Modeling Functional Roles Dynamics in Small Group Interactions

Modeling Functional Roles Dynamics in Small Group Interactions

Wen Dong, *IEEE Member*, Bruno Lepri, Fabio Pianesi, *IEEE Member*, and Alex Pentland, *IEEE Member*

***Abstract*—The paper addresses the automatic recognition of social and task-oriented functional roles in small-group meetings, focusing on several properties: a) the importance of non-linguistic behaviors, b) the relative time-consistency of the social roles played by a given person during the course of a meeting, and c) the interplays and mutual constraints among the roles enacted by the different participants in a social encounter. In particular, this paper proposes that the Influence Model framework can address these properties of functional roles, and compares the performance obtained by this framework to the performances of models that consider only property (a) (SVM), and to those that address both (a) and (b) (HMM). The results obtained confirm our expectations: the classification of social functional roles improves if models account for temporal dependencies among the roles played by the same subject, for the time properties of the roles played by each individual, and for the mutual constraints among the roles of different group members. The two versions of the Influence Model (IM and newIM), which encode all three properties together, outperform both the SVM and the HMM on most of the figures of merit used. Of particular interest is the capability of the Influence Model to obtain good or very good results on the less-populated classes – Orienteer and Seeker for the task area, and Attacker and Supporter for the socio-emotional area.**

***Index Terms*—Functional roles, influence model, non-linguistic behavior, multimodal analysis**

I. Introduction

SMALL-group interactions, such as meetings, are increasingly important in structuring our daily work life inside organizations. For example, according to a survey in [1], executives spend an average of 40% to 50% of their working hours in meetings, 50% of which is unproductive, and up to 25% of which is spent discussing irrelevant issues. These problems are often due not only to task-related factors (e.g., difficulties in choosing the right items for the agenda, or in focusing attention on relevant issues), but also to the complexity of group dynamics and social behaviours, which can hinder the team's performance. Different means and tools can be put at work to support dysfunctional teams, ranging from facilitation to training sessions conducted by experts. The availability of rich multimodal information makes it possible to explore providing some of these services automatically or semi-automatically.

For instance, in [2], the usefulness and the acceptability of a functionality inspired by the practice of coaching is investigated. This consists of a report about the relational/social behaviour of individual participants, generated from multimodal information extracted during the meeting and privately delivered after the meeting is over. A notable finding in [2] is that recipients of these reports found no significant difference in terms of usefulness, reliability, appropriateness, completeness, or clarity between those produced by an automatic system and those produced by a human expert.

In order to implement such functionality, an automatic system must be able to observe and understand people's social behavior, although without necessarily taking exact notice of what people are saying. Indeed, reports from human coaches are not minutes, but are based on a more qualitative and meta-level interpretation of the social dynamics of the group. These man-made reports do not contain information such as "In the first part of the meeting you talked for ten minutes about machine-learning techniques that would be useful for solving the problem," but rather "In the first part of the meeting you provided the group with background information," or "You prevented others from intervening in the discussion." In practice, both coaches and the systems aiming to reproduce them must abstract over low-level (visual, acoustic, etc.) information in order to produce medium/coarse-grained data about people's social behaviours.

Information that is demonstrably very useful for the purpose of coaching is the roles people play in the course of the interaction. Who is the leader or protagonist? Who is most involved in the discussion? Least involved? There are at least three perspectives on such social roles in the sociological and psychological literature. The first considers roles to be expectations regarding the behavior of a specific individual. The second emphasizes the behaviors associated with a particular position in a group or in an organization [3], [4]; group managers or other more-or-less officially-appointed roles belong to this class. The third view, in turn, considers roles as dependent on the specific interaction context, consisting of concrete behaviors enacted by people [5], [6]; here we find consideration of who currently (in a given meeting portion) plays the protagonist, who just sits and watches, who shows aggressive behavior, and so on. Because they are highly-situated in specific social encounters, and because they pay attention to group dynamics, the roles addressed by this third perspective – also called social functional roles [6] – are more directly useful for the purpose of coaching.

W. Dong is affiliated jointly to the MIT Media Lab, Cambridge (MA), USA, and Northeastern University, Boston (MA), USA (e-mail: wdong@media.mit.edu); B. Lepri is affiliated jointly to the MIT Media Lab, Cambridge (MA), USA, and FBK, Trento, Italy (e-mail: lepri@fbk.eu); A. Pentland is affiliated to the MIT Media Lab, Cambridge (MA), USA (e-mail: sandy@media.mit.edu); F. Pianesi is affiliated to FBK, Trento, Italy (e-mail: pianesi@fbk.eu).

More in general, an important motivation for the investigation of role recognition is the contribution to the efforts that are being done towards the automatic understanding of social interactions (see [7] for an extensive survey on this topic). Moreover, as pointed out by [8], roles can be used to enrich the content description of multimedia data in retrieval applications, can enhance media browsers for data like meeting recordings (e.g., [9]), can allow summarization approaches to identify media segments particularly rich in information [10], [11], can be used to segment the data into semantically coherent segments [12], [8].

This paper addresses the automatic recognition of social functional roles in small-group meetings, focusing on properties of theirs such as: a) the importance of non-linguistic behaviors, b) the relative time-consistency of the social roles played by a given person during the course of a meeting, and c) the interplays and mutual constraints among the roles enacted by the different participants in a social encounter. By considerably extending previous works on a similar topic [13], [14], [2], this paper compares the performance of models that consider only property (a) (SVM) to those that address both (a) and (b) (HMM), and to those that target all three through Influence Modeling [15].

In particular, we considerably improve on our previous works by: (i) exploiting a different and larger set of acoustic features. In addition to the sole speaking activity, we use the number of voiced segments per second (speaking rate), the duration of speaking turns in seconds, the number of voiced segments per speaking turns, the number of speakers with overlapping turns; (ii) introducing, based on the literature, a number of hypotheses concerning the dynamics of social roles and showing in detail how the data from our corpus confirm them. In more detail, we delve into: the existence of constraints on the possible combinations of task and socio-emotional roles played by a given individual at any given time; the relative stability in time of functional roles; the existence of constraints on the distribution of roles among meeting participants at any given time; (iii) proposing the Influence Model (IM) as a suitable way to exploit those interdependencies and investigate this possibility by means of two variants of the IM: the basic one and a new version in which influence matrices are not time-homogenous but change over time. The comparison is completed by investigating the performance of Hidden Markov Models (HMMs) and Support Vector Machines (SVMs). Taken together, the four models (SVM, HMM, basic and extended IM) capture increasing levels of multiparty interaction complexity and mutual dependencies among group participants.

Points (ii) and (iii) are particularly important. The first provides empirical grounds to commonly held assumptions; moreover, it motivates the exploitation of modeling approaches that capture different levels of data complexity, as discussed in point (iii). The experimental results we obtained seem to confirm the fruitfulness of this approach that, starting from hypotheses about phenomena, tries to first root them in data and then uses them for the purpose of automatic recognition.

The remainder of this paper is organized as follows. A detailed discussion of related works is presented in the next section. Section III then discusses Mission Survival Corpus I (MSC-I) used in our experimental analysis, and also the techniques employed to automatically extract both acoustic and visual non-linguistic features. Section IV discusses a number of characteristics of our data that can provide important insights for the automatic classification of functional roles. Section V proposes the Influence Model as a possible framework for the automatic recognition of functional roles. Section VI describes the features and the other algorithms (HMM and SVM) used for role classification. Section VII present and discuss our results. Finally, Section VIII draws conclusions.

II. Related Works

The view of roles we are considering sees them as abstractions over the actual behaviors of people in the course of an interaction, rather than as the outcomes of social expectations, position in a hierarchy, status, and so on.

In this section we review key works closely related to our perspective, from two distinct fields: social psychology and social computing.

A. Social Psychology Approaches

Much research in the social and psycho-social literature addresses this topic [16]. Of particular interest for our purposes are functional roles [6] defined in terms of the behavior enacted in a given situation and in a particular context – by allowing us to focus on what is actually happening, functional roles reduce the need for knowledge related to organizational aspects, history, the status of group members, and other organizational qualities.

Benne and Sheats [17] provided a list of social functional roles and collected them into three classes: task-oriented, maintenance-oriented, and individual-oriented. The first two kinds of roles relate to the group's needs. Task-oriented roles concern facilitation and coordination activities for task accomplishment, while maintenance-oriented roles contribute to social structure and to interpersonal relations. The third type of roles – individual roles – focuses on single members and their goals and needs, rather than on the group. Importantly, Benne and Sheats's definitions emphasize the dynamic nature of social functional roles, allowing for one and the same person enact more than one role in each of the three classes during the course of the same interaction episode.

