International Doctorate School in Information and Communication Technologies

DIT - University of Trento

MULTIMODAL RECOGNITION OF SOCIAL BEHAVIORS AND PERSONALITY TRAITS IN SMALL GROUP INTERACTION

Bruno Lepri

Advisor:

Dr. Fabio Pianesi

Foundation Bruno Kessler

Abstract

In recent years, the automatic analysis of human behaviour has been attracting an increasing amount of attention from researchers because of its important applicative aspects and its intrinsic scientific interest. In many technological fields (pervasive and ubiquitous computing, multimodal interaction, ambient assisted living and assisted cognition, computer supported collaborative work, user modelling, automatic visual surveillance, etc.) the awareness is emerging that system can provide better and more appropriate services to people only if they can understand much more of what they presently do about users' attitudes, preferences, personality, etc., as well as about what people are doing, the activities they have been engaged in the past, etc. At the same time, progress on sensors, sensor networking, computer vision, audio analysis and speech recognition are making available the building blocks for the automatic behavioural analysis. Multimodal analysis—the joint consideration of several perceptual channels—is a powerful tool to extract large and varied amounts of information from the acoustical and visual scene and from other sensing devices (e.g., RFIDs, on-body accelerometers, etc.).

In this thesis, we consider small group meetings as a challenging example and case study of real life situations in which the multimodal analysis of social signals can be used to extract relevant information about the group and about individuals. In particular, we show how the same type of social signals can be used to reconstruct apparently disparate and diverse aspects of social and individual life ranging from the functional roles played by the participants in a meeting, to static characteristics of individuals (personality traits) and behavioural outcomes (task performance).

Keywords
Automatic Behavior Analysis, Social Signal Processing, Multimodality, Machine Learning

Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

1.1. Motivation

In recent years, the automatic analysis of human behaviour has been attracting an increasing amount of attention from researchers because of its important applicative aspects and its intrinsic scientific interest. In many technological fields (pervasive and ubiquitous computing, multimodal interaction, ambient assisted living and assisted cognition, computer supported collaborative work, user modelling, automatic visual surveillance, etc.) the awareness is emerging that system can provide better and more appropriate services to people only if they can understand much more of what they presently do about users' attitudes, preferences, personality, etc., as well as about what people are doing, the activities they have been engaged in the past, etc. At the same time, progress on sensors, sensor networking, computer vision, audio analysis and speech recognition are making available the building blocks for the automatic behavioural analysis.
Multimodal analysis—the joint consideration of several perceptual channels—is a powerful tool to extract large and varied amounts of information from the acoustical and visual scene and from other sensing devices (e.g., RFIDs, on-body accelerometers, etc.). It has found major applications to advanced interaction modalities, providing flexible and efficient ways to enhance human-computer interaction (see for example [Oviatt, 2002]). Language apart, multimodal analysis has so far targeted low level behaviours, only recently attempting to put its power to the service of the reconstruction and understanding of high level aspects of human behaviour. Within this moving picture, an important subfield is emerging that attempts to understand social behaviour by exploiting so called 'social signals' [Pentland, 2008; Gatica-Perez, 2009; Vinciarelli et al., 2009], a multiplicity of non-verbal cues including prosodic features, facial expressions, body postures and gestures, whose correct manipulation is a prerequisite of social and emotional intelligence, hence for one of the most important aspects for human life [Goleman, 2006]. The possibility of building computer applications able to understand, simulate or manipulate social signals opens new scenarios in human-computer and human-human interaction, by producing a new generation of socially-aware machines and interfaces, built for humans and based on models of human behaviour. In particular, an attracting possibility is that the analysis of social signals provides direct access to high level aspects — be they dynamic, such as social roles or emotions, or static, such as personality traits — without the necessity of going through intermediate concepts, such as those commonly employed in a coarse grained description of social behaviour (what was the meaning of X's statement; what did X reply to Y's statements, etc.). This minimalist path to high level behavioural aspects finds sound empirical justifications in psychological studies that demonstrate how humans can form judgments about complex aspects of social life — other people's traits, preferences, and dispositions [Ambady et al., 2000] — outcomes of tasks and social processes — teaching and jobs performances, negotiations, outcomes of employment interviews

[Ambady et al., 2000] — and intelligence just by considering short behavioural sequences of low level signals, the so called *thin slices* of behaviour. 'Thin slices' are brief excerpts (less than five minutes) of expressive behaviours sampled from the behavioural stream [Ambady and Rosenthal, 1992; Ambady et al., 2000] that reveal and convey information related with a wide spectrum of psychological constructs and phenomena including internal states, personality traits, social relationships and social behaviors. Their potential has recently stimulated the interest of computer scientists, who have started investigating their usage for the automatic prediction of business negotiations outcomes [Curhan and Pentland, 2007] and of social verticality [Jayagopi et al., 2008].
For many of the task of interest, 'thin slices' are based on *honest signals*, signals that are reliable indicators of internal states, dispositions as well as of social dynamics because of their being too costly to fake thanks to their roots in our brain structure and biological nature [Pentland, 2008].

1.2. The Problem

In this thesis, we consider small group meetings as a challenging example and case study of real life situations in which the multimodal analysis of social signals can be used to extract relevant information about the group and about individuals. In particular, we show how the same type of social signals can be used to reconstruct apparently disparate and diverse aspects of social and individual life ranging from the functional roles played by the participants in a meeting, to static characteristics of individuals (personality traits) and behavioural outcomes (task performance).

1.2.1 The Automatic Recognition of Personality

Personality is the complex of all the attributes — behavioral, temperamental, emotional and mental — that characterize individual dispositions. Humans have the tendency to understand, explain and predict other humans' behavior in terms of stable properties — personality traits — that are variously assorted on the basis of the observation of everyday behavior. In this sense, the attribution of personality traits and their usage to infer about the others is a fundamental property of our folk psychology [Andrews, 2008] and an important determinant of social interaction. In this respect, it is remarkable that most of those trait attributions are based on impressions and judgments made from 'thin slices' of the very first moments of acquaintance with previously unknown people [Ambady et al, 2000]. Given that humans are quite good at social cognition and that trait attribution and their usage to explain and predict behaviour is part and parcel of their folk psychology, one might argue that:

1. computers could deploy similar capabilities, where they are able to summarize people's dispositions into traits. For instance, social network websites could try to increase the chances of a successful relationship based on knowledge about the traits of the network members [Donnellan et al., 2004]. Tutoring systems could be more effective if they could adapt themselves to the learner's personality [Komarraju and Karau, 2005]. Some studies proved that users' evaluation of conversational agents depends on their own personality ([Reeves and Nass, 1996]; [Cassell and Bickmore, 2003]); consequently, a requirement

for such systems to adapt to the users' personality, like humans do, is emerging ([Funder and Sneed, 1993]; [McLarney-Vesotski, 2006]). Because of its relevance in social settings, information on user' personality could be useful in personalized support to group dynamics.

2. impressionistic trait attribution based on thin slices might play with computers a role similar to that played with humans; that is, they might suffice to provide the machine with enough information about people's disposition to make the first task feasible. Our work addresses this second task.

In folk-psychological practice, the personality of a person is assessed along several dimensions: we are used to talk about an individual as being (non-)open-minded, (dis-)organized, too much/little focused on herself, etc. Several existing theories have formalized this folk-psychological practice to model personality by means of multi-factorial models, whereby an individual's 'objective' personality is described in terms of a number of more fundamental dimensions known as traits. A well known example of a multi-factorial model is the Big Five [John and Srivastava, 1999] which owes its name to the five traits it takes as constitutive of people's personality:

1. Extraversion vs. Introversion (sociable, assertive, playful vs. aloof, reserved, shy);
2. Emotional stability vs. Neuroticism (calm, unemotional vs. insecure, anxious);
3. Agreeableness vs. Disagreeable (friendly, cooperative vs. antagonistic, faultfinding);
4. Conscientiousness vs. Un-conscientiousness (self-disciplined, organized vs. inefficient, careless);
5. Openness to experience (intellectual, insightful vs. shallow, unimaginative)

Despite some known limits ([Eysenck, 1991]; [Paunonen and Jackson, 2000]), over the last 50 years the Big Five has become a standard in Psychology. At least three groups of researchers have worked independently on this problem and have identified the same Big Five factors: Goldberg at the Oregon Research Institute [Peabody and Goldberg, 1989], Cattell at the University of Illinois [Cattell, 1957; Cattell and Mead, 2007], and Costa and McCrae at the National Institutes of Health [Costa and McCrae, 1992; McCrae and John, 1992]. Despite the different methodologies exploited, the different names and sometimes the different internal constitutions of the five factors, the consensus is high on their meaning and on their breadth of coverage [Grucza and Goldberg, 1988]. Moreover, the Big Five model has found wide application is such areas as consulting, clinical psychology, personnel selection, orientation, etc.
Moreover, experiments show that personality traits influence many aspects of task-related individual behaviour such as leadership [Hogan et al., 1994], attitude toward machines [Sigurdsson, 1991], attitude toward some basic dimensions of adaptivity [Graziola et al., 2005].
In our thesis, we limit ourselves to the extraversion-introversion dimension of the Big Five. The choice of the this trait was due to the fact that of the Big Five traits, Extraversion is the one that shows up more clearly in, and has the greater impact on, social behaviour [Funder, 2001].

In Eysenck's theory [Eysenck and Eysenck, 1964], extroverts are sociable, have many friends, like parties, need to have people talk to, and do not like reading or studying by themselves. They take chances, crave excitement, act on the spur of the moment, and are generally impulsive. They are fond of practical jokes, always have a ready answer, and generally like change; they are carefree, easygoing, optimistic, and like to 'laugh and be merry'. They prefer to keep moving and doing things, tend to be aggressive and lose their temper quickly; altogether their feelings are not kept under tight control, and they are not always reliable persons.
Introverts, in turn, are quiet, retiring persons, introspective, fond of books rather than people; they are reserved and distant except to intimate friends; they tend to plan ahead, 'look before they leap', and distrust the impulse of the moment. They do not like excitement, take matters of everyday life with proper seriousness, and like a well-ordered mode of life. They keep their feelings under close control, seldom behave in an aggressive manner, and not lose their temper easily. They are reliable but somewhat pessimistic.
Moreover, correlation has been shown between extraversion and speech behavior, in particular with prosodic features: higher pitch and higher variation of the fundamental frequency [Scherer, 1978; Scherer, 1979], fewer and shorter silent and filled pauses, and higher voice quality and intensity [Mallory and Miller, 1958]. Other studies on the differences between the communication styles of introverts and extroverts suggest that the latter speak more and more rapidly, with fewer pauses and hesitations [Furnham, 1990].
Besides models that, as the Big Five, attempt to provide a comprehensive assessment of people personality, others have privileged specific dimensions, possibly useful to characterize specific dispositions in specific domains. An example is the so-called Locus of Control (LoC) [Rotter, 1965], which measures whether causal attribution [Weiner, 1974] for one's behavior or beliefs is made to oneself or to external events or circumstances. It consists of a stable set of belief about whether the outcomes of one's actions are dependent upon what the subject does (internal orientation) or on events outside of her control (external orientation) [Rotter, 1965]. For example, college students with a strong internal locus of control may believe that their grades were achieved through their own efforts, while students with a strong external locus of control may believe that their grades are the result of good or bad luck; hence, they are less likely to expect that their own efforts will result in successes and are less likely to work hard for high grades.
According to Rotter [Rotter, 1965], internals exhibit two essential characteristics: high achievement motivation and low outer-directedness. Weiner suggested considering also differences in the stability of the internal and external causes [Weiner, 1974], arguing that attributions could be done to own ability (an internal stable cause), own effort (an internal unstable cause), task difficulty (an external stable cause) or luck (an external, unstable cause).
LoC has been used as an empirical tool in several domains; for instance, it was shown to play a major role in determining the social agency that people attribute to computers, with internals inclining towards seeing the computer as a tool that they can control and use to extend their capabilities, and externals much more prone to regard computers as an autonomous social entity they are forced to interact with [Johnson et al., 2002].

1.2.2 Automatic Recognition of Social Behaviors

Small group interactions, such as meetings, are more and more important in structuring our daily work life inside organizations. For example, according to a survey in [Doyle and Straus, 1993] executives spend on average 40%-50% of their working hours in meetings and 50% of that time is unproductive and up to 25% of it is spent discussing irrelevant issues. The problems are not only due to task related factors (e.g., a difficult of choosing the right items for the agenda, and/or of focusing the attention on relevant issues), but most often than not by the complexity of group dynamics and social behaviours, which hinders the team's performance. Different means and tools can be put at work to support dysfunctional teams, ranging from facilitation to training sessions conducted by experts. The availability of rich multimodal information makes it possible to explore the possibility of providing some of these services automatically or semi-automatically.
For instance, in [Pianesi et al., 2006], the usefulness and the acceptability of a functionality inspired by the practice of *coaching* [Bloom et al., 2003] were investigated; it consisted of a report about the relational/social behaviour of individual participants, which were generated from multimodal information extracted during the meeting, and privately delivered after the meeting was over. A notable finding reported in [Pianesi et al., 2006] was that people who were given the report did not find any significant difference in terms of usefulness, reliability, appropriateness, completeness and clarity according to whether it was produced by an automatic system or by a human expert.
Obviously, to implement such functionality an automatic system should be able to observe the meeting as a coach would; this means that system does not keep trace of exactly what people said and what people did during the meeting. These reports, in fact, are not minutes but represent a more qualitative and meta-level interpretation of the social dynamics of the group. They do not contain information as "in the first part of the meeting you have talked for ten minutes about machine learning techniques useful to solve the problem" but rather "in the first part of the meeting you have provided the group with background information" or "you have prevented others from intervening in the discussion" In practice, the system have to abstract over low level (visual, acoustic, etc.) information to produce medium-/coarse-grained one about social behaviours of the members, for example about the roles that members play in the group (who is the protagonist/the leader? who is the person less involved in the discussion?, and so on) . This latter is, in fact, the kind of information that most coaches and group facilitators use implicitly or explicitly while doing their job.
The term role has been treated in various ways in the sociological and psychological literature. Three perspectives on roles became important inside the small group research. The first considers roles as the expectations regarding the behavior of a specific individual [Bormann, 1990]. The second perspective emphasizes the behaviors associated with a particular position in a group or in an organization [Katz and Kahn, 1978; McGrath, 1984]. Finally, the third views roles as behavior enacted by individuals in a particular context [Biddle, 1979; Salazar,1996].
In contrast to viewing roles as the expectations of others, several authors [Katz & Kahn, 1978; McGrath, 1984] have taken the roles as equivalent to positions in a larger system or organization. In particular, Katz and Kahn [Katz and Kahn, 1978] define roles as the activities or behaviors expected from a person in a particular office. Hence, according to this view a role is "a set of ex-

pected activities associated with the occupancy of a given position" (p. 200). Similarly, for McGrath [1984] a role "is not characteristic of a particular person, but rather is a characteristic of the behavior of the incumbent of a particular position" (p. 249).

We needed a definition of roles based on information about what actually happened in the course of the interaction, and which reduces the resort to knowledge about the group' structure, history, position in the organization, etc. For this reason, we considered those approaches to social dynamics that focus on the roles members play inside the group, as opposed to approaches that define roles according to the social expectations associated with a given position (as in [Katz and Kahn, 1978]). This kind of roles—called *functional roles* [Salazar, 1996]— are defined in terms of the behaviour enacted in a particular context.

Benne and Sheats [Benne and Sheats, 1948] provided a list of functional roles for working groups, and collected them into three classes:

- *Function roles aimed to a task*: generally, these roles tend to facilitate and coordinate the group effort in selecting, defining and carrying out a particular task. The roles in this category are: initiator-contributor, information seeker, opinion seeker, information giver, opinion giver, elaborator, coordinator, orienteer, evaluator-critic, energizer, procedural technician, recorder.
- *Group maintenance oriented roles*: these roles are oriented to sustain the group performance, in particular "those participations which have for their purpose the building of group-centered attitudes and orientation among the members of the group or the maintenance and perpetuation of such group-centered behavior" [Benne and Sheats, 1948, p. 44]. These roles are: encourager, harmonizer, compromiser, gate-keeper, standard setter, group-observer, follower.
- *Individual roles*: these roles tend to satisfy personal needs, also if irrelevant for the group goals and non-oriented or negatively oriented to group building. These roles are: aggressor, blocker, recognition-seeker, self-confessor, playboy, dominator, help seeker, special interest pleader.

Benne and Sheats model is interesting because it provides functional roles adaptable to different interaction contexts. However, those concepts are not enough rigorous for a systematic interaction analysis.

Building on the work of Benne and Sheats, Bales [Bales, 1970] focused his research on the interaction among group members. In particular, Bales classified behaviors into two categories: behaviors which were instrumental or task related and behaviors which were expressive or socio-emotional related. Bales believed that groups have a natural tendency towards equilibrium and, therefore, move through cycles of instrumental and expressive behavior [Bales, 1970]; moreover, he argued that groups with members who play both task and socio-emotional roles tend to be more cohesive.

Building on Benne and Sheats's functional roles and on Bales' two dimensional approach, and drawing on observations performed on a set of face-to-face meetings, the Functional Role Coding Scheme (FRCS), consisting of five labels (Orienteer, Giver, Seeker, Procedural Technician, and Follower) for the Task Area and five labels (Attacker, Gate-Keeper, Protagonist, Supporter, and Neutral) for the Socio-Emotional Area, was produced (see Chapter 3 for a more detailed introduction to the coding scheme).

