Honest Signals in the Recognition of Functional Relational Roles in Meetings*

Bruno Lepri[1], Ankur Mani[2], Alex (Sandy) Pentland[2], Fabio Pianesi[1]

[1] FBK-Irst [2]The MIT Media Laboratory
[1]Via Sommarive 38050 Povo-Trento, Italy
[2]20 Ames Street 02139-4307 Cambridge, MA, USA
{lepri, pianesi}@fbk.eu
{amani, pentland}@mit.edu

Abstract

In this paper we focus on the usage of *honest signals* for the automatic recognition of functional relational roles. We consider 16 speech honest signal features, along with three fidgeting visual features, and investigate their potential for role classification comparing: a) different labeling convention for relational roles, b) the contribution of honest signals' class to classification, and c) independent vs. joint classification of task and social roles, in order to understand the extent to which the redundancy between the task and socio roles can be exploited for classification

Introduction

During a meeting, participants may play different functional roles such as leading the discussion or deflating the status of others. The effectiveness of a meeting is often directly related to the roles participants play, and to their distribution during the meeting. Professional facilitators team coachers are often used to identify dysfunctional role patterns and help the team to re-organize their roles' distribution (Hall and Watson 1970).

The availability of multimodal information about what is going on during the meeting makes it possible to explore the possibility of providing various kinds of support to dysfunctional teams, from facilitation to training sessions addressing both the individuals and the group as a whole. Empirical studies, e.g., (Pianesi et al. 2008b), provide some initial evidence that people's attitudes towards those services might be similar to those towards reports produced by human experts, this way encouraging efforts in that direction. Clearly, crucial to any automatic system aiming to provide facilitation or coaching is that it be capable of understanding people social behavior, e.g., by abstracting over low level information to produce medium-/coarse-grained one about the functional roles members play.

In this paper we pursue the automatic detection of task and social functional relational roles (Benne and Sheats 1948) by means of *honest signals* (Pentland 2008)—that is, "behaviors that are sufficiently expensive to fake that they can form the basis for a reliable channel of communication" (Pentland 2008). Honest signals have already been shown to be capable of predicting and explain the human behavior in social interactions—e.g., the outcomes of business negotiations (Cuhran and Pentland 2007). More recently, combined with three visual features (hand, body, and head fidgeting), they were used to automatically detect two personality traits (Extraversion and Locus of Control) (Pianesi et al. 2008a).

After introducing 16 speech honest signals, grouped into five classes (*Consistency*, *Spectral Center*, *Activity*, *Mimicry*, and *Influence*), and two body gestures (hand and body fidgeting), we will turn to a number of classification experiments, comparing: a) different labeling convention for relational roles, b) the contribution of each class of honest signals class to classification and c) independent vs. joint classification of task and social roles, in order to understand the extent to which possible relationships between the two role classes can be exploited for classification. In our experiments we will exploit the Influence Model (Dong 2006), because of its nice capability of modeling the mutual influences among group members and of naturally adapting to groups of different sizes.

Our results prove that honest signals are indeed promising cues for the automatic detection of functional relational roles, and that joint classification of task and socio role can take full advantage of them.

Previous and Related Work

Multimodal analysis of group behavior is a relatively recent research area. McCowan et al. (2005) 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. For example, "discussion" is a group action which can be recognized from the verbal activity of individuals.

Other works have addressed the automatic detection of higher level properties of social interaction, such as dominance or the roles played by meeting participants. For example, Rienks and Heylen (2006) used Support Vector Machines to detect dominance, by relying on few nonverbal—e.g., speaker turns, floor grabs, speaking length—and verbal—e.g, number of spoken words—audio features. More recently, Hung et al. (2008) tried to estimate the most dominant person on non-scripted meetings using audio and visual cues, while Jayagopi et al. (2008) addressed the automatic estimation of two aspects of social verticality (role-based status and dominance) using non-verbal features (vocalic and visual activity, and visual focus of attention). The detection of functional roles was pursued through social network analysis (Vinciarelli 2007) or by means of a combination of social networks and lexical choices (Garg et al. 2008).

Functional relational roles (social and task roles) were addressed by Zancanaro et al. (2006) 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 et al. (2007) extended this work by comparing SVM to a HMM- and IM-based approaches.

Other work, still, e.g., (Pianesi et al. 2008a), 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 Extraversion and Locus of Control, using speech features that are provably *honest signals* for social behavior, and visual fidgeting features.