Drawing on Benne's and Sheats's work, Bales [18] proposed Interaction Process Analysis (IPA), a framework for the study of small groups based on the classification of individual behavior in a two-dimensional role space consisting of a task and of a socio-emotional area. The roles related to the socio-emotional area embody activities that support, enforce, or weaken interpersonal relationships. For example, complementing another person is a positive socio-emotional behavior in that it increases group cohesion and mutual trust among members; conversely, insulting another participant is a negative socio-emotional behavior that can undermine social relationships. Task area roles, on the other hand, concern behavioral manifestations that impact the management and solution of the problem(s) that the group is addressing. Giving

or requesting information, or providing personal opinions and suggestions regarding the task, are examples of task-oriented activities.

B. Social Computing Approaches

In the computational camp, most attention has been devoted to the conception of roles relative to the first and second perspectives discussed above: roles as expectations of people's behaviors, or roles as defined on the basis of organizational matters (managers vs. clerks), status, and so on [7].

For instance, Weng et al. [12] applied Social Network Analysis (SNA) to identify the hero, the heroine, and their respective friends in three movies, based on the co-occurrences of roles in different scenes. Barzilay et al. [19] exploited the keywords, the durations of speaking turns, and explicit speaker introduction segments to identify the anchor, journalists, and guest speakers in a radio program. They obtained 80.5% classification accuracy on human transcripts, and 77% accuracy on automatically-recognized transcripts (ASR data), in both cases using the Maximum Entropy algorithm.

Banerjee and Rudnicky [20] proposed a simple taxonomy of participant roles (presenter, information provider, participator, and information consumer), and then trained a decision tree classifier to learn them from simple speech-based features. The classifier took as input the feature representation of a short time window (meeting history) to classify the roles at the end of the window. The method used seven features, all of which were manually extracted: the number of speaker changes, the number of speakers, the number of overlaps in speech, the average length of those overlaps in seconds, the total amount of speech by a given participant in seconds, the number of overlaps initiated by that participant, and the number of overlaps initiated by other participants. Different experiments on the same data set produced a best classification accuracy of 53%.

Vinciarelli [21] exploited audio recordings of radio news shows to address the recognition of roles such as the primary and secondary anchormen, the guest, the interviewee, the headline announcer, and the weatherperson. In this data set, conversations are usually dyadic and the show follows a regular structure consisting of sections, each managed by one of the roles; by reducing role recognition to section recognition this setup considerably limits the complexity of the task that can be analyzed, as compared to other settings such as meetings. By using features based on basic concepts of social network analysis and on the duration of each role segment, Vinciarelli reported up to 85% frame-based classification accuracy on 96 bulletins. Additional experiments with a variant of this approach and a different source of radio shows (talk-shows) were discussed in Favre et al. with similar performance [22].

Using the same data and targeting the same roles as above, Salamin et al. [23] exploited specific acoustic features, such as who talks when and how much (turn-taking), and statistical properties of pitch, formant, energy, and speaking rate, reporting an accuracy of 89%.

Finally, Favre et al. [24] applied Hidden Markov Models and n-gram language models for the recognition of role sequences in 90 hours of both broadcasts and meetings data. The roles exploited for the broadcast data were the same as in [21], whereas meeting roles were mainly of the organizational type, including project manager, marketing expert, industrial designer, and user-interface expert.

Favre et al. found that classification performance depended on the degree of role formality – that is, on how much a given role corresponded to a function specific to the specific interaction setting (for example, the moderator in a debate), and how strong the constraints were that the role imposed on participant behavior. Informal roles, which in their terminology correspond to a position in a specific social system (for example, the manager in a company), and which do not impose specific constraints on the behavior of people, were harder to model – though still recognizable with a performance higher than chance. In another work, Favre et al. [22] attempted to recognize the project manager, the marketing expert, the user-interface expert, and the industrial designer in a large portion of the AMI corpus (138 meetings, 45 hours). They employed acoustic features extracted from the whole meeting rather than from only "thin slices" thereof, using a simple Bayesian classifier. For the four-role task, the authors reported a best performance of 44% classification accuracy. Addressing the same problem with the same data set, Garg et al. [25] combined non-linguistic and linguistic information – that is, words derived from manual or automatic speech transcripts. The joint usage of the two types of features significantly improved accuracy over the usage of only non-linguistic information, with a frame-based classification accuracy of 68% for the four roles.

Jayagopi et al. [26] addressed the recognition of role-based status in small groups. In the workplace, status often corresponds to a person's position in the group's or organization's hierarchy, and it is often defined by a role – a project manager has a different status than his assistant. Using 5 hours of meeting data (AMI corpus) divided into time slices of 5 minutes, Jayagopi et al. studied the detection of role-based status (the project manager of the team) using several automatically-extracted non-linguistic features that characterize speaking activity, visual activity, and visual attention. This research showed that the best non-linguistic cues (the total number of a speaker's turns, and the total number of times each subject speaks first after another speaker) can correctly predict the project manager with 66.7% accuracy.

Most of the research discussed so far addresses meetings where people get together to achieve common objectives, and where coordination among participants is important for success. Raducanu et al. [27], in turn, investigated role analysis in competitive meetings. Their data set consisted of videos from a popular US reality TV show, wherein participants aim at earning a real job at a firm. In each episode, contestants participate in a business-related task in one of two opposing teams. During a subsequent group meeting, led by a strongly-minded boss, one of the participants is fired based on his or her performance. Using 90 minutes of meeting data, Raducanu et al. considered simple approaches based on manually-extracted cues related to social status (speaking time and turns, interruptions, and centrality) and reported performance accuracy

for the estimation of both the meeting chairman and the fired person of 85% and 92%, respectively.

In all the works discussed so far, roles are defined according to either the behavior expectation perspective, the status/organizational perspective, or a mixture thereof. In only a few cases is the notion of roles as discussed by Benne and Sheats [17] and by Bales [18] addressed in the computational literature. In particular, Zancanaro et al. [13] exploited SVM to classify a simplified version of Bales's [18] roles, exploiting acoustic and visual features and reporting accuracy figures above 65%. Dong et al. [14] compared SVM, HMM, and a version of the Influence Model, reporting 75% accuracy for the latter. Our present paper capitalizes on, and considerably extends, these works by exploiting a different feature set and by introducing a new version of the Influence Model that allows for additional flexibility.

More recently, Wilson and Hofer [28] and Valente and Vinciarelli [29] annotated the AMI corpus using the coding scheme proposed by Zancanaro et al. [13]. In their paper, Wilson and Hofer [28] investigated whether automatically derived linguistic subjectivity and expressive prosodic features can be used to improve social role recognition of participants in meetings. They found that combining these expressiveness features with speech activity ones social role recognition is improved over the usage of the sole speech activity. Valente and Vinciarelli [29] used turn statistics and prosodic features in order to automatically recognize the socio-emotional roles played by the AMI meeting participants. At first, turn-taking statistics and prosodic features were integrated into a single generative conversation model which achieved a role recognition accuracy of 59%. This model was then extended to explicitly account for dependencies (or influence) between speakers yielding an accuracy of 65%. Finally, they investigated the statistical dependencies between the formal (roles that does not change during the recording; i.e. Project Manager) and the social roles and then integrated these information sources achieving a 68% accuracy.

III. The Mission Survival Corpus I

In our research, we used the Mission Survival Corpus I, a multimodal annotated corpus based on the audio and video recordings of eight meetings that took place in a lab setting equipped with cameras and microphones. Each meeting consisted of four people engaged in the solution of the Mission Survival Task (MST). This task was originally designed by the National Aeronautics and Space Administration (NASA) to train astronauts before the first Moon landing, and it proved to be a good indicator of group decision-making processes[1]; since then, it has been frequently used in experimental and social psychology to elicit decision-making processes in small groups. The exercise consists of promoting group discussion by asking participants to reach a consensus on how to survive in a disaster scenario, like a moon landing, or a plane crashing in mountain wilderness. The group must rank fifteen items according to importance for the survival of crew members. A consensus decision-making scenario is appropriate to our needs because of the intensive engagement required for groups to reach agreement. This affords the opportunity to observe a large set of social dynamics and attitudes.

In consensus decision-making processes, each participant is asked to express her or his opinion. The group is then encouraged to discuss each individual contribution by weighting and evaluating its quality. In our case, consensus was enforced by establishing that any participant's proposal would become part of the common sorted list only if that participant managed to convince the others of the proposal's validity. We also added an element of competition by awarding a prize to the individual who proposed the greatest number of correct and consensually-accepted items.

These sessions were recorded in a specially-equipped room by means of 4 firewire cameras in the corners of the room. Four wireless close-talk microphones (one for each participant) and one omni-directional microphone in the middle of the table were used to record speech activity.

The Mission Survival Corpus I consists of 8 meetings and 32 subjects: each subject took part only in one meeting. The average length of the collected meetings is around 18 minutes and the total length of the corpus is 2 hours and 49 minutes. The longest meeting lasts around 26 minutes while the shortest lasts around 13 minutes.

