Modeling Group Discussion Dynamics

Wen Dong, Ankur Mani, Alex Pentland
20 Ames Street, E15-383
Cambridge, MA 02142
{wdong,amani,sandy}@media.mit.edu

Bruno Lepri, Fabio Pianesi
FBK-Irst
Via Sommarive 38050 Povo-Trento, Italy
{lepri, pianesi}@fbk.eu

***Abstract*—In this paper, we present a formal model of group discussion dyanmics. An understanding of the face-to-face communications in a group discussion can provide new clues about how humans collaborate to accomplish complex tasks and how the collaboration protocols can be learned. It can also help us to evaluate and facilitate brainstorming sessions. We will discuss the following three findings about the dynamics: Meetings in different languages and on different topics could follow the same form of dynamics; The functional roles of the meeting participants could be better understood by inspecting not only their individual speaking and activity features but also their interactions with each other; The outcome of a meeting could be predicted by inspecting how its participants interact.**

I. Introduction

THE focus of the paper is about the *form* of the group discussion process in which participants in the group present their opinions, argue on an issue, and try to reach a consensus. The dynamics of group discussion depends on the meeting participants' tendencies to maximize the outcome of the discussion. People "instinctively" know how to cope with each other in many different common situations to make an effective discussion. As a result, we could expect some invariant structures from one discussion to another, and could by just watching the discussion dynamics answer the following four questions: (1) how the opinions of the meeting participants differ from each other, (2) what are the psychological profiles of the participants, (3) how the discussion progress, and (4) whether the discussion is effective. We do not concern ourselves with the content of the discussion, and thus we cannot answer questions concerning the content, yet we can still tell a lot of things from the just the form (i.e., the container).

A. A Thought Experiment

Let us abstractly inspect by a thought experiment an imaginary group discussion involving the participants putting together their opinions on an open problem in order to reach a consensus. The quantitative analysis corresponding to the following description will be given in the following sections.

A discussion is normally driven by one person at a time, and it can be driven by different persons at different times. The person who drives the discussion normally has longer uninterrupted speaking time, steadier intonation and speaking speed, and more attention from the other persons who normally show their attentions by turning their bodies to and looking at the speaker.

The persons who do not drive the discussion (listeners) can from time to time and individually support the one who drives the discussion (turn-taker) by briefly showing their agreement or by briefly adding in supporting materials to the turn-taker's argument. The listeners can sometimes request clarifications on the turn-taker's argument, and the requested clarification can subsequently be provided by either the turn-taker or the other listeners.

One or more listeners can infrequently show their disagreement with the turn-taker's opinion/argument and initiate an "attack", which may consequently pull more listeners into the "battle". The intensity of the battle is indicated by the significantly less body/hand movement of the person who initiate the attack, the significantly more body/hand movement of the others in response of the attack-initiator (who speak and turn to each other), and the large number of simultaneous speakers.

The turn taker and a listener can from time to time engage in a series of back-and-forth negotiation to fill the gap between their understandings or opinions. If the negotiation takes too long, the other discussion participants will jump in and terminate the negotiation. When the turn-taker finishes his turn, he can either simply stops speaking or explicitly hands over his turn to a listener, and the next turn-taker will continue to drive the discussion appropriately.

In many of the discussions, there is a distinctive Orienteer, who has the "charisma" to drive the discussion on when it comes into a halt or a chaos. The charisma is reflected by the capability of the Orienteer to quickly seize the attention of the others: When the Orienteer takes on the orientation role, all other speakers quickly turn their body towards the Orienteer, and the other existing (normally multiple) speakers quickly stops speaking.

B. Influence Modeling

What we have just described can be expressed as an "influence model", in which each participant randomly chooses to maintain his role (e.g., turn-taker, supporter, attacker, Orienteer) for some duration or chooses to make a transition to another role depending on each other's roles: The duration that a turn-taker drives the discussion depends among many factors on the option and arguments of the turn-taker, his styles of him and the responses of the other participants. When one person expresses his opinion and arguments, the other persons normally listen to his statement attentively and patiently, and they show their agreements and doubts unobtrusively. The transition from one turn-taker to another depends on how the latter's opinion is related to the former's and how the latter wants to drive the discussion. The way and the likelihood for a participant to express his support, doubt or disagreement depend both on his judgment about the importance of making an utterance and on his personal style.

An observer who watches the group discussion dynamics — the turn-takers at each time, the transitions of the turns, the responses of the listeners and the dyadic back-and-forth negotiations — can often get a precise understanding of them (the testing data set) by pattern-matching them with the past group discussion dynamics (the training data set) stored in his memory. The reason is that the multi-person face-to-face interactions normally take on a small number of regular patterns among a huge number of possibilities. For example, if each of the four participants in a discussion can take one of the four roles — protagonist, attacker, supporter and neutral, there will be $4^4 = 256$ role combinations. However, in an efficient group discussion, only a few combinations exist most of the time. A corollary of the regularity of face-to-face interactions is that we can evaluate the effectiveness of a group discussion by inspecting how likely its interaction pattern is an efficient one.

Since the dynamics of a discussion is dependent on the purpose of it, we can either imagine the characteristics of an efficient discussion of a certain purpose, or compute the characteristics based on simplified mathematical models representative of the discussion purpose. We can also use our intuition either to guide our experimental designs or to help interpreting the experimental results. In specifics, since we normally get a more comprehensive range of perspectives by listening to more people in an open problem and since we can normally pay attention to a single person at a time, we could imagine that most part of an effective discussion is driven by a single person. Since discussing a topic normally requires a considerable amount of set-up time and summarizing time, we shouldn't see frequent transitions among topics. The topics can be separated from each other by different amount of participation from the individuals and different interaction dynamics. Since a back-and-forth dyadic negotiation normally involves the interests and attention of only two individuals, it shouldn't last long generally in an effective group discussion. Thus the effectiveness of a group discussion could be studied using stochastic process models and statistical learning methods.

Different group-discussion purposes require different types of dynamics, yet there are invariants in interpersonal communications: The cognitive loads of individuals has a statistical distribution; different types of turn-taking dynamics statistically result in different performances conditioned on the meeting purposes and the individual parameters; Similar kinds of group-discussion issues such as blocking and social loafing [1], [2] may exist in different types of discussions. As a result, while we use the same stochastic model to fit all group-discussions, different purposes may require different parameters. We should take special care of the compatibility of two group discussions when we fit a dynamic model to the former with appropriate parameters and apply the fitted model to the latter.

C. Plan for the Paper

The current paper models the group discussion dynamics as interacting stochastic processes, with each process representing a participant. The paper also identifies the different functional roles that the participants take at each time in a group discussion and evaluates the discussion efficiency within the framework of the stochastic process. The rest of the paper is organized in the following way. In section II, we review the previous work in understanding group dynamics, *influences*, meeting progression, and the effectiveness of a meeting. The study on the non-verbal aspects of a face-to-face group discussion is not a new one, yet our approach gets better accuracy in estimating the participants' functional roles and the discussion outcome than the previous ones by taking into account the interaction features. In Section III, we describe several data sets and give our new results on their interaction statistics. The statistics both motivate our new formulation of *influence* and provide intuitions on why the new formulation could give better estimation results. The old and new formulations on *influence* are compared in Section V. The estimation results on functional roles and discussion outcomes are given and analyzed in Section VI, and comparisons with our previous results are given when the latter exist. We conclude this paper by briefly describing the experiences and lessons we have learned in our efforts to understand the non-verbal aspects of group discussion, as well as future directions in our opinions.

II. Literature Review

Our main concern in this paper is the automated recognition of the turn-taking dynamics in a discussion. We hope to draw a connection between the turn-taking dynamics on the one hand and the discussion performance on the other hand. In this section, we will review the work that we know of about the discussion dynamics and the discussion performance.

Various approaches have been applied to detect the roles in a news bulletin based on the distinctive characteristics of those roles. Vinciarelli [3], [4] used Bayesian methods to recognize the anchorman, the second anchorman, the guest and the other roles based on how much they speak, when in the bulletin they speak (beginning, middle, or end), and after who they speak. The same social network analysis

(SNA) idea was adopted by Weng et al. [5] 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. [6] exploited the keywords used by the roles, the durations of the roles' speaking turns and the explicit speaker introduction segments in the identification of the anchor, the journalists and the guest speakers in a radio program.

Different meeting states and roles have been defined, and their characteristics and estimation algorithms have been studied: Banerjee and Rudnicky [7] defined three meeting states (discussion, presentation and briefing) and correspondingly four roles (discussion participators, presenter, information provider, and information giver). They subsequently used the C4.5 algorithm to estimate the meeting states and the roles based on four features (number of speaker turns, number of participants spoken, number of overlaps, and average length of overlaps). McCowan et al. [8] developed a statistical framework based on different Hidden Markov Models to recognize the sequences of group actions starting from audio-visual features concerning individuals' activities—e.g., "discussion", as a group action recognizable from the verbal activity of individuals. Garg et al. [9] discussed the recognition of the project manager, the marketing expert, the user interface expert and the industrial designer in a simulated discussion on the development of a new remote control. His recognizer is based on when the participants speak and what keywords the participants use.

Dominance detection aroused much interest perhaps because the dominant person is believed to have large influence on a meeting's outcome. Rienks et al. [10], [11] used various static and temporal models to estimate the dominance of the participants in a meeting, and concluded that the automated estimation is compatible with the human estimation. The features they used include several nonverbal — e.g., speaker turns, floor grabs, speaking length — and verbal — e.g., number of spoken words used by the next speaker — audio features retrieved from the discussion transcription. Jayagopi et al. [12], [13], [14] extended the work of Rienks et al. and estimated dominance using features directly computed from the audio and video recordings — e.g., total speaking energy, total number of times being unsuccessfully interrupted.