1.3. Proposed Approach

As said, in this thesis we deal with the multimodal (audio and visual) detection and analysis of persistent and dynamic individual traits and social behaviors during small group interactions.
The methodology used is the statistical machine learning that integrate multiple features extracted from audio and video. More precisely, in this work we deal with classification and regression tasks using various learning approaches: data-driven methods, such as Support Vector Machines [Vapnik, 1995; Cristianini and Shawe-Taylor, 2000], and generative (graphical) models, such as Hidden Markov Models [Rabiner, 1989] and Influence Models [Dong, 2006; Dong and Pentland, 2007].
Our tasks involve individual and social constructs with different characteristics: for example, roles and personality traits have different temporal properties, with traits being stable over time and a definitional property of the behavior of a given individual; while roles change dynamically inside a given face-to-face interaction. These dynamic aspects were accommodated by means of Influence Models, Hidden Markov Models, and sliding windows with Support Vector Machines.
In our approach, we proposed of exploiting simple audio-visual features, the so-called 'honest signals' from social interaction to automatically extract knowledge about the participants' behaviour and the participants' traits. In fact, the non-verbal behaviour is an important and continuous source of signals conveying information about emotions, feeling, mood, personality, and in general traits and behaviours of people [Vinciarelli et al. 2009].
Another pillar of our work is the focus on short behavioral sequences or 'thin slices', recently popularized in Malcolm Gladwell's best-seller, *Blink*, wich emphasises "the ability of our inconscious to find patterns in situations and people based on very narrow slices of experience" [Gladwell, 2005, p. 23].
Hence, we designed our experiments to understand and verify if the 'minimalist' approach to human behavior understanding discussed here could be suited to design and implement automatic systems. Our results suggest that both dynamic and transitory properties as well as persistent properties, such as personality factors, can be extracted with a reasonably high accuracy, showing that machines can indeed take advantage of thin slices of behavior in a fashion that closely resembles their usage by humans [Ambady and Rosenthal, 1992].
Another important issue we addressed in our work is the importance of the social context in modeling individual and social behaviors: in fact, each of our tasks (recognition of functional roles, recognition of personality traits and recognition of performance) can be pursued in, at least, two different manners, each corresponding to a different hypothesis about the way in which these dimensions, as manifested in social interaction, can be assessed.
According to the first, the sole consideration of the target subject' behaviour (her thin slices) is enough: the way she/he moves, the tone and energy of her/his voice, etc., are sufficiently informative. The second view maintains that, the appreciation of these constructs requires information not only about the target's behaviour, but also about the social context: the same behaviour

might have a different import for the assessment if it is produced in a given social environment than in another. Hence, thin slices of the other group members are needed as well.

1.4. Structure of the Thesis

This thesis is organized as follows: the Chapter 2 discusses related works in automatic behavior analysis inside small group interactions. In particular, we focus on previous works on automatic detection of roles played by meeting participants, and previous works on the automatic recognition of personality traits and other relevant individual characteristics, such as dominance.

Chapter 3 introduces our approach to the automatic detection of social behaviors and individual traits. Here we introduce our assumptions concerning the notions of 'thin slices', 'social signals', and the importance of the context for predicting behaviors in small group interactions. Then, an introduction of the history and the importance of 'thin slices' concept and of 'social signal' concept in social psychology is given. Finally, we introduce two corpora, Mission Survival I and Mission Survival II (MSC-I and MSC-II, respectively), that were used for our experiments, and describe the acoustic (five classes of acoustic cues: Conversational Activity, Emphasis, Spectral Centre, Mimicry, and Influence) and visual features (hand, body, and head fidgeting) that were extracted from them.

In Chapter 4, we turn to presenting the first task that we targeted in this thesis: the automatic classification of functional (task and socio-emotional) roles. We modeled role assignment as a multi-class classification problem on a relative large and very unbalanced dataset and we compared three machine learning approaches to this task: Support Vector Machines [Vapnik, 1995; Cristianini and Shawe-Taylor, 2000] with sliding windows (the windows are used to take in account the time dimension) using Radial Basis Function (RBF) and linear kernels, Hidden Markov Models [Rabiner, 1989], and Influence Model (a team-of-observers approach to complex and highly structured interacting processes [Dong, 2006; Dong and Pentland, 2007]). Moreover, we analyzed the predictive power of five classes of acoustic and visual features, the so called 'honest signals', by means of three different measures, namely the misclassification error, the Bayes error and the covariance matrix. To this end we also run some machine learning experiments to verify the power of these classes of signals in classifying functional roles. Finally, we compared classification performances obtained on two different experimental conditions: (i) using as features only the behavioral data of the target subject or (ii) using features from all participants.

In Chapter 5, we report our works on the automatic prediction of two personality traits. In this thesis we focus on two traits related to the personality dimension: Extraversion-Introversion and Locus of Control. In particular, we design two different tasks: a classification one and a regression one. The problem was given the following form: on the basis of 1-minute-long behavioral sequences, the system had to assign the subjects to the right class (classification) or to the scores (regression) obtained by the participants in filling out the standard questionnaires for the two traits. Then, we execute the regression and the classification studies addressing two hypotheses: (a) that some simple feature selection procedures (ANOVA-based for classification and ANOVA-based and Correlation-based approach for regression) could provide a smaller, but still

effective, subset of features, and (b) that the encoding of the social contexts (in the form of the other group members' features) could contribute to regression performance.
Chapter 6 describes an ongoing work on the automatic classification of performance scores obtained by the meeting participants in solving the Mission Survival Task. This experiment addresses the following aspects: (a) whether subsets of the original audio-visual features could do any better than the full set; (b) whether consideration of the context of interaction, encoded by means of the audio-visual features of the other members of the group, provided any advantage; (c) whether the predictive power of our thin slices differs according to their temporal position (beginning of the meeting vs. central part vs. final part).
Finally, Chapter 7 draws the conclusions and outlines future works and research avenues.

1.5. Publications

The work presented in this thesis has been partially previously published in the following papers.

- B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi. "Automatic Prediction of Individual Performance for Thin Slices of Human Behavior". In Proceedings of ACM International Conference on Multimedia 2009 (ACM-MM 2009).

- B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. "Modelling the Personality of Participants during Group Interactions". In Proceedings of User Modelling, Adaptation, and Personalization 2009 (UMAP 2009).

- B. Lepri, A. Mani, A. Pentland, and F. Pianesi. "Honest Signals in the Recognition of Functional Roles". In Proceedings of AAAI Spring Symposium on Human Behavior Modelling 2009.

- F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. "Multimodal Recognition of Personality Traits in Social Interactions". In Proceedings of 10^{th} International Conference on Multimodal Interfaces (ICMI 2008). Special Session on Social Signal Processing and Behavioral Analysis

- W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. "Using the Influence Model to Recognize Functional Roles in Meetings". In Proceedings of 9^{th} International Conference on Multimodal Interfaces (ICMI 2007)

- N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. "Multimodal Corpus of Multi-party Meetings for Social Behavior Analysis and Personality Traits Detection". In Proceedings of ICMI Workshop on Tagging, Mining, and Retrieval of Human-Related Activity Information, 2007

- M. Zancanaro, B. Lepri, and F. Pianesi. "Automatic Detection of Functional Group Roles in Face to Face Interactions". In Proceedings of 8^{th} International Conference on Multimodal Interfaces (ICMI 2006).

Chapter 2

2. State of the Art

2.1. Modeling Social Behaviors

A lot of research has been conducted in social psychology and sociology on the modeling, analysis and understanding of social behaviors and social relations, in particular on the analysis and definitions of roles in small groups. The sociologist Hare proposed the following definition of role ‘‘associated with a position in a group (or status) with rights and duties to one or more group members [. . .] for formal group roles that members perform consciously” [Hare, 1994] (p. 434). He distinguished between: *functional roles* based on differences existing or emerging among group members, which include functions like control of others, status, access to and use of resources, etc.; *sociometric roles*, where people can occupy central, friendly, or isolated roles inside the group’s network; *dramaturgical roles*, involving traditional roles played in social drama, including protagonists, antagonists, or audience members; and *emotional roles*, which involve roles that are not consciously acted or realized by a group, and can include prototypical roles such as hero or clown.
From the computational side, Banerjee and Rudnicky [Banerjee and Rudnicky, 2004] proposed a simple taxonomy of participant roles (presenter, information provider, participator, and information consumer), and then trained a decision tree classifier to learn them from simple speech-based features. The classifier takes as input a feature representation of a short time window (meeting history) and classifies the roles at the end of the window. The method used seven features (the number of speaker changes, the number of speakers, the number of overlaps in speech, the average length of these overlaps in seconds, the total amount of speech spoken by a given participant X in seconds, the number of overlaps initiated by the participant X, and the number of overlaps initiated by some other participant). All features were manually extracted. Different experiments on the same data set produced a best classification accuracy of 53%.
Various approaches have been applied to detect the roles in a news bulletin based on the distinctive characteristics of those roles. Weng et al. [Weng, 2007] used Social Network Analysis (SNA) to identify the hero, the heroine, and their respective friends in three movies based on the co-occurrences of roles in different scenes. Barzilay et al. [Barzilay et al., 2000] exploited the keywords used by the roles, the durations of 3 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. They obtained a best performance of 80.5% classification accuracy on human transcripts using the Maximum Entropy algorithm and a best performance of 77% classification accuracy on automatically recognized transcripts (ASR data) again using the Maximum Entropy algorithm.

Moreover, Vinciarelli studied deeply the problem of role recognition in audio recordings of professional radio news shows [Vinciarelli, 2007]. Six roles corresponded to the different sections that people are responsible for or part of in a news show, including the primary and the secondary anchorman, the guest, the interviewee, the headline person, and the weather person. In this data set the conversations are usually dyadic and the sections of the show follow a regular structure, which facilitates the role recognition problem compared to other settings, such as meetings. The method uses features based on basic concepts of social network analysis and on the duration of each of the role segments. The reported performance was 85% frame-based classification accuracy on 96 bulletins (with an average duration of 12 minutes). Additional experiments with a variation of the approach and another source of radio shows (talk-shows) was presented by Favre et al. with similar performance scores [Favre et al., 2008]. Finally, Favre et al. [Favre et al., 2009] applied Hidden Markov Models and n-gram language models for the recognition of role sequences underlying the sequence of speakers in conversation. The experiments were performed on different kinds of data (around 90 hours of broadcast and meetings data) and they showed that the performance depends on the degree of formality of these roles. The approach, in fact, is particularly suitable for the recognition of formal roles, i.e. those that correspond to specific functions in a given interaction setting (e.g. the moderator in a debate) and impose rigorous constraints on the behavior of people. Informal roles, i.e. those that correspond to a position in a specific social system (e.g., the manager in a company) and do not impose constraints on the behavior of people, are harder to model, but still recognized with a performance higher than chance.

Jayagopi et al. [Jayagopy et al., 2008a] addressed a problem related to the roles, the recognition of a role-based status in small groups. The notion of status can be defined as ‘‘an ascribed or achieved quality implying respect or privilege, [but] does not necessarily include the ability to control others or their resources” [Hall et al., 2005] (p. 898). In the workplace, status often corresponds to a person’s position in a group or an organization’s hierarchy, and it is often defined by a role (e.g. a project manager has a different status in comparison with the assistant). Using 5 hours of meeting data (AMI corpus) divided into time slices of 5 minutes, Jayagopi et al. presented a study on detection of role-based status (the project manager of the team) using various automatically extracted nonverbal features that characterize speaking activity, visual activity, and visual attention. The work showed that the best nonverbal cues (the total number of speaker turns and the total number of times speaking first after a speaker) can correctly predict the project manager with 66.7% of classification accuracy.

Using only acoustic cues Favre et al. [Favre et al., 2008] attempted the recognition of the project manager, the marketing expert, the user interface expert, and the industrial designer in a larger portion of the AMI corpus (138 meetings, 45 hours). Using the data from full meetings, and not only from “thin slices”, the approach extracts features of each person’s occurrence on a set of temporal windows, as well as the proportion of speaking time, and uses a simple Bayesian classifier. For the four-role task, they reported a best performance of 44% classification accuracy. Garg et al. [Garg et al., 2008] 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. In particular, they combined a non-verbal approach with one that uses verbal information (more precisely, words derived from manual or automatic speech

transcripts). The results using verbal features showed a significant improvement over the use of nonverbal information only, with frame-based classification accuracy of 68% for the four roles.
A very recent work has started to examine cases where the main goal is the competition, rather than cooperation or coordination among meeting participants. Raducanu et al. [Raducanu et al., 2009] proposed to investigate the case of role analysis in competitive meetings coming from a popular US reality TV show, where participants aim at getting a real job in a firm. In each episode, after participating in a business-related assigned task among two opposing teams, one participant is fired based on his/her performance in a group meeting led by a strong-minded boss. Raducanu et al. investigated simple approaches based on manually extracted cues related to high social status (speaking time and turns, interruptions, and centrality), and reported performance accuracy for the estimation of both the meeting chairman and the fired person of 85% and 92%, respectively, using 90 min of meeting data corresponding to a full season of the TV show.

2.2. Modeling Personality Traits

The first work addressing the automatic recognition of personality was [Argamon et al., 2005], who used the relative frequency of function words and of word categories based on Systemic Functional Grammar, to train Support Vector Machines with linear kernel for the recognition of Extraversion and Emotional Stability. The data concerning the two personality traits were based on self-reports. Oberlander and Nowson [Oberlander and Nowson, 2006] trained Naive Bayes and Support Vector Machines with linear kernel for four (Neuroticism, Extraversion, Agreeableness, and Conscientiousness) of the Big Five traits on a corpus of personal weblogs, using n-gram features extracted from the dataset. Also their personality data were obtained through self-reports. A major finding of theirs is that the model for Agreeableness was the only one to outperform the base-line. Finally, Mairesse et al. [Mairesse and Walker, 2006; Mairesse et al., 2007] applied classification, regression and ranking models to the recognition of the Big Five personality traits. They also systematically examined the usefulness of different sets of (acoustic and textual) features suggested by the psycholinguistic and psychosocial literature. As to the personality data, they compared self-reports with observed data. Mairesse et al. could show that Extraversion is the easiest personality trait to model from spoken language, that prosodic features play a major role, and that their results were closer to those based on observed personality than on self-reports.
In a recent work, Olguín and colleagues [Olguin Olguin et al., 2009] collected various behavioral measures of the daily activities of 67 professional nurses in a Hospital. The data were collected by means of the *sociometer badge* [Choudhury and Pentland, 2003], a wearable device integrating a number of sensors (an accelerometer, a microphone, and an infrared sensor) measuring aspects such as physical and speech activity, number of face-to-face interactions with other people, level of proximity to relevant objects (people, but also beds, etc.) and social networks parameters. Although the authors' goal was not that of predicting personality traits from those signals, by exploiting simple correlation analysis they were able to prove that the signals they targeted can provide quite a lot of information about people personality.

In general, the field of human computer interaction has shown a recurring interest in the notion of personality. For instance, the latter has found a place in the repertoire of features that a lifelike character should possess in order to improve its believability; the underlying assumption is that a virtual agent would appear more realistic, understandable, and, ultimately, human-like, if, as a human, it exhibited a personality through consistent behaviors that the interacting humans could use in order to understand its goals, form expectations about future behaviors, etc. [André et al., 1999; Castelfranchi and DeRosis, 1999]. In the user modeling literature, information about the personality has been used to help inferring people's goals and motivations from their behavior, as in the work of Zhou and Conati [Zhou and Conati, 2003] in the context of a tutoring system.
Again, at a more general level theoretical frameworks have been studied to provide principled links between people personality and a number of technology-related variables, such as attitudes towards and acceptability of technology. A notable example is the CASA—Computer as Social Agents—framework [Reeves and Nass, 1996] positing that, in certain conditions, the relationship between humans and technology may be modeled in terms of social relations. One might therefore expect that personality plays a role in the way people use and experience technology, an intuition that Goren-Bar et al. [Goren Bar et al., 2006] demonstrated to be true for adaptive systems: strong external orientation correlates with a preference for non-adaptive systems over adaptive ones: people who are highly sensitive to the social facets of technology because of their external LoC are not comfortable with adaptivity, or other forms of control delegation, in technology.
Dominance is a different individual dimension, related to the people personality, that aroused much interest in automatic behavior analysis community: in fact, one component of the so-called vertical dimension of social relationships [Hall et al., 2005]. Dominance is usually seen in two different ways: (i) "as a personality characteristic" (a trait) [Schmid Mast, 2002 p. 421]; or (ii) a sign of "a person's hierarchical position within a group" (a state) [Schmid Mast 2002 p. 421]. In particular, the dominant person is believed to have large influence on a meeting's outcome. Basu et al. [Basu et al., 2001] described an approach to estimate the most influential participant in a debate. The Influence Model was applied to automatically detect the degree of influence a person has on the others. Features related to the speaking activity (manually labeled, such as the speaker turns, and automatically extracted, such as the voicing information and the speaker energy) and the visual activity (region-based motions derived from skin-color blobs) were used. Rienks and Heylen [Rienks and Heylen, 2005] proposed a supervised learning approach based on Support Vector Machines in order to address a three-class classification task in which meeting participants were labeled as having high, normal, or low dominance. They used for the classification task a number of manually produced audio-only cues, both nonverbal (speaking time, number of speaker turns, number of successful floor grabbing attempts) and verbal (number of spoken words). They used a data set containing meetings from two different corpora (M4 and AMI) and they obtained 75% of classification accuracy. More recently, Rienks et al. [Rienks et al., 2006] compared the approaches from [Rienks and Heylen, 2005] and the Influence Model for the same three-classes dominance-level task, on a data set obtained from the AMI corpus larger than the one used in [Rienks and Heylen, 2005] but with similar audio features. The SVM approach outperformed the Influence Model, reporting 70% classification accuracy as best result. In [Rienks et al, 2007] was conducted an analysis of participants influence using the same data as in [Rienks

et al., 2006] and some additional verbal information, such as manually annotations of dialog acts and argumentation categories. The authors reported that the use of the argumentation did not produce influence predictions better than a naive assumption that assigned the most frequent class to all test instances.