A. Functional Role Coding Scheme

We employ Bales's categories, while interpreting their functions as (functional) roles in terms of Benne and Sheats's approach. This move is motivated by the expectation that, though dynamic, the behaviour of each participant would not change too much or too often during a meeting, such that the slightly-more-static concept of functional role would be at least as appropriate as the more dynamic concept of function. Section IV details evidence that these assumptions are tenable. Finally, we further adapt the resulting two-dimensional scheme, adjusting the roles according to observations performed on a number of face-to-face meetings.

Our coding scheme – the Functional Role Coding Scheme (FRCS) – consists of five labels for the Task Area and five labels for the Socio-Emotional Area [30]. The Task Area includes functional roles related to the facilitation and coordination of the tasks in which the group is involved, as well as roles related to the technical skills of members as they are deployed through the course of the meeting.

- The Orienteer (o) orients the group by introducing the agenda, defining goals and procedures, and keeping the group focused and on track. He or she summarizes the main ideas of the group, recording the most important arguments in the discussion, the minutes, and the group's decisions. The Orienteer spells out suggestions in terms of examples, offers a rationale for suggestions, and tries to deduce how an idea would work if adopted by the group. From a behavioral point of view, this is often the first person to speak, and this person tends to look at all members rather than at one specific person (as opposed to

[1]This task was supposedly created for training purposes by a Mark Wanvig, former U.S. Army survival instructor for the Reconnaissance School of the 101^{st} Division.

the Giver, for example, who focuses on the interlocutor). The Orienteer has a major role in structuring the discussion ("Okay, let's move on"), and in planning future efforts.
- The Giver (g) provides factual information and answers questions. This person states his or her beliefs and attitudes about an idea, expresses personal values, and offers factual information. From a behavioral point of view, the Giver usually speaks only if consulted by another person, and his or her gaze is directed primarily towards that interlocutor.
- The Seeker (s) requests suggestions, information, and clarifications in order to promote effective group decision-making. This role can be mistaken with the Orienteer; however, while the latter's questions are meant mostly to help the group to reach objectives ("What about moving to the next agenda item?"), the Seeker's questions relate to the task under discussion ("What's the status of project?" or "What do you think about adding a new functionality to the system?").
- The Recorder (r) manages the resources available to the group. The most apparent (and useful) manifestation of this role consists of keeping track of discussions and decisions for the group. In this respect, the Recorder should not be mistaken for the Follower, who might take notes but does so only for his or her own sake.
- The Follower (f) only listens, and possibly takes notes for personal use, but does not participate actively.

The Socio-Emotional Area includes roles oriented toward the formal relationships among group members, and roles oriented toward the functioning of the group as a social entity.

- The Attacker (a) may work in many ways: deflating the status of others; expressing disapproval of the values, acts, or feelings of others; attacking the group or the problem it is working on; joking aggressively. This person consistently reacts negatively to the ideas of other group members. The behavioral indicators that signal this role include an aggressive tone of voice, looking elsewhere, making noise, and moving nervously.
- The Gate-keeper (gk) is the moderator within the group. This person effects communication and attempts to keep channels open by encouraging and facilitating participation. The Gatekeeper mediates the differences between other members, attempts to reconcile disagreements, and relieves tension in conflict situations.
- The Protagonist (p) takes the floor (the right to speak) and drives the conversation. This person assumes a personal perspective, asserting his or her authority or superiority by virtue of either special status or performing a particular important task.
- The Supporter (su) shows a cooperative attitude – manifesting attention, understanding, and acceptance in addition to providing technical and relational support to other members of the group. This person also helps to maintain a collaborative atmosphere by endorsing common objects.
- The Neutral (n) passively accepts the idea of others, serving as an audience in a group discussion.

The reliability of our proposed coding scheme is assessed on a subset of the corpus consisting of 130 minutes for the Socio-Emotional Area and 126 minutes for the Task Area. Two trained annotators code five participants on the Socio-Emotional Area and five participants on the Task Area. Cross-judge consistency is measured by means of Cohen's κ [31]. The results are the following:

- Task Area: $\kappa = 0.70$ (N = 758, p < 0.001; confidence interval with $\alpha = 0.05$: 0.67-0.75).
- Socio-Emotional Area: $\kappa = 0.60$ (N = 783, p < 0.001; confidence interval with $\alpha = 0.05$: 0.56-0.65).

According to Landis and Koch's criteria [32], the agreement in the Task Area is good ($0.6 < \kappa < 0.8$), but is borderline between good and moderate ($0.4 < \kappa < 0.6$) in the Socio-Emotional Area.

B. Acoustic Non-linguistic Cues

Previous research suggests that the non-linguistic information conveyed by speech can be very informative regarding different social behaviors and characteristics, such as roles [14], [22], [24], task performance [33], [34], personality traits [35], [36], dominance status [26], [37], and emergent leadership [38].

Starting from Bales's definitions of roles and from our coding scheme, we automatically extract the following set of cues for each meeting participant:

- Speaking activity: presence/absence of speech. This information helps to identify who speaks first, and to recognize subjects playing the Orienteer and the Protagonist roles.

- Speaking rate: we defined this feature as the number of voiced segments (essentially, syllables) per second. More precisely, the speaking rate is computed only over speaking segments, i.e. pausing between phrases does not affect this number [39].

- Duration of speaking turns in seconds where a speaking turn is a sequence of voice segments bounded by 1-second-or-longer pauses. Given that we are using close-talk microphones, our definition of turn is applied to an audio channel capturing the voice of one participant only. We hypothesize that people with longer speaking turns might be playing the Protagonist role.

- Speaking rate per speaking turn: we defined this feature as the number of voiced segments (essentially, syllables) per speaking turn [39].

- Number of speakers with overlapping turns. Normally, such overlaps are 'backchannel' comments such as 'ok' or 'really?' and so useful to identify the Supporter role. More precisely, we use the term 'backchannel' to denote the short messages that the person holding the turn receives without relinquishing the turn [40]. These short messages are different from short answers because a short answer by a listener B assume that speaker A ended his/her turn. However, sometimes we can have long overlaps similar to interruptions and they could be a relevant cue for identifying the Attacker role.

Concerning speaking activity, the output from the microphones is automatically segmented at a 330ms frame rate, and labeled by means of a VAD – Voice Activity Detector

[41]. For each frame, the VAD produces an output of the form '<temporal frame; label-S1; label-S2; label-S3; label-S4>,' where <temporal frame> is the frame's identifier, and <label-*> takes on the values '0' and '1' corresponding to 'non-speech' and 'speech' respectively, for each participant (speakers S1, S2, S3, and S4).

Voiced segments are extracted from the four audio tracks at every 10^{-3} seconds using the 9-parameter Boersma's algorithm [42]. This algorithm has been reported to have minimal pitch determination error and large resolution of harmonics-to-noise ratio due to its method of computing the short-term autocorrelation function of continuous-time audio time series.

Voiced segments are clustered into speaking turns using a Hidden Markov Model (HMM). Two latent states indicate whether there is a turn boundary, and two independent observations correspond to the existence of a speaker change and to the time interval between the current and the past-voiced segment.

Finally, we compute the number of voiced segments inside a given turn, the duration of a turn, and the number of speakers with overlapping turns.

C. Visual Cues

Fidgeting is defined as "a condition of restlessness as manifested by nervous movements" [43] and it can reveal important clues about the emotional state of an individual such as boredom, nervousness, impatience, and more in general social anxiety [44]. Using visual means, fidgeting can be localized by employing techniques such as optical flow or Motion History Images (MHIs) [45]. However, due to ambiguities between the actions say of fiddling with a pen and actually writing with it, it is impossible to be sure which is taking place. Consequently, following [46] any action where repetitive, localized motion is being observed will be labeled as fidgeting. In sum, fidgeting refers to localized repetitive motions, such as rhythmically tapping fingers on the table, playing with water glasses, or adjusting clothing. Fidgeting was automatically annotated by means of MHI [45], a technique that uses skin-region features and temporal motions to detect repetitive motions in the images. An energy value is then associated with these motions in such a way that the higher the value, the more pronounced the motion. All fidgeting values for a given person were normalized to the fidgeting activity of that person during the entire meeting. We compute two separate features, hand fidgeting and body fidgeting, to capture the differences between the amount of movement produced by the hands and that produced by the rest of the body. Indeed, we hypothesize that these differences could be useful to pin point different social and task roles. Both these features were generated at a frame rate of 330ms and so are synchronized with the output from the microphones.

Fidgeting was extensively used in works dealing with the automatic recognition of social functional roles [13], [14] and with the prediction of personality traits [47], [35]. So, although there are other visual features (e.g. focus of attention) of potential relevance to characterize task and socio-emotional roles, the visual features presented here are robust.

IV. Role-taking Behaviors

Based on this analysis of the Mission Survival Corpus, we now provide some support that for a number of hypotheses on which our research is based. In particular, we will argue that a) there are constraints on the possible combinations of task and socio-emotional roles that a given individual can be play at at any given time; b) at the individual level, functional roles have a certain degree of stability in time; and c) the distribution of roles among meeting participants at any given time is constrained. Section V details how these properties of functional roles can be leveraged for the purpose of automatic classification.