The historical work on social psychology, especially that related to the structures and the performances of small group discussions, provides useful observations, insights, and challenges for us to work on with automated computer algorithms. In particular: Conversation and discourse analysis provide helpful observations and examples [15], [16], [17], [18], so that the features and structures of conversational group processes and be figured out by experiments and simulations. Bales investigated the phases (e.g., giving opinion, showing disagreements, asking for suggestion) and the performances of group discussions, as well as the different roles that the discussion participants play [19], [20], [21]. McGrath on the other hand inspected meetings based on their different tasks [22], [23]. The usefulness of group brainstorming has been widely argued [24], [17], and production blocking and social loafing have been identified as two drawbacks of group brainstorming [1], [25], [26]. Hall [27] and Wilson [28] systematically analyzed their respective group brainstorming experiments, and answered why a group in their respective cases could outperform its individuals.

The work related to the Mission Survival corpus that we discuss in this paper includes the following. Identifying functional relational roles (social and task roles) were addressed by Zancanaro et al. [29], [30] through an SVM that exploited speech activity (whether a participant is speaking at a given time) and the fidgeting of each participant in a time window. Dong [31] extended this work by comparing SVM-based approach to HMM- and IM-based approaches. Pianesi et al. [32] have exploited social behavior to get at individual characteristics, such as personality traits. The task consisted in a three-way classification (low, medium, high) of the participants' levels in extroversion and locus of control, using speech features that are provably honest signals for social behavior, and visual fidgeting features.

According to our knowledge, the current paper is the first to discuss the features and the modeling issues of the turn-taking behavior and the personal styles in an unconstrained group discussion that can be extracted with computer algorithms from the audio and video recordings. The paper also gives our initial findings on the correlation between the discussion turn-taking behavior and discussion performance. Our discussion is based on *Mission Survival Corpus I*. The difficulty in the current work is that we are studying an unconstrained group discussion. Thus there aren't any pre-defined agenda and keywords such as in a news bulletin to exploit, nor are there any visual cues such as a whiteboard or a projector screen. A person who is dominant in one part of a discussion may be non-dominant in another part. We will nevertheless show that, although the predefined macro-structure does not exist in an unconstrained discussion, the micro-structures at different parts of the discussion are based on the instantaneous roles of the meeting participants, and the statistics about the micro-structures are related to the discussion performance.

The influence modeling that we use in this paper captures interactions and temporal coherence at the same time, and it has a long history of development. The coupled hidden Markov models was first development to capture the interactions and temporal coherence of two parts based on audio and visual features [33], [34], [35]. Asavathiratham introduced the influence model to study the asymptotic behavior of many individual power plants in a network [36], [37]. The approximation use by Asavathiratham is that the probability measure of a power plant's state is a linear functional of the probability measures of all power plants' states in the network. The similar idea was exploited by Saul and Boyen [38], [39]. Choudhurry noted that individuals have their characteristic styles in two-person face-to-face conversations, and the overall style of a two-person face-to-face conversation looks more like the style of the more influential person. Choudhurry et al. subsequently used the influence modeling to study the structures of discussions and organizations [40], [41], [42]. Dong developed several versions of multi-agent dynamic Bayesian networks using the same name which are better fitted with the probability measures of group processes [43], [44].

III. The Mission Survival Corpus I

For the experiments discussed in this paper, we have used the Mission Survival Corpus [30], a multimodal annotated corpus based on the audio and the video recordings of eight meetings that took place in a lab setting appropriately equipped with cameras and microphones. Each meeting consisted of four people engaged in the solution of the "mission survival task". This task is frequently used in experimental and social psychology to elicit decision-making processes in small groups. Originally designed by National Aeronautics and Space Administration (NASA) to train astronauts, the Survival Task proved to be a good indicator of group decision making processes [27]. The exercise consists in promoting group discussion by asking participants to reach a consensus on how to survive in a disaster scenario, like moon landing or a plane crash in Canada. The group has to rank a number (usually 15) of items according to their importance for crew members to survive. In our setting, we used the plane crash version. This consensus decision making scenario was chosen for the purpose of meeting dynamics analysis mainly because of the intensive engagement requested to groups in order to reach a mutual agreement, thus offering the possibility to observe a large set of social dynamics and attitudes. In our setting, we retained the basic structure of the Survival Task with minor adjustments: a) the task was competitive across groups/team, with a price being awarded to the group providing the best survival kit. b) the task was collaborative and based on consensus within the group, meaning that a participant's proposal became part of the common sorted list only if he/she managed to convince the other of the validity of his/her proposal.

The recording equipment consisted of five FireWire cameras—four placed on the four corners of the room and one directly above the table— and four web cameras installed on the walls surrounding the table. Speech activity was recorded using four close-talk microphones, six tabletop microphones and seven T-shaped microphone arrays, each consisting of four omni-directional microphones installed on the four walls in order to obtain an optimal coverage of the environment for speaker localization and tracking. Each session was automatically segmented labeling the speech activity recorded by the close-talk microphones every 330ms [45]. The fidgeting—the amount of energy in a person's body and hands—was automatically tracked by using skin region features and temporal motion [46]. The values of fidgeting for hands and body were extracted for each participant and normalized on the fidgeting activity of the person during the entire meeting.

Fig. 1. A picture of the experimental setting.

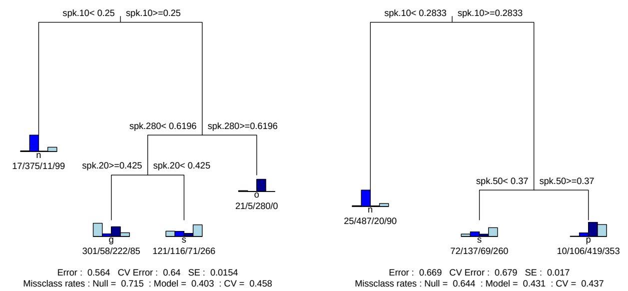
The Functional Role Coding Scheme (FRCS) was partially inspired by Bales' Interaction Process Analysis [20]. It consists of ten labels that identify the behavior of each participant in two complementary areas: the Task Area, which includes functional roles related to facilitation and coordination tasks as well as to technical experience of members; the Socio Emotional Area, which is concerned with the relationships between group members and the functional roles "oriented toward the functioning of the group as a group". Below we give a synthetic description of the FRCS (for more information, see [30]. The Task Area functional roles consist of: the Orienteer (o),who orients the group by introducing the agenda, defining goals and procedures, keeping the group focused and on track and summarizing the most important arguments and the group decisions; the Giver (g), who provides factual information and answers to questions, states her beliefs and attitudes about an idea, and expresses personal values and factual information; the Seeker (s), who requests information, as well as clarifications, to promote effective group decisions; the Procedural Technician (pt), who uses the resources available to the group, managing them for the sake of the group; the follower (f), who just listens, without actively participating in the interaction. The Socio-Emotional functional roles consist of: the Attacker (a), who deflates the status of others, expresses disapproval, and attacks the group or the problem; the Gate-keeper (gk), who is the group moderator, mediates the communicative relations, encourages and facilitates the participation and regulates the flow of communication; the Protagonist (p), who takes the floor, driving the conversation, assuming a personal perspective and asserting her authority; the Supporter (su), who shows a cooperative attitude demonstrating understanding, attention and acceptance as well as providing technical and relational support; the Neutral Role (n), played by those who passively accept the ideas of the others, serving as an audience in group discussion. Of course, participants may—and often do—play different roles during the meeting, but at a given time each of them plays exactly one role in the Task Area and one role in the Socio-Emotional one. The FCRS was showed to have a high inter-rater reliability (Cohen's statistics $\kappa = 0.70$ for the Task Area; $\kappa = 0.60$ for the Socio-Emotional Area).

IV. Social Signals

An individual in a group discussion has his characteristic style on the frequency, the durations and the functional (i.e., task and socio-emotional) roles of his speaking turns. In *Mission Survival Corpus I*, some individuals take certain functional roles consistently often, while some other individuals take these roles consistently rarely. The functional roles have their respective characteristics, durations in particular, and interactions with other functional roles, independent of who take them. As a result, the functional roles of a speaker turn can be inferred from the characteristics of the turn and the characteristics of the turn taker.

Fig.2 gives the decision trees that tell a meeting participant's functional roles at a specified moment only from the amounts of time he speaks in the time windows of different sizes around the moment. The C4.5 algorithm is used to generate the decision trees from four discussions of *Mission Survival Corpus I* as training data, and it correctly captures the characteristics of the functional roles: An information giver speaks more than an information seeker in a short time window, a protagonist speaks more than a supporter in a long time window, and a neutral role (i.e., a listener or a follower) speaks much less than the other roles in time windows of up to several minutes. The C4.5 algorithm, like many other modern statistical learning algorithms, is guarded against overfitting by a mechanism. The trained decisions trees can attain an accuracy of around 55%. (As a comparison, the inter-rater reliability has Cohen's statistics $k = .70$ for the Task Area and $\kappa = 0.60$ for the Socio-Emotional area.) Further accuracy can be achieved by considering the speaker characteristics and more functional role characteristics: Since the participant who spends more time in giving information often spends more time in seeking information ($R^2 = .27$, $F = 12.4$ on 1 and 30 degrees of freedom, $p = .0014$), the total amount of time that a participant has spent in giving information can be used to determine whether a short speaking turn of his corresponds to a seeker role or a neutral role; Due to the way that an auxiliary role such as seeker/supporter and a major role such as giver/protagonist co-occur, the amounts of speaking time of a participant in time windows of different sizes can be contrasted with those of the other participants to disambiguate the roles of the participant; Since an attacker is relatively quiet by himself and arouses significant agitation from the others, and since a neutral role is often less paid attention to, the intensities of hand/body movements can be taken as the characteristics of those roles.