Hung et al. [Hung et al., 2007] addressed the task of estimating the most-dominant person in a group using automatically extracted speech features (speaking time and energy) from headset microphones, and kinesic cues (coarse visual activity measures) computed from compressed-domain video recorded using close-up view cameras. The more predictive feature was the speaking time providing a classification accuracy of 85% over 5 hours of the AMI corpus divided in meeting segments of 5 minutes. Recently, a further analysis was conducted by Jayagopi et al. [Jayagopy et al. 2009]. The study included a larger set of nonverbal activity cues, an additional SVM-based approach, and two classification tasks (most-dominant person and least dominant person) divided into two conditions, each of which evaluated data with a different degree of variability with respect to human perception of dominance. The results suggested that, while audio is the most informative modality, visual activity also carried some discriminative power (e.g. the best performance obtained was 79% of classification accuracy for the most-dominant task), and also that nonverbal cue fusion in the supervised setting was beneficial in some cases (e.g. the best performance obtained was 91% of classification accuracy for the most-dominant task). Furthermore, more challenging data in terms of higher variability of dominance judgment by people did translate into a consistent decrease of performance for the automatic methods.

Using the same data, Hung et al. [Hung et al., 2008] investigated the automation of the visual dominance ratio studied in social psychology [Exline et al., 1975], extending it to the multi-party case, revisiting the ‘‘looking-while-speaking” definition to include all people whom a person looks at when she/he talks, and the ‘‘looking-while-listening” case to include all cases when a person does not talk and looks at any speaker. Using visual attention automatically estimated from monocular video [Ba and Odobez, 2008], and speaker turns derived from close-talk microphones, the results for estimating the most dominant person showed that the visual dominance ratio outperformed both its individual components and the total amount of received attention, but also that despite this good performance, certain audio-only cues were still the most discriminant ones. In a different research line, Jayagopi et al. [Jayagopy et al., 2008b] applied the same methodology to the task of classifying the dominant clique (i.e, the subgroup of people who are most dominant) in a conversation, achieving similar performance levels (the best performance obtained was 90% of classification accuracy), and observing similar trends regarding the discrimination of single and fused features.

Finally, Hung et al. [Hung et al., 2008a] studied the problem of estimating the most-dominant person for the single distant microphone case, using a fast speaker diarization algorithm and the speaking time as only feature. The results showed a decreased performance in the estimation of the most-dominant person if compared to one obtained using the close-talk microphone signals. These results are not surprising given the difficulty of accurately segmenting speaker turns from a single audio source. An important problem when using a single audio channel and diarization is the lack of direct ways of associating people identities with the speaker clusters produced by diarization.

Chapter 3

3. The Proposed Approach

The main aspect in which this work is focused on is the multimodal (audio-visual) detection and analysis of persistent and dynamic individual traits during small group interactions.
In particular, we face three challenging tasks: (i) the automatic recognition of functional relational roles (Socio-Emotional Area roles and Task Area roles) played by the meeting participants, (ii) the automatic prediction of two personality traits (Extraversion and Locus of Control) of these participants, and (iii) the automatic prediction of individual performance during the meeting.
Our approach to these challenges is characterized by a common framework based on the following assumptions:

(i) Observing ‘thin slices’ of behavior is enough to understand high-level aspects of humans. In practice, our tasks are similar to ones we, as humans, are routinely involved in when judging and inferring some beliefs about the goals, the intentions, the personality, etc. starting from very short behavioral sequences.

(ii) The importance of the social context in modeling individual and social behaviors: the same behavior might have a different import if produced in a given social environment than in an

(iii) other. In this thesis, we limit ourselves to the consideration of meetings as a challenging example and case study of real life situation in which the multimodal analysis of social signals can be used to extract relevant information about the group and the individuals.

(iv) The predictive power of simple audio-visual cues not related to the semantic contents of the interactions (‘honest signals’).

Regarding the point (i) we used for our learning experiments features vectors obtained concatenating very short behavioral sequences (from some seconds, for predicting the functional roles, to 1 minute, for predicting the personality traits and the individual’s performance).
About the point (ii) we used for predicting the behavior of subject X also information related to the behavior of the other meetings participants.
Finally, regarding the point (iii) we analyzed the predictive power of different subset of audio-visual features. In particular, we used in our task some classes of simple acoustic honest signals (Conversational Activity, Emphasis, Influence, and Mimicry) and three visual features related to the amount of energy in participants’ bodies (head, hands, and body fidgeting).
In this chapter, we introduce two central concepts of our approach, social ‘honest signals’ (Section 4.1) and ‘thin slices’ (Section 4.2). Then, we introduce two corpora (Section 4.3), MSC-I and MSC-II, that were used for our experiments, and describe the acoustic and visual features that were extracted from them.

3.1. Social Signals

In the evolutionary biological literature, the term 'signal' is defined in opposition to the term 'cue': the word 'signal' usually is used for those cues that are meant to serve as communication, either because they have evolved for that purpose or because they are intentionally communicative. Instead, the term "cue" is used to refer to all the things we perceive that indicate some other hidden state or intention [Hasson 1997; Maynard Smith and Harper 2003]. Signals are cues that are meant to indicate some quality and to have a communicative goal. A signal is a perceivable feature or structure that is intended to indicate an otherwise unperceivable quality about the sender. Hence, the purpose of a signal is communication and its goal is to alter beliefs or behaviour of the receiver in ways that benefit the sender [Bradbury and Vehrenkamp, 1998].
Another fundamental quality of the signals is the 'honesty'. The concept of 'honesty' in animal communication is controversial: in fact, this term honesty is used metaphorically because no assumption is made that an animal has 'meanings' that are true or false. The term is used only as a convenient and simple way to describe animal communicative behaviour and this does not at all suppose that the animal has 'intentions' or 'meanings' in any psychological sense [Scott-Phillips, 2008]. A guarantee of the 'honesty' of a signal are the costs paid by the sender: for example, the Handicap Principle is a hypothesis, proposed by biologist Amotz Zahavi [Zahavi, 1975; Zahavi, 1997], to explain how evolution may lead to 'honest' or reliable signaling between animals who have an obvious motivation to deceive each other. This principle suggests that reliable signals must be costly to the signaler: the paradigmatic example is the peacock's tail. Bigger tails leave the peacock less dexterous and less agile, and hence appear to be evolutionarily costly. However, peahens choose to mate with the peacocks with the biggest tails.
In Pentland's view [Pentland, 2008] that we use as reference framework in this thesis, some human social signals are reliable because they are too costly to fake. In human-human interaction the non verbal behavior is a great source of reliable signals which give information about emotions, mental states, personality, attitudes, preferences, and other traits of people [Richmond and McCroskey, 1995]. De Paulo [De Paulo, 1992] affirmed that these expressive non-verbal behaviors are more difficult to suppress and to fake in comparison with verbal behaviors and are more accessible to the external observers. Hence, the lack of control and of accessibility of expressive behavior implies that such behavior provides observer with a relatively sound source of information regarding the internal states and the dispositions of the other subjects. A related implication is that the attempts of intentionally manipulating and faking these expressive behaviors during self-presentations are usually unsuccessful. In fact, expressive behavior could be more revealing of communicative intentions and internal states than what is being consciously and verbally communicated [Ekman and Friesen, 1969]. In this way, by sampling the expressive behavior the 'thin slices' can capture reliable psychological information not subject to a conscious monitoring [De Paulo, 1992; Ekman and Friesen, 1969].
In their survey on Social Signal Processing, Vinciarelli et al. [Vinciarelli et al., 2009] reported a number of expressive behaviors that the research in social psychology has recognized as being the most important for the formation of human judgments about the social behavior. For example, the prosody conveys a large spectrum of socially relevant cues: the pitch influences the perception of extraversion by other observers [Scherer, 1979]; while the speaking fluency gives

good insights about the perceived competence of a given speaker [Scherer, 1979]. Instead, the linguistic vocalizations (also called *segregates*), including all the non-words that are used as words (e.g., “ehm”, “uhm”, etc.), have mainly two functions. The first one is replacing words not founded during a verbal interaction. In this case, they are usually called *disfluencies* and are a signal of a situation of embarrassment [Glass et al., 1982]. The second one is the back-channeling and in this sense these variations can express attention, agreement but also the attempt of grabbing the floor [Shrout and Fiske, 1981].
Another important aspect of the non-verbal behaviour is the silence. It is usually considered as simple non-speech but there are three different kinds of silence in speech: psycholinguistic silence, interactive silence, and hesitation silence [Richmond and McCroskey, 1995].
Hesitation silence takes place when a given speaker has some problems in talking, for example she/he is expressing a difficult notion. The psycholinguistic silence, instead, usually happen at the beginning of a verbal intervention (the speaker is thinking about the words to use). Finally, the interactive silence conveys messages about the actual interaction: it could be work as a way to attract attention, as a way of ignoring interlocutors or as a signal of respect for people we want to listen to.
Another important aspect of the vocal non-verbal behavior is the turn-taking [Psathas, 1995]. In particular, the turn-taking includes two components: (i) the coordination during the speaker changes and transitions [Burgoon et al., 1995], and (ii) the regulation of the conversations [Yule, 1996]. In particular, regarding the coordination role conversations where the latency times between speaker turns are too long sound uncomfortable. Instead, as concerns the regulation role of the turn-taking, it encompasses behaviors aimed at maintaining, requesting, yielding, and denying the conversational turns: gaze and quality of voice are used to signal the so-called transition relevant points [Yule, 1996]. Finally, the overlapping speech is another important aspect that signals disputes and is a display of status and dominance [Smith-Lovin, and Brody, 1989].
As the human voice, the human face also provides a lot of signals useful and essential for interpersonal communication in our daily life: for example, the face is used to regulate the conversation by means of gaze and head nods. Moreover, the face is an important, maybe the preeminent, mean of communicating and interpreting the affective states and the intentions of a given subject on the basis of the shown facial expression [Keltner and Haidt, 1999]. In the same way, also the personality can be seen from the someone’s face [Ambady and Rosenthal, 1992].
Again, postures conveys reliable signals about the attitude of a subject towards a given social situation. An important classification of postural behaviors proposes three different criteria for assessing their social meaning [Scheflen, 1964]. The first distinction is between inclusive and non-inclusive postures and accounts for how much a given posture takes in account the presence of others (facing in the opposite direction of the conversational partners is signal of non-inclusion). The second criterion is the distinction between face-to-face and parallel body orientation: face-to-face interactions are more active and engaging, whereas people sitting in parallel to each other tend to be less mutually interested. The third and last criterion is congruence: symmetric postures are usually a signal of a deep psychological involvement.

3.2. Thin Slices

In social psychology literature, a 'thin slice' is defined as a brief excerpt of expressive behavior sampled from the behavioral stream of a given subject. In particular, by 'brief' it is meant any excerpt with dynamic information but with a duration less than 5 minutes. These 'thin slices' can be sampled from any available channel of communication, for example the face, the body, the voice, the speech, and combinations of them. In fact, the way in which people talk, move, make gestures, and their facial expressions, posture, speech contribute to the formation of impressions about them.

In our daily life, the on-line and dynamic social cognition usually starts with the identification of the expressive behavior. We, as human beings, are able to form immediate impressions and evaluations from the ongoing behavior. Numerous studies in social psychology have shown that social information processing is schema-and-expectancy driven and that following judgments are strongly influenced by initial immediate impressions of expressive behavior.

So, 'thin slices' reveal, conveys and contains information related with a wide spectrum of psychological constructs and phenomena including internal states, personality traits, social relationships and social behaviors. 'Thin slices' of behavior conveys also sounded and reliable information about temporary emotions and affect. Moreover, 'thin slices' are useful also for providing information about chronic and long-lasting affective states as depression and anxiety [Waxer, 1976; Waxer, 1977].

Regarding the personality dimension, observable traits as extroversion and sociability have been studied and recognized starting from brief exposures more successfully than more internal traits such as perseverance and openness to experience [Albright, Kenny, and Malloy, 1988; Paunonen, 1991, Watson, 1989]. However, it is possible that the social context within which these slices are sampled may strongly moderate the extent to which a given trait is manifested [Dabbs, Strong, Milun, Bernieri, and Campo, 1999].

Some personality and dispositional traits and dimensions can be judged rapidly from brief observations [Borkenau and Leibler, 1992; Kenny, 1994, Funder and Sneed, 1993]. In a study, 148 participants were video recorded while they entered in a room walking over to a seated female experimenter who greeted them and then took their seat and begun a brief interview [Dabbs and Bernieri, 1999]. From these tapes only the first 30 seconds were extracted and so this slice contained a little bit more than the entry, the meeting, the greeting and the seating. All the participants had been previously assessed by filling the big five personality traits [Costa and McCrae, 1995]. Some naïve observers judged each of the 148 participants on each of the big five traits. The result of this experiment [Dabbs and Bernieri, 1999] is that judgments of extraversion, agreeableness, conscientiousness, and openness did correlate significantly with targets' traits assessed by the NEO-PI [Costa and McCrae, 1995], while the only trait for which these traits did not correlate was the neuroticism.

Also, interpersonal roles and goals can be revealed by observing 'thin slices' of behavior. For example, interpersonal goals such as forming an impression of the partner or managing the impression of one can be assessed starting from 'thin slices' [Richeson and Ambady, 1999a].

Regarding the dimension of the social relations, two standardized measures that ask to the examiners to draw judgments regarding social relations are composed of a series of 'thin slices': (i)

the Profile of Nonverbal Sensitivity (PONS) [Rosenthal, et al., 1979], and (ii) the Interpersonal Perception Task (IPT) [Costanzo and Archer, 1989].
PONS is composed of 220 video clips, each with a duration of less than 2 seconds. In particular, each scene is extracted from a brief scene in which a woman portrays herself in a number of different social and interpersonal situations (for example, returning an item purchased at a store, admonishing her children, etc.). After removing the verbal content from each of these clips, the accuracy level of the judgments is over the chance and it seems that this 2 seconds of behavior really conveys some diagnostic information of social relations [Rosenthal et al., 1979].
IPT is composed of clips ranging from 30 seconds to 60 seconds. In this task, the observer makes judgments regarding the level of romantic involvement, the status, the winners and the losers in sports competitions [Costanzo and Archer, 1989].
Moreover, "thin slices" have been used also to study interpersonal relationships domain such as power and status hierarchy [Costanzo and Archer, 1989], dominance, kinship, and acquaintanceship [Costanzo and Archer, 1989], and level of romantic involvement [Gada, Bernieri, Grahe, Zuroff, and Koestner, 1997].

3.3. Mission Survival Corpora

In order to provide for as much uniform context as possible, two corpora (MSC-I and MSC-II) of groups engaged in the solution of the Mission Survival Task were collected. The Mission Survival [Hall and Watson, 1970] is a task often used in experimental and social psychology to elicit decision-making processes in small groups. The exercise consists in promoting group discussion by asking participants to reach a consensus on how to survive in a disaster scenario, like a moon landing or a plane crashing in Canadian mountains. The group has to rank a number (15) of items according to their importance for crew members survival. A consensus decision making scenario was chosen and enforced, because of the intensive engagement it requests to groups in order to reach agreement, this way offering the possibility to observe a large set of social dynamics and attitudes. The Mission Survival task was originally designed by the National Aeronautics and Space Administration (NASA) to train astronauts before the first Moon landing and it proved to be a good indicator of group decision making processes[1].
In consensus decision making processes, each participant is asked to express her/his opinion and the group is encouraged to discuss each individual contribution by weighting and evaluating their quality. In our case, consensus was enforced by establishing that any participant's proposal would become part of the common sorted list only if she/he managed to convince the others of the validity of her proposal. We also added an element of competition by awarding a prize to the individual who proposed the greatest number of correct and consensually accepted items.

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

Figure 1. A picture of the experimental setting

Both for MSC-I and for MSC-II, the sessions were recorded in a specially-equipped room at FBK-Irst (see Figure 1 and Figure 2) by means of 4 fire-wire cameras placed in the corners of the room, and 4 actively driven web cameras (PTZ IP cam) installed on the walls surrounding the table. Four wireless close-talk microphones (one for each participant) and one omni-directional microphone placed on tabletop around which the group sat were used to record speech activity.