In this work we address the same relational roles as considered in (Zancanaro, Lepri and Pianesi 2006) and (Dong et al., 2007). However, we: a) exploit a much larger, and principled set of 16 audio features grouped into five classes of *honest signals* (Pentland 2008) computed over 1-minute windows; b) propose and test a labeling strategy to project frame-based role labeling into 1-minute windows, which avoid missing relevant information about rarer (hence more important) roles, while providing for data and results that are useful for applicative scenarios such group coaching and facilitation; c) extend our investigation to the joint classification of task and socio roles, addressing a much more complex problem that in (Zancanaro, Lepri and Pianesi 2006) and in (Dong et al. 2007).

The Mission Survival Corpus I

For the experiments discussed in this paper, we have used the Mission Survival Corpus I (Pianesi et al. 2008b), 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 (Hall and Watson 1970). 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 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.

The Functional Role Coding Scheme

The Functional Role Coding Scheme (FRCS) was partially inspired by Bales' Interaction Process Analysis (Bales 1970). Its eight labels 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 expertise of members; the Socio Emotional Area, which is concerned with the relationships between group members and the "functioning of the group as a group". We give now a synthetic description of the FRCS; for more details, see (Pianesi et al. 2008b).

The Task Area functional roles include: the Orienteer, who orients the group by introducing the agenda, defining goals and procedures, keeping the group focused and summarizing the most important arguments and decisions; the Giver, who provides factual information and answers to questions, states her beliefs and attitudes, and expresses personal values and factual information; the Seeker, who requests information, as well as clarifications, to promote effective group decisions; the Follower, who just listens, without actively participating in the interaction.

The Socio-Emotional functional roles include: the Attacker; who deflates the status of others, expresses disapproval, and attacks the group or the problem; the Protagonist; who takes

the floor, driving the conversation, fostering a personal perspective and asserting her authority: the Supporter, who shows a cooperative attitude demonstrating understanding, attention and acceptance as well as providing technical and relational support; the Neutral, played by those who passively accept the ideas of the others, serving as an audience in group discussion.

Participants usually 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 data reported in Tables 1(a) and 1(b) show that the social roles and task roles can be highly correlated so that a person's social role suggests one or two most probable task roles, and vice-versa (the percentage of redundancy between the two role classes is 32.1%). It might be, therefore, worth exploring the possibility of taking advantage of this fact by pursuing the joint classification of task and social roles.

Table 1: (a) Probabilities of task roles given a social role, (b) Probabilities of social roles given a task role.

(a)	Orient.	Giv.	Seek.	Foll.
Supp.	0.340	0.377	0.054	0.229
Prot.	0.067	0.780	0.039	0.115
Att.	0	0.403	0.372	0.225
Neut.	0.012	0.119	0.018	0.850

(b)	Supp.	Prot.	Att.	Neut.
Orient.	0.585	0.238	0.000	0.177
Giv.	0.126	0.538	0.004	0.332
Seek.	0.183	0.272	0.039	0.506
Foll.	0.030	0.032	0.001	0.937

Audio-Visual signals

Speech Features

Existing works suggests that speech can be very informative about social behavior. For instance, (Pentland 2008) 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 (Pentland 2008) 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 (Rienks and Heylen 2006).

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 (Pentland 2008), 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.

Given that some of the speech features (features 1, 2, 3, 4, 7, 8, 9 and 10) are defined only over voiced segments; each one-minute window was segmented into voiced and unvoiced segments as in (Basu 2002) and the features were computed over the voiced segments. Other features, related to the patterns of the voicing segments (5, 6, 11, 12, 13 and 14), where computed directly from the outcome of the segmentation.

Body Gestures

Body gestures have been successfully used to predict social and task roles (Dong et al. 2007). 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 (Chippendale 2006). 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.

Modeling and Prediction

We modeled role assignment as a multi-class classification problem and used the Influence Model (IM) as classifier (Dong 2006). The IM assumes that people influence each other, accounting for the intuition that the current role of a person is influenced by those of other participants. For example, it can be expected that if a person acts as a giver, other participants might be listening to her, hence acting as followers.

The IM is a team-of-observers approach to complex and highly structured interacting processes: different observers look at different data, and can adapt themselves according to different statistics in the data. The different observers find other observers whose latent states, rather than observations, are correlated, and use these observers to form an estimation network. In this way, we effectively exploit the relationships among the underlying interacting processes, while avoiding the risk of over-fitting. Another advantage of the IM is the modularity of the influence matrix, which allows generalizing the influence model to groups with different numbers of participants.