We start by considering the association of task with socio-emotional roles at the individual level. Table I reports the number of instances observed in the corpus for each task*socio-emotional role combination. Notice that the Recorder role of the task area and the Gate-keeper role of the socio-emotional area are not reported – no instances were present in the corpus.

As one might expect, the Neutral and the Follower roles are the most common ones, followed by Giver for the task area and Protagonist for the socio-emotional area. The Attacker role has very few instances, most probably because the nature of the meetings – which are based on consensus reaching – prevents and makes useless the display of aggressive behavior.

The significance of an association between task and socio-emotional roles association was tested by means of a χ^2 test that rejected the null (no association) hypothesis ($\chi^2 = 71.083$, df=9, p<.0001). A more detailed analysis of the task*socio-emotional association was then pursued by means of the Pearson adjusted residuals, setting the significance threshold at residual$\geq$2. These results are reported in Table II, where the '+' and the '-' signs indicate significantly high and low co-occurrences, respectively.

According to Table II, the significant task*socio-emotional association detected by the χ^2 test results from role co-occurrences: the Follower with the Neutral, the Giver with the Protagonist, the Orienteer with the Supporter, and the Seeker with the Attacker. On the other hand, the Giver and the Neutral tend to mutually exclude, as do the Giver and the Supporter, the Follower and the Protagonist, and the Orienteer and the Protagonist. The most likely role patterns are, therefore, those of people who a) play a neutral socio-emotional role and just follow the discussion without a strong personal participation, b) provide information while playing the protagonist, c) orient the discussion and are cooperative, and d) seek information in an aggressive fashion.

Besides associating the roles of the two areas in characteristic ways, participants execute them with different characteristic durations, as shown in Table III. We submitted the data to a Generalized Estimating Equations [48] analysis, using the duration of each task*socio-emotional combination as the dependent variable, and the task and socio-emotional areas as factors. Given the highly-skewed data, a gamma distribution with identity link was exploited.

The results show the primary effects of both task (Wald $\chi^2 = 22.841$, df=3, p<.001) and socio-emotional (Wald $\chi^2 = 25.842$, df=3, p<.001) roles on duration, as well as a significant

Table I
NUMBER OF TIMES IN WHICH EACH TASK*SOCIO-EMOTIONAL ROLE COMBINATION IS OBSERVED IN THE CORPUS. THE COLUMN 'TOTAL' CONTAINS THE NUMBER OF TIMES EACH ROLE IS OBSERVED IN THE CORPUS.

	Attacker	Neutral	Protagonist	Supporter	Total
Giver	5	487	270	116	878
Follower	9	677	251	173	1110
Orienteer	0	64	19	48	131
Seeker	5	85	29	16	135
Total	19	1313	569	353	2254

Table II
SIGNIFICANT ASSOCIATIONS AMONG BETWEEN TASK AND SOCIO-EMOTIONAL AT THE INDIVIDUAL LEVEL.

	Attacker	Neutral	Protagonist	Supporter
Giver		-	+	-
Follower		+	-	
Orienteer			-	+
Seeker	+			

Table III
AVERAGE DURATION (IN SECONDS) OF THE DIFFERENT TASK*SOCIO-EMOTIONAL ROLES COMBINATIONS.

Task area roles \ Socio-emotional area roles	Attacker	Neutral	Protagonist	Supporter	Marginal
Giver	8	7	**21**	10	11
Follower	2	**33**	3	4	11
Orienteer	N/A	4	10	**17**	10
Seeker	7	6	9	9	8
Marginal	6	12	11	10	

task*socio-emotional interaction (Wald $\chi^2 = 296.645$, df=8, p<.001). As can be seen in Table III, the main effect of task roles is due to the lower average duration of the Seeker role with respect to the others. The main effect of socio-emotional roles is due to the lower duration of the Attacker role. That is, episodes involving Seekers or Attackers have a lower duration. The task*socio-emotional interaction, in turn, is due to the longer average duration of the (Giver, Protagonist), (Follower, Neutral), and (Orienteer, Supporter) combinations – see the bold-marked figures in Table III. Interestingly, these three combinations are among the most frequent, as we saw while discussing Table I and Table II.

In summary, in the Mission Survival Corpus, people often enter specific combinations of social function roles, in particular (Giver, Protagonist), (Follower, Neutral), and (Orienteer, Supporter), and they do so for longer on each occasion. Other combinations are significantly infrequent: (Giver, Neutral), (Giver, Supporter), (Follower, Protagonist), and (Orienteer, Protagonist).

As anticipated, individuals not only execute their roles with a characteristic length and combine them in specific manners, but also coordinate their own roles with those of the others in a restricted number of ways.

As shown by in Table IV, there is always at least one person playing the Follower, and the most frequent task area configurations are those where at least one person in the group plays the Giver and the other members are Followers – an

Table IV
FREQUENCY OF TASK AREA CONFIGURATIONS. ROWS LABELED F, G, S, AND O REPORT THE NUMBER OF MEETING MEMBERS SIMULTANEOUSLY PLAYING THE GIVEN ROLE. ONLY CONFIGURATIONS WITH A FREQUENCY $\geq$1% ARE REPORTED.

F	1	2	2	2	3	3	3	4
G	2	1	1	2	0	0	1	0
S	1	0	1	0	0	1	0	0
O	0	1	0	0	1	0	0	0
%	1	4	5	20	13	2	36	11

Table V
FREQUENCY OF SOCIO-EMOTIONAL AREA CONFIGURATIONS. ROWS LABELED N, P, S, AND A REPORT THE NUMBER OF MEETING MEMBERS SIMULTANEOUSLY PLAYING THE GIVEN ROLE. ONLY CONFIGURATIONS WITH A FREQUENCY $\geq$1% ARE REPORTED.

N	1	1	1	2	2	2	3	3	4
P	1	2	3	0	1	2	0	1	0
S	2	1	0	2	1	0	1	0	0
A	0	0	0	0	0	0	0	0	0
%	1	1	1	3	7	11	18	36	21

occasion totaling 56% of the number of task configurations. Moreover, participants rarely distribute over more than two roles. Similar considerations could can be made for the socio-emotional area – see Table V. The considerations made and a visual inspection of Tables IV and V demonstrate the existence of important interdependences among the roles simultaneously played by the various group members. The potential utility of exploiting these interdependences for the purpose of role classification is clear.

Another important characteristic of social functional roles is their relative stability in time. In the socio-emotional area, the (split-half) correlation coefficients for the amount of time spent playing a given role during the first and the second halves of each meeting are 0.85, 0.59, and 0.27 for the Neutral, Protagonist, and Supporter roles, respectively; that is, people tend to play the Neutral and the Protagonist roles in quite a consistent fashion over time. In the task area, time consistency is even higher, with split-half correlation coefficients of 0.80, 0.90, 0.55, and 0.40 for the Giver, Follower, Orienteer, and Seeker roles, respectively. This data shows that our expectation of functional roles being a little more static than Bales's functions was on the right track. As a consequence, it seems possible to exploit knowledge about the past role-taking behavior of an individual in the Mission Survival Corpus to improve the predictions about future behavior.

We conclude this discussion of the data offered by the Mission Survival Corpus with a few considerations about the ways in which role dynamics depend on and affect group behavior. One interesting case is the intensity level of the discussion, as measured for example through the number of simultaneously-speaking people. The odds that roles such as Attacker, Supporter, or Seeker will appear increase significantly with the number of people speaking at the same time; for example, the odds of an Attacker being present when two or more individuals speak together simultaneously are 40:1 over the case when only one individual speaks. For task roles, when the speech of two individuals overlaps, the odds of them both being Givers are 4:1 over the odds of one being a Giver and the other a Seeker , and 8:1 over the odds of

Table VI
AVERAGE INTENSITY VALUE OF BODILY MOVEMENT (HAND/BODY FIDGETING), MEASURED ON THE 'TARGET' PERSON AND THE OTHER MEMBERS IN 10-SECOND WINDOWS, WITH RESPECT TO THE SOCIO-EMOTIONAL ROLE PLAYED BY THE TARGET. USING 'TARGET PERSON' WE DENOTE THE PERSON UNDER ANALYSIS WHILE USING 'OTHER MEMBERS' WE DENOTE ALL THE OTHER PARTICIPANTS EXCEPT THE 'TARGET' PERSON.

	Attacker	Neutral	Protagonist	Supporter
target hand	11	18	18	16
other hand	20	16	19	17
target body	11	20	21	18
other body	23	19	22	19

them being a Giver and an Orienteer. In comparison, when only one individual speaks, his or her odds of being a Giver are 17:1 over those of his/her being a Seeker, and 4:1 over being an Orienteer. Similar effects of role dynamics can be detected in visual behavior. For example, the presence of an Attacker is associated with increased hand and body activity in the others participants (20/23 vs. 11/11); see Table VI.