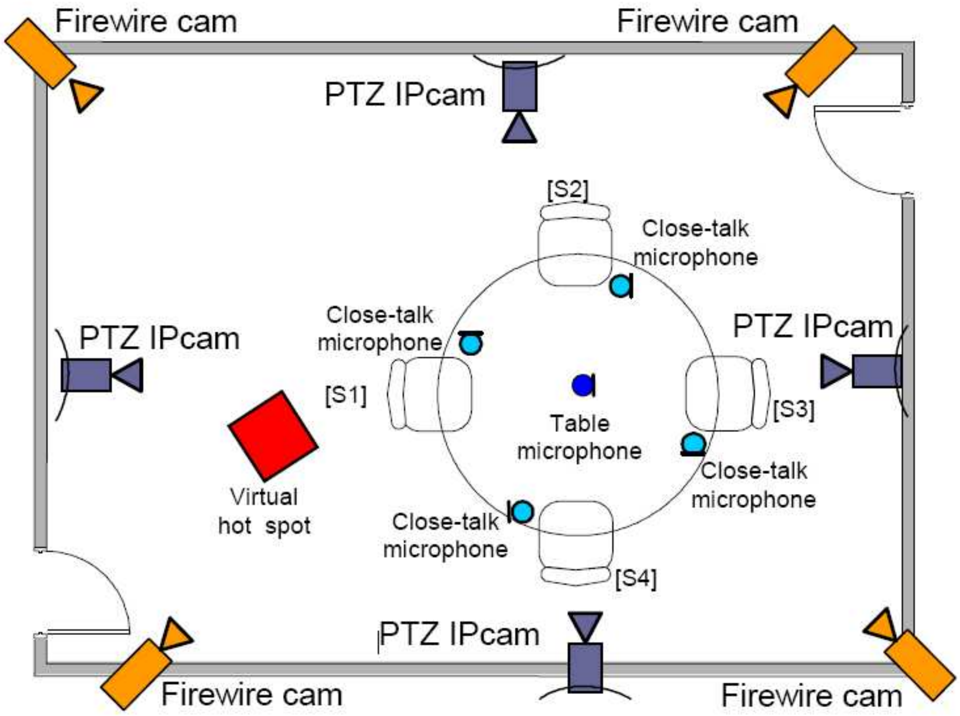

Figure 2. A map of the experimental setting

3.3.1 Speech Features

In our first experiments, the only acoustic feature we considered for MS-I was the speech activity of each participant. Speech activity refers to the presence/absence of human speech. The audio channels from the microphones were automatically segmented at a 500ms frame rate and labeled by means of a VAD - Voice Activity Detector [Carli and Gretter 1992]. For each session, the VAD detected participant's speech activity and produced an output of the form '<temporal frame; label-S1; label-S2; label-S3; label-S4>', where <temporal frame> corresponds to a 500ms interval and <label-*> takes on the values '0' and '1', in correspondence to 'non-speech' and 'speech' respectively, for each participant (speakers S1, S2, S3, and S4).

In more recently experiments with MS-I and in all our experiments with MS-II, we considered a more large set of speech features: in fact, a speech analysis of the recorded audio was conducted in order to extract 22 acoustic features (see Table 1) using the extraction toolbox developed by the Human Dynamics group at Media Lab[2] [Pentland, 2006].

[2] http://groupmedia.media.mit.edu/data.php

LABELS	ACOUSTIC FEATURES
F1	Mean of Formant Frequency (Hz)
F2	Mean of Confidence in formant frequency
F3	Mean of Spectral Entropy
F4	Mean of Largest Autocorrelation Peak
F5	Mean of Location of Largest Autocorrelation Peak
F6	Mean of Number of Autocorrelation Peaks
F7	Mean of Energy in Frame
F8	Mean of Time Derivative of Energy in Frame
F9	Standard Deviation of Formant Frequency (Hz)
F10	Standard Deviation of Confidence in formant frequency
F11	Standard Deviation of Spectral Entropy
F12	Standard Deviation of Value of Largest Autocorrelation Peak
F13	Standard Deviation of Location of Largest Autocorrelation Peak
F14	Standard Deviation of Number of Autocorrelation Peaks
F15	Standard Deviation of Energy in Frame
F16	Standard Deviation of Time Derivative of Energy in Frame
F17	Average length of voiced segment (seconds)
F18	Average length of speaking segment (seconds)
F19	Fraction of time speaking
F20	Voicing rate
F21	Fraction speaking over
F22	Average number of short speaking segments

Table 1. Acoustic features

These 22 acoustic features can be divided in four classes: ‘Activity’, ‘Emphasis’, ‘Mimicry’, and ‘Influence’ [Pentland, 2008]:

- **Activity**, meant as conversational activity level, indicates interest and excitement. Activity level is measured from the following features: energy in frame, length of voiced segment, length of speaking segment, fraction of time speaking, and voicing rate, the number of voiced regions per second speaking. In order to compute all these features except the energy in frame, the acoustic stream of each participant is first segmented into voiced and non-voiced segments, and then the voiced ones are split into speaking and non-speaking segments.
- **Emphasis** is an indication of how strong is the speaker’s motivation. Moreover, the consistency of emphasis signals mental focus, while higher variability signals an openness to influence from other people. Emphasis is measured by the variation in prosody, i.e. pitch and amplitude. The emphasis class encompasses two different sets of features: *consistency* and *spectral centre*. The features for determining consistency are related to the variations in spectral properties and prosody of speech: the less the variations, the higher the consistency. The relevant features are: confidence in formant frequency, spectral entropy, number of autocorrelation peaks, time derivative of energy in frame. The features for de-

termining the spectral centre are formant frequency, value of largest autocorrelation peaks, and location of largest autocorrelation peaks.

- **Mimicry**, meant as the un-reflected copying of one person by another during a conversation (i.e. gestures and prosody of one participant are "mirrored" by another one), is expressed through short interjections (e.g. "uh-huh", "yup") or back-and-forth exchanges consisting of short words (e.g. "OK?", "done!"). Usually, more empathetic people are more likely to mimic their conversational partners: for this reason, mimicry is often used as an unconscious signal of empathy. Mimicry is a complex behavior and therefore difficult to measure computationally. A proxy of its measure is given by the z-scored frequency of short utterances (< 1 second).
- **Influence** is the amount of influence, hence dominance, each person has on another one in a social interaction, and is measured by calculating the number of overlapping speech segments. Influence strength in a conversation can serve as an indicator of attention; it is difficult, in fact, for a person to maintain the rhythm of the conversational turn-taking without paying attention to it.

Activity, emphasis, mimicry and influence signals are all honest ones [Pentland, 2008], in that they refer to "behaviors that are sufficiently expensive to fake that they can form the basis for a reliable channel of communication"; hence, they can be reliably used to predict and explain human behavior in social interactions.
For our analysis and prediction tasks discussed in the following chapters, 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.

3.3.2 Visual features

Both for MSC-I and MSC-II the only visual features considered were the amount of energy in participants' bodies (fidgeting). Fidgeting refers to localized repetitive motions such as when the hand remains stationary while the fingers are tapping on the table, or playing with glasses, etc. Fidgeting was automatically annotated by means of MHI (Motion History Images) [Chippendale, 2006], a technique that uses skin region features and temporal motions to detect repetitive motions in the images; an energy value is then associated to such a motion in such a way that the higher the value, the more pronounced the motion. In the corpora, the annotation for fidgeting consists of an absolute timestamp, followed by the values of the fidgeting energy, all normalized to the fidgeting activity of the person during the entire meeting.
In particular, MSC-II contains an additional fidgeting feature respect to MSC-I: the MSC-I, in fact, contains hand and body fidgeting; while the MSC-II contains hand, body, and head fidgeting.
All these visual features were extracted and tracked for each frame at a frequency of three hertz.

3.3.3 Mission Survival Corpus 1 (MSC-I)

MSC-I consists of the audio-visual recordings of 11 groups of 4 people involved in the Mission Survival task. The participants (40% males and 60% females) were clerks from the administrative services of our research centre. In all cases, they knew each other, and had often been involved in common group activities. The average age was 35 years. All the groups were mixed gender.

3.3.4 Functional Role Coding Scheme (FRCS)

The goal of recognizing and analyzing social behaviors and relations suggested to us of considering those approaches in social psychology that focus on the roles that members plays inside the group, as opposed to other approaches that defined roles according to the social expectations associated with a given position. Our kinds of roles, called functional roles [Salazar, 1996], are defined in terms of the behaviour enacted in a given situation and particular context and so allow to exploit the information about what is actually happening in the course of the interaction, while reducing the need of knowledge related to the group structure, history and to the position of the group members, etc.

In social psychology and in the studies on small group interactions, Benne and Sheats provided a list of functional roles for groups and collected them into three classes: (i) task-oriented, (ii) maintenance-oriented, and (iii) individual-oriented. The first two kinds of roles are related to the group's needs: in fact, task-oriented roles provide facilitation and coordination for the task accomplishment, while maintenance-oriented roles contribute to social structure and interpersonal relations in order to reduce tensions. The third type of roles, the 'individual roles', is focused on the individual and his/her goals and needs rather than on the group. Another interesting aspect of the Benne and Sheats proposal is that during the interaction, each person can enact more than one role.

More recently, drawing on Benne and Sheats Bales [Bales, 1970] proposed the Interaction Process Analysis—IPA, a framework to study small groups by classifying individual behaviour in a two dimensional role space consisting of a Task and of a Socio-Emotional area. So, the roles pertaining to the Socio-Emotional area stem from activities that support, enforce or weaken interpersonal relationships. For example, complimenting another person is a positive socio-emotional behaviour in that it increases the group cohesion and the mutual trust among the members; while on the other side insulting another meeting's participant can undermine social relationships. The other six categories, so called task area roles, pertain to task-oriented activities, that is, behavioural manifestations relating to management and solution of the problem(s) the group is addressing. Giving and asking for information, opinions, and suggestions related to the problem at hand are examples of task-oriented activities.

We decided to employ the Bales' categories, given the wide acceptance of the Interaction Process Analysis, while interpreting his functions as (functional) roles in terms of Benne and Sheats' approach. This move was motivated by the expectation that the behaviour of each participant would not change too often during the meeting, hence the more static concept of a functional role should be more appropriate than the dynamic concept of function. Finally, we further adapted the resulting two-dimensional scheme, adjusting the roles according to observations performed on a number of face-to-face meetings. Our coding scheme—the Functional Role Coding Scheme (FRCS)—consists of five labels for the Task Area and five labels for the Socio Emotional Area. The Task Area includes functional roles related to the facilitation and coordination of the tasks the group is involved in, as well as to the technical skills of the members as they are deployed in the course of the meeting. The Socio Emotional Area involves roles oriented toward the relationships between group members and roles oriented toward the functioning of the group as a group.

Task area functional roles

- The *Orienteer* (o) orients the group by introducing the agenda, defining goals and procedures and keeping the group focused on track. S/He summarizes the main ideas of the group, recording the most important arguments in the discussion, the minutes, and the group decisions. S/He spells out suggestions in terms of examples, offers a rationale for suggestions previously made and tries to deduce how an idea would work out if adopted by the group. From a behavioural point of view, s/he is often the first person to speak, tends to look at all the audience, rather than at one specific person (as opposed to the giver who focuses on the interlocutor); s/he has a major role in structuring the discussion ("ok, let's move on"), and in planning the future works.
- The *Giver* (g) provides factual information and answers to questions. She/He states her/his beliefs and attitudes about an idea, expresses personal values and factual information. From a behavioural point of view, she/he usually speaks if is consulted by another person, then her/his look will mainly directed towards the interlocutor.
- The *Seeker* (s) requests suggestions and information, as well as clarifications, to promote effective group decisions. This role can be mistaken with the Orienteer; however, whereas the latter's questions are mostly meant to help the group reaching the objectives (for example, "what about moving to the next agenda item?"), the Seeker's ones are related to the task under discussion (e.g. "what's the status of project?", "what do you think about adding a new functionality to the system?").
- The *Recorder* (r) uses the resources available to the group, managing them for the sake of the group. The most apparent manifestation (and useful) function of this role consists of keeping tracks of the discussions and the decision for the group. In this respect, it should not be mistaken with the Follower (see below) who takes notes only for his/her own sake.
- The *Follower* (f) only listens, possibly takes notes for personal use, but does not participate actively.

Socio-Emotional area functional roles

- The *Attacker* (a) may work in many ways, deflating the status of others; expressing disapproval of the values, acts or feelings of others; attacking the group or the problem it is working on; joking aggressively and so on. S/He consistently reacts negatively to other's ideas: makes very critical comments, uses humor and so on. The behavioural indicators that signal this role are, among others, an aggressive tone of voice, looking elsewhere, making noise, moving nervously.

- The *Gate-keeper* (gk) is the moderator within the group. S/He mediates the communicative relations and attempts to keep communication channels open by encouraging and facilitating the participation. S/He mediates the differences between other members, attempts to reconcile disagreements, and relieves tension in conflict situations.
- The *Protagonist* (p) is the participant that takes the floor and drives the conversation. She/He assumes a personal perspective asserting his/her authority or superiority because of her/his status or because of the particular task she/he is performing.
- The *Supporter* (su) shows a cooperative attitude indicating understanding, attention and acceptance as well as providing technical and relational support to other members of the group. S/He also keeps a collaborative atmosphere sharing the common objects and trying to make them available to each member.
- The *Neutral* (n) passively accepts the idea of others, serving as an audience in group discussion.

The coding scheme was applied to a corpus consisting of the video and audio recordings of 9 group meetings, for a total of 12.5 hours. Its reliability was assessed on a subset of the corpus consisting of 130 minutes of meetings for the Socio-Emotional Area and 126 minutes for the Task Area. Five people were coded on the Socio-Emotional Area and five in the Task Area by two trained annotators. The annotations were sampled every 10 seconds to get a timed sequence of events which is more suitable for data analysis [Gottman and Roy, 1989]. Then, two confusion matrices were built, one for the Task Area (see Table 2) and one for the Socio-Emotional Area (see Table 3), to measure the cross-judge consistency of class membership using the Cohen's k [Cohen, 1960].

		JUDGE2					
		G	F	O	R	S	Total
JUDGE1	G	115	55	13	3	0	186
	F	3	140	15	18	1	177
	O	2	18	231	0	16	267
	R	1	7	0	81	0	89
	S	0	8	3	0	28	39
Total		121	228	262	102	45	758

Table 2. Confusion matrix for the roles in the Task Area (758*10 secs = 126 minutes): g = Giver; n= Follower; o= Orienteer; r = Procedural Technician; s= Seeker

In the Task Area, Cohen's statistics was $\kappa = 0.70$ (N=758, SE=0.02; p<.001; confidence interval for α=.05: 0.67-0.75). According to Landis and Koch's [Landis and Koch, 1977] criteria, the agreement on the task area is good ($0.6 < \kappa < 0.8$). For the Socio-Emotional the inter-annotator agreement was κ =0.60 (N=783, SE=0.02, p<.001; confidence interval for α=.05: 0.56-0.65).

According to Landis and Koch's (1977) criteria, the agreement on the Socio-Emotional roles is at the borderline between good ($0.6 < \kappa < 0.8$) and moderate ($0.4 < \kappa < 0.6$).

		JUDGE2				
		A	N	P	S	Total
JUDGE1	A	26	1	5	0	32
	N	3	241	29	105	378
	P	0	32	233	12	277
	S	0	14	7	75	96
Total		29	288	274	192	783

Table 3. Confusion matrix for the roles of the Socio-Emotional area (783*10 secs = 130 minutes): a = Attacker; n= Neutral; p= Protagonist; s = Supporter; g= Gate-Keeper (not present)

The FRCS was used to manually annotate the MSC-I: for each participant units spanning 5 seconds were considered and then re-sampled every 330ms to align them with the acoustical features. The corpus was quite unbalanced: *Follower* and *Neutral*—as expected—were the most frequent roles while *Attacker* was quite rare (the participants knew they were observed and perhaps they tended to avoid aggressive or uncooperative behavior). The *Recorder* and the *Gate-Keeper* roles were never observed.

Task Roles			Socio-Emotional Roles		
Follower	71147	66.12%	**Neutral**	78427	72.88%
Orienteer	5458	5.07%	**Gate-Keeper**	0	0%
Giver	28214	26.22%	**Supporter**	9401	8.74%
Seeker	2789	2.59%	**Protagonist**	19487	18.11%
Recorder	0	0%	**Attacker**	293	0.27%
107608			107608		

Table 4. Distribution of the categories in the corpus (330ms time stamp)

The corpus was then reduced by considering only time units where the relevant participant was speaking. This lowered the impact of the Follower and Neutral roles even if the datasets remained quite unbalanced (see Table 5).

Task Roles (reduced)			Socio-Emotional Roles (reduced)		
Follower	10462	31.74%	**Neutral**	78427	44.74%
Orienteer	3567	10.82%	**Gate-Keeper**	0	0%
Giver	17659	53.57%	**Supporter**	5579	16.93%
Seeker	1275	3.87%	**Protagonist**	12460	37.80%
Recorder	0	0%	**Attacker**	177	0.54%
	32963			32963	

Table 5. Distribution of the categories in the reduced corpus by considering speech events only (330ms time stamp)

The final annotated MSC-I consisted of 107608 rows each reporting the speech activity of one of the participants during a 330ms interval, his/her hands and body fidgeting, the number of people speaking during that time, and the functional roles that the person is playing.

The data reported in Tables 6 and 7 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.

