entire meeting for each participant.
- *Total Successful Interruptions* (*TSI*) The feature is defined by the cumulative number of times that speaker $i \in \{1, 2, 3, 4\}$ starts talking while another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn before i does, i.e. only interruptions that are successful are counted. Though such a definition does not perfectly capture successful interruptions, nevertheless, it is a computationally efficient proxy.
- *Total Unsuccessful Interruptions* (*TUI*) The feature is defined by the cumulative number of times that while speaker $i \in \{1, 2, 3, 4\}$ is speaking, another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn before i does, i.e. only interruptions that are unsuccessful or 'potential' backchannels by another participant are counted.

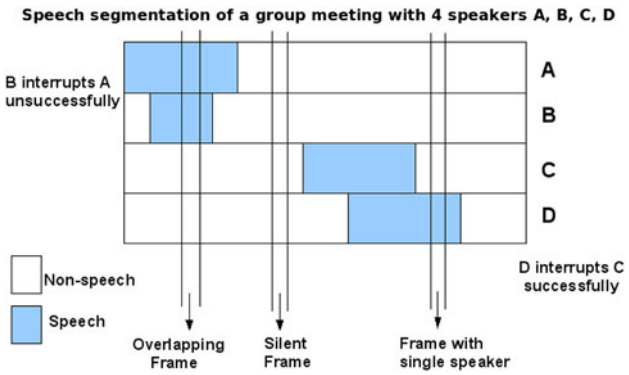

Fig. 5 Nonverbal cues extracted from speech segmentation

Figure 5 illustrates the individual nonverbal cues.

3.3 Group nonverbal cue extraction

Different groups differ in the way they speak. Some groups speak a lot. Some groups are silent. While some groups are more egalitarian either in nature or due to the performed task, some other groups have status differences leading to differences in the level of participation. Some groups could have lots of overlapped speech due to the nature of the participants or the social situation, while other groups do not. Our group cues capture these differences.

Three types of group cues are extracted. Figure 6 summarises the cue extraction process. A first set of cues characterize the participation rates of the group by accumulating it over the participants. Let D denote the duration of the meeting. We compute the following six cues—Group Speaking Length (GSL), Group Speaking Turns (GST), Group Successful Interruptions (GSI), Group Unsuccessful Interruptions (GUI), Group Successful Interruptions-to-Turns Ratio (GIT), Group Unsuccessful Interruptions-to-Turns Ratio (GUT)—from speaking length, turns, and interruptions of each of the participants:

$$\mathrm{GSL} = \frac{\sum_{i=1}^{P} \mathrm{TSL}(i)}{D} \tag{1}$$

$$\mathrm{GST} = \frac{\sum_{i=1}^{P} \mathrm{TST}(i)}{D} \tag{2}$$

$$\mathrm{GSI} = \frac{\sum_{i=1}^{P} \mathrm{TSI}(i)}{D} \tag{3}$$

$$\mathrm{GUI} = \frac{\sum_{i=1}^{P} \mathrm{TUI}(i)}{D} \tag{4}$$

$$\mathrm{GIT} = \frac{\sum_{i=1}^{P} \mathrm{TSI}(i)}{\sum_{i=1}^{P} \mathrm{TST}(i)} \tag{5}$$

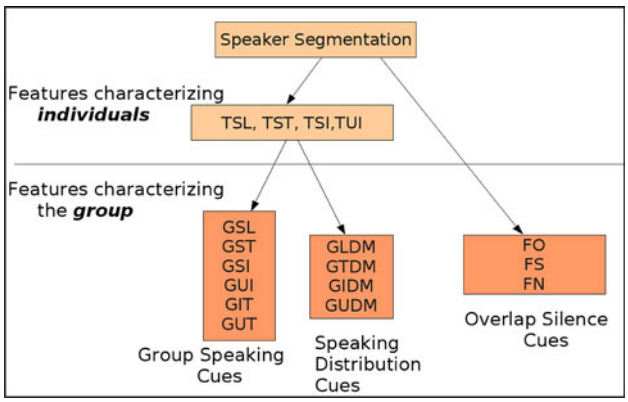

Fig. 6 Group cues extracted from individual cues

$$\mathrm{GUT} = \frac{\sum_{i=1}^{P} \mathrm{TUI}(i)}{\sum_{i=1}^{P} \mathrm{TST}(i)} \tag{6}$$