In this section we have discussed a number of characteristics of our data that can provide important insights for the automatic classification of functional roles. We have seen that in the Mission Survival Corpus the combination of roles that an individual simultaneously plays are is highly constrained; that the role played by one member of the group is affected by, and affects, those the roles played by other members; and that social functional role dynamics are moderated by a certain stability in time. Finally, we have exemplified how specific behaviors are associated with specific role-playing. Although it is not possible to generalize these results away from the corpus we used, these findings confirm reasonable assumptions regarding the dynamics of functional roles. The next section applies these findings to automatic role classification.

V. INFLUENCE MODELING

In the previous section we discussed evidence that an individual's role-taking dynamics are affected by those of the other members of the group. For example, when a person expresses his opinions, the others usually listen to these statements and show either agreements and or doubts. Moreover, multi-person face-to-face interactions normally take on a small number of regular patterns among the huge number (256) of available possibilities. We propose to exploit these interdependencies by adopting the Influence Model as a framework for the automatic recognition of roles in small-group interactions.

The Influence Model has been applied in the past to different social systems. The first application of the Influence Model [49] attempted to infer influence networks from audio recordings of a group discussion session with five individuals. The researchers used the model to infer the underlying pattern of interpersonal influence from the noisy signals measured directly from each individual and their interactions on turn taking. Another set of researchers applied the influence model to conversational data from sociometric badges on 23 individuals and showed that the influence strength between individuals learned by the model correlated extremely well with individual centrality in their social networks [50]. The model has also been applied to the Reality Mining [51] cell-phone sensor data. Using information from 80 MIT affiliates as observations and constraining the latent space of each individual to be binary "work" and "home", researchers found that the influence matrix learned from this data matches well with the organizational relationship between individuals [15]. More recently, the influence model has been extended to a variety of systems, including traffic patterns [52] and flu outbreaks [53].

In the following subsections, we describe two variants of the Influence Model: (i) the latent structure Influence Model introduced by Dong and Pentland [15] and (ii) a new latent structure version in which the influence matrices are not time-homogenous but change over time.

A. Influence Model with State-independent Influence

An influence model is a stochastic process that captures how nodes in a network (e.g., people in a meeting in the Missions Survival Corpus) signal role-taking to one another (e.g., "I am taking a Giver role; please switch your roles accordingly") and change states to complement it. In a network with C nodes and m_c discrete states $\{1, \cdots, m_c\}$ for each node $c \in \{1, \cdots, C\}$, the states of the nodes evolve in the following way:

$$\begin{aligned} P(S_{t=0}^{(c)} = i) &= \pi_i^{(c)}, \\ P(S_{t+1}^{(c)} = i | S_t^{(0)}, \ldots, S_t^{(C)}) &= \sum_{c'} h_{c' \to c} \times a_{S_t^{(c')} \to i}^{c' \to c}, \end{aligned}$$

where $i \in \{1, \ldots, m_c\}$ and $\pi_i^{(c)}$ is a categorical distribution, i.e.$\sum_i \pi_i^{(c)} = 1$.

In this network, each node c signals node c' with rate $h_{c \to c'}$ (i.e., node c has an influence $h_{c \to c'}$ over node c'). Node c in state $s_t^{(c)} = i$ at time t can signal node c' to change state to $s_{t+1}^{(c')} = j$ at time $t+1$ with rate $h_{c \to c'} \times a_{i \to j}^{c \to c'}$, where $\sum_j a_{i \to j}^{c \to c'} = 1$ and $\sum_c h_{c \to c'} = 1$. That is, node c can distribute its influence $h_{c \to c'}$ according to how much it wants node c' to change into each of the $m_{c'}$ states. Hence, the signaling rate of the whole network is $\sum_{c,c'} h_{c \to c'}$, and the probability of a specific event is $h_{c \to c'} \times a_{s_t^{(c)} \to j}^{c \to c'} / \sum_{c,c'} h_{c \to c'}$ when any signaling event happens at time t.

Since we deal with noisy data, unknown node states, and unknown network structure, we use the latent structure Influence Model to infer node states and network structure from incomplete and noisy observations $(Y_t^{(c)})_{t \in N}^{c \in \{1,\ldots,C\}}$, assuming conditional probability distributions $P(Y_t^{(c)} | S_t^{(c)}; \theta)$ parameterized by θ and the following stochastic process of updating note states and observations:

$$\begin{aligned} &P\left(S_{t=0}^{(c)} = i, Y_{t=0}^{(c)}\right) = P\left(Y_{t=0}^{(c)} | S_{t=0}^{(c)} = i\right) \cdot \pi_i^{(c)}, \\ &P\left(S_{t+1}^{(c)} = i, Y_t^{(c)} | S_t^{(0)}, \cdots, S_t^{(C)}\right) = \\ &P\left(Y_{t+1}^{(c)} | S_{t+1}^{(c)} = i\right) \cdot \sum_{c'} h_{c' \to c} \cdot a_{S_t^{(c')} \to i}^{c' \to c}. \end{aligned}$$

A detailed discussion of the latent structure Influence Model, as well as its algorithms, can be found in [15].

B. Influence Model with State-dependent Influence

We also designed and tested a new latent structure State-dependent Influence Model – henceforth referred to as newIM – in which each individual node $c \in \{1, \ldots, C\}$ can decide how much it wants to vote on the latent states of another node $c' \in \{1, \ldots, C\}$. For example, in our scenario if a person takes on the Giver role, that person can vote vote for another person to take the Neutral or the Seeker role. On the other hand, if a person takes the Neutral role, he or she can choose not to vote on another person's task and social role:

$$P\left(S_{t=0}^{(c)} = i, Y_{t=0}^{(c)}\right) = P\left(Y_{t=0}^{(c)} | S_{t=0}^{(c)} = i\right) \cdot \pi_i^{(c)},$$

$$P\left(S_{t+1}^{(c)} = i, Y_{t+1}^{(c)} | S_t^{(0)}, \ldots, S_t^{(0)}\right) =$$

$$\frac{1}{Z} \cdot P(Y_{t+1}^{(c)} | S_{t+1}^{(c)} = i) \sum_{c'} h_{c' \to c} \cdot a_{S_t^{(c')} \to i}^{c' \to c},$$

$$\text{where } Z = \sum_i P(Y_{t+1}^{(c)} | S_{t+1}^{(c)} = i) \cdot \sum_{c'} h_{c' \to c} \cdot a_{S_t^{(c')} \to i}^{c' \to c},$$

$$\forall c, c', i, \sum_j a_{i \to j}^{c \to c'} \leq 1, \text{ and } \forall c', \sum_c h_{c \to c'} = 1.$$

The major difference between the new model and the old model is that the "influence" of node c on node c' is no longer a constant, $\sum_j a_{i \to j}^{c \to c'} = 1, \forall c, c', i$, as in the previous model, but instead is $\sum_j a_{i \to j}^{c \to c'} \leq 1, \forall c, c', i$. Hence, the influence matrix is not time-homogeneous but instead changes over time (Figure 1). In particular, the parameters α in the new formulation of the Influence Model are assigned to the number of times a given transition happens divided by the total number of times the transition could happen (see for more details [53])

The algorithms for the latent state inference and the parameter learning can be derived from this new definition.

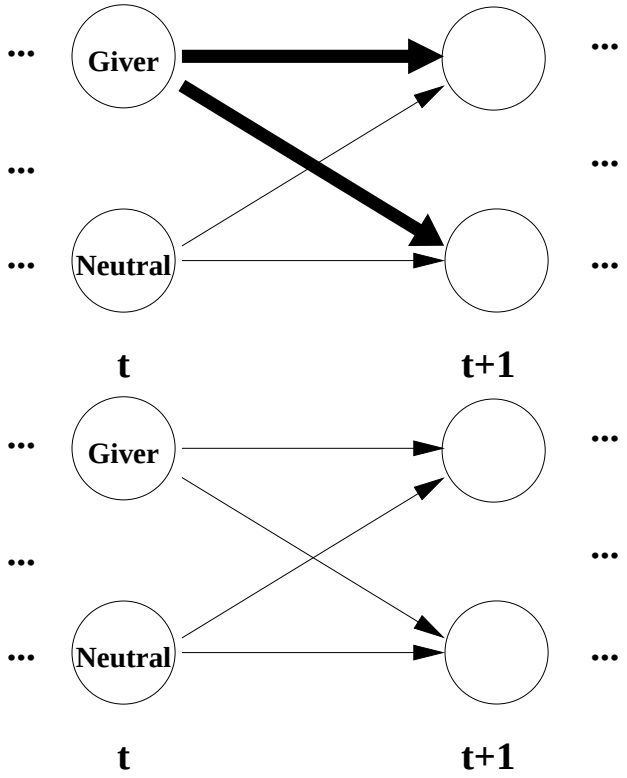

Figure 1. An influence model with state-dependent influence (above) assigns different amount of influences to a node according to his current state, while an influence model with state-independent influence (below) assigns the same amount of influence to a node irrespective of its current state. State-dependent influence fits functional-role dynamics better because for instance a giver has more influence than a neutral person. For sake of readability, the figure represents the interaction only between 2 subjects.