	Orienteer	Giver	Seeker	Follower
Supporter	0.340	0.377	0.054	0.229
Protagonist	0.067	0.780	0.039	0.115
Attacker	0	0.403	0.372	0.225
Neutral	0.012	0.119	0.018	0.850

Table 6. Probabilities of task roles given a social role

	Supporter	Protagonist	Attacker	Neutral
Orienteer	0.585	0.238	0.000	0.177
Giver	0.126	0.538	0.004	0.332
Seeker	0.183	0.272	0.039	0.506
Follower	0.030	0.032	0.001	0.937

Table 7. Probabilities of social roles given a task role

Another interesting aspect discovered analyzing the MS-1 corpus is that some individuals take certain functional roles consistently often, while some other individuals take these roles consistently rarely. Moreover, the functional roles have their respective characteristics, for example related to the durations, and interactions with other functional roles, independent of who take them. 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 speaks much less than the other roles in time windows of up to several minutes.

Mean (Std)	Attacker	Neutral	Protagonist	Supporter	Marginal
Giver	8(6)	10(16)	23(24)	11(7)	19(20)
Follower	2(2)	52(79)	4(6)	5(5)	34(45)
Orienteer	n/a	4(8)	10(9)	18(16)	17(14)
Seeker	7(4)	6(4)	9(7)	10(5)	9(5)
marginal	9(4)	56(85)	26(27)	15(14)	7(50)

Table 8. Durations in seconds of task roles, socio-emotional roles and their combination (each entry gives the mean and the standard deviation)

Table 8 shows 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 seconds vs. 26 seconds). This seems highlighting the fact that a Protagonist is the main role inside a small group interaction and a Supporter takes a secondary importance. An Attacker role takes an average duration of 9 seconds; this reflects a role's strategy to show his contrasting ideas concisely, so that he can make constructive utterances and avoid conflicts at the same time. A participant asks questions, when he takes an Seeker's role more shortly than he provides information, when he takes an Giver's role. A Protagonist role is on average 37% longer: a discussion is usually driven by one person and thus has a single protagonist at a time. The durations of the neutral (Follower and Neutral) roles in the task and in the social areas 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.

Mean(Std)	Attacker	Neutral	Protagonist	Supporter	Total
Giver	5	316	233	112	666
Follower	9	426	185	147	767
Orienteer	0	67	21	53	141
Seeker	5	74	27	17	123
Total	19	883	466	329	1697

Table 9. Number of instances that a person takes a task role, a social-role and a their combinations.

The Table 9 shows the amount of time that the meeting participants take the different task roles, social roles and the combinations of task and social roles. In Mission Survival Corpus I, the configuration 1g3n0o0s, which denotes the configuration of with 1 Giver, 3 Neutrals, 0 Orienteer, and 0 Seeker, takes the majority (36%) of the discussion time, and the configurations 2g2n0o0s, 0g3n1o0s, 0g4n0o0s, 1g2n0o1s, 3g1n0o0s, 1g2n1o0s, 0g3n0o1s and 2g1n0o1s take respectively 20%, 13%, 11%, 5%, 5%, 4%, 2% and 1% of the discussion time. For the Socio-Emotional roles area, the different distributions 0a3n1p0s, 0a4n0p0s, 0a3n0p1s, 0a2n2p0s, 0a2n1p1s, 0a2n0p2s, 0a1n2p1s, 0a1n3p0s and 0a1n1p2s take respectively 36%, 21%, 18%, 11%, 7%, 3%, 1%, 1% and 1% of the discussion time.
Finally, in the Table 10 and in the Table 11 it is shown how the meeting participants change their roles as a function of the number of simultaneous speakers.

		Giver	Follower	Orienteer	Seeker
Number of Speakers	0	0.162	0.777	0.043	0.018
	1	0.251	0.675	0.049	0.025
	2	0.325	0.591	0.054	0.031
	3	0.358	0.536	0.070	0.037
	4	0.329	0.572	0.076	0.023

Table 10. Distributions of task roles conditioned on number of simultaneous speakers

		Attacker	Neutral	Protagonist	Supporter
Number of Speakers	0	0.001	0.817	0.104	0.078
	1	0.002	0.740	0.177	0.081
	2	0.004	0.680	0.220	0.096
	3	0.005	0.620	0.238	0.137
	4	0.008	0.581	0.305	0.107

Table 11. Distributions of socio-emotional roles conditioned on number of simultaneous speakers

We could consider 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.

3.3.5 Mission Survival Corpus 2 (MSC-II)

Twelve groups of 4 members each (male: 51.9%; females: 48.1%; average age: 35 years) participated in the data collection for MSC-II. They were recruited outside our research center and their participation took place on a voluntary basis. Besides involving them in the Mission Survival task, we also asked participants to fill two standard questionnaire for measuring personality traits: the Italian version of Craig's Locus of Control of Behavior scale (LCB) [Farma and Cortivonis, 2000], and the part of the Big Marker Five Scales (BFMS) that measures the Extraversion dimension [Perugini and Di Blas, 2002].

3.3.6 Personality Traits

The personality questionnaires, filled in by the participants before the meetings, were the Italian version of Craig's Locus of Control of Behavior scale (LCB) , and part of the Big Marker Five Scales (BFMS) related to the Extraversion dimension [Perugini and Di Blas, 2002].
For each personality dimension, we conducted two tests in order to verify if the participants' scores distribution is normal. In particular, we used the Kolmogorov-Smirnov statistic with a Lilliefors significance level [Lilliefors, 1967] and the Shapiro-Wilk statistic [Shapiro and Wilk, 1965]. For both tests, if the p-values of the two tests are greater than 0.05, the distribution is normal.
In particular, the LoC questionnaire was composed of 17 questions/items with a rating scale from 0 to 5 points [Farma and Cortivonis, 2000]. The mean and the standard deviation for the LoC raw scores are 27.6 and 8.8 respectively. These values are consistent with the population mean (27) and population standard deviation (9.2) reported in [Farma and Cortivonis, 2000]. As depicted in Table 12, the distribution is normal.

(*) This is a lower bound of the true significance.
(a) Lilliefors Significance Correction

Kolmogorov-Smirnov (a)			**Shapiro-Wilk**		
Statistic	Df	Sig.	Statistic	Df	Sig.
0.077	52	0.200(*)	0.971	52	0.230

Table 12 Locus of Control: Tests of Normality

Instead, for the Extraversion dimension the mean and the standard deviation of the raw scores are 43.6 and 10.3. Again in this case, the results of the tests of normality, reported in the Table 13, show us that the distribution is normal.

(*) This is a lower bound of the true significance.
(a) Lilliefors Significance Correction

Kolmogorov-Smirnov (a)			**Shapiro-Wilk**		
Statistic	Df	Sig.	Statistic	Df	Sig.
0.068	52	0.200(*)	0.980	52	0.535

Table 13. Extraversion: Tests of Normality

3.3.7 Individual Performance

The Individual performance have been measured by scoring the individual solutions provided by the participants.

The individual solutions of each participant were assessed by assigning each of the first five items in the subject's final list a score corresponding to its position in the correct solution (e.g. if "wool blanket" is one of the five for subject X and it is the seventh in the correct solution list, then it gets 7), and then summing up all the scores. The best performance is achieved by the lowest score; the resulting range is 15 (best) - 50 (worst). In the MS-2 corpus, the raw scores of the 48 subjects ranged from 20 to 39, with mean 28.33 and standard deviation 5.3.

(*) This is a lower bound of the true significance.
(a) Lilliefors Significance Correction

Kolmogorov-Smirnov (a)			**Shapiro-Wilk**		
Statistic	Df	Sig.	Statistic	Df	Sig.
0.131	52	0.026	0.940	52	0.011

Table 14. Individual Performance: Tests of Normality

Concerning the tests of Normality, as showed in Table 14, in this case the p-values obtained do not let us to assume that the distributions are normal.

Chapter 4

4. The Automatic Detection of Functional Roles

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 and 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. 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.
The structure of the chapter is the following: first we describe our preliminary results (Section 4.1) obtained using a very simple set of multimodal features (hand and body fidgeting on the visual side and only the speaking activity on the acoustic side) and SVMs as classifiers. Precisely, we used SVMs with sliding windows to take in account the dynamic nature of this task (the participants change often their roles during group interactions such as meetings). Again, in these experiments we use two different arrangements of the feature vectors: the first includes the information about the speech and fidgeting activity of the target subject, as well as the number of simultaneous speakers, during the window time; whereas the second one includes all the above information plus additional information about the speaking activity and the fidgeting of all the other participants.
Then, we compare the performance on this task of three classifiers: SVMs with sliding windows, HMMs and Influence Model (Section 4.2). Moreover, we turn our attention to the analyses (using three different techniques: Misclassification error and the covariance matrix. performed for understanding the relevance of our classes of social 'honest signals' (following the Pentland's definition, Conversational Activity, Spectral Centre, Consistency, Mimicry, and Influence for the acoustic side; while the only Body Energy for the visual side) in predicting the different functional roles (Section 4.4). To this end, we also run some classification experiments to verify the predictive power of various classes of signals.
Finally, we exploited the relationships between social and task roles by training a joint classifier on these roles (Section 4.5).

4.1. Preliminary Experiments

In these experiments, we modeled the role assignment task as a multi-class classification problem on a relative large and very unbalanced dataset, and used SVMs as classifier [Vapnik, 1995; Cristianini and Shawe-Taylor, 2000].
SVMs are a powerful discriminative learning method. In fact, these algorithms try to find a hyper-plane that not only discriminate the classes but also maximizes the margin between these classes [Cristianini and Shawe-Taylor, 2000].

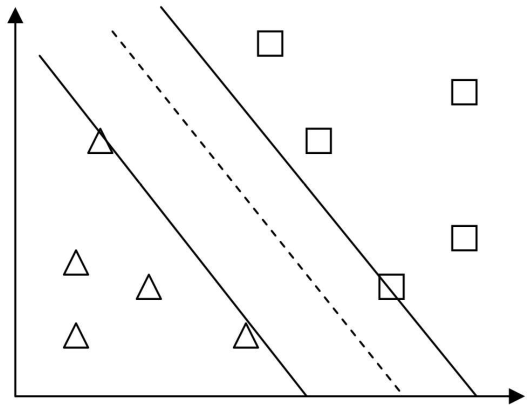

Figure 3. Classifier hyperplane and margins for a training set of two classes (△ and □)

One of the advantages of using SVMs is that it produces models that only depend on a subset of the training data, the so-called *Support Vectors* (SV). The SV are members of the set of training input data that outline an hyper-plane in the feature space. This l-dimensional hyper-plane, where l is the number of features in the input vectors, defines the boundary among the different classes. In practice, the classification task consist in simply determining on which side of the hyper-plane the testing vectors reside.
Mathematically speaking, given a set of instance-label pairs (x_i, y_i), i = 1 … l where $x_i \in \mathbf{R}^d$, $y_i \in \{-1, 1\}$ (two class problem) the SVMs require the solution to the following optimization problem

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$s.t.\ \ y_i(w \cdot \Phi(x) + b) \geq 1 - \xi_i, \forall x_i$$
$$\xi_i \geq 0$$

Then, the SVM finds a linear separating hyper-plane with the maximum margin in this higher dimensional space (see Figure 3). C>0 is the penalty parameter of the error term.
In order to take into account the time dimension in the classification task, sliding windows were used: the classifier considers all the data in the time window to assign a Task area role and a Socio-Emotional area role only at the end of the window [Dietterich, 2002]. We considered win-

dows of varying size, from 0 to 14 seconds (0 to 42 instances). For each window size, a dataset was built by adding to each row all the features of the rows before included in the window width. For a given time and a given participant, the information that the classifiers had available to classify his/her roles was the information about his/her speech and fidgeting activity, as well as the number of simultaneous speakers, during the window time. Each dataset was then split in two equal parts for training and testing.

More precisely, in these experiments the bound-constrained SV classification algorithm with a RBF kernel $K(x,y) = \exp(-\gamma||x-y||^2)$ was used. The cost parameter C and the kernel parameter γ ($\gamma>0$) were estimated through the grid technique by means of cross-fold validation using a factor of 103. Given the computational costs of this procedure, we estimated the parameters for the windows 0 instances, 21 instances and 32 instances only, and used the parameters estimated for the window 0 also for the windows from 1 to 3; the parameters estimated for the window of 27 also for the windows from 4 to 27; and parameters estimated for the window 32 also for the windows from 30 to 42. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size.

The “one-against-one” method [Kressel, 1999] was used whereby each training vector undergoes a number of binary comparisons, corresponding to the number of class pairs available (12 for each area in our case), each time 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.

By way of comparison, we used two baselines: the *trivial classifier*—that assigns all instances to the most frequent class—and the *equidistributed classifier*—that distributes the instances assigning them equal prior probabilities. Accuracy is known to be somewhat inadequate for unbalanced datasets, because the trivial classifier always has very high accuracy. Therefore, we used both accuracy and F-score as figures of merit, where the latter is computed as the harmonic means of the macro-averaged one-class precisions and recalls (macro F-score). We also considered average F-score computed as the average of the one-class F-scores.

4.1.1 Task area roles with left-only windows

Fifteen datasets were built considering windows from 0 to 14 seconds (from 0 to 42 instances) to the left of the time point to classify. The number of features varied accordingly, from 4 for the 0 seconds window to 173 for the 14 seconds window. Figure 4 and Figure 5 plots accuracy and macro F-scores comparing them to the baselines.

[3] We used the BSVM tool (Hsu and Lin, 2002) available at http://www.csie.ntu.edu.tw/~cjlin/bsvm/.

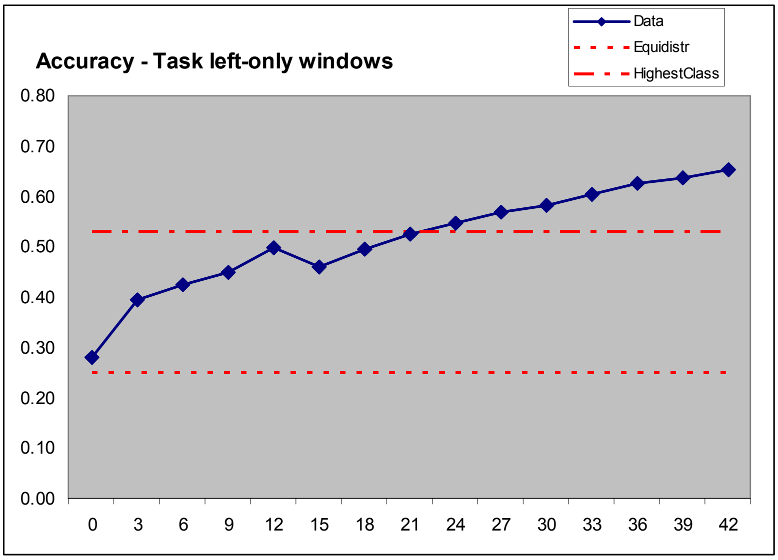

Figure 4. Accuracy for Task area roles' classification with left-only windows.

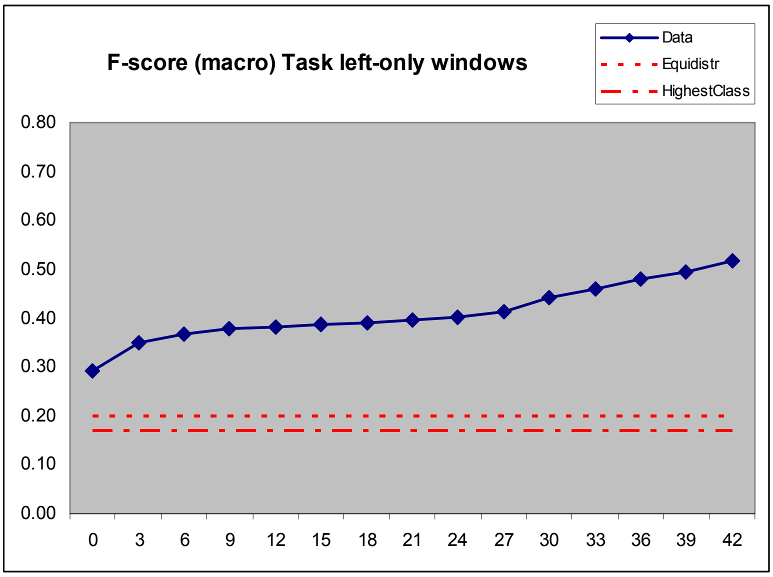

Figure 5. Accuracy for Task area roles' classification with left-right windows.

The best performance is obtained by the window of 14 seconds (see the confusion matrix depicted in the Table 15).

Task Roles 14 secs. left	Follower	Orienteer	Giver	Seeker	total
Follower	3592	167	1399	76	5236
Orienteer	289	550	942	3	1784
Giver	1673	569	6511	87	8840
Seeker	222	46	237	141	646
Total	5776	1332	9089	309	16506

Table 15. Confusion matrix for Task roles at left-only window width 14 seconds.

For this window the macro Precision reaches 0.55 while macro Recall is 0.48. Table 15 reports the precision, recall and F-scores for the individual roles.

Task Roles 14 secs. left	Follower	Orienteer	Giver	Seeker
Precision	0.62	0.41	0.72	0.46
Recall	0.69	0.31	0.73	0.22
F-score	0.65	0.35	0.73	0.30

Table 16. Precisions, Recalls and F-scores for Task roles at left-only window width 14 seconds.