The current section is organized into two subsections. In Subsection IV-A, we analyze the durations of each functional roles and the likelihood that different functional roles co-occur. In Subsection IV-B, we analyze who is more likely to take which roles from their individual honest signals.



(a) Task roles. (b) Socio-Emotional roles.

Fig. 2. Decision trees trained with the C4.5 algorithm for functional role detection: If a person takes the Neutral/Follower role at a moment, he speaks noticeably less in the 10-second window around the moment; The Giver speaks more than the Seeker in a 20-second window; The Protagonist speaks more than the Supporter in a 50-second window; The Orienteer on average speaks 62% of time in a 280-second window; The speaking/non-speaking signal seems to be insufficient to detect Attacker.

A. Turn-taking behavior

The patterns in the functional roles, social signaling and turn taking behavior, and their relations are given as follows. Any effective heuristics and statistical learning methods that model the group discussion behavior should exploit these patterns.

We will first inspect the (Task Area and Socio-Emotional Area) role assignments of the subjects in *Mission Survival Corpus I*. The role assignments reflect how the observers understand the group processes.

Table I gives the durations in seconds of social roles, task roles and their combinations. In this table, an instance of a supporter role has a significantly less average duration than that of a protagonist role (15 vs. 26). This coincides with the fact that a protagonist is the main role to drive the conversation and a supporter takes a secondary importance. An attacker role takes an average duration of 9 seconds, which is equivalent to 10˜20 words and around one sentence assuming conversations are around 100˜200 words per minute. This reflects a person's strategy to show his contrasting ideas concisely, so that he can make constructive utterances and avoid conflicts at the same time. A person asks questions (when he takes an information-seeker's role) more shortly than he provides information (when he takes an information-giver's role). This reflects our natural tendency to make task-oriented discussions more information-rich and productive. A protagonist role is on average 37% longer. This indicates the fact that the social roles happen on a different time scale compared with the task roles, and they are not totally correlated statistically: A discussion is normally driven by one person and thus has a single protagonist at a time. The protagonist can ask another person questions, and the latter generally gives the requested information briefly, avoiding assuming the speaker's role for too long.

TABLE I
Durations in seconds of social roles, task roles and their combinations. Each table entry $\mu(\sigma)$ gives the mean and standard deviation for a specific case.

$\mu(\sigma)$	a	n	p	s	marginal
g	8(6)	10(16)	23(24)	11(7)	**19**(20)
n	2(2)	52(79)	4(6)	5(5)	34(45)
o	N/A	4(8)	10(9)	18(16)	17(14)
s	7(4)	6(4)	9(7)	10(5)	**9**(5)
marginal	9(4)	56(85)	**26**(27)	**15**(14)	7(50)

TABLE II
Number of instances and amount of time in seconds that a person takes a task role, a social-role and a task/social role combo.

	a	n	p	s	total
g	5(39)	316(3k)	233(5k)	112(1k)	666(9k)
n	9(22)	426(22k)	185(747)	147(718)	767(24k)
o	0(0)	67(323)	21(432)	53(1k)	141(2k)
s	5(36)	74(471)	27(253)	17(170)	123(930)
total	19(97)	883(26k)	466(6k)	329(3k)	1697(36k)

TABLE III
Average percentage of speaking time in 10-second windows (spk) around different social roles and task roles, average body movement (bdy) and hand movement (hnd) of the self (self) and the others (othr) in the 10-second windows around the shifts into different social roles. Each table entry $\mu(\sigma)$ gives the mean and standard deviation for a feature-role combination.

$\mu(\sigma)$		Social-Emotional Area Roles			
		Attacker	Neutral	Protagonist	Supporter
features	self spk	.47(.16)	.20(.21)	.62(.21)	.54(.21)
	othr spk	.30(.09)	.32(.12)	.25(.14)	.25(.15)
	self hnd	11(14)	18(21)	18(21)	16(19)
	othr hnd	20(14)	16(13)	19(14)	17(12)
	self bdy	**11**(19)	20(22)	21(22)	18(20)
	othr bdy	**23**(14)	19(14)	22(14)	19(14)

$\mu(\sigma)$		Task Area roles			
		Giver	Follower	Orienteer	Seeker
features	self spk	.58(.21)	.16(.18)	**.62**(.21)	.41(.19)
	othr spk	.25(.14)	.33(.12)	**.23**(.14)	.30(.12)
	self hnd	17(21)	18(20)	17(20)	15(17)
	othr hnd	19(14)	16(13)	18(12)	18(14)
	self bdy	19(22)	20(22)	20(20)	18(21)
	othr bdy	21(14)	18(13)	21(14)	19(14)

The protagonist can seldom be interrupted by questions, and the questioner generally seeks additional information in a brief and collaborative way. The durations of the neutral roles in the task role "dimension" and the social role "dimension" are less than twice the durations of the giver's role and the protagonist's role respectively. This indicates that the participants do not passively listen when they take listeners' roles.

Table II shows the number of speaking-turns and the amount of time that the meeting participants take different task roles, social roles and combinations of task and social roles. This table complements Table I on how individual participants take different functional roles in a discussion. The group process can also be viewed as a Markov process with different distributions of functional roles at different time: In *Mission Survival Corpus I*, the configuration *1g3n0o0s*, which denotes the configuration of the discussion with 1 Giver, 3 Neutrals, 0 Orienteer, and 0 Seeker, takes 36% discussion time, and the configurations *2g2n0o0s*, *0g3n1o0s*, *0g4n0o0s*, *1g2n0o1s*, *3g1n0o0s*, *1g2n1o0s*, *0g3n0o1s* and *2g1n0o1s* take 20%, 13%, 11%, 5%, 5%, 4%, 2% and 1% discussion time respectively. In the same data set, the different socio-emotional role distributions *0a3n1p0s*, *0a4n0p0s*, *0a3n0p1s*, *0a2n2p0s*, *0a2n1p1s*, *0a2n0p2s*, *0a1n2p1s*, *0a1n3p0s* and *0a1n1p2s* take 36%, 21%, 18%, 11%, 7%, 3%, 1%, 1% and 1% discussion time respectively. Each different role distribution tends to last for some duration. Considering a group process in terms of role distribution makes the group process model speaker-independent, and thus effectively compresses the number of states of the group process. Since we can model a group process in a better way in terms of "influence", we will not discuss the distribution of roles any further.

We will proceed to analyze the turn-taking behavior, the body movements, and the hand movements corresponding to different roles. The analysis will show that the roles reflect a set of essential features of the group processes, rather than being artificially imposed to the group processes.

In a discussion involving multiple persons, the individuals normally orient their bodies to the locus of the discussion, which can be the protagonist or the information giver. On the other hand, the protagonist and the information giver normally make non-verbal communications with the listeners by turning his body to them. The attention shifts consist of a significant fraction of hand and body movement. In the mission survival data set, the correlation between the change-of-speaker and the body/hand movement intensity is greater than 0.50.

Table III shows how the meeting participants execute their Task Area Roles and Socio-Emotional Area Roles in terms of how much to speak and to whom should the attentions be given. While the patterns are weak and might not be sufficient for constructing good role classifiers, they nevertheless exist and coincide with our intuition: An attacker provokes a significant amount of attention, hand movements and body movements from the others, while he shows significant less hand and body movements. The neutral roles, the supporter role and the seeker role attract less average attention from the others compared with the giver role and the protagonist role. In the 10 second window when a person takes either a supporter role or a seeker role, he has less hand and body movements. This may be due to the fact that he has already paid attention to the locus of the discussion when he takes those roles. When a person takes an Orienteer role, on average only 23% of the time in the 10-second window do the other three participants speak, and the Orienteer speaks 62% of the time in this window. This indicates that the one task of an Orienteer is to keep the brainstorming in track.

Table IV shows how the meeting participants shift their

TABLE IV
DISTRIBUTION OF SOCIAL ROLES AND TASK ROLES CONDITIONED ON NUMBER OF SIMULTANEOUS SPEAKERS.

		Socio-Emotional Area Roles				
		Attacker	Neutral	Protagonist	Supporter	$\sum$
#. of spks	0	**.001**	.817	.104	.078	15k(1.0)
	1	**.002**	.740	.177	.081	60k(1.0)
	2	**.004**	.680	.220	.096	26k(1.0)
	3	**.005**	.620	.238	.137	6k(1.0)
	4	**.008**	.581	.305	.107	656(1.0)
$\sum$		293	78k	19k	9k	11k

		Task Area Roles				
		Giver	Follower	Orienteer	Seeker	$\sum$
#. of spks	0	.162	.777	.043	.018	15k(1.0)
	1	.251	.675	.049	.025	60k(1.0)
	2	.325	.591	.054	.031	26k(1.0)
	3	.358	.536	.070	.037	6k(1.0)
	4	.329	.572	.076	.023	656(1.0)
$\sum$		28k	71k	5k	3k	11k

roles as a function of the number of simultaneous speakers. We intuitively view the number of simultaneous speakers as an indicator of the intensiveness of a discussion. The tables indicate that for 80% time in *Mission Survival Corpus I*, there are only from one to two simultaneous speakers, less than $.22\times4 = .88$ protagonist (who drives the meeting) and from $.251 \times 4 = 1.004$ to $.325 \times 4 = 1.3$ information-givers . This is determined by the participants' mental loads and their conscious or subconscious attempts to increase the efficiency of the discussions. On the other hand, the fraction of the secondary roles, such as attackers, increases significantly.