A second set of cues attempts to capture the overlap and silence patterns of a group as a whole. Let $T = D \times \mathrm{Fps}$ be the total number of frames in a meeting, S be the number of frames when no participant speaks, M be the number of frames when only one participant is speaking, and O be the number of frames when more than one participant talks. Fps being frames-per-second, the rate at which speaking status is available. Then we define the following three cues—Fraction of Silence (FS), Fraction of Nonoverlapped Speech (FN), Fraction of Overlapped Speech (FO)—defined as follows:

$$\mathrm{FS} = \frac{S}{T} \tag{7}$$

$$\mathrm{FN} = \frac{M}{T} \tag{8}$$

$$\mathrm{FO} = \frac{O}{T} \tag{9}$$

A third set of cues characterizes which meeting is more 'egalitarian' with respect to the use of the speaking floor, i.e. everyone gets equal opportunities. Let **TSL** denote the vector composed of P elements, whose elements are $\frac{\mathrm{TSL}(i)}{\sum_i \mathrm{TSL}(i)}$ for the ith participant. Employing an analogous notation for **TST**, **TSI**, and **TUI**, these vectors are first ranked ($\mathbf{p}$) and then compared with the uniform (i.e. "egalitarian") distribution, i.e. a vector of the same dimension with values equal to $\frac{1}{P}(\mathbf{q})$. The comparison is done using the Hellinger distance, a measure useful to compare probability distributions and bounded between 0 and 1. The Hellinger distance is defined in terms of the Bhattacharya coefficient as follows:

$$\mathrm{HD}(\mathbf{p}, \mathbf{q}) = \sqrt{1 - \mathrm{BC}(\mathbf{p}, \mathbf{q})} \tag{10}$$

where the Bhattacharya coefficient is defined below:

$$\mathrm{BC}(\mathbf{p}, \mathbf{q}) = \sum_i \sqrt{p(i) \times q(i)} \tag{11}$$

For our case Hellinger distance of 0 would correspond to a egalitarian meeting and closer to 1 corresponds to a one-man show. This results in four cues:

- Group Speaking Length Distribution Measure (GLDM)
- Group Speaking Turns Distribution Measure (GTDM)
- Group Successful Interruption Distribution Measure (GIDM)
- Group Unsuccessful Interruptions Distribution Measure (GUDM)

These group cues do not take into account individual contributions and so do not contain the identity of each person. Table 2 summarizes the group cues.

3.4 Group conversational context classification

We used two supervised models to classify the group conversational context type. The first is a Gaussian Naive-Bayes classifier [32], which assumes that the features are independent given the class, and that the conditional densities are univariate Gaussians. Let A and B denote the class labels. Also, let $f_{1:N} = (f_1, f_2, \ldots, f_N)$ denote the feature set and f_1, f_2, …, f_N the individual features. Then the log-likelihood ratio is given, by using Bayes' theorem and cancelling the common terms as follows:

$$\log\left(\frac{P(A|f_{1:N})}{P(B|f_{1:N})}\right) = \log\left(\prod_{k=1}^{N} \frac{P(f_k|A)}{P(f_k|B)}\right) + \log\left(\frac{P(A)}{P(B)}\right) \tag{12}$$