4.1.2 Task area roles with left-and-right windows

Twelve datasets were built considering windows from 0 to approx 14 seconds, split to the left and the right of the time point to classify (0 to 22 instances on each side). The dataset contained a number of features variable from 4 to 173. Figure 6 and Figure 7 plot accuracy and macro F-scores comparing them to the baselines.

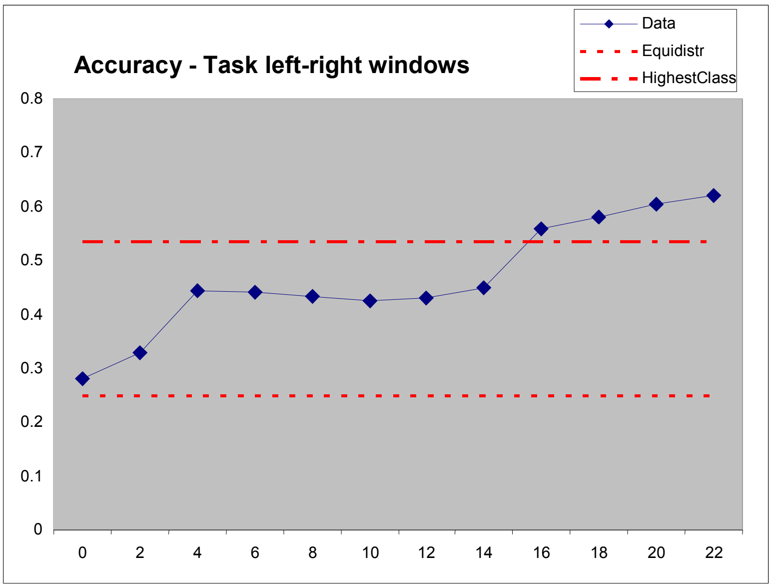

Figure 6. Accuracy for Task area roles' classification with left-right windows.

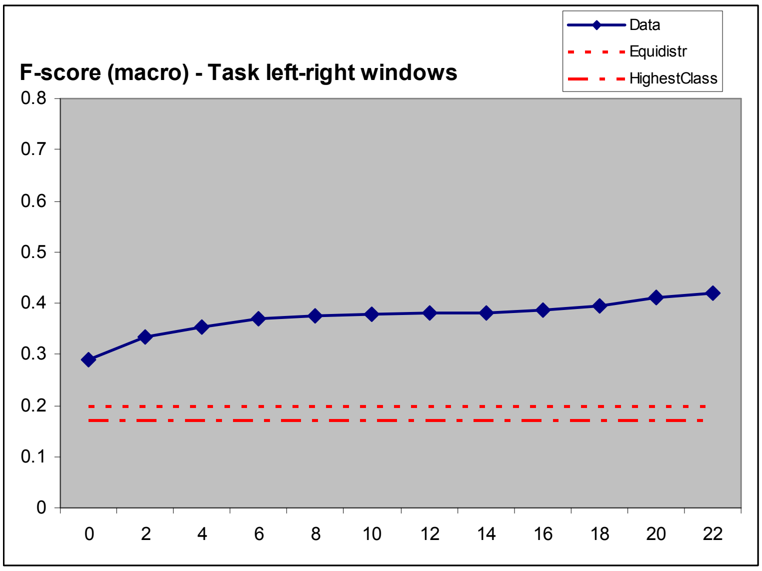

Figure 7. Macro F-scores for Task area roles' classification with left-right windows.

Again, the window width with the best accuracy and macro F-score is the longest, 22 time points per side (approx 14.5 seconds). It outperforms the *equidistributed classifier* and the accuracy of the *trivial classifier*. Accuracy reaches 0.62 while macro F-score reaches the value of 0.42 (averaged F-score = 0.41). Therefore it does not reach the same performance of the left only windows of equivalent size, at least in terms of F-score.

4.1.3 Socio area roles with left-only windows

As already for the task area, fifteen datasets were built considering windows from 0 to 14 seconds to the left of the time point to classify. Figure 8 plots accuracy while Figure 9 plots macro F-scores comparing them to the baselines.

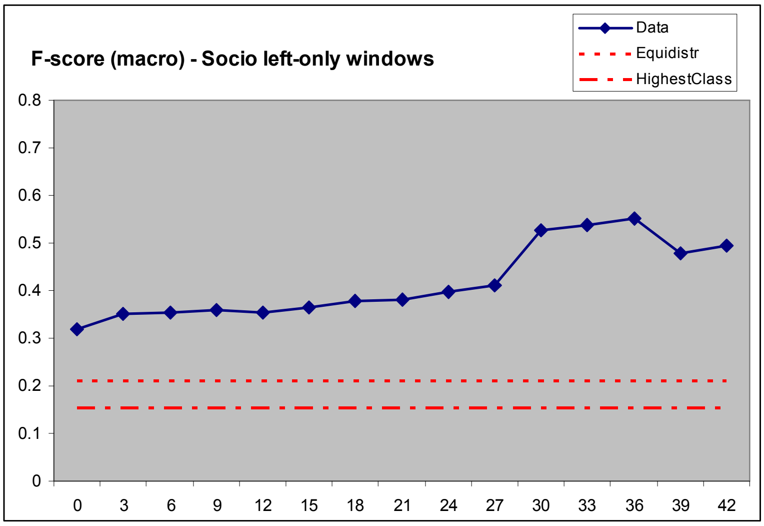

Figure 8. F-scores for Socio area roles' classification with left-only windows.

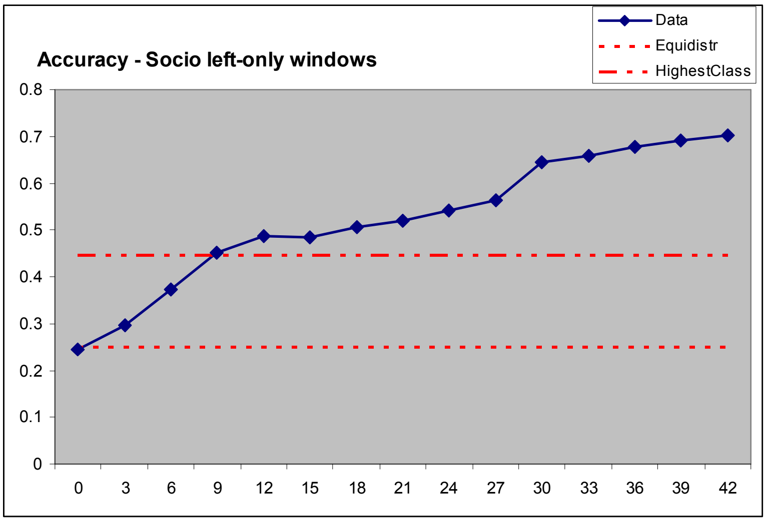

Figure 9. Accuracy for Socio area roles' classification with left-only windows

Because of the small number of examples of the *Attacker* role, accuracy reaches its maximum—0.70—at the greater window width, at the expenses, however, of missing all the instances of the *Attacker* role. A better performance is at 12 seconds where the highest macro F-score is reached (macro F-score = 0.55, macro precision = 0.75, macro recall = 0.43). Yet, the classifier still under-classified the Attacker role, see the confusion matrix for window of 12 seconds in Table 17.

Social Roles 12 secs. left	Neutral	Supporter	Protagonist	Attacker	total
Neutral	5886	95	1399	0	7380
Supporter	1086	395	1314	0	2795
Protagonist	1208	114	4922	0	6244
Attacker	58	1	27	1	87

Total	8238	605	7662	1	16506

Table 17. Confusion matrix for Socio roles at window of 12 seconds.

Table 18 and Table 19 show the precisions, the recalls and the F-scores for the individual roles for window of 12 seconds and for window of 14 seconds, respectively. .

Social Roles 12 secs. left	Neutral	Supporter	Protagonist	Attacker
Precision	0.71	0.65	0.64	1
Recall	0.80	0.14	0.79	0.01
F-score	0.75	0.23	0.71	0.02

Table 18. Precisions, recalls, and F-scores for Socio roles at left-only window of 12 seconds.

Social Roles 14 secs. left	Neutral	Supporter	Protagonist	Attacker
Precision	0.74	0.84	0.66	0.00
Recall	0.82	0.12	0.83	0.00
F-score	0.78	0.21	0.74	0.00

Table 19. Precisions, recalls, and F-scores for Socio roles at left-only window of 14 seconds.

4.1.4 Socio area roles with left-and-right windows

Twelve datasets were built considering windows from 0 to approx 14 seconds split to the left and the right of the time point to classify (0 to 22 instances on each side).
Figure 10 and Figure 11 plot accuracy and macro F-scores comparing them to the baselines.

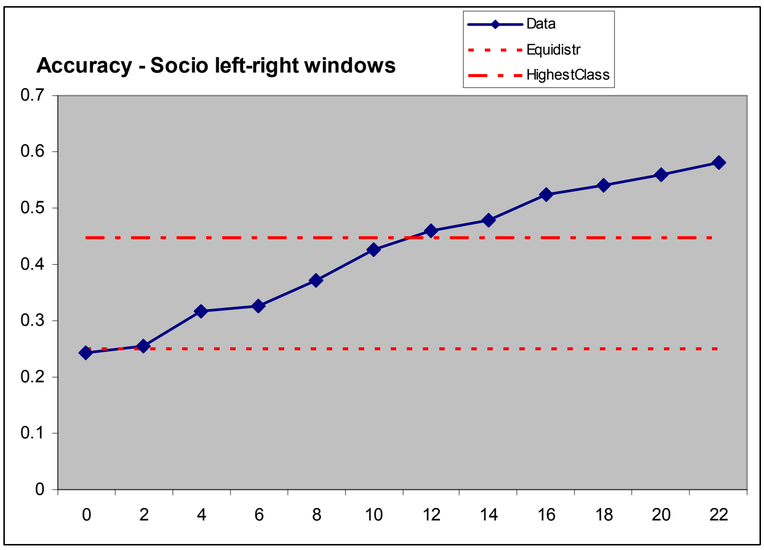

Figure 10. Accuracy for Socio area roles' classification with left and right windows.

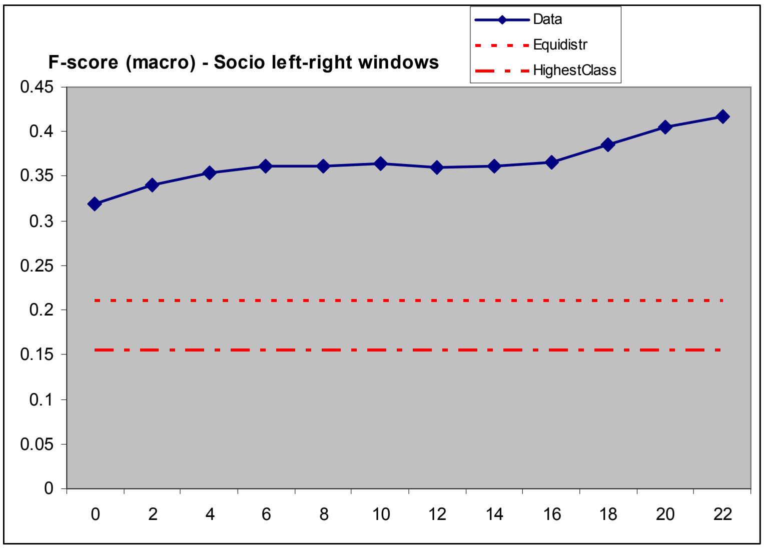

Figure 11. Macro F-score for Socio area roles' classification with left and right windows.

Again, the window width with the best accuracy and F-score is the longest. Both the *equidistributed classifier* and the *trivial classifier* are over performed with accuracy of 0.58 and macro F-score of 0.42 (averaged F-score = 0.40). Yet left-and-right windows do not reach the performance of the left only window of 12 seconds.

4.1.5 Using information of other meetings participants

In these additional experiments we built two datasets for each window size. For a given time and a given participant, the first included the information about his/her speech and fidgeting activity, as well as the number of simultaneous speakers, during the window time. The second one included all the above plus the information about speaking activity and the fidgeting of all the other participants. For these experiments we report only the results obtained using left windows, because centered windows are showed less effective.

The accuracy values for the different windows in the two datasets are compared in Figure 12 to the baselines (*trivial classifier* and *equidistributed classifier*).

Turning the attention to the task roles, the classifier trained on the minimal dataset (i.e. the one containing the participant's features only) improves over both baseline from windows of 7 seconds up, while the classifier trained on the features for all the participants is always above the baseline.

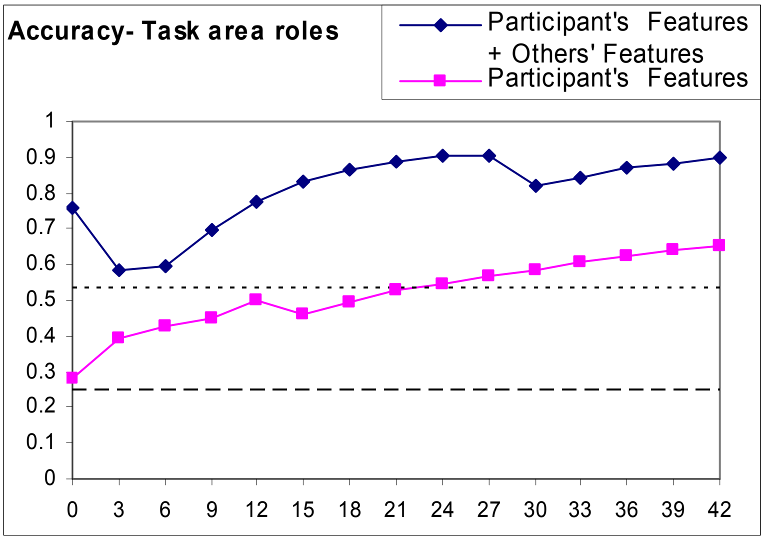

Figure 12. Accuracy for the roles in Task Area

Focusing on the latter, performance starts from quite high values for window of 0 seconds (accuracy = 0.76, F-score = 0.69), then drops until window of 4 seconds (accuracy = 0.78, F-score = 0.69), where the values of both figures are stably higher than for the 0-sized window. One might conjecture that contextual time information is only useful when enough temporal context is considered. An alternative explanation starts from the similar (though smaller) drop of performance for windows of size 10 seconds at a possible effect of the way parameters were estimated and applied; the 3 seconds sized and 10 seconds sized windows are, in fact, the lower bounds of the window intervals to which parameters estimated with windows 7 seconds and 14 seconds are applied, respectively.

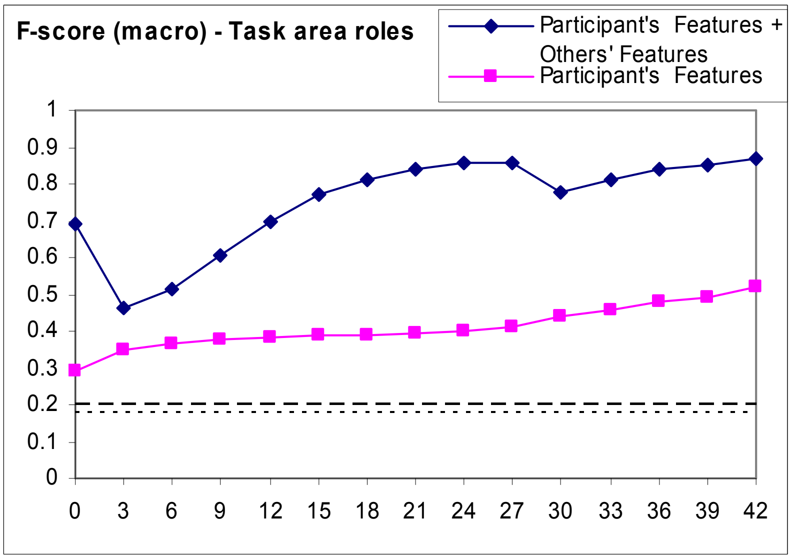

Figure 13. Macro F-score for the roles in Task Area

The highest accuracy is reached with window of 9 seconds with a value of 0.90. The max value of macro F is 0.87 and is reached at the largest window size (14 seconds). We prefer to consider window of 14 seconds since we value macro F-score as a better measure of accuracy on our corpus. Table 20 summarizes the precision, recall and F-score values for the Task Area roles on that window.

Task Roles 14 secs. Left	Follower	Orienteer	Giver	Seeker
Precision	0.89	0.89	0.91	0.83
Recall	0.92	0.81	0.91	0.74
F-score	0.91	0.85	0.91	0.78

Table 20. Precision and Recall Values for Task Roles values on window 14 seconds.

Turning our attention to the Socio-Emotional Area role, we found a very similar pattern as plotted in Figure 14. The classifier trained on the minimal data set exceeds the baseline on accuracy from window size of 4 seconds while the accuracy of the classifier trained on the augmented dataset is always higher.