VI. Automatic Recognition of Roles

In Section IV we have showed that at the individual level roles have a certain degree of stability in time. Besides allowing the encoding of relationships between the roles played by different group members, the Influence Model (IM) also considers time dependence. In order to separate the two effects, we are going to compare the two influence models with simple Hidden Markov Models (HMMs). Finally, we will further extend these comparisons by exploiting multi-class Support Vector Machine, a model that uses none of the properties discussed above.

These three classifiers incrementally use more information for classification. The SVM considers each sample as independent and identically distributed, and the prior probability of each class as constant, for each sample. The HMM considers the temporal correlation between the samples and the prior probability of the classes in the current sample as dependent upon the posterior probabilities of the classes in the last sample; these properties of HMMs enables them it to capture the time consistencies we discussed above. Finally, IM captures the idea that people influence each other, and that the current role of a person is influenced by the roles of other participants in addition to being influenced by that person's previous roles.

We use the following ten groups of features to estimate task roles and socio-emotional roles: the fractions of time that a subject is (1) speaking, (2) showing hand fidgeting, and (3) showing body fidgeting; the fractions of time in similar time windows that when the counterparts of the target subject in the same meeting are (4) speaking, (5) showing hand fidgeting, and (6) showing body fidgeting; (7) the target subject's average "speaking turn" length and (8) voicing rate; (9) the average "speaking turn" length; and (10) the voicing rate of the subject target's counterparts. These features were computed in time windows of different durations, from 1 to 120 seconds, with 10-second steps. For example, for the fraction of speaking time of subject X at the 60th second, we computed (i) the fraction of X's speaking time in the window from second 60 to second 61 (a 1-second. window), (ii) the fraction of X's speaking time in the window from second 55 to second 66 (10-sec. window), (iii) the fraction of X's speaking time in the window from second 50 to second 71 (20-sec. window), and so on until (xii) the fraction of X's speaking time in the window from second 0 to second 121 (121-second window). Once extended to all seconds and all features, this procedure yields 12 measures at different granularity for each feature, from very fine (a 1-sec. window) to very coarse (a 120-sec. window), allowing the capture of behaviors having both a short duration (e.g., asking a question) and those spanning longer time intervals (e.g., answering a question).

Turning to SVM classification, we estimate roles through a majority vote of 16 independently-trained support vector classifiers, one for each task*socio-emotional roles combination. Each support vector classifiers was trained using a one-versus-one scheme. We used a Gaussian radial basis function (RBF) kernel, $k(x, y) = \exp(-|x - y|^2/2\sigma^2)$, and performed model selection by choosing $\sigma = median(|x - x'|)$ [54], with a training set of 1024 records evenly sampled from four different roles. Our choice of Gaussian RBF kernel was based on preliminary experiments in which we compared its performance to the performance of a linear kernel: the latter performed 5% worse on average accuracy, and much worse with infrequent roles.

Roughly speaking, the support vector classifier estimates the class label of a sample point based on its distances from the "prototype" points (e.g. support vectors) with known labels. If the sample point is on average closer to the "prototype" points of class A than of class B, it will be estimated as belonging to class A. However, if the dimensions of the sample point are correlated, the distance used for estimating the classes becomes illusive, and two sample points that are actually far away could appear to be close in our "conceptual space." In particular, the Gaussian RBF kernel function treats all the dimensions in the same manner, and requires large samples to achieve good classification accuracy. Given the limited size of the Mission Survival Corpus, we resorted to Principle Component Analysis (PCA) to adjust the distance function by using the principal components of the features (for each feature we have 12 values, corresponding to the 12 window sizes) to estimate the roles. For each feature we have 12 values, corresponding to the 12 window sizes. We use matrices $X^{(i)}$, where $i \in \{1, \cdots, 12\}$, to represent these features' values according to the each of the 12 measures. Each row of $X^{(i)}$ corresponds to a specific meeting, to a specific time in a that meeting, and to a specific meeting participant. Each column of $X^{(i)}$ corresponds to the a specific behavior statistic (a given feature) for a participant in a particular meeting at a specific time, for the windowing index by superscript i. This usage of principle components we yields an improved accuracy by of 5% on average, and accuracy rises to as much as 100% for infrequent roles.

As to HMMs and IMs, we first train support vector classifiers as described, fit the trained support vector classifiers with logistic regression models, and then employ the logistic regression models to find the normalized prior likelihoods of being in different roles corresponding to each observations. The output of the probabilistic SVM gives to us the best prior probability distribution of roles-taking behaviors.

We represent our HMM as follows: t, time; $y(t)$, the feature vector; $x(t)$, the role at time t; $p(x)$, the priors for role x; $p\left(x\left(t\right)|x\left(t-1\right)\right)$, the role transitions probabilities; $p\left(y\left(t\right)|x\left(t\right)\right)$, the conditional distribution of the observed feature vector given role x. We assume speaker independence, and all the feature sequences (one per subject) from all the eight meetings are used to train a single HMM by means of the standard Expectation Maximization (EM) algorithm. For prediction, each person is represented by an independent instantiation of the same Markov process. Thus, four independent HMMs represent the four people in the meeting. For classification, we use the standard Viterbi algorithm to compute the most likely sequence of roles.

Table VII
GLOBAL ACCURACY SCORES

	IM	SVM	HMM	new IM
Task	0.772	0.752	0.742	0.744
Socio-emotional	0.778	0.777	0.754	0.804

Going back to the two Influence Models, for each of them we used $2n$ interacting processes to model the task roles and the socio-emotional roles of the n individuals in the meeting. Doing this, we train each of the two IMs jointly on social and task roles, this way capturing another insight from the Section IV: the mutual dependencies between the task and the socio-emotional roles played by a given person at a given time. More generally, thanks to the two IMs, the task role played by subject X can interact with the socio-emotional role played by any other subject, including X. The observations for the individual processes are the corresponding acoustic and visual raw features (speaking/non-speaking, body movement, hand movement, and number of simultaneous speakers) averaged over short fixed-length time windows centered around the observation times, as described above. The latent states for the individual processes are the role labels. In the training phase, we compute the observation statistics of different functional role classes, as well as the interaction of different speakers by means of the EM (expectation maximization) algorithm. In the application phase, we infer each individual's task and socio-emotional roles based on the observations about the individual, as well as on their interactions with others, using the parameters previously trained.

We estimated the generalization capability of the trained classifiers by leave-one-meeting-out cross-validation. At each iteration, seven meetings were used for training and the one left-out for testing. Given that each subject participated in only one meeting, subjects are rigorously separated between training and test set.

VII. RESULTS

We start our discussion of the experiment results from the global accuracy scores, reported in Table VII.

The baseline classifier that relies only on priors (Table I) can at most attain 0.40 global accuracy for the task roles and 0.43 for the socio-emotional roles. If we compare those results with the global accuracies reported in Table VII, it is clear that all our models largely outperforms the baseline classifier's performance.

Then we perform two series of binomial tests on accuracy values – the first on Task Area values, and the second on Socio-emotional values – comparing each to the global average accuracy. In both cases, we set the significance level to $p<.05$, with Bonferroni adjustment for multiple comparisons.

In the Task Area, IM performs significantly better than the global average accuracy, whereas HMM and newIM do significantly worse. In the Socio-Emotional Area, newIM's accuracy is significantly better than the criterion, and HMM's significantly worse.

Table VIII
RECALL VALUES

		IM	SVM	HMM	newIM
Task	G	0.69	0.73	0.71	0.73
	F	0.84	0.81	0.80	0.77
	O	0.48	0.36	0.34	0.47
	S	0.48	0.21	0.44	0.65
		0.62	0.53	0.57	0.66

		IM	SVM	HMM	newIM
Socio-emotional	A	0.58	0.51	0.00	0.33
	N	0.85	0.85	0.86	0.88
	P	0.62	0.62	0.52	0.67
	S	0.39	0.39	0.21	0.31
		0.61	0.59	0.40	0.55

Table IX
PEARSON RESIDUAL FOR RECALL VALUES

		IM	SVM	HMM	newIM
Task	G	**-3.54**	2.16	-0.62	2.00
	F	**7.13**	1.01	-1.02	**-7.13**
	O	**4.76**	**0.36**	**0.34**	**0.47**
	S	1.98	**-12.59**	-0.51	**11.16**

		IM	SVM	HMM	newIM
Socio-emotional	A	**5.17**	**3.57**	**-8.12**	-0.62
	N	-1.95	-2.13	-0.24	**4.32**
	P	1.15	0.96	**-10.05**	**7.94**
	S	**6.56**	**6.59**	**-11.46**	-1.70

As a measure of performance, accuracy tends to overprize classifiers that perform very well on highly-populated classes, and to penalize those that produce better results on low-populated ones. That something similar is happening in our case is witnessed by the analysis of macro-recall figures in Table VIII, where, for example, the advantage of newIM in terms of accuracy in the socio-emotional area seems to be due to its performances on the two most-populated classes, Neutral and Protagonist. In order to better investigate this point and allow for statistically-significant comparisons, we convert the recall figures of Table VIII into their standardized residual counterparts – Pearson residuals, which are N(0,1) – and allow for straightforward comparisons, including testing the significance of a cell value. Pearson residuals have been computed with respect to the same baseline as above – namely, a classifier yielding global average recall. This way, our comparisons amount to determining how much the different classifiers do better (positive sign) or worse (negative sign) than the baseline in terms of recall. Table IX reports the residuals, marking in bold those that are significant (greater than three standard deviations).