The representation of the model is similar to the HMMs with a small difference. Each Markov process independently is non-stationary and the transition probabilities $p(\mathbf{x}_i(t)|x_i(t-1))$ for a chain i is given as

$$p\left(x_i\left(t\right)|\, x_i\left(t-1\right)\right)=\sum_{j=1}^{C_N}\left(d_{j,i}\sum_{x_j=1}^{X_N}a\left(x_j,x_i\right)p\left(x_j\left(t\right)\right)\right),$$

where $d_{j,i}$ represents the influence between processes j and i, and $a(x_j,x_i)$ represents the influence between the states x_j and x_i of the interacting processes j and i. We used four interacting Markov processes to model the evolution of task roles and four to model the evolution of social roles of the four participants. The observations for the individual processes are the participants' raw features. The latent states for the individual processes are the role labels. In the training phase of influence modeling, we find out the observation statistics of different functional role classes, as well as the interaction of different participants with the EM (expectation maximization) algorithm, based on the training data. In the prediction phase, we infer the individual participant's social/task roles based on observations about her, as well as on observations about the interactions with other participants.

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, were computed on 1-minute windows. Hence a decision must be taken as to how the frame-based annotation should be projected at the level of the 1-minute window. As to the relational roles, the usage of a straightforward frequency criterion (call it *Heuristic 1*) resulted in highly unbalanced data, with most of the windows labeled as neutral/follower. Tables 2(a) and 2(b) show the distributions of social and task roles obtained through Heuristic 1.

Table 2. Distributions of social (a) and task (b) roles after the application of *Heuristic 1*.

(a)

Supporter	Protagonist	Attacker	Neutral
0.034	0.158	0	0.808

(b)

Orienteer	Giver	Seeker	Follower
0.038	0.210	0.003	**0.749**

In the following we will take as benchmark the classifier assigning the most common role.

Table 3 shows the accuracy of predicting social and task roles using visual features, speech features, and their combination on the data obtained through Heuristic 1. The training of the IM was performed using a leave-one-out procedure on the meetings.

Table 3: Accuracy for social and task roles prediction with *Heuristic 1*.

	Socio	Task
Visual	0.71	0.68
Audio	0.75	0.74
Joint	0.77	0.74

The results are comparable to those in (Dong et al. 2007), and show that better accuracy is obtained with audio features than with visual features. However, these figures do not do any better than the baseline; see bold figures in Table 2. Moreover, as already pointed out, Heuristic 1 makes for a task of low interest because it inflates the contribution of the most frequent roles. To provide for a more balanced and more interesting data set and not to miss rarer roles, we have exploited a slightly different labeling heuristic, whereby a one-minute window is given the most frequent among the non-neutral (or non-follower) labels if that label occupied at least one fourths of the window; otherwise the window is labeled as neutral (follower). This strategy (*Heuristic 2*) avoids inflating frequent roles, missing non-neutral/non-follower ones, and provides for data that are more useful in the automatic facilitation/coaching scenarios described at the beginning of this paper, where finding out about non-neutral/non-follower roles is crucial. Table 4 reports the resulting

distribution of roles in the corpus. As can be seen, a more balanced distribution emerges.

Table 4. Distributions of social (a) and task (b) roles after the application of *Heuristic 2*

(a)

Supporter	Protagonist	Attacker	Neutral
0.149	0.326	0.002	**0.522**

(b)	Orienteer	Giver	Seeker	Follower
	0.070	**0.517**	0.017	0.395

We trained two independent classifiers, one for social roles and one for task roles, using visual features, speech features and their combination. Table 5 reports the accuracy scores.

Table 5: Accuracies for social and task roles using independent classifiers on Heuristic 2data.

	Social roles	Task roles
Visual	0.51	0.43
Audio	0.57	0.61
Joint	0.57	0.58

The results obtained by means of the sole speech features are always better than those obtained by means of the visual features and those obtained using a combination of speech and visual features (multimodal features). 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.

We also considered the contribution of the various audio feature classes . Table 6 shows the accuracy values obtained using independent classifier.

Table 6: Accuracies for social and task roles (independent classifiers) with the different classes of speech features on Heuristic 2 data.