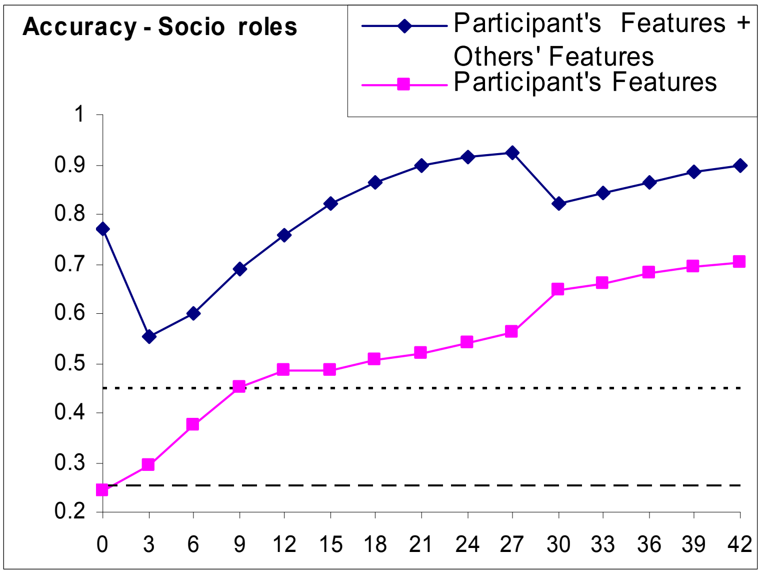

Figure 14. Accuracy for the roles in Socio-Emotional Area

The pattern of the augmented classifier is virtually identical to that discussed for the task area, including the drop around the 3 seconds and 10 seconds sized windows, and the maximal values, which are reached with window of size 9 seconds for accuracy (0.92), and windows size of 14 seconds for macro-F score (0.86).

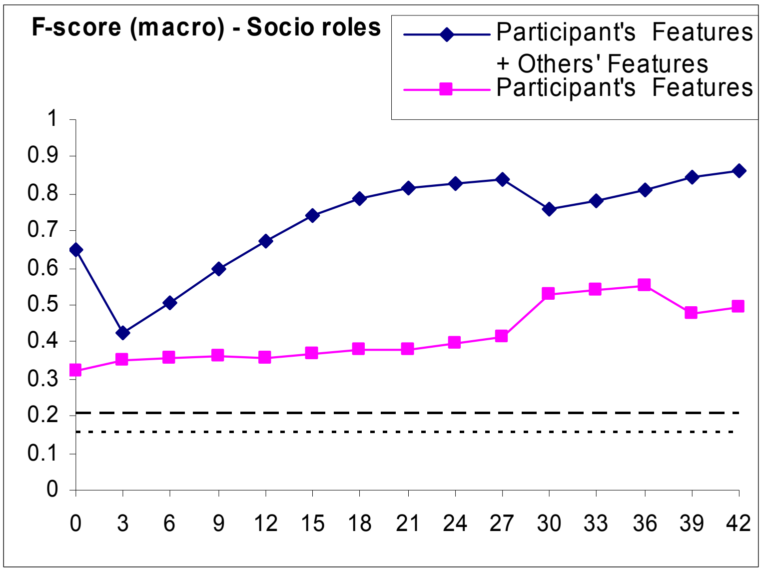

Figure 15. Macro F-score for the roles in Socio-Emotional Area

The Table 21 summarizes the precision and recall values for the Socio-Emotional area roles on the window of 14 seconds when the macro-F score reaches its maximum.

Socio Roles 14 secs. left	Neutral	Supporter	Protagonist	Attacker
Precision	0.89	0.89	0.91	0.83
Recall	0.92	0.81	0.91	0.74
F-score	0.91	0.85	0.91	0.78

Table 21. Precision, Recall and F-scores values for Socio Roles on the window of 14 seconds.

4.2. Using Influence Model

In this Section we compare the results obtained from three learning approaches: SVMs with linear kernel, HMMs and Influence Model.

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 simple multi-class SVM approach, although powerful, has several limitations in its generalization capability. The first issue is related to finding general features applicable to all speakers. Different speakers might have different ways to fulfil their functional roles in a group discussion. Having a speaker specific implementation is a nontrivial task for support vector classifiers. The second issue is related to the curse of dimensionality. When we make use of the observations of other speakers for our classification task, the length of the observation vector grows linearly with a large multiplication constant. For instance, using 9-seconds windows the length of the observation vector is 432 (9 seconds x 3 samples/second x 4 features x 4 speakers). The third issue is about the assessment of the trained classifier, as well as how the speakers interact with each other. Extracting an intuitive understanding of group interactions among the speakers from the trained support vector classifiers is not easy. A final issue is generalizability to different numbers of speakers. The SVM approach is not modular in the number of participants, whereas a network approach can be scaled to different sizes of groups. As a result, a natural step is to use a Bayesian hierarchical dynamic model, and compare the performance with that of a standard multi-class SVM.

We used four meetings for train-set and the other meetings for test-set. Moreover, in our SVMs experiments we used a linear kernel $K(\mathrm{x}i,\mathrm{x}j) = \mathrm{x}i^{T}\mathrm{x}j$ in order to reduce the risk of overfitting. The RBF kernel, in fact, might have an infinite Vapnik-Chernovenkis dimension and might be subject to over-fitting. The highest accuracy score obtained is 70%. The macro precision and the macro recall for Task area roles are 48% and 52%. The performance is worst for Socio-emotional area roles: the macro precision is 39% and the macro recall 48%. Table 22 and Table 23 show the confusion matrices for Task area roles and Socio-Emotional area roles respectively. The observation vector is composed of the smoothed version of speaking/non-speaking, hand movement, and body movement of the speaker under investigation, as well as the number of simultaneous speakers in a fixed-length window around the moment of interest. We take this window to be from 10 seconds before until 10 seconds after the moment of interest.

		SVM classification on test data				
		Giver	Follower	Orienteer	Seeker	Total
Ground truth	Giver	8468	4049	1624	635	14776
	Follower	2517	29304	520	899	33240
	Orienteer	1385	527	205	74	2191
	Seeker	35	18	535	717	1305
	Total	13571	28364	2416	7161	51512

Table 22. Confusion matrix between the ground truth and the typical classification result for task roles with Support Vector Machine and linear kernel

		SVM classification on test data				
		Attacker	Neutral	Protagonist	Supporter	Total
Ground truth	Attacker	74	70	20	21	185
	Neutral	460	32766	3936	1309	38471
	Protagonist	322	2463	5777	818	9380
	Supporter	146	1351	1748	231	3476
	Total	124	35386	8048	7954	51512

Table 23. Confusion matrix between the ground truth and the typical classification result for socio-emotional roles with Support Vector Machine and linear kernel (A=attacker, N=neutral, P=protagonist, and S=supporter)

Turning to the Influence Model, this approach is a method developed in the tradition of the N-heads dynamic programming on coupled hidden Markov models [Oliver et al., 2000], the observable structure influence model [Asavathiratham et al., 2001; Cristani et al., 2009], and the partially observable influence model [Basu et al., 2001]. It extends, though, these previous models by providing greater generality, accuracy, and efficiency (see [Dong, 2006] and [Dong and Pentland, 2007] for a detailed introduction to this technique). 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 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 interaction of the underlying interacting processes, while avoiding the risk of overfitting and the difficulties of observations with large dimensionality.

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

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

where dj,i represents the influence between processes j and i, and a(xj,xi) represents the influence between the states xj and xi of the interacting processes j and i.

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

$$\Pr(S_t^{(c)}=s)=\pi_s^{(c)}$$

$$\Pr(S_{t+1}^{(c)}=s)=\sum_{c_1=1}^{C}\sum_{s_1=1}^{m_{c_1}}h_{s_1,s}^{(c_1,c)}\Pr(S_t^{(c)}=s_1)$$

where $1 \le s \le m_c$ and $h_{s_1,s}^{(c_1,c)} = d^{(c_1,c)} a_{s_1,s}^{(c_1,c)}$. The observations $\vec{Y}_c = (Y_t^{(1)},...,Y_t^{(C)})$ are coupled with the latent states $\vec{S}_c = (S_t^{(1)},...,S_t^{(C)})$ through a memory-less channel:

$$P(\vec{S}_c)P(\vec{Y}_c \mid \vec{S}_c) = \prod_{c=1}^{C} P(S_t^{(c)})P(Y_t^{(c)} \mid S_t^{(c)})$$

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 Expectation Maximization (EM) algorithm, based on the training data. In the prediction phase, we infer the individual participant's social/task roles based on observations about her/his, as well as on observations about the interactions with other participants. In the influence modeling of the speakers' functional roles, we used $2n$ number of interacting processes to model the task roles and the social roles of the n individual speakers in a meeting. The observations for the individual processes are the corresponding speakers' raw features (speaking/non-speaking, body and hands fidgeting, and number of simultaneous speakers) averaged over short fixed-length time windows centered around the observation times. The latent states for the individual processes are the corresponding 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 speakers with the EM (expectation maximization) algorithm, based on the training data. Each iteration of the EM algorithm consists of two processes: the E-step and the M-step. In the expectation, or E-step, the missing or hidden data are estimated given the observed data and the current estimate of the model parameters. This is achieved using the conditional expectation. In the M-step, the likelihood function is maximized under the assumption that the missing or hidden data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing or hidden data. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

In the application phase, we infer the individual speakers' social/task roles based on the observations about the individual speakers, as well as their interactions, using the parameters previously trained.

We partitioned the data set of the eight meetings into two parts, as in the SVM experiments, and estimated the generalization capability of the trained classifier by 2 fold cross validation.

With the influence modeling, we can generally get 75% accuracy in classifying both the task roles and the social roles.

		Influence model classification on test data				
		Giver	Neutral	Orienteer	Seeker	total
Ground truth	Giver	8059	4225	1858	634	14776
	Follower	2535	29362	406	937	33240
	Orienteer	1304	500	320	67	2191
	Seeker	526	714	64	1	1305
	Total	12424	34801	2648	1639	51512

Table 24. Confusion matrix between the ground truth and the typical classification result for task roles obtained using Influence Model

		Influence model classification on test data				
		Attacker	Neutral	Protagonist	Supporter	Total
Ground truth	Attacker	74	72	19	20	185
	Neutral	341	32767	3521	1842	38471
	Protagonist	269	2536	5290	1285	9380
	Supporter	127	1281	1455	613	3476
	Total	811	36656	10285	3760	51512

Table 25. Confusion matrix between the ground truth and the classification results for socio-emotional roles obtained using Influence Model

The macro precision and the macro recall of the Influence Model are 40% and 39% for Task area roles. The performance is better for Socio-Emotional area roles: the macro precision is 41% and the macro recall is 50%.

Turning to the HMM, it is a standard model for modeling partially observable stochastic processes and was originally developed for speech understanding [Rabiner, 1989]. HMMs have more representational power than SVMs because they can model some of temporal dependencies of functional roles. More precisely, the representation of the model is the following: t, time; $\mathbf{y}(t)$, the feature vector; $x(t)$, the role; $p(x)$, the priors for the roles; $p(x(t)|x(t-1))$, the role transition probabilities; $p(\mathbf{y}(t)|x(t))$: conditional distribution of observed feature vector given the current role. In our task, 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 $p(\mathbf{y}(t)|x(t))$. Thus, all four-feature sequences (one per subject) from all eight meetings are used to train a single HMM (in practice, we train four single HMMs for each meeting). As for the Influence Model, the training is done using the EM algorithm.

For classification, the Viterbi algorithm is used to compute the most likely sequence of roles. HMM typically yields 60% accuracy for task roles, 70% accuracy for social roles, and 65% overall accuracy. The typical confusion matrices for task/social roles are given in Table 26 and Table 27.

		HMM per speaker classification on test data				
		Giver	Follower	Orienteer	Seeker	total
Ground truth	Giver	7126	3637	1370	2643	14776
	Follower	4728	23740	809	3963	33240
	Orienteer	1212	310	195	474	2191
	Seeker	505	677	42	81	1305
	total	13571	28364	2416	7161	51512

Table 26. Confusion matrix between the ground truth and the typical classification result for task roles with one hidden Markov model per speaker

		HMM per speaker classification on test data				
		Attacker	Neutral	Protagonist	Supporter	Total
Ground truth	Attacker	54	91	40	0	185
	Neutral	20	30874	3017	4560	38471
	Protagonist	0	2825	4695	1860	9380
	Supporter	50	1596	296	1534	3476
	Total	124	35386	8048	7954	51512

Table 27. Confusion matrix between the ground truth and the typical classification result for socio-emotional roles with one hidden Markov model per speaker

Putting these results together, it can be seen that by including the influence modeling to capture connections between speaker roles, we can achieve approximately 10% increase in accuracy, to about 75% overall accuracy. This is similar to the inter-rater accuracy of the human labeling of this corpus. By comparing the confusion matrices for the influence model and the hidden Markov models, one can see that most of the improvements are in the majority classes, and are due to the fact that influence modeling uses the functional roles of other speakers. However, for the Socio-Emotional area roles the macro precision (52%) and the macro recall (51%) of the hidden Markov model are better than the macro precision (41%) and macro recall (50%) of the Influence Model. These results of the Influence Model are due to two different reasons: the high number of false positives in the Attacker role classification and the not good performance at classifying the Supporter role.

4.3. Relevant Honest Signals for different functional roles

In this Section we report some our preliminary analysis and our machine learning experiments for understanding the relevance of five classes of honest signals for predicting the different functional roles. We computed the means and variances of the features' distributions for each social and task role. For sake of simplicity, we assumed that the distributions were 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 two different measures, namely the misclassification error and the covariance matrix.

These analyses show that 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 other. For example, consistency features are good at predicting the Supporter role but bad at predicting the Protagonist role.

4.3.1 Misclassification error

We define the misclassification error 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 as

$$err_i = 1 - \int_{\phi_i} p(y;i)$$

$$\phi_i = \{y : p(y;i) > p(y;j); \forall j \neq i\}$$

where *p(y;i)* is the conditional distribution of a feature *y* given class (role) *i* and ϕ_i is the set of the values of the feature for which a Bayesian classifier predicts the 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. The misclassification errors for the different set of features (C: Consistency, SC: Spectral center, A: Activity, M: Mimicry, I: Influence, BG: Body Gestures), assuming equal priors for the roles, are shown in Figure 16 (social roles) and in Figure 17 (task roles).

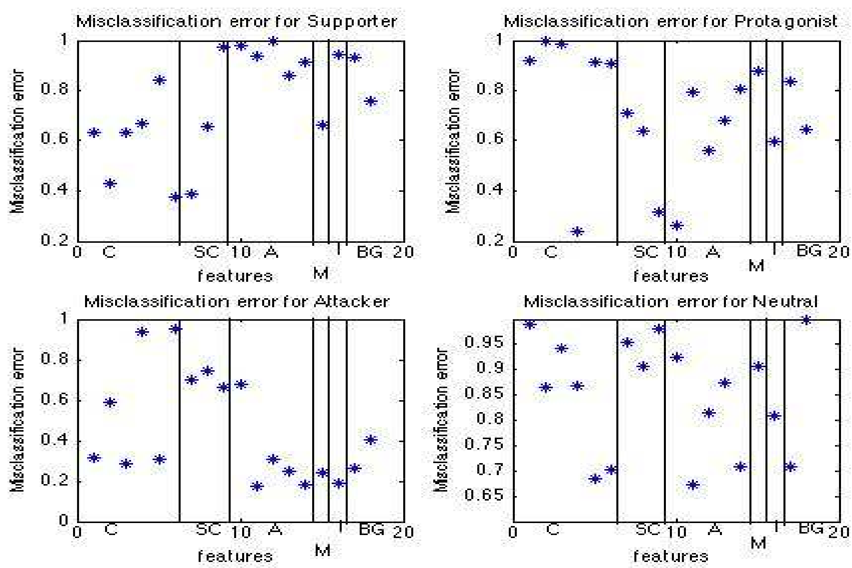

Figure 16. Misclassification errors for each socio role while using different features

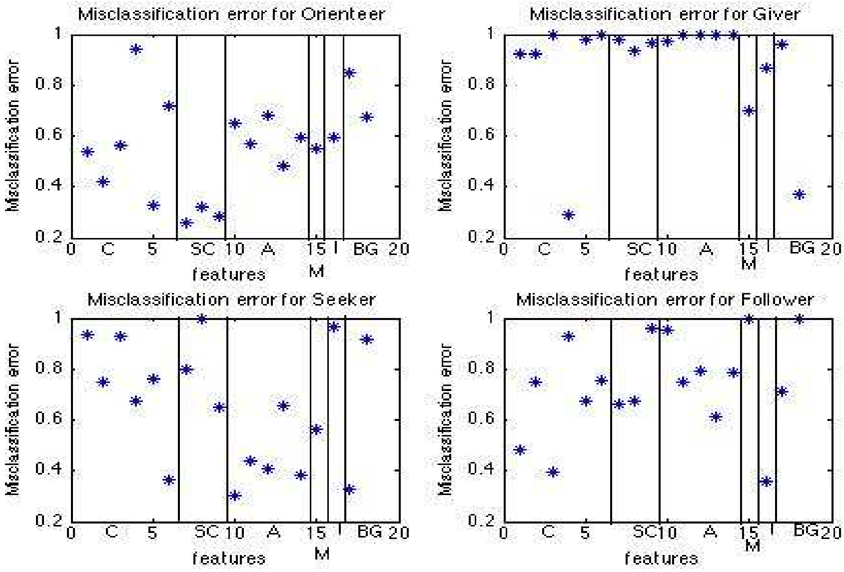

Figure 17. Misclassification error for each task role while using different features.