According to Pearson residuals, the high accuracy value of IM in the task area is due to its very good results with the Follower and Orienteer roles – the most and one of the least populated role classes, respectively. The lower global accuracy of newIM, in turn, is due to trading very good performances on the two least-populated classed – Orienteer and Seeker – with low recall on the most populated one, Follower.

Turning to the socio-emotional area, newIM's higher accuracy stems from the high recall values on the two most-populated classes – Neutral and Protagonist – and from a close-to-criterion performance on the other two classes. IM and SVM yield substantially identical results, performing quite well on the less-populated roles, Attacker and Supporter.

Table X
PRECISION

		IM	SVM	HMM	newIM
Task	G	0.79	0.74	0.67	0.70
	F	0.94	0.90	0.92	0.94
	O	0.47	0.37	0.29	0.38
	S	0.10	0.05	0.11	0.14
	Macro	0.58	0.52	0.50	0.54

		IM	SVM	HMM	newIM
Socio-emotional	A	0.11	0.09	0.00	1.00
	N	0.97	0.97	0.88	0.91
	P	0.51	0.50	0.52	0.57
	S	0.26	0.26	0.17	0.35
	Macro	0.46	0.46	0.39	0.71

Table XI
MACRO F SCORES

	IM	SVM	HMM	new IM
Task	0.566	0.508	0.508	0.552
Socio-emotional	0.491	0.481	0.394	0.584

Concerning precision, IM and newIM produce higher values than the competitors both in the task and in the socio-emotional area; see Table X. In the latter, newIM reaches high precision values, especially in the less-populated classes, Attacker and Supporter.

In conclusion, as the summary Macro F-scores in Table XI show, the IM and the newIM models perform better than the competitors in both role areas. Upon direct comparison, IM yields better performances in the task area, whereas newIM does better in the socio-emotional area.

VIII. DISCUSSIONS AND CONCLUSIONS

This paper has dealt with the automatic recognition of task-oriented and socio-emotional functional roles in small group meetings, exploiting several of their properties: a) the importance of non-linguistic behaviors, b) the relative time-consistency of the social roles played by a given person during the course of a meeting, and c) the interplays and mutual constraints among the roles enacted by the different participants in a social encounter. After providing empirical evidence in favor of those properties, we have proposed Influence Modeling as an approach capable of addressing and exploiting all three of these properties, and compared its performances with those of models that consider only property (a) (SVM), or both (a) and (b) (HMM).

The results discussed in the previous section seem to confirm our expectations: social functional role classification improves if models are exploited to account for the dependencies among roles played by the same subject, for the temporal dynamics of roles, and for the mutual constraints among the roles of different group members. The two versions of the Influence Model, which encode all three properties together, outperform both SVM and HMM on most of the figures of merit used. Of particular interest is the capability of the Influence Models to obtain good or very good results for the less-populated classes – Orienteer and Seeker for the task area, and Attacker and Supporter for the socio-emotional area. In no case, in fact, are the standardized recall values of the IMs significantly negative,

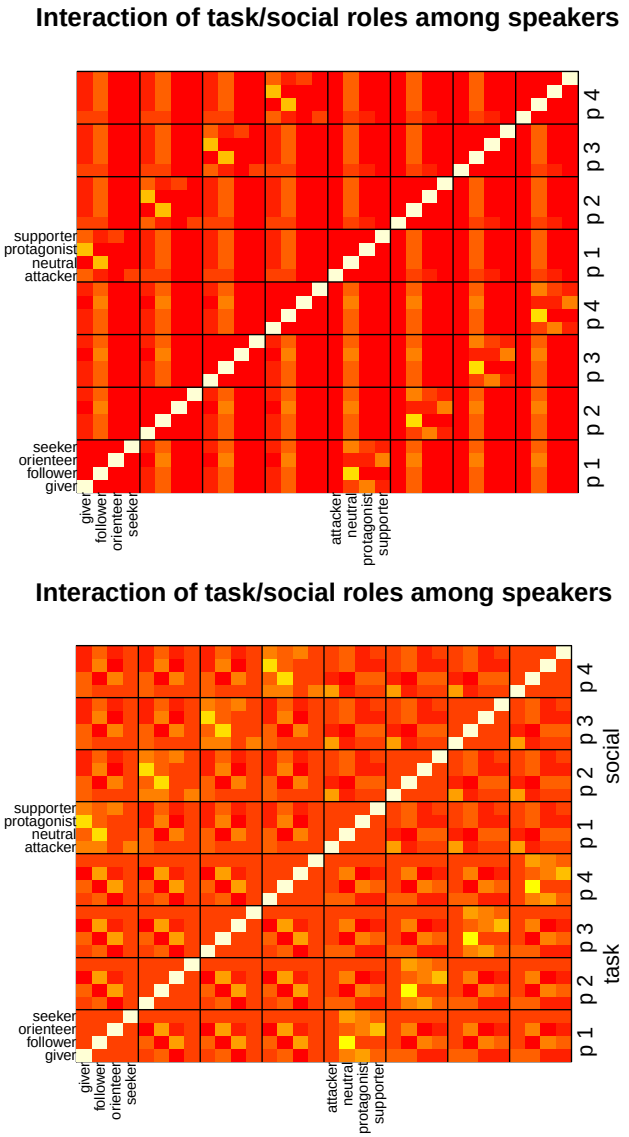

Figure 2. Influence matrices of the IM (above) and of the newIM (below). An entry at row x and column y indicates how large the state corresponding to row x is indicative of the state corresponding to column y (brighter color, e.g. yellow, means more indicative and darker color, e.g. red, means less indicative).

while they are often significantly positive; see Table IX. This is in sharp contrast with the behavior of the two competing classifiers: SVM has scores significantly lower on the less-populated classes of the task area, while HMM does the same for both the task and the socio-emotional areas. The results with HMM also show that the mere consideration of roles' temporal properties at the individual level may not suffice. Although further studies with different corpora are needed in order to reach firmer conclusions, these results are at least indicative of the superiority of Influence Models in capturing more of the social functional roles' structure and dynamics, and using them for the sake of classification.

There are slight differences in the behavior of the two Influence Models exploited in this paper that deserve some discussion: 1) newIM yields a higher overall precision (0.63) than IM (0.52) against a substantial balance on recall (0.60 for newIM and 0.62 for IM); and 2) newIM and IM are somewhat complementary, the first doing best in the socio-emotional area and the second in the task area. The difference between the two Influence Models consists of the fact that, whereas the basic version IM assumes a fixed amount of influence that a node exerts on another, newIM fixes only an upper limit, allowing the actual amount of influence to be adjusted in the training process. It seems that this greater flexibility allows newIM to capture the mutual constraints among the roles of different people in a better way, maintaining very high precision rates in the less-populated classes of the socio-emotional area without losing much in terms of recall.

The better ability of the newIM to capture the mutual constraints among the roles played by different people is also confirmed by comparing the trained influence matrix of the newIM with the one of the standard IM (see Figure 2). In these matrices, an entry at row x and column y indicates how the state corresponding to row x is indicative of the state corresponding to column y. As shown in the plots, the influence matrix of the newIM captures in a more clear way the interactions captured by the influence matrix of the standard IM (a large number of brighter entries are shown in the plot of the newIM). Moreover, a larger number of interactions is discernible in the influence matrix of newIM: for instance, this matrix is sensitive to the interactions between a giver role and a follower role and to the interactions between attacker roles.

IX. Acknowledgment

This work was partially supported by the UE under the CHIL (FP6) project. Bruno Lepri's research is funded by PERSI project inside the Marie Curie COFUND Ð 7th Framework.

References

[1] M. Doyle and D. Straus, *How to make meetings work : the new interaction method*. New York: Berkley Publishing Group, 1982,1976.

[2] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal support to group dynamics," *Personal Ubiquitous Comput.*, vol. 12, pp. 181–195, January 2008.

[3] D. Katz and R. L. Kahn, *The social psychology of organizations*. New York: Wiley, 2d ed. ed., 1978.

[4] J. E. McGrath, *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[5] B. J. Biddle, *Role theory : expectations, identities, and behaviors*. New York: Academic Press, 1979.

[6] A. Salazar, "An analysis of the development and evolution of roles in the small group," *Small Group Research*, vol. 27, no. 4, pp. 475–503, 1996.

[7] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vision Comput.*, vol. 27, no. 12, pp. 1775–1787, 2009.