Table 2 Glossary of abbreviations for the group cues

Glossary of feature acronyms	
Group Speaking Length	GSL
Group Speaking Turns	GST
Group Successful Interruptions	GSI
Group Unsuccessful Interruptions	GUI
Group Successful Interruptions-to-Turns Ratio	GIT
Group Unsuccessful Interruptions-to-Turns Ratio	GUT
Fraction of Silence	FS
Fraction of Nonoverlapped Speech	FN
Fraction of Overlap	FO
Group Speaking Length Distribution Measure	GLDM
Group Speaking Turns Distribution Measure	GTDM
Group Successful Interruptions Distribution Measure	GIDM
Group Unsuccessful Interruptions Distribution Measure	GUDM

The probabilities $P(f_k|A)$ or $P(f_k|B)$ are estimated by fitting a Gaussian to the data from the respective class and the ratio of the priors are inferred from the data. When this ratio is greater than zero, the test data is assigned to class A. Otherwise to class B.

The second model is an SVM classifier, employing a linear kernel, using $(f_1, f_2, \ldots, f_N)$ as features [2].

4 Experiments

The speaking status was obtained by thresholding the speech variation data collected by the sociometer and then downsampling to Fps = 10 frames per second. As described in Sect. 3.1, we have 24 participant groups, solving two "Twenty-questions" games, one in collocated and the other in distributed settings. Each game involved a brainstorming phase followed by a decision-making phase. To model the difference between brainstorming and decision-making interactions, we define the following four data sets and three binary classification tasks.

1. Data set A: consists of 24 brainstorming meetings in collocated scenario.
2. Data set B: consists of 24 decision-making meetings in collocated scenario.
3. Data set C: consists of 24 brainstorming meetings in distributed scenario.
4. Data set D: consists of 24 decision-making meetings in distributed scenario.

Based on the data sets we define three classification tasks.

Task 1 The first task is to distinguish between brainstorming and decision-making meetings during the collocated setting. We classify Data set A versus Data set B. Each class has 24 datapoints.

Task 2 The second task is to distinguish between brainstorming and decision-making meetings during the distributed setting. We classify Data set C versus Data set D. Each class has 24 datapoints.

Task 3 The third task is to distinguish between brainstorming and decision-making meetings. We classify Data set A+C versus Data set B+D. Each class has 48 datapoints.

Group Adaptation Step To account for the feature variations among the 24 groups, we perform z-normalization on the group nonverbal cues before using it for classification as follows: $\hat{f^s} = (f^s - \mu_f)/(\sigma_f), \forall s \in A, B, C, D$ where $\hat{f}$ and f are the values of the feature in a particular scenario s before and after z-normalization, respectively.

In all cases, we use a leave-one-out approach for evaluation, to maximize the size of the training data for each data split.

5 Results

In this section, first we document the results on the above-mentioned three tasks using just single cues. This helps understand the discriminative power of each of the cues. Later, we combine the cues to see the complementarity between them.

Single cues Figure 7 shows for Task 1 (collocated setting). Random performance for all the tasks is 50%. Though we experimented with two different classifiers, as described in Sect. 3.4, we report the results using the Gaussian Naive Bayes classifier only as the results are similar when a linear SVM is employed. FS, GSL, and GUI were the top performing cues with a performance of 81.3%, 81.3%, and 79.1%, respectively. Fig. 8 shows the performance of the group cues for Task 2 (distributed setting). FS, Fraction of Overlap (FO), and GSL were the top performing cues with an accuracy of 79.2%. For Task 3, a similar trend was observed. FS, GSL, and FO gave the best classification result with an accuracy of 80.2, 78.1, and 74% (Fig. 9). All these results are statistically significant compared to the random performance at 5% level using a standard binomial test.

The results suggest that some of the investigated features indeed have discriminating power. Also, it is interesting to observe the following trend: Most groups have higher FS during brainstorming; and higher GSL and FO while making decisions. A possible reason may be that during brainstorming, groups tend to have higher cognitive load and hence speak less as compared to decision-making interactions. Figure 10 shows the normalized histogram of two best perfoming cues in collocated setting—FS and GSL—in both brainstorming and decision-making scenarios.