	Social roles	Task roles
Consistency	0.50	0.47
Spectral Center	0.50	0.51
Activity	0.60	0.62
Mimicry	0.50	0.37
Influence	0.54	0.52

Interestingly, the Activity class yields accuracy values (slightly) higher than those produced through the usage of the entire set of audio features, cf. Table 5. Hence using the sole set of Activity 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 7 reports the distribution of the joint roles in the corpus, while Table 8 the classification results.

Table 7. Distribution of social and task roles with Heuristic 2.

	Sup.	Prot.	Attack.	Neut.
Orient.	0,011	0,023	0,000	0,037
Giv.	0,077	0,169	0,001	**0,270**
Seek.	0,003	0,006	0,000	0,009
Foll.	0,059	0,129	0,001	0,206

Table 8: Accuracy of joint prediction of social and task roles.

	Social roles	Task roles
Visual	0.47	0.41
Audio	0.58	0.60
Joint	0.59	0.56

The results are interesting. Notice, first of all, that the accuracies are always much higher than the baseline, see the bold figure in Table 7. 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.

Conclusions

In this paper we have investigated the prospects of extending the use of honest signals to the prediction of functional relational roles. In doing so, we have compared them to some visual features that previous researched (Zancanaro, Lepri, and Pianesi 2006; Dong et al. 2007) have exploited for a similar task. Finally, we have compared the independent classification task consisting in the separate prediction of task and socio roles, to the corresponding joint classification task. The results show that a) our honest signals are, overall, superior to the visual features; b) the sole activity features provide for as much (if not more) classificatory power than the whole set of speech features; c) the joint classification task, exploiting the redundancy between the task and the socio roles, provides results that overall compare very well with those obtained by means of independent classifiers.

References

Basu S.; Conversational Scene Analysis, PhD thesis, MIT, 2002

Benne K.D.; Sheats P. 1948. Functional Roles of Group Members, *Journal of Social Issues* 4, 41-49.

Chippendale P. 2006. Towards Automatic Body Language Annotation. In 7th International Conference on Automatic Face and Gesture Recognition - FG2006 (IEEE) Southampton, UK.

Curhan J.R.; and Pentland A. 2007. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, vol. 92, n° 3, pp. 802-811.

Dong W. 2006. Influence Modeling of Complex Stochastic Processes, Masters thesis, MIT.

Dong W.; Lepri B.; Cappelletti A.; Pentland A.; Pianesi F.; Zancanaro M. 2007. Using the influence model to recognize functional roles in meetings. ICMI07.

Garg N.P.; Favre S.; Salamin H.; Hakkani Tur D.; and Vinciarelli A. 2008. Role Recognition for Meeting Participants: an Approach based on Lexical Information and Social Network Analysis. In Proceedings of ACM International Conference on Multimedia, Vancouver (Canada).

Hall J. W.; Watson W. H. 1970. The Effects of a normative intervention on group decision-making performance. In *Human Relations*, 23(4), 299-317.

Hung H.; Huang Y.; Friedland G.; and Gatica-Perez D. 2008. Estimating the Dominant Person in Multi-Party Conversations using Speaker Diarization Strategies. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas.

Jayagopi D.; Ba S.; Odobez J-M.; and Gatica Perez D. 2008. Investigating two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In Proceedings International Conference on Multimodal Interfaces (ICMI), Special Session on Social Signal Processing.

McCowan I.; Bengio S.; Gatica-Perez D.; Lathoud G.; Barnard M.; and Zhang D. 2005. Automatic Analysis of Multimodal Group Actions in Meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27 (3), pp. 395-317.

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

Pianesi F., Mana N., Cappelletti, A., Lepri, B., and Zancanaro. 2008a. Multimodal Recognition of Personality Traits in Social Interactions. In Proceedings of International Conference on Multimodal Interfaces (ICMI). Special Session on Social Signal Processing.

Pianesi F.; Zancanaro M.; Not E.; Leonardi C.; Falcon V.; and Lepri B. Multimodal Support to Group Dynamics. 2008b. In *Personal and Ubiquitous Computing*. Vol. 12, No. 2.

Rienks R.; and Heylen D. 2006. Dominance Detection in Meetings Using Easily Obtainable Features. In Revised Selected Papers of the 2nd MLMI Workshop. Edinburgh, Scotland.

Vinciarelli A. 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9 (6).

Zancanaro M.; Lepri B.; Pianesi F. 2006. Automatic Detection of Group Functional Roles in Face to Face Interactions. In Proceedings of International Conference of Multimodal Interfaces, ICMI06.