[8] A. Vinciarelli and S. Favre, "Broadcast news story segmentation using social network analysis and hidden markov models," in *ACM Multimedia* (R. Lienhart, A. R. Prasad, A. Hanjalic, S. Choi, B. P. Bailey, and N. Sebe, eds.), pp. 261–264, ACM, 2007.

[9] L. Matena, A. Jaimes, and A. Popescu-Belis, "Graphical representation of meetings on mobile devices," in *Mobile HCI* (G. H. ter Hofte, I. Mulder, and B. E. R. de Ruyter, eds.), ACM International Conference Proceeding Series, pp. 503–506, ACM, 2008.

[10] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," in *INTERSPEECH*, ISCA, 2003.

[11] A. Vinciarelli, "Sociometry based multiparty audio recordings segmentation," in *ICME*, pp. 1801–1804, IEEE, 2006.

[12] C. Y. Weng, W. T. Chu, and J. L. Wu, "Movie analysis based on roles' social network," in *Proceedings of IEEE International Conference on Multimedia & Expo*, pp. 1403–1406, 2007.

[13] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *ICMI*, pp. 28–34, 2006.

[14] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *ICMI*, pp. 271–278, 2007.

[15] W. Dong and A. Pentland, "Modeling influence between experts," *Artificial Intelligence for Human Computing*, vol. 4451, no. 170-189, 2007.

[16] A. P. Hare, "Types of roles in small groups — a bit of history and a current perspective," *Small Group Research*, vol. 25, no. 3, pp. 433–448, 1994.

[17] K. D. Benne and P. Sheats, "Functional roles of group members," *Journal of Social Issues*, vol. 4, pp. 41–49, 1948.

[18] R. F. Bales, *Interaction Process Analysis: a Method for the Study of Small Groups*. Addison-Wesley Press, 1950.

[19] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *Proceedings of AAAI/IAAI*, pp. 679–684, 2000.

[20] S. Banerjee and A. I. Rudnicky, "Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants," in *Proceedings of the 8th International Conference on Spoken Languate Processing*, vol. 2189-2192, 2004.

[21] A. Vinciarelli, "Role recognition in broadcast news using bernoulli distributions," in *Proceedings of IEEE International Conference on Multimedia & Expo*, pp. 1551–1554, 2007.

[22] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliation networks and discrete distributions," in *ICMI*, pp. 29–36, 2008.

[23] H. Salamin, A. Vinciarelli, K. Truong, and G. Mohammadi, "Automatic role recognition based on conversational and prosodic behaviour," in *ACM Multimedia* (A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, eds.), pp. 847–850, ACM, 2010.

[24] S. Favre, A. Dielmann, and A. Vinciarelli, "Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models," in *ACM Multimedia*, pp. 585–588, 2009.

[25] N. P. Garg, S. Favre, H. Salamin, D. Z. Hakkani-Tür, and A. Vinciarelli, "Role recognition for meeting participants: an approach based on lexical information and social network analysis," in *ACM Multimedia*, pp. 693–696, 2008.

[26] D. B. Jayagopi, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *ICMI*, pp. 45–52, 2008.

[27] B. Raducanu, J. Vitrià, and D. Gatica-Perez, "You are fired! nonverbal role analysis in competitive meetings," in *ICASSP*, pp. 1949–1952, IEEE, 2009.

[28] T. Wilson and G. Hofer, "Using linguistic and vocal expressiveness in social role recognition," in *IUI* (P. Pu, M. J. Pazzani, E. André, and D. Riecken, eds.), pp. 419–422, ACM, 2011.

[29] A. Vinciarelli, F. Valente, S. H. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus," in *SMC*, pp. 374–379, IEEE, 2011.

[30] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of concensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2008.

[31] J. A. Cohen, "Coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[33] J. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes," *Journal of Applied Psychology*, vol. 92, no. 3, pp. 802–811, 2007.

[34] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi, "Automatic prediction of individual performance from "thin slices" of social behavior," in *ACM Multimedia* (W. Gao, Y. Rui, A. Hanjalic, C. Xu, E. G. Steinbach, A. El-Saddik, and M. X. Zhou, eds.), pp. 733–736, ACM, 2009.

[35] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," in *UMAP* (G.-J. Houben, G. I. McCalla, F. Pianesi, and M. Zancanaro, eds.), vol. 5535 of *Lecture Notes in Computer Science*, pp. 114–125, Springer, 2009.

[36] B. Lepri, S. Ramanathan, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Employing social gaze and speaking activity for automatic determination of the *extraversion* trait," in *ICMI-MLMI* (W. Gao, C.-H. Lee, J. Yang, X. Chen, M. Eskenazi, and Z. Zhang, eds.), p. 7, ACM, 2010.

[37] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[38] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, In press.

[39] S. Basu, *Conversational Scene Analysis*. PhD thesis, MIT, 2002.

[40] V. Yngve, "On getting a word in edgewise," in *6th Regional Meeting of the Chicago Linguistic Society*, pp. 567–577, 1970.

[41] G. Carli and R. Gretter, "A start-end point detection algorithm for a real-time acoustic front-end based on dsp32c vme board," in *In Proceedings of ICSPAT*, pp. 1011–1017, 1992.

[42] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[43] "The american heritage® dictionary of the english language, fourth edition," Apr 2012.

[44] E. A. Heerey and A. M. Kring, "Interpresonal consequences of social anxiety," *Journal of Abnormal Psychology*, vol. 116, no. 1, pp. 125–134, 2007.

[45] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 39–46, 2001.

[46] P. Chippendale, "Towards automatic body language annotation," in *Proceedings of the 7th International conference on Automatic Face and Gesture Recognition*, pp. 487–492, 2006.

[47] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *ICMI* (V. Digalakis, A. Potamianos, M. Turk, R. Pieraccini, and Y. Ivanov, eds.), pp. 53–60, ACM, 2008.

[48] J. W. Hardin and J. Hilbe, *Generalized estimating equations*. Boca Raton, Fla.: Chapman & Hall/CRC, 2003.

[49] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," tech. rep., MIT Media Laboratory Vision & Modeling Technical Report #539, 2001.

[50] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed-memory markov process," in *NIPS*, 2004.

[51] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Journal of Personal and Ubiquitous Computing*, 2005.

[52] W. Dong and A. Pentland, "A network analysis of road traffic with vehicle tracking data," in *AAAI Spring Symposium 2009 on Human Behavior Modeling*, 2009.

[53] W. Pan, W. Dong, M. Cebrian, T. Kim, J. Fowler, and A. Pentland, "Modeling dynamical influence in human interactions: Using data to make better inferences about influence within social systems," in *IEEE Signal Processing Magazine* (29, ed.), vol. 2, pp. 77–86, March 2012.

[54] B. Caputo, I. Sim, F. Furesjo, and A. Smola, "Appearance-based object recognition using SVMs: Which kernel should i use?," in *Proceedings of NIPS workshop on statistical methods for computational experiments in visual processing and computer vision*, 2002.

Wen Dong focuses on modeling human interaction dynamics with stochastic process theory. He has published dozens of papers on the network influence model, group dynamics, and human problem-solving, and has earned recognition from several big businesses for constructing successful data products. Wen received his Ph.D. from the MIT Media Laboratory, with the thesis Modeling the Structure of Collective Intelligence.

Bruno Lepri received the PhD degree at the Department of Information Engineering and Computer Science in 2009, from the Trento University. He is currently a Marie Curie post-doctoral fellow in MITs Human Dynamics Laboratory, Cambridge, US, and in FBK research centre, Trento, Italy. His research interests are human behavior understanding, social signal processing, computational social science, social network analysis, and multimodal interaction.

Fabio Pianesi is the Vice Director for Research of Trento RISE, a joint partnership between FBK and the University of Trento, and Co-location Manager of the Trento Node of EIT ICT Labs. In the past, he served as head of the Cognitive and Communication Technologies Division of ITC-irst and as head of Computational Cognition Laboratory of FBK. Fabio has been very active in many of the disciplines that contribute to human computing, including computational and formal linguistics, human computer interaction, multimodal interaction, social signal processing and human behavior understanding. He is the chair of the Advisory Board of ACM-ICMI (the International Conference on Multimodal Interaction).

Alex ‘Sandy’ Pentland directs MIT’s Human Dynamics Laboratory and the MIT Media Lab Entrepreneurship Program, and advises the World Economic Forum, Nissan Motor Corporation, and a variety of start-up firms. He has previously helped create and direct MIT’s Media Laboratory, the Media Lab Asia laboratories at the Indian Institutes of Technology, and Strong Hospital’s Center for Future Health. Sandy is among the most-cited computational scientists in the world, and a pioneer in computational social science, organizational engineering, mobile computing, image understanding, and modern biometrics. His research has been featured in Nature, Science, the World Economic Forum, and Harvard Business Review, as well as being the focus of TV features including Nova and Scientific American Frontiers. His most recent book is ‘Honest Signals,’ published by MIT Press.