Multiple cues Later, we also combined the cues to investigate if there is complementarity among them. Figure

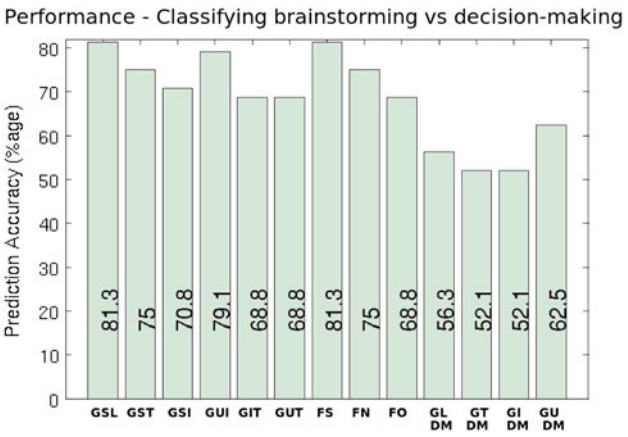

Fig. 7 Performance of the group cues on classifying the brainstorming and decision-making meetings during collocated setting (Task 1)

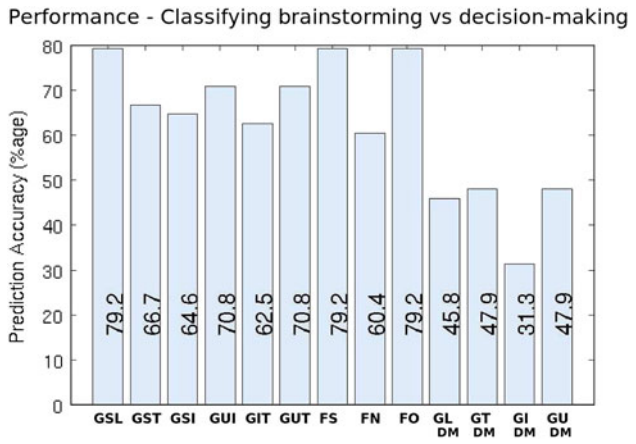

Fig. 8 Performance of the group cues on classifying the brainstorming and decision-making meetings during distributed setting (Task 2)

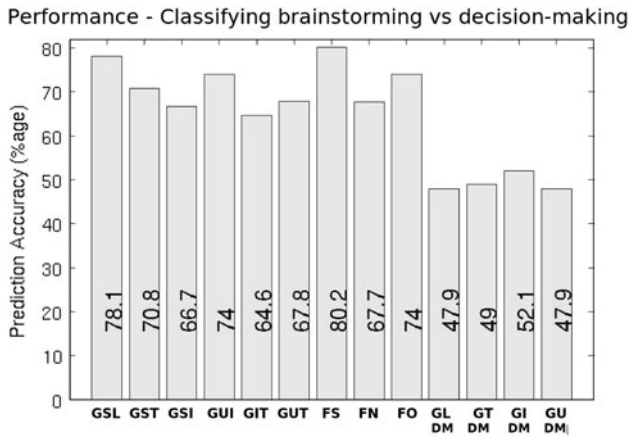

Fig. 9 Performance of the group cues on classifying the brainstorming and decision-making meetings (Task 3)

11 shows the classification performance of some combinations using the Gaussian-Naive Bayes classifier for each of the three tasks. The combination of FS and FO improves the classification accuracy to 83.3% in the collocated case (Task 1). Figure 12 illustrates the classification in this joint space. When GSL, GST, and GU) were added the accuracy improved to 87.5%. The combination of FS and GSL improves the classification accuracy to 81.3% in the distributed setting (Task 2). For the combined data set (Task 3), the combination of FS, GSL, and GUI improved the classification accuracy to 81.3%.