We note that the statistical learning theory does not guarantee the learnability of features, and thus we cannot treat a statistical learning method as a magical black box that takes training data as input and generate working models about the training data. What the theory provides is instead mathematically rigor ways to avoid overfitting in statistical learning. As a result, our introspection into how we solve problems by ourselves and attain efficiency provides good intuitions on how our individual and collective mental processes can be "learned" and simulated by machines.

B. Individual Honest Signals

In this section, we introduce the individual honest signals, their relationships to role-taking tendencies, their relevance with the roles, and their correlations.

1) Speech Features : Existing works suggests that speech can be very informative about social behavior. For instance, Pentland [47] singled out four classes of speech features for one-minute windows (emphasis, activity, mimicry and influence), and showed that those classes are informative of social behavior and can be used to predict it. In Pentland's [47] view, these four classes of features are honest signals, "behaviors that are sufficiently hard to fake that they can form the basis for a reliable channel of communication". To these four classes, we add spectral center, which has been reported to be related to dominance [10].

Emphasis is usually considered a signal of how strong is the speaker's motivation. In particular, its consistency is a signal of mental focus, while its variability points at openness to influence from other people. The features for determining **emphasis consistency** are related to the variations in spectral properties and prosody of speech: the less the variations, the higher consistency. The relevant features are: (1) confidence in formant frequency, (2) spectral entropy, (3) number of autocorrelation peaks, (4) time derivative of energy in frame, (5) entropy of speaking lengths, and (6) entropy of pause lengths.

The features for determining the **spectral center** are (7) formant frequency, (8) value of largest autocorrelation peak, and (9) location of largest autocorrelation peak.

Activity (=conversational activity level) is usually a good indicator of interest and engagement. The relevant features concern the voicing and speech patterns related to prosody: (10) energy in frame, (11) length of voiced segment, (12) length of speaking segment, (13) fraction of time speaking, (14) voicing rate (=number of voiced regions per second speaking).

Mimicry allows keeping track of multi-lateral interactions in speech patterns can be accounted for by measuring. It is measured through (15) the number of short reciprocal speech segments, (such as the utterances of 'OK?', 'OK!', 'done?', 'yup.').

Finally, **influence**, the amount of influence each person has on another one in a social interaction, was measured by calculating the overlapping speech segments (a measure of dominance). It can also serve as an indicator of attention, since the maintenance of an appropriate conversational pattern requires attention.

For the analysis discussed below, we used windows of one minute length. Earlier works [47], in fact, suggested that this sample size is large enough to compute the speech features in a reliable way, while being small enough to capture the transient nature of social behavior.

2) Body Gestures : Body gestures have been successfully used to predict social and task roles [31]. We use them as baselines to compare the import of speech features for socio and task roles prediction. We considered two visual features: (17) hand fidgeting and (18) body fidgeting. The fidgeting—the amount of energy in a person's body and hands—was automatically tracked by using the MHI (Motion History Image) techniques, which exploit skin region features and temporal motion to detect repetitive motions in the images and associate them to an energy value in such a way that the higher the value, the more pronounced is the motion [46]. These visual features were first extracted and tracked for each frame at a frequency of three hertz and then averaged out over the one-minute window.

3) Relationship between honest signals and role-taking tendencies: We note that different people have different yet consistent styles in taking functional roles, and their styles are reflected in their honest signals.

We compared the frequencies that the 32 subjects (eight sessions times four persons per session) take each of the eight functional roles (four task roles and four socio-emotional roles) in the first half of their respective discussions with those in the second half, taking the fact that it is hard to collect a data set in which the same persons participant in many different types of discussions. The frequencies of people in taking the Neutral/Follower roles, the Giver role, the Protagonist role, the Supporter role and the Seeker role in the first half of their discussions are predictive of the frequencies in the second half ($R^2 \geq .8, p < .001$). As a comparison, we compared the frequencies in the first half with those in the second half but randomly permuted (within discussion groups and within the whole data set). The permutations destroy both the correlations and the normality ($R^2 \leq .1, p \geq .3$). The consistencies of taking the other roles in both half of the discussions are weaker, but they nevertheless exist.

We proceed to test the hypotheses that the frequencies of taking different roles are linearly dependent on the honest signals with normal random noise. It turns out that the rate that a person takes the Seeker role is linearly dependent on his rate of short speaking segments ($R^2 = .47, p = .006$), the rate of taking the Supporter role is linearly dependent on the fraction of speaking and the fraction of voicing ($R^2 \geq .46, p \leq .007$), and the rates of taking the Giver role, the Protagonist role and the Neutral/Follower roles are linearly dependent on most of the honest signals in the **activity** category.

4) Relevance of the features for different roles : We computed the means and variances of the features' distributions for each social and task role. For simplicity the distributions were assumed to be in the Gaussian family. Given that the sample spaces of some of the features are bounded, the support of the distributions should be bounded too; we obtained this result by restricting the support and normalizing the distribution. Using the distributions, we analyzed the importance of the various features by means of the misclassification error.

We define the misclassification error (ME) for a given role as the probability that a Bayesian classifier will make an error while classifying a sample window with the given role assuming equal prior probabilities for each role. Thus, the misclassification error for a class i is given by:

$$\begin{aligned} err_i &= 1 - \int_{\phi_i} p(y;i)dy, \\ \phi_i &= \{y : p(y;i) \geq p(y;j); \forall j \neq i\}, \end{aligned}$$

where $p(y;i)$ is the conditional distribution of a feature y given class (role) i and i is the set of the feature values for which a Bayesian classifier predicts class i. The misclassification error gives a theoretical estimate of the separation of the feature distributions for different roles. A feature with distributions that are widely separated for the different classes can predict well and have small misclassification error.

As can be seen from Fig. 3(a), some certain features have a small ME for some socio roles, hence are distinctive for them. In details: Consistency and Spectral Center features have small ME with Supporter; energy and the

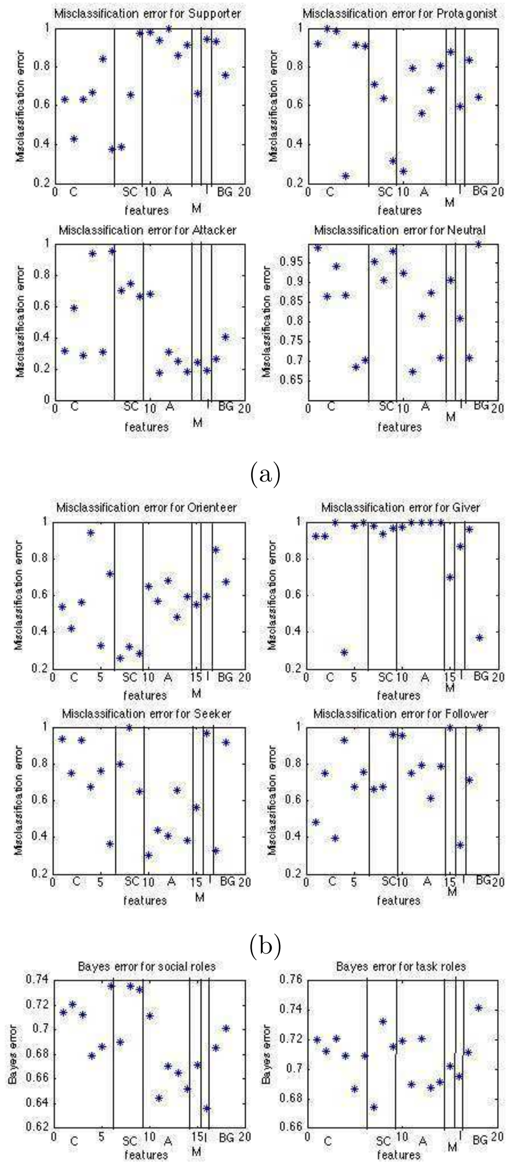


Fig. 3. Misclassification errors for predicting (a) each socio roles, (b) each task roles and (c) the combined misclassifications for predicting social and task roles. C: Consistency, SC: Spectral center, A: Activity, M: Mimicry, I: Influence, BG: Body Gestures.

location of the largest autocorrelation peak (which are features for Consistency) have low ME for Protagonist; Activity, Mimicry, Influence and Body Gestures for Attackers, accounting for the observation that attackers often utter small questioning sounds and fidget their hands and body in discomfort. Finally, MEs are always quite high for Neutral, suggesting a strong variation of all the features with this role.