Figure 16 suggests that certain features are key for predicting certain socio roles. The key features are individuated by their small misclassification errors and hence are distinct for that socio role. Supporters have distinct consistency and spectral center in their speech; hence, the consistency and spectral center features give small misclassification error while predicting Supporters. Protagonists have a distinct energy and the distinct location of the largest autocorrelation peak (in other words a particular energy and pitch) in their speech. Attackers have a distinct speech activity, mimicry, influence and body gestures. This is intuitive because, the attackers are often observed to utter small questioning sounds and fidget their hands and body in discomfort. The misclassification error for Neutrals is high for all the features suggesting that the neutrals have a wide variation in all the features.

Figure 17 suggests that certain features are key for predicting the certain task roles. Orienteers are marked by distinct consistency and spectral center. Givers have a distinct variation in their energy and distinctive body fidgeting. The high error rate for Giver using most of the other features is mainly due to the fact that Givers have very similar features to those of Followers. Seekers have distinct activity patterns and mimicry. Followers are marked by influence. This is intuitive because the Followers often nod or speak over, or ask questions to the Giver.

4.3.2 Correlation among features

We now see the redundancy of information among acoustic features in the light of the covariance matrix. The covariance matrix is shown in Figure 19. 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 clubbing 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.

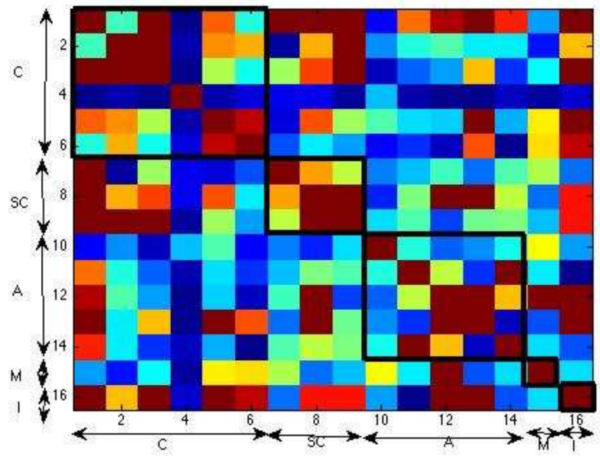

Figure 18. Covariance of the feature set. Blue suggests small value and red suggests large value and other values lie in between.

4.3.3 Classification results

In the MS-I corpus, the visual features were extracted on a frame base of 0.33 seconds. Similarly, the relational roles were manually annotated on a frame base. Instead, the acoustic 'honest signals' 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. Table 28 and table 29 show the distributions of social and task roles obtained through *Heuristic 1*.

Supporter	Protagonist	Attacker	Neutral
0.034	0.158	0	**0.808**

Table 28. Distributions of social roles after the application of *Heuristic 1*.

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

Table 29. Distributions of task roles after the application of *Heuristic 1*.

Table 30 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 Influence Model was performed using a leave-one-out procedure on the meetings.

	Socio	Task
Visual	0.71	0.68
Audio	0.75	0.74
Joint	0.77	0.74

Table 30. Accuracy for social and task roles prediction with *Heuristic 1*.

The results show that better accuracy is obtained with audio features than with visual features. However, these figures do not do any better than the baseline (the classifier assigning the most common role); see bold figures in Table 28 and Table 29. 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 where finding out about non-neutral/non-follower roles is crucial. Table 31 and Table 32 report the resulting distribution of roles in the corpus. As can be seen, a more balanced distribution emerges.

Orienteer	Giver	Seeker	Follower
0.070	**0.517**	0.017	0.395

Table 31. Distributions of task roles after the application of *Heuristic 2*

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

Table 32. Distributions of social roles after the application of *Heuristic 2*

We trained two independent classifiers, one for social roles and one for task roles, using visual features, speech features and their combination. Table 33 reports the accuracy scores.

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

Table 33. Accuracies for social and task roles using independent classifiers on *Heuristic 2* data.

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 acoustic *honest signals* seems to be the more predictive and informative features. Hence, we also considered the contribution of the various audio feature classes. Table 34 shows the accuracy values obtained using independent classifier.

	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

Table 34. Accuracies for social and task roles (independent classifiers) with the different classes of speech features on *Heuristic 2* data.

Interestingly, the Activity class yields accuracy values (slightly) higher than those produced through the usage of the entire set of audio features, see Table 34. Hence using the sole set of Activity features emerges as a promising strategy.

4.4. Joint prediction of social and task roles

Finally, using the dataset obtained applying the *Heuristic 2* we explored the extent to which the relationships between task and social role 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 x task roles; a more difficult task than the ones considered so far. Table 35 reports the distribution of the joint roles in the corpus, while Table 36 the classification results.

	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

Table 35. Distribution of social and task roles with *Heuristic 2*.

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

Table 36. Accuracy of joint prediction of social and task roles.

The results are interesting. Notice, first of all, that the accuracies are always much higher than the baseline, see the bold figure in Table 35. Moreover, the only 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 Influence Model is capable of scaling up to larger multi-class tasks without performance losses.

Chapter 5

5. Automatic Prediction of Personality Traits

We exploited the MS-II corpus to automatically predict personality traits from the observation of low-level features automatically extracted from the acoustical and visual scene. In particular, we designed two different tasks: a classification one (Section 5.1) and a regression one (Section 5.2). For the classification task, on the basis of 1-minute-long behavioral sequences, the system had to assign the subjects to the right class on two personality traits, Extraversion (one of the dimensions of Big Five) and Locus of Control (LoC); to this end, the continuous distributions of Extraversions and LoC were be turned into discrete ones (Low, Medium and High) by assigning to the Medium class the score comprised between ±1 standard deviation from the average; the Low and the High classes collected scores below -1 standard deviation and above +1 standard deviation, respectively. The 1-minute-long sequences were our 'thin slices'.
For the regression task, our goal was to predict the raw scores that subjects obtained by filling the questionnaires.
The features used were the acoustic honest signals and the three features related to the fidgeting. However, in regression task we applied two feature selection approaches (Correlation-based feature selection on audio-visual features and ANOVA-based feature selection only on audio features) while in classification task we applied only ANOVA-based feature selection on audio features.
The automatic detection (regression and classification) of personality can be pursued in (at least) two different manners, each corresponding to a different hypothesis about the way personality, as manifested in social interaction, can be assessed. According to the first, the only consideration of the target subject' behavior (her/his 'thin slices') is enough: the way she/he moves, the tone and energy of her/his voice, etc., are informative to get at her/his personality. The second view maintains that, the appreciation of personality requires information not only about the target's behavior, but also about the social context: the same behavior might have a different import for personality assessment if produced in a given social environment than in another. To put things differently and assuming (as it seems natural) that manifest behavior is causally affected by personality, the first hypothesis has it that such a causal relation is enough to obtain accurate personality estimates. The second hypothesis, in turn, acknowledges that the way personality manifests in behavior is modulated by the social context—that is, by the behaviour of the other group members. Hence, 'thin slices' of the other group members are needed as well.

5.1. Classification

For our classification task we used SVMs. The multi-class nature of the problem was dealt with through the "one-against-one" method plus a voting strategy. We used the bound-constrained SVM classification algorithm with a RBF kernel. The cost parameter C and the kernel parameter

γ were estimated through by cross-fold validation using a factor of 10. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size.

5.1.1 Feature Selection

A feature selection was performed only on the acoustic features by comparing their means through ANOVA: each feature was treated as a dependent variable in two between subject analysis of variance, with factor Extraversion (3 levels: Low, Medium, High) and LoC (3 levels: Low, Medium, High); significance level was set at $p<.05$. No adjustment for multiple comparisons was performed, in order to have a more liberal test. Only the features for which the analysis of variance gave significant results were retained, for the given factor: the mean value of the formant frequency, the mean value of the confidence in formant frequency, the mean value of the number of autocorrelation peaks, and the standard deviation of the number of autocorrelation peaks, a subset of the Emphasis class, and the fraction speaking over, the Influence feature, for Extraversion, and the mean value of the formant frequency, the mean value of the number of autocorrelation peaks, the standard deviation of the number of autocorrelation peaks, the same subset of the Emphasis class apart for the mean energy, and the average number of short speaking segments, the Mimicry feature, for LoC.

5.1.2 Experimental design

In our experiments we try to test two hypothesis: (i) the consideration of the social context improves personality assessment; and (ii) the selected subsets of features improve the performance of the classification.

In order to test these hypothesis, and focusing only on the acoustic features, we designed a between-subject experiment with factors ‘target’ and ‘others’, each relating to different arrangements of the target subject’s (target) and of the other group members’ (others) features.

- Target has two levels: all acoustic features+ visual features (ALL) vs. selected acoustic features + visual features (SEL).
- Others has three levels: no acoustic features + visual features (No-Feat); all acoustic features+ visual features (ALL); selected acoustic features + visual features (SEL).

A given experimental combination—e.g., (ALL, No-Feat)—corresponds to a specific arrangement of the feature vectors used to train and test the classifiers—in the example, all the acoustic plus the visual features of the subject, and, for each of the other group members, only the visual features—and to a specific combination of the hypothesis dimensions discussed above—in the example, that it is enough to consider thin slices of the sole target subject, and that the whole set of acoustic features are needed. The result is a 2×3 design. For each experimental condition, the training instances included the average values of the relevant acoustic and visual feature, computed over a 1-minute window; this way, the total number of generated instances corresponded to the total meetings’ duration in minutes (i.e. 366 minutes).

The analysis was conducted by means of 15-fold stratified cross-validation, with the same 15 training/test sets pairs being used in all the design’s 6 conditions. Stratification was conducted in order to closely reproduce in the training and test sets the distribution of Extraversion and LoC in the whole corpus.

5.1.3 Results and Discussion

The Table 37 and the Table 38 report the results in terms of accuracy, while the Table 39 and the Table 40 report the average macro-F figures. Here, we will limit our discussion to accuracy, comparing our results with those of the trivial classifier that always assigns the most frequent class to each instance (Accuracy=0.6667). Both for Extraversion and for LoC, the global average values of accuracy are well above the performance of the trivial classifier (0.8914 and 0.8698, respectively).

Two analysis of variance, one for Extraversion and one for LoC, showed that all the main effects are significant (p<.0001); interaction effects were not significant (p>.05). with reference to the marginal means, both for Extraversion and LoC the usage of all the features for the target subjects yields much better results in terms of accuracy, the advantage being even more marked for LoC (0,9116 vs. 0.8713 for Extraversion and 0.9197 vs. 0.8199, for LoC).

		Others			
		No-Feat	ALL	SEL	
Target	ALL	0.8889 (.029)	0.9021 (.028)	**0.9438** (.021)	0.9116 (.035)
	SEL	0.8493 (.024)	0.8611 (.036)	0.9035 (.026)	0.8713 (.037)
	Total	0.8691 (.033)	0.8816 (.038)	0.9237 (.031)	0.8914 (.041)

Table 37. Means and standard deviations of accuracy for Extraversion

		Others			
		No-Feat	ALL	SEL	
Target	ALL	0.9014 (.026)	0.9090 (.021)	**0.9486** (.016)	0.9197 (.030)
	SEL	0.7847 (.040)	0.8278 (.061)	0.8472 (.039)	0.8199 (.054)
	Total	0.8431 (.068)	0.8684 (.061)	0.8979 (.059)	0.8698 (.066)

Table 38. Means and standard deviations of accuracy for LoC

		Others			
		No-Feat	ALL	SEL	
Target	ALL	0.8399 (.048)	0.8529 (.055)	**0.9198** (.037)	0.8708 (.058)
	SEL	0.7774 (.039)	0.7837 (.038)	0.8630 (.039)	0.8081 (.054)
	Total	0.8087 (.053)	0.8183 (.058)	0.8914 (.047)	0.8395 (.064)

Table 39. Means and standard deviations of macro-Fscore for Extraversion

		Others			
		No-Feat	ALL	SEL	
Target	ALL	0.9404 (.016)	0.9488 (.012)	**0.9722** (.012)	0.9538 (.019)
	SEL	0.8740 (.023)	0.7628 (.063)	0.9059 (.019)	0.8476 (.073)
	Total	0.9072 (.039)	0.8558 (.104)	0.9390 (.03)	0.9007 (.075)

Table 40. Means and standard deviations of macro-Fscore for LoC

Concerning the effect of the context, as captured through the factor ‘others’, contrast analysis shows that the usage of acoustic features yields better results for both Extraversion (contrast value=0.067, p<.0001) and LoC (contrast value=0.080, p<.0001). Moreover, the best results are obtained when the social context is capture by means of the selected features (condition SEL), both for Extraversion (contrast value=0.097, p<.0001) and LoC (contrast value=0.084, p<.0001). Contrary to our expectations, the features selected according to the ANOVA-based approach are not effective: when applied to the target subject they constantly yield worst results, as the summary curves in Figure 20 and Figure 21 show. Clearly, the feature selection procedure was not that effective, as far as the target subject is concerned.

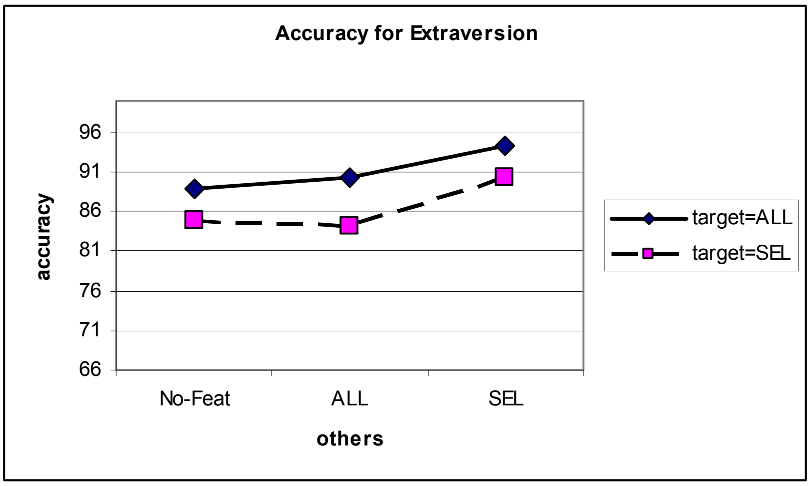

Figure 19. Accuracy for Extraversion

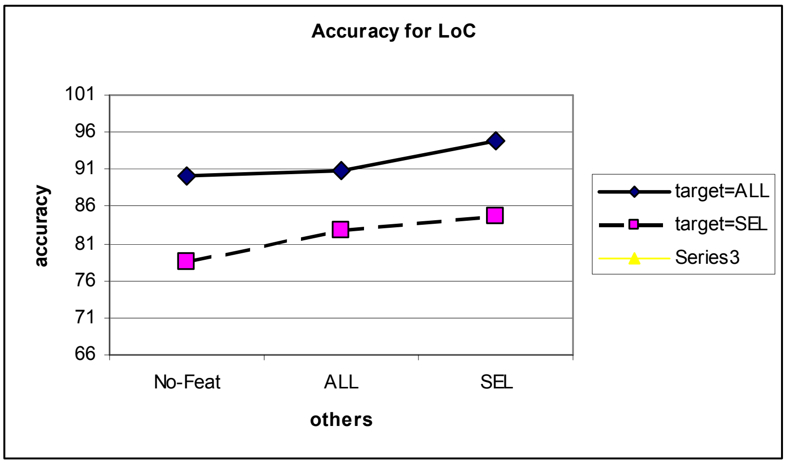

Figure 20. Accuracy for LoC

Concerning the other hypothesis, it is confirmed that the encoding of the social context (what the other members of the group do) improves personality classification. It is important to emphasize, however, that even in the absence of any attempt to (acoustically) capture the social context, the performance obtained are all much higher than the baseline provided by the trivial classifier: 0.8691 for Extraversion and 0.8431 for LoC. Considering that the baseline is 0.6667, the relative improvement is 0.607 and 0.529, respectively. Hence, thin slices of the sole target subject's behaviour are enough to obtain quite a good automatic classification of the two personality traits we are considering.

Our results show that the way the social context is encoded matters too: the best performances are obtained when the selected features are used. A more detailed analysis shows that when all the acoustic features are used for the target subject, the usage of the same features for the social context does not improve the accuracy over the No-Feat condition (comparison between (ALL, No-Feat) and (ALL, ALL)), both for Extraversion and LoC. The improvement is only due to the usage of the selected features (comparisons between (ALL, SEL), and (ALL, No-Feat) and

(ALL, ALL) both significant at p<.0001), again for both personality traits. The role of the selected features for capturing the context is striking and unexpected, given that a) the choice procedure aimed to improve the recognition of personality on the target subject , and b) they are inefficacious to that purpose.

5.1.4 Using Functional Roles for Predicting Personality Traits

As we affirmed, personality is expressed in social situations through relational behaviours. Yet, listing several types of social behaviours may turn out to be cumbersome and, possibly, not interesting. For this reason, we took into account for our task a kind of abstraction at a lower level than personality—namely, functional relational roles. By abstracting over complex behaviours spanning time segments of varying length, relational roles capture both the individual behaviour and its dependence on those of the others in the same group.

Relational roles were automatically annotated in the MSC-II using a SVM-based approach. However, the performance, in terms of accuracy, of the model was not very high: 0.54 for the task area roles and 0.53 for the socio emotional area: in particular, for some classes (e.g., the attacker) we obtained very low precision and recall scores. Moreover, in the Mission Survival 2 