To conclude, we could discriminate these interactions with an accuracy of up to 87.5 and 81.3% in the collocated and distributed setting, respectively. The group adaptation i.e z-normalization step helps in improving performance and also tackling inter-group differences (as the mean behavior is subtracted out). As compared to the work in [25], where cooperative and competitive interactions were discriminated, this work improved on three aspects of the

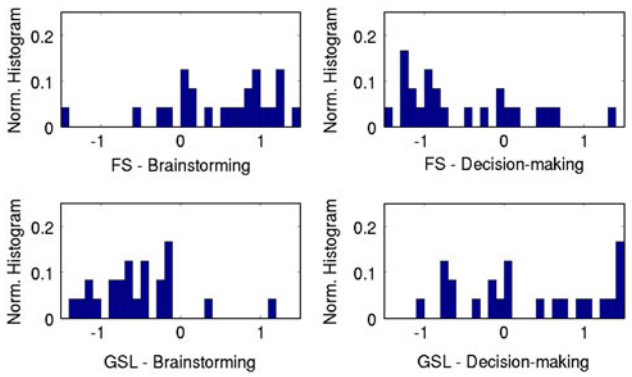

Fig. 10 Normalized histogram of two features in collocated setting: Fraction of Silence (FS) and Group Speaking Length (GSL)—in brainstorming and decision-making scenarios. Brainstorming interactions have higher Fraction of Silence and lower Group Speaking Length for most groups

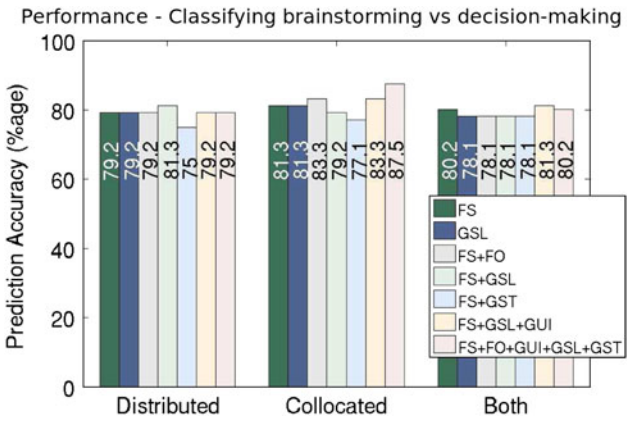

Fig. 11 Performance of combination of group features on classifying the brainstorming and decision-making meetings

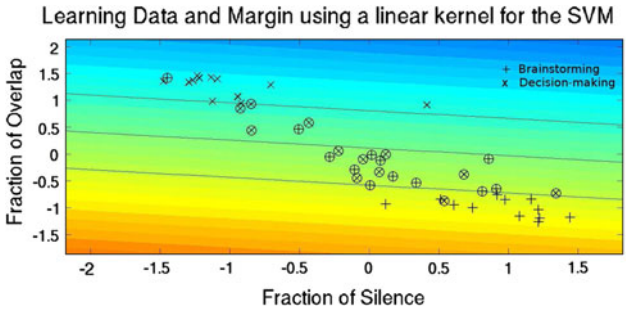

Fig. 12 Illustration: shows that in the joint space of Fraction of Silence (FS) and Fraction of Overlap (FO), brainstorming and decision-making interactions in the collocated scenario can be classified better using a linear kernel SVM

experimental design. First, the work used a larger data set. Second, two scenarios—collocated and distributed were explored. This shows that the results are indeed robust across conditions. Third, the same group of people interact

in both the conditions to be discriminated—brainstorming and decision-making, resulting in the control of one important variable of the experiments, i.e. the variation due to the individuals in the group.

6 Potential applications

In this section, we discuss briefly about two potential applications for the techniques presented to model group conversational context and better understand groups.