Similar considerations can be made for the task role (cf. Fig. 3(b)): Orienteer is marked by low ME with Consistency and Spectral Center; Giver by low ME with energy (Consistency) and Body Fidgeting; Seeker by Activity and Mimicry; Follower by Influence, accounting for the observation that followers often nod or speak over, or ask

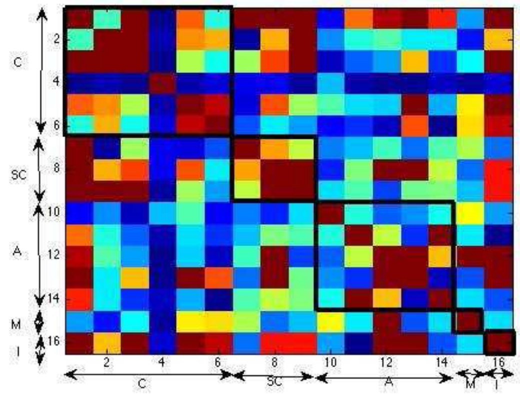


Fig. 4. Covariance of the feature set. Blue suggests small value and red suggests large value and other values lie in between.

questions to the Giver.

Unfortunately, no subset of features emerges that performs uniformly well for predicting all (task and/or socio) roles; some features are good at predicting one role, while others are good for other roles. In other words, some features may be better for binary classification, but they can be expected to work poorly in a multi-class classification like ours. Hence, in the experiments of the following section we use the full set of audiovisual features discussed above.

5) Correlation among features : We now see the redundancy of information among features in the light of the covariance matrix for the feature set. The covariance matrix for the audio feature set is shown in Fig.4. The covariance matrix is normalized to accommodate for different units of the features. From the figure, we see that most of the energy in the covariance matrix lies close to the diagonal and within the blocks shown in the figure. Thus the activity features are highly correlated with each other and less with other features. Similarly, the mimicry feature does not show any strong correlation with any other feature. This provides a good empirical justification for the clustering of the low level features into higher level type of activity and keeping the mimicry and influence separate. We also notice that there is strong correlation between the consistency features and spectral center features especially between (a) the confidence in formant frequency and the spectral center features and (b) the location of the largest autocorrelation peak and consistency features related to the speech spectrum. This is because we separated the original macro feature of emphasis into two features (consistency and spectral center) that are more intuitively separate. The correlation among features of the macro type persisted as the cross feature correlation. In summary, the activity, mimicry and influence features are more correlated within their types and less correlated outside their types and the consistency and spectral center features are more correlated within their type and with the features of the other but less correlated with any other feature type.

The above analysis explains why a linear combination of the low level features to define the measures of higher level features works well for predicting social outcomes [48]. For the purposes of this paper, we have used all the low-level features. There are two reasons for that. Firstly, there is no subset of the low-level features that performs uniformly well for predicting all roles. Some features are good for predicting one role while others are good for predicting others. For example, consistency features are good at predicting the supporter role but bad at predicting the protagonist role. Hence, though some features may be better for binary classification, they work poorly for multi-class classification of roles. Secondly, all computation was done offline and accuracy was a more important criterion than computational efficiency. Hence, we did not cluster the low level features into higher-level features. However, for online prediction, high-level, features can be computed as a linear combination of the low-level features (the coefficients can be obtained using the principal component analysis) and the prediction can be done using only the high-level features.

The above analysis identifies important speech features for prediction and also provides error bounds for Bayesian classification. However, the empirical results may be different, because we assumed the form of feature distributions to be Gaussian, so far. The distribution of the observed data in the experiments may not be Gaussian, but the true form of distribution is easy to estimate with limited data. In section VI, we will discuss prediction accuracies.

V. Model

We used three types of classifiers for predicting the roles, namely support vector machines, hidden Markov models and the influence model [44]. The three classifiers incrementally use more information for classification. The SVM considers each sample to be independent and identically distributed and the prior probability of each class is constant for each sample. The HMM considers the temporal correlation between the samples and the prior probability of the classes in the current sample depends upon the posterior probabilities of the classes in the last sample. It is intuitive that people do have some continuity in the roles and the roles do not change randomly within a small time. The influence model assumes that people influence each other and the current role of a person is influenced by the roles of other participants. For example, it can be expected that if a person acts as a giver, providing information, other participants might be listening to her, hence acting as followers. Thus the influence model presents a much richer representation of data. However, the extra richness comes at the additional cost of sample complexity. Thus, a much bigger training corpus is needed for training the more complex classifier.

A. Support Vector Machines

We modeled role assignment as a multiclass-classification problem on a relative large and very unbalanced dataset, and used Support Vector Machines as classifier, because of their robustness with respect to over-fitting. In fact, SVMs try to find a hyperplane that not only discriminate the classes but also maximizes the margin between these classes [49], [50].

SVM were originally designed for binary classification but several methods have been proposed to construct multi-class classifier. The “one-against-one” method [15] was used whereby each training vector is compared against two different classes by minimizing the error between the separating hyper-plane margins. Classification is then accomplished through a voting strategy whereby the class that most frequently won is selected.

B. Hidden Markov Models

The Hidden Markov model is a standard model for modeling partially observable stochastic processes and was originally developed for speech understanding [51]. In our earlier work we used HMMs to model meeting data and predict social and task roles. HMMs have more representational power than SVMs because they can model some of temporal dependencies of roles.

The representation of the model is as follows: t, time; y(t), the feature vector; x(t), the role; p(x), the priors for the roles; p(x(t)|x(t-1)), the role transition probabilities; p(y(t)|x(t)): conditional distribution of observed feature vector given the current role. We assume speaker independence; i.e., the Markov process determining the roles, the speech features and the hand and body fidgeting of each person have the same parameters, p(x), p(x(t)|x(t-1)) and p(y(t)|x(t)). Thus, all four-feature sequences (one per subject) from all eight meetings are used to train a single HMM. The training is done using the EM algorithm. For prediction, each person is represented by an independent instantiation of the same Markov process. Thus, four independent HMMs represent the four different people in the meeting. For classification, the Viterbi algorithm is used to compute the most likely sequence of roles.

C. Influence dynamics of a group of influencing people

The influence modeling approach is a method that can effectively deal both with the curse of dimensionality and the over-fitting problem. It has been developed in the tradition of the N-heads dynamic programming on coupled hidden Markov models [35], the observable structure influence model [36], and the partially observable influence model [40]. It extends though these previous models by providing greater generality, accuracy and efficiency. The influence modeling is a team-of-observers approach to complex and highly structured interacting processes. In this model, different observers look at different data, and can adapt themselves according to different statistics in the data. The different observers find other observes whose latent state, rather than observations, are correlated, and use these observers to form an estimation network. In this way, we effectively exploit the interaction of the underlying interacting processes, while avoiding the risk of overfitting and the difficulties of observations with large dimensionality.

Specifically, a latent structure influence process is a stochastic process $\{S_t^{(c)}, Y_t^{(c)} : c \in \{1, \cdots, C\}, t \in \mathbb{N}\}$. In this process, the latent variables $S_t^{(1)}, \cdots, S_t^{(C)}$ each have finite number of possible values $S_t^{(c)} \in \{1, \cdots, m_c\}$ and their (marginal) probability distributions evolve as the following:

$$\begin{aligned}
\mathbb{P}(S_1^{(c)} = s) &= \pi_s^{(c)}, \\
\mathbb{P}(S_{t+1}^{(c)} = s) &= \sum_{c_1=1}^{C} \sum_{s_1=1}^{m_{c_1}} h_{s_1,s}^{(c_1,c)} \mathbb{P}(S_t^{(c)} = s_1),
\end{aligned}$$

where $1 \le s \le m_c$ and $h_{s_1,s}^{(c_1,c)} = d^{(c_1,c)} a_{s_1,s}^{(c_1,c)}$ ($a_{s_1,s}^{(c_1,c)}$ represent the relations of different states for the interacting processes, and $d^{(c_1,c)}$ represent the influence among the processes). The observations $\vec{Y}_t = (Y_t^{(1)}, \cdots, Y_t^{(C)})$ are coupled with the latent states $\vec{S}_t = (S_t^{(1)}, \cdots, S_t^{(C)})$ through a memory-less channel,

$$\mathbb{P}(\vec{S}_t)\mathbb{P}(\vec{Y}_t|\vec{S}_t) = \prod_{c=1}^{C} \mathbb{P}(S_t^{(c)})\mathbb{P}(Y_t^{(c)}|S_t^{(c)}).$$

The algorithms for the latent state inference and the parameter learning of the influence model follows from the above definition and maximum likelihood estimation algorithm. A detailed discussion of this model, as well as its algorithms, can be found in [43], [44].

The following imaginary group dynamics example (Fig.5) illustrates the interacting dynamic processes we are referring to, and how the influence modeling simultaneously exploring and exploiting the structure among them. In this example, we have a network of six people who may be engated in one, two, or more separate discussions, and our task is to estimate whether a person is neutral or excited from noisy observations of him. A natural approach is to make estimations of the states of a person from the observations of not only this person, but also the related people, in a short window around the time of the estimated states. The estimation is a chicken and egg problem: the more we know about the structure (i.e., who is involved with what discussion), the better we can estimate the latent states, and vice versa.

An important consideration in choosing a multi-class classifier is whether the classifier, after it is trained from a training data set, can generalize to future applications. With increased dimensionality and without regularization, even a linear classifier, which is considered stable, can overfit. The latent structure influence modeling of interacting processes avoids the curse of dimensionality problem with the team of observers approach. In this approach, the individual observers only look at the latent states of the other related observers, rather than looking at the raw observations, and thus are less likely to be overfit and more likely to be generalizable.

Fig.6 compares the performances of several dynamic latent structure models (the influence model, the hidden Markov model with 16 latent states and 10-dimensional

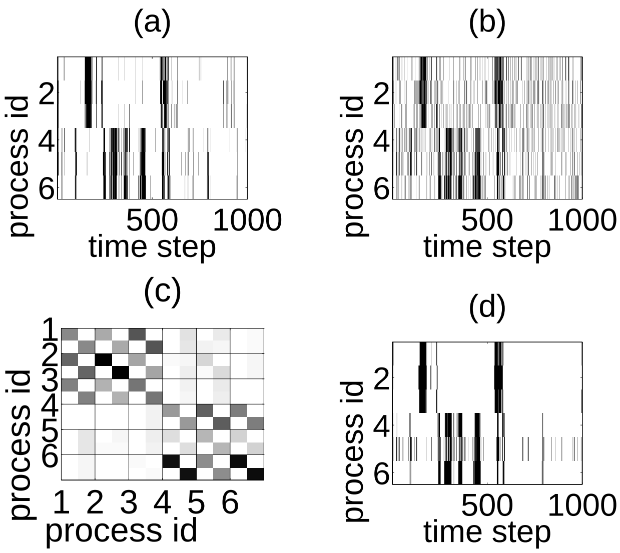


Fig. 5. Estimating network structure and latent states simultaneously from noisy observations with the influence model. The task is to estimate the true states, as well as the interaction, from noisy observations shown in (b). The recovered interaction structure in (c) has 90% accuracy, and the estimated the latent states in (d) have 95% accuracy.

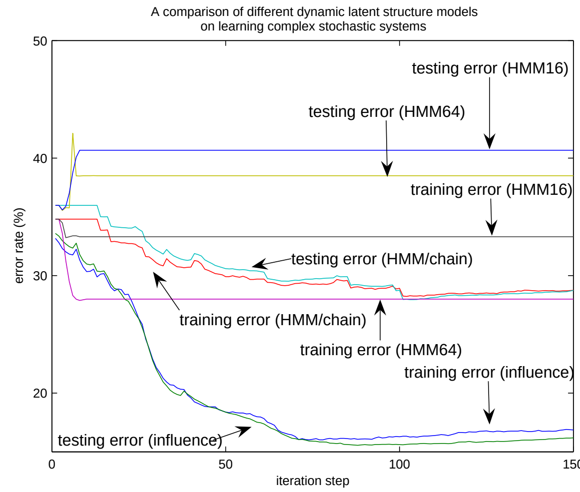


Fig. 6. Latent state influence process is immune to overfitting.

Gaussian observations, the hidden Markov model with 64 latent states and 10-dimensional Gaussian observations, and 10 hidden Markov models, each on one dimensional data). Of the 1000 samples, we use the first 250 for training and the rest 750 for validation.

Judged from Fig.6, the logarithmically scaled number of parameters of the influence model allows us to attain high accuracy based on a relatively small number of observations. This is because the influence model preserves the asymptotic marginal probability distributions of the individual "bits", as well as the linear relationship among them. Hence, the influence model shrinks the number of parameters of the original hidden Markov model logarithmically and in an efficient way, while still preserving the principal dynamics of the process.

In our recent efforts to improve the descriptive power of the influence modeling, we realized that we could give the individual sites ($c \in \{1, \cdots, C\}$) the volition on how much it wants to vote on the latent states of another site ($c' \in \{1, \cdots, C\}$). (In the group discussion example, if a person takes the giver role, he can vote another person to take the neutral or seeker role. On the other hand and in contrast to our previous influence modeling, if a person takes the neutral role, he can choose not to vote on another person's task and social roles.) As a result, the latent state space of an individual site consists of the summed votes from other sites on the site's next possible state, with a probability measure associated with the space. Specifically,

$$\begin{aligned} S_1^{(c)}(s) &= \pi_s^{(c)}, \\ S_{t+1}^{(c)}(s) &= \sum_{c_1=1}^{C} \sum_{s_1=1}^{m_{c_1}} h_{s_1,s}^{(c_1,c)} \mathbb{P}(S_t^{(c)} = s_1), \end{aligned}$$

$$\mathbb{P}(\vec{Y}_t, \vec{S}_t) = \mathbb{P}(\{S_t^{(c)}(s^{(c)})\}_{c \in 1,\cdots,C}^{s^{(c)} \in 1,\cdots,m_c}) \cdot \prod_{c=1}^{C} \mathbb{P}(Y_t^{(c)} | S_t^{(c)}).$$

The algorithms for the latent state inference and the parameter learning of the influence model could be derived from the new definition.

VI. Experimental Results

In this section, we will present our experimental results in recognizing the roles of discussion participants with their honest signals and turn-taking behavior.

A. *Performance of Individual Honest Signals*

In the Mission Survival I corpus, the visual features are extracted on a frame(=0.33 seconds) base. Similarly, the relational roles were manually annotated on a frame base. The audio features, as we have seen, are computed on one minute windows. Hence a decision must be taken as to how the frame-based annotation should be projected at the level of the one minute window. We applied two heuristics for this projection. Heuristic 1 uses the straightforward frequency criterion. This heuristic resulted in highly unbalanced data, with most of the windows labeled as neutral/follower. Table V shows the distributions of functional roles, the accuracy of predicting social roles for different classifiers with different feature sets, and the accuracy for task roles[1]. The results are comparable to those in [31], and show that better accuracy is obtained with audio features than with visual features. The SVM tends to perform better that the other two models, whereas the influence model performs better than HMM. As pointed in [31]

[1] We used the BSVM tool [52] available at http://www.csie.ntu.edu.tw/~cjlin/bsvm/. The bound-constrained SVM classification algorithm with a RBF kernel was used. The cost parameter C and the kernel parameter g were estimated through the grid technique by cross-fold validation using a factor of 101. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size.

TABLE V
Distributions of social (a) and task (b) roles after the application of Heuristic 1, and the corresponding prediction accuracy (c) for SVM/HMM/IM with visual/speech "honest signals"/all features.

(a)	Supporter	Protagonist	Attacker	Neutral
	.034	.158	.000	.808

(b)	Orienteer	Giver	Seeker	Follower
	.038	.210	.003	.749

	Social			Task		
(c)	SVM	HMM	IM	SVM	HMM	IM
Visual	.70	.73	.71	.68	.73	.68
Audio	.82	.71	.75	.76	.71	.74
Joint	.92	.75	.77	.79	.73	.74

TABLE VI
Distributions of social (a) and task (b) roles after the application of Heuristic 2, and the corresponding prediction accuracy (c) for SVM/HMM/IM with visual/speech "honest signals"/all features.

(a)	Supporter	Protagonist	Attacker	Neutral
	.149	.326	.002	.522

(b)	Orienteer	Giver	Seeker	Follower
	.070	.517	.017	.395

	Social			Task		
(c)	SVM	HMM	IM	SVM	HMM	IM
Visual	.21	.54	.51	.19	.44	.43
Audio	.45	.57	.57	.52	.54	.61
Joint	.46	.56	.57	.51	.53	.58

the accuracies are comparable to any human predicting the roles. A classifier that classifies all observations as the class with highest frequency always classifies them as neutral and follower. The accuracy of this classifier is the probability of the class with highest frequency and we call this the benchmark accuracy. However, as already pointed out the frequency of non-neutral and non follower labels in the used corpus were extremely small due to the labeling procedure, and the accuracy figures mostly reflect the high accuracy on the neutral and follower roles. To provide some more balance to the data set and to not miss the important and rarer roles, we used a slightly different labeling mechanism. In heuristic 2, a one-minute window was given the most frequent of the non-neutral (or non-follower) labels if that label was present for at least in one fourths of window's frames; otherwise the window was labeled as neutral. The distribution of social and task roles in the data after applying heuristic 2 is shown in Table VI. Besides avoiding missing non-neutral/non-follower roles, this strategy has its justification in the application scenarios described at the beginning of this paper, where finding out about non-neutral/non-follower roles is of the utmost importance for facilitation and/or coaching. The overall accuracy was lower than with the other labeling method, as shown in Table VI, with the influence model and the HMM performing similarly, and better than the SVM, a result that can be attributed to the capability of the Influence model and of the HMM to model temporal relationships . The details concerning the precision and recall figures for each methods are reported in Table VII.

The results obtained by means of the sole speech features are almost always superior to those attained by means of the video features and to those obtained by combining the two. Hence, we will restrict the discussion to the speech features.

When considering the average values of precision and recall for the speech features, a slight advantage for the HMM emerge in the socio-role area, and a similar slight advantage of the SVM in the task area. All the classifier completely miss the attacker role; the HMM and the IM miss the seeker; in addition, the IM misses the orienteer.

TABLE VII
(Precision, Recall) of task/social-emotional roles using SVM/HMM/IM with body gestures, speech "honest signals" and both.

		Supporter	Protagonist	Attacker	Neutral
SVM	Visual	.13, .72	.05, .04	.00, .00	.54, .23
	Audio	.09, .33	.29, .17	.00, .00	.74, .61
	Joint	.12, .38	.34, .22	.00, .00	.73, .60
HMM	Visual	.00, .00	.07, .03	.00, .00	.58, .91
	Audio	.11, .10	.38, .65	.00, .00	.79, .63
	Joint	.15, .16	.37, .50	.00, .00	.73, .67
IM	Visual	.06, .01	.35, .19	.00, .00	.59, .77
	Audio	.15, .07	.44, .43	.00, .00	.68, .75
	Joint	.13, .06	.41, .34	.00, .00	.66, .80

		Orienteer	Giver	Seeker	Follower
SVM	Visual	.02, .34	.12, .15	.03, .33	.62, .28
	Audio	.09, .39	.64, .40	.11, .11	.69, .67
	Joint	.07, .22	.63, .40	.08, .06	.69, .68
HMM	Visual	.00, .00	.38, .10	.00, .00	.47, .86
	Audio	.15, .25	.56, .67	.00, .00	.79, .54
	Joint	.16, .17	.57, .63	.00, .00	.69, .57
IM	Visual	.03, .05	.44, .37	.00, .00	.47, .57
	Audio	.00, .00	.64, .60	.00, .00	.66, .74
	Joint	.00, .00	.63, .55	.00, .00	.62, .75

In a way, the IM seems to be less sensitive to rarer roles. Moreover, they are now better than the benchmark. In the end, the five classes of honest signals seems to be the more predictive and informative features. Since the classification accuracies are the best when we use IM with speech features, we now investigate what honest signals in speech are important for classification.

We considered the contribution of the various audio feature classes. Table IX shows the accuracy values obtained using independent classifier. Interestingly, the Activity class yields accuracy values (slightly) higher than those produced through the usage of the entire set of audio features, cf. Table VI. Hence using the sole set of Activity

TABLE VIII
Average precision and recall figures for speech features.

	Social		Task	
	Mean P	Mean R	Mean P	Mean R
SVM	.28	.28	.38	.39
HMM	.32	.35	.38	.37
IM	.32	.31	.33	.34

TABLE IX
Accuracies for social and task roles (independent classifiers) with the different classes of speech features on Heuristic 2 data.

	Social roles	Task roles
Consistency	.50	.47
Spectral Center	.50	.47
Activity	.60	.62
Mimicry	.50	.37
Influence	.54	.52

TABLE X
Distribution of social × task roles with Heuristic 2.

	supporter	protagonist	attacker	neutral
orienteer	0,011	0,023	0,000	0,037
giver	0,077	0,169	0,001	**0,270**
seeker	0,003	0,006	0,000	0,009
follower	0,059	0,129	0,001	0,206

features emerges as a promising strategy.

Finally, we explored the extent to which the relationships between task and social role discussed above can be exploited, by training a joint classifier for social and task roles—that is, a classifier that considers the whole set of the 16 combinations of social × task roles; a more difficult task than the ones considered so far. Table X reports the distribution of the joint roles in the corpus, while Table XI the classification results.

The results are interesting. Notice, first of all, that the accuracies are always much higher than the baseline, see the bold figure in Table X. Moreover, the sole audio features produce results that are comparable to those obtained by means of independent classifier, despite the higher complexity of the task. These results show a) that it makes sense to try to take advantage of the relationships between task and social role through the more complex task of joint classification; b) that the IM is capable of scaling up to larger multi-class tasks without performance losses.

B. Performance of Turn-taking Signals

Modern statistical methods are normally guarded against overfitting by some mechanisms, and can normally attain comparable performances by careful selection of features and careful formulation of problems. On the other hand, some methods may be easier to use and more intuitive to understand for some problems. This subsection discusses the one-person features and the interaction features that statistical learning methods (the support vector method and the influence model in particular) should utilize to get good performances, the different ways the methods use the features, and the resulting performances.

TABLE XI
Accuracy of joint prediction of social and task roles.

	social roles	task roles
visual	.47	.41
audio	.58	.60
joint	.59	.56

The turn-taking behavior related to the functional roles of a speaker at a specific moment includes: (1) his amounts of speaking time in time-windows of different sizes around the moment; (2) whether other persons turned their bodies to the speaker at the beginning/end of his speaking turn; (3) the amounts of speaking time of the other persons in time windows of different sizes around the moment, in particular the amounts of speaking time of the persons that speak the most in those time windows; and (4) the psychological profile of this person, e.g., his extrovertedness, his tendency to take control and his level of interest in the discussion topic.

The influence model formulates the group process in terms of how an individual takes his functional role based on the functional roles of the others on the one hand, and how an individual presumes the others' possible functional roles based on everyone's current roles on the other hand: When a person is taking the giver role, he prefers the others to take the neutral/follower role or the seeker role at least for a while; In comparison, when a person takes the neutral role, he does not quite mind who is going to take which role next; When all participants take the neutral role, the overall preference of the whole group could be quite weak, so that the individuals could wander about their role-taking states until some individual takes a "stronger" role. Specifically, an influence model can tell a participant's maximum likelihood functional role at a moment by comparing how likely different roles correspond to his amounts of speaking time in time-windows of different sizes around this moment. When there are doubts on whether a person is shifting to the giver role or the seeker role, for example, the influence model will look at the intensity of the other participants' body movements: The giver role is associated with more body movements at the beginning of the corresponding speaking-turn and more attention from the others. The participants' roles a moment ago can be exploited by the influence model to generate a "vote" for different roles for the participant under investigation, and the vote can be subsequently used to bias the model's Bayesian estimation. The psychological profile of the participants can be further used for generating the votes (for the participants to take certain roles).

The support vector method (SVM) on the other hand does not involve probability distribution in the training phase and the application phase, although SVM can be inspected in the Bayesian framework. SVM also requires the model observations to be points in a (possibly high-dimensional) Euclidean space. In terms of utilizing the amounts of speaking time corresponding to different window sizes, the support vector method performs as good as any Bayesian method, and the latter requires appropriate probability estimations of the observations conditioned on the functional roles. We sorted the amounts of speaking time over all participants in every time-window of different sizes up to some upper bound (two minutes in our experiments), and use the sorted amounts of speaking time over all speakers and corresponding to all window sizes around the moment of inspection (among other features) for functional-role classification. The arrangement by sorting makes the corresponding feature permutation independent. The support vector method can subsequently disambiguate among the possible roles of a person by comparing his amounts speaking time with those of the others, and with those of the person who speaks the most in particular. The hand/body movements involved with role-shifting is the hardest feature for the support vector method, since the boundaries of the role assignments are unknown. The trained functional-role classifiers with SVM seem to indicate that SVM uses the body/hands movement corresponding to the current moment for disambiguating the roles.

Table XII compares the performance of the influence model and the performance of SVM using the best construction of features we can think of. We can see that the performances of both SVM and the influence model are improved compared with our previous work [31], especially for the infrequently appearing roles. The improvement is because we have a better understanding of the group process. The influence model has a slightly higher performance than the SVM, while previously the former is slightly lower than the latter. This is due to our new perspective and corresponding EM algorithm for the influence modeling. The latent state space of an influence model is the summed voting of the individuals' future states (e.g., participants' next functional roles) associated with a probability space, while previously it was the marginal probability distributions among those states. A direct consequence of this new perspective is that a neutral role couldn't waive his votes previously, and now he can.

TABLE XII
Performances of classifying task/social-emotional roles using SVM/IM with interaction signals.

		SVM				
		Giver	Follower	Orienteer	Seeker	%
ground truth	G	**10758**	1944	982	1092	.72
	F	2581	**26924**	334	3401	.80
	O	796	153	**795**	447	.36
	S	362	639	27	**277**	.21
	%	.74	.90	.37	.05	

		influence model				
		Giver	Follower	Orienteer	Seeker	%
ground truth	G	**10173**	1441	880	2282	.68
	F	1542	**28000**	246	3542	.84
	O	708	260	**1045**	178	.48
	S	426	219	32	**628**	.48
	%	.79	.94	.47	.10	

		SVM				
		Attacker	Neutral	Protagonist	Supporter	%
ground truth	A	**95**	63	15	12	.51
	N	452	**32779**	3932	1308	.85
	P	320	813	**5789**	2458	.62
	S	143	251	1738	**1344**	.39
	%	.09	.97	.50	.26	

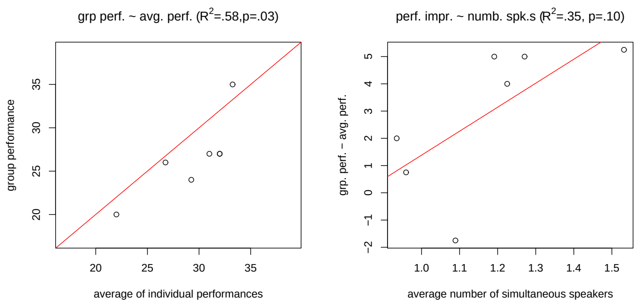
		influence model				
		Attacker	Neutral	Protagonist	Supporter	%
ground truth	A	**108**	54	5	18	.58
	N	450	**32813**	3916	1292	.85
	P	318	804	**5804**	2454	.62
	S	142	258	1733	**1343**	.39
	%	.11	.97	.51	.26	

VII. Functional Roles and Performances

One reason for us to inspect the group processes and the Task/Socio-Emotional Area roles is that we want to design automated tools to improve the group performance. In the Mission Survival Corpus I, the initial individual scores and the final group scores of seven discussions out of a total of eight are available, and they are in terms of how the individual/group rankings of 15 items are different from the standard expert ranking. ($f(r_1 \cdots r_{15}) = \sum_{i=1}^{15} |r_i - r_i^{(0)}|$, where f is the score function, $r_1 \cdots r_{15}$ is an individual/group ranking, $r_1^{(0)} \cdots r_{15}^{(0)}$ is the expert ranking, $r_i \neq r_j$ and $r_i^{(0)} \neq r_j^{(0)}$ for $i \neq j$, and $r_i, r_i^{(0)} \in \{1, \cdots, 15\}$.) Thus the corpus provides us laboratory data to inspect how individuals with different initial performances and psychological profiles interact with each other and incorporate their individual information to reach a better performance. Our preliminary findings are given below.

The post-discussion group performance is linearly and positively correlated to the average of its participants' pre-discussion individual performances, with the former being slightly better than the latter. (group score = .93$\times$ average of individual scores -.74, $R^2 = .58$, $p = .03$.) The relationship is shown in Fig.7 (a), and can be explained with a probabilistic model on how individuals combine their results: Prior to the discussion, pieces of information for solving the ranking problem of the mission survival task are probabilistically distributed among the participants, and different individuals have the correct/best rankings for different items. During the discussion, the individuals merge their information through a group process that is probabilistically dependent on their initial performances, their interactions with each other, and many other factors. When the individuals disagree on the ranking of an item, they can either choose one from their repository of rankings that results in minimal disagreements, or find a creative and better ranking by further information sharing and an 'aha' experience. Previous experiments show that groups make use of their resources to a probabilistically similar extent, and experienced discussion groups outperform inexperienced groups by generating more item rankings that are creative and better through better information sharing [27]. Thus the relationship between the group performance and the average of the individual performances follows from the fact the different groups are more or less similar.

The improvement of a group's performance over the average of the individuals' is positively correlated to the average number of simultaneous speakers over the discussion, which reflects the intensity of the discussion and is used by us as a measure of influence ($R^2 = .35$, $p = .10$). The relationship is shown in Fig.7 (b). The improvement is also positively correlated to the frequency that meeting



(a) (b)

Fig. 7. (a) Group score is linearly related with the average of pre-discussion individual scores. (b) Group performance is linearly related with the average number of simultaneous speakers.

participants take the protagonist role ($R^2 = .35$, $p = .10$) and the rate that meeting participants take the giver role ($R^2 = .28$, $p = .12$). Since the average length of a continuous protagonist role segment is almost 50% longer than the average length of a continuous giver role segment (26 vs. 19), we speculate that the improvement is more dependent on the longer utterances of the individuals. The improvement doesn't seem to correlate with either the length of a discussion or the rates that the participants take other roles. These correlations are again compatible with the observation from previous experiments that an experienced discussion group encourages better problem solving and thorough information sharing [27]: An inexperienced discussion group worries more on whether it could reach a consensus. As a result, its members treat consensus-reaching as the goal, rather than the natural result of sufficient information sharing. They either argue for their own rankings without paying attention to the others' arguments, or give up their rankings easily. In either way, they feel their importance in the discussion are not sufficiently recognized, quickly lose their motivation for participation. In contrast, an experienced group encourages different opinions, and views conflicts as insufficient information sharing. Its participants solve a conflict by a thorough discussion and a win-win problem solving strategy rather than by cheap techniques such as majority voting and coin-flipping. The participants also give sufficient suspicion on easy initial agreements. Different group discussion process characteristics and different role playing behavior can be dependent on the different opinions on what is a fruitful group discussion between an experienced group and an inexperienced group.

In many of the discussions, one or a few individuals take certain task and social roles twice as much as the other roles. The fraction of speaking time of an individual and the rates that an individual takes the giver role and the protagonist role do not seem to correlate with the initial performance of the individual. Based on the small number of discussions and the fact that the meeting corpus is in Italian, we can only speculate that role-taking is related to the individuals' personal styles, motivation, interaction with each other, and many other factors.

While it is our belief that the symptoms of group process problems could be found by automated tools and the prescriptions could be accordingly given, we note that facilitating the group process is trickier. For one reason, an inappropriately phrased prescription may inadvertently block the participants' attentions to the real goal of the discussion, and shift the attentions from one unimportant thing (e.g., the pressure of reaching a consensus) to another (e.g., the "appropriate" speaking-turn lengths and the "appropriate" number of simultaneous speakers).

VIII. Conclusions

This paper discussed the turn-taking dynamics and the changing individual role assignments of several group brainstorming sessions in *Mission Survival Corpus I*. It also discussed their modeling and learnability issues using several statistical learning methods (in particular, the support vector method and the influence model). We model the group discussion dynamics by first introspecting how such discussions should work to suit their purposes and then by applying the appropriate statistical learning methods. We have several future directions in our minds: Firstly we are simulating the behavior and performances of different types of brainstorming sessions with stochastic processes and simplified assumptions, and comparing the simulated results with the experimentally collected results. The simulation could provide insight in understanding the collective intelligence. Secondly we would like to use our modeling to improve the multi-person interaction efficiency. Thirdly we would like to know whether our turn-taking and role-assignment modeling could be suitable for other types of multi-person interactions with appropriately tuned parameters.

IX. Acknowledgment

This work was partially supported by the UE under the CHIL (FP6) project. We would like to thank Massimo Zancanaro for helping us with the data collection.

References

[1] M. Diehl and W. Stroebe, "Productivity loss in idea-generating groups: tracking down the blocking effect," *Journal of Personality and Social Psychology*, vol. 61, pp. 392–403, 1991.

[2] J. M. Jackson and S. G. Harkins, "Equity in effort: An explanation of the social loafing effect," *Journal of Personality and Social Psychology*, vol. 49, pp. 1199–1206, 1985.

[3] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

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

[5] 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*, 2007, pp. 1403–1406.

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

[7] 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.

[8] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, 2005.

[9] 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*, 2008, pp. 693–696.

[10] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features," in *MLMI*, 2005, pp. 76–86.

[11] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *ICMI*, 2006, pp. 257–264.

[12] 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*, 2008, pp. 45–52.

[13] H. Hung, D. B. Jayagopi, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Investigating automatic dominance estimation in groups from visual attention and speaking activity," in *ICMI*, 2008, pp. 233–236.

[14] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating the dominant person in multi-party conversations using speaker diarization strategies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 2197–2220.

[15] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simple systematics for the organization of turn-taking for conversation," *Languate*, vol. 50, no. 4, 1974.

[16] J. M. Atkinson, Ed., *Structures of Social Action (Studies in Conversation Analysis)*. Cambridge: Cambridge University Press, 1985.

[17] L. R. Frey, D. S. Gouran, and M. S. Poole, *The Handbook of Group Communication Theory and Research*. SAGE, 1999.

[18] E. A. Schegloff, *Sequence Organization in Interaction: A Primer in Conversational Analysis*. Cambridge: Cambridge University Press, 2007, vol. 1.

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

[20] ——, *Personality and Interpersonal Behavior*. Holt, Rinehart and Winston, 1969.

[21] R. F. Bales and S. P. Cohen, *Symlog, A System for the Multiple Level Observation of Groups*. Free Press, 1979.

[22] J. McGrath and D. Kravitz, "Group research," *Annnual Review of Psycology*, vol. 33, pp. 195–230, 1982.

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

[24] A. F. Osborn, *Applied Imagination: Principles and Procedures of Creative Problem Solving*. New York, NY: Charles Scribner's Sons, 1963.

[25] S. J. Karau and K. D. Williams, "Social loafing: A meta-analytic review and theoretical integration," *Journal of Personality and Social Psychology*, vol. 65, pp. 681–706, 1993.

[26] B. A. Nijstad, W. Stroebe, and H. F. M. Lodewijkx, "Production blocking and idea generation: Does blocking interfere with cognitive processes?" *Journal of Experimental Social Psychology*, vol. 39, pp. 531–548, 2003.

[27] J. Hall and W. H. Watson, "The effects of a normative intervention on group decision-making performance," *Human Relations*, vol. 23, no. 4, pp. 299–317, 1970.

[28] D. S. Wilson, J. J. Timmel, and R. R. Miller, "Cognitive cooperation: When the going gets tough, think as a group," *Human Nature*, vol. 15, no. 3, pp. 1–15, 2004.

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

[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] 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*, 2007, pp. 271–278.

[32] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th International Conference on Multimodal Interfaces: Special Session on Social Signal Processing*, 2008, pp. 53–60.

[33] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[34] M. Brand, "Coupled hidden markov models for modeling interacting processes," MIT Media Laboratory Vision & Modeling Technical Report #405, Tech. Rep., 1996. [Online]. Available: http://vismod.media.mit.edu/tech-reports/TR-405.ps.Z

[35] N. M. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22(8), 2000, pp. 831–843.

[36] C. Asavathiratham, "The influence model: A tractable representation for the dynamics of networked markov chains," Ph.D. dissertation, MIT, 1996.

[37] C. Asavathiratham, S. Roy, B. C. Lesieutre, and G. C.Verghese, "The influence model," *IEEE Control Syst. Mag.*, no. 12, 2001.

[38] L. K. Saul and M. I. Jordan, "Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning*, no. 37, pp. 75–86, 1999.

[39] X. Boyen and D. Koller, "Exploiting the architecture of dynamic systems," in *AAAI/IAAI*, 1999, pp. 313–320.

[40] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," MIT Media Laboratory Vision & Modeling Technical Report #539, Tech. Rep., 2001. [Online]. Available: http://vismod.media.mit.edu/tech-reports/TR-539.pdf

[41] T. Choudhury and A. Pentland, "The sociometer: A wearable device for understanding human networks," *Computer Supported Cooperative Work - Workshop on Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*, 2002.

[42] T. Choudhury, "Sensing and modeling human networks," Ph.D. dissertation, MIT, 2003.

[43] W. Dong, "Infuence modeling of complex stochastic processes," Master's thesis, MIT, 2006.

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

[45] G. Carli and R. Gretter, "A start-end point detection algorithm for a real-time acoustic front-end based on dsp32c vme board," in *Proceedings of ICSPAT*, 1997, pp. 1011–1017.

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

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

[48] ——, "Socially aware computation and communication," *Computer*, vol. 38, no. 3, pp. 33–40, 2005.

[49] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[50] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[51] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[52] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, 2002.