1. Behavior-based support of individuals and groups: Social inference machines could be part of relevant applications including self-assessment, training, and educational tools [37], and of systems to support offline [41] and online group collaboration [7]. As the expectations on individuals and groups are different in different contexts, automatic perception estimators could benefit from the knowledge of the type of interaction. For example, critically commenting about ideas is an expected behavior in the decision-making interactions but not in brainstorming interactions.
2. Behavior-based media retrieval: Research on performance of groups has shown that both individual and group behavior affects group performance. For example, the results in [26] showed that dominance, an individual behavior, had an interesting effect on performance: having a dominant person in the group had a significant negative effect on brainstorming, i.e. groups with dominant people tended to generate fewer ideas. Also, a group behavior like equal distribution of turns increases the collective intelligence of the group and therefore improves the performance of the group on a variety of tasks including brainstorming and decision-making [46]. From a human resource perspective, analyzing group behavior of interaction corpuses could signal the need for a team-building exercise or a leadership change to improve group performance. Jointly analyzing group conversational context and individual behavior could retrieve the potential individuals and groups for a behavioral training exercise.

7 Conclusion

In this work we investigated the problem of characterizing group conversational context using nonverbal turn taking behavior. Specifically, we presented a supervised learning approach that works at two layers, with the first layer capturing individual behavior and the second layer capturing group behavior. We apply our framework to classify brainstorming versus decision-making interactions. Our methods produce an accuracy of up to 87% in the collocated case, which is encouraging and suggests that the characterization of entire groups by the aggregation (both temporal and person-wise) of their nonverbal behavior is promising. The most effective features for classifying brainstorming versus decision-making interactions were FS, FO, and GSL.

As the size of the data set is relatively modest, many of the observed performance differences between the best cues are not statistically significant at 5% level although the difference between the best cues and random performance is statistically significant. Our work shows the promise of characterizing group behavior using just an instance of brainstorming and decision-making interaction. Further studies need to be done with varied and larger data sets to understand the generality of the results. This represents in itself a challenge as collecting such data is an expensive task which involves mobilization of participants, as traditionally done in social psychology.

Future work should use an expanded feature set to include prosodic cues and temporal aspects of cues as well as explore generative models that might characterize brainstorming and decision-making interactions more accurately. Second, other group conversational contexts, could also be interesting to study. As more of these contexts are studied and understood, an online detection of group interaction context in real situations would also be a possibility in the future. Third, building social inference machines which consider the conversational context could be explored. A certain behavior like 'judging a team member' in a brainstorming scenario would be perceived differently as compared to a decision-making scenario.

Finally, investigating the group behavior of high-performing groups in both brainstorming and decision-making scenarios could be an interesting study. The effect of an individual behavior like dominance on the group performance is an open question, though initial research has shown that dominance affects performance in brainstorming groups, i.e. groups having a dominant person in the group had a significant negative effect on brainstorming resulting in the generation of fewer ideas [26]. A general relationship between dominance, group behavior in brainstorming and decision-making groups, and performance has not yet been firmly established.

Acknowledgments This research was supported by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). Taemie Kim would like to thank Anges Chang for her support in the experiments.

References

1. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.S.: Towards measuring human interactions in conversational settings. In:

Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR) Workshop on Cues in Communication, Kauai, USA (2001)
2. Bishop, C.M.: Pattern recognition and machine learning, vol. 4. Springer, New York (2006)
3. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L.: Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A Math. Theor. **41**, 224015 (2008)
4. Carletta, J. et al.: The AMI meeting corpus: a pre-announcement. In: Renals, S., Bengio, S. (eds.) Machine Learning for Multimodal Interaction. Lecture Notes in Computer Science, vol. 3869, pp. 28–39. Springer, Berlin (2006)
5. Choudhury, T., Pentland, A.S.: The sociometer: a wearable device for understanding human networks. In: CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments (2002)
6. Dielmann, A., Renals, S.: Automatic meeting segmentation using dynamic Bayesian networks. IEEE Trans. Multimed. **9**(1), 25–36 (2007)
7. DiMicco, J., Pandolfo, A., Bender, W.: Influencing group participation with a shared display. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), New York, USA, pp. 614–623 (2004)
8. DiMicco, J.M., Hollenbach, K.J., Bender, W.: Using visualizations to review a group's interaction dynamics. In: Proceedings of the International Conference on Human Factors in Computing Systems, New York, USA, pp. 706–711 (2006)
9. Dong, W., Lepri, B., Cappelletti, A., Pentland, A.S., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI), New York, USA, pp. 271–278 (2007)
10. Dong, W., Mani, A., Pentland, A.S., Lepri, B., Pianesi, F.: Modeling group discussion dynamics. IEEE Trans. Auton. Mental Dev. (2011) (in press)
11. Eagle, N., Pentland, A.S.: Reality mining: sensing complex social systems. Pers. Ubiquitous Comput. **10**(4), 255–268 (2006)
12. Eagle, N., Pentland, A.S.: Eigenbehaviors: identifying structure in routine. Behav. Ecol. Sociobiol. **63**(7), 1057–1066 (2009)
13. Farrahi, K., Gatica-Perez, D.: What did you do today? Discovering daily routines from large-scale mobile data. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), Vancouver, Canada, pp. 849–852 (2008)
14. Farrahi, K., Gatica-Perez, D.: Probabilistic mining of socio-geographic routines from mobile phone data. IEEE J. Sel. Top. Signal Process. **4**(4), 746–755 (2010)
15. Garg, N.P., Favre, S., Salamin, H., Hakkani-Tur, D., Vinciarelli, A.: Role recognition for meeting participants: an approach based on lexical information and social network analysis. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), Vancouver, Canada, pp. 693–696 (2008)
16. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis. Comput. **27**(12), 1775–1787 (2009)
17. Gatica-Perez, D., McCowan, I., Zhang, D., Bengio, S.: Detecting group interest-level in meetings. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, USA, pp. 489–492 (2005)
18. Gips, J.P.: Social motion: mobile networking through sensing human behavior. Master's Thesis, MIT Media Laboratory (2006)
19. Grudin J. (1994) Computer-supported cooperative work: history and focus. Computer 27(5):19–26
20. Hassin R.R., Uleman J.S., Bargh J.A.: The New Unconscious. Oxford University Press, USA (2005)
21. Huynh, T., Fritz, M., Schiele, B.: Discovery of activity patterns using topic models. In: Proceedings of the International Conference on Ubiquitous Computing (UbiComp), Seoul, South Korea, pp. 10–19 (2008)
22. Jayagopi, D., Gatica-Perez, D.: Mining group nonverbal conversational patterns using probabilistic topic models. IEEE Trans. Multimed. **12**(8), 790– 802 (2010)
23. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations using nonverbal activity cues. IEEE Trans. Audio Speech Lang. Process (Special issue on multimodal processing in speech-based interactions) **17**(3), 501–513 (2009)
24. Jayagopi, D., Kim, T., Pentland, A.S., Gatica-Perez, D.: Recognizing conversational context in group interaction using privacy-sensitive mobile sensors. In: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM), Limassol, Cyprus, pp. 8:1–8:4 (2010)
25. Jayagopi, D., Raducanu, B., Gatica-Perez, D.: Characterizing conversational group dynamics using nonverbal behaviour. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), New York, US, pp. 370–373 (2009)
26. Kim, T., Chang, A., Holland, L., Pentland, A.S.: Meeting mediator: enhancing group collaboration using sociometric feedback. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), San Diego, USA, pp. 457–466 (2008)
27. Laughlin, P.R., Ellis, A.L.: Demonstrability and social combination processes on mathematical intellective tasks. J. Exp. Soc. Psychol. **22**(3), 177–189 (1986)
28. Lepri, B., Mana, N., Cappelletti, A., Pianesi, F.: Automatic prediction of individual performance from thin slices of social behavior. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), Beijing, China, pp. 733–736 (2009)
29. Madan, A., Pentland, A.S.: Vibefones: socially aware mobile phones. In: 2006 10th IEEE International Symposium on Wearable Computers, pp. 109–112. IEEE, New York (2006)
30. McGrath, J.E.: Groups: Interaction and Performance. Prentice Hall, New Jersey (1984)
31. Miluzzo, E., Cornelius, C.T., Ramaswamy, A., Choudhury, T., Liu, Z., Campbell, A.T.: Darwin phones: the evolution of sensing and inference on mobile phones. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, pp. 5–20. ACM, New York (2010)
32. Mitchell, T.M.: Machine Learning. Mc Graw Hill, New York (1997)
33. Olguín, D.O., Pentland, A.S.: Assessing group performance from collective behavior. In: Proceedings of the Computer Supported Collaborative Work, Workshop on Collective Intelligence In Organizations, Savannah, USA (2010)
34. Olguín, D.O., Waber, B.N., Kim, T., Mohan, A., Ara, K., Pentland, A.S.: Sensible organizations: Technology and methodology for automatically measuring organizational behavior. IEEE Trans. Syst. Man Cybern. Part B Cybern. **39**(1), 43–55 (2009)
35. Otsuka, K., Sawada, H., Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? From gaze, head gestures, and utterances. In: Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI), Nagoya, Japan, pp. 255–262 (2007)
36. Otsuka, K., Yamato, J., Takemae, Y., Murase, H.: Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) Extended Abstract, Montreal, Canada, pp. 1175–1180 (2006)
37. Pentland, A.S.: Socially aware, computation and communication. Computer **38**(3), 33–40 (2005)

38. Pentland, A.S.: Honest Signals: How They Shape our World. MIT Press, Cambridge (2008)
39. Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., Zancanaro, M.: Multimodal recognition of personality traits in social interactions. In: Proceedings of the ACM International Conference on Multimodal interfaces (ICMI), Chania, Greece, pp. 53–60 (2008)
40. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. Lang. Resour. Eval. **41**(3), 409–429 (2007)
41. Pianesi, F., Zancanaro, M., Not, E., Leonardi, C., Falcon, V.: Multimodal support to group dynamics. Pers. Ubiquitous Comput. **12**(3), 181–195 (2008)
42. Sanchez-Cortes, D., Aran, O., Schmid-Mast, M., Gatica-Perez, D.: Identifying emergent leadership in small groups using nonverbal communicative cues. In: Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI-MLMI), Beijing, China, pp. 39:1–39:4 (2010)
43. Vinciarelli, A.: Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Trans. Multimed. **9**(6), 1215–1226 (2007)
44. Waber, B.N., Pentland, A.S.: Recognizing expertise. In: Winter Conference on Business Intelligence, University of Utah, Utah, USA (2009)
45. Wilson, D.S., Timmel, J.J., Miller, R.R.: Cognitive cooperation: when the going gets tough, think as a group. Hum. Nat. **15**(3), 225–250 (2004)
46. Woolley, A.W., Chabris, C.F., Pentland, A.S., Hashmi, N., Malone, T.W.: Evidence for a collective intelligence factor in the performance of human groups. Science **330**(6004), 686–688 (2010)
47. Wrede, B., Shriberg, E.: Spotting hotspots in meetings: human judgments and prosodic cues. In: Proceedings of European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, pp. 2805–2808 (2003)
48. Wyatt, D., Choudhury, T., Kautz, H.: Capturing spontaneous conversation and social dynamics: a privacy-sensitive data collection effort. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007, vol. 4, p. IV–213. IEEE, New York (2007)
49. Xu, X., Tang, J., Liu, X., Zhang, X.: Human behavior understanding for video surveillance: recent advance. In: IEEE International Conference on Systems Man and Cybernetics (SMC), 2010, pp. 3867–3873. IEEE, New York (2010)
50. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Modeling individual and group actions in meetings with layered HMMs. IEEE Trans. Multimed. **8**(3), 509–520 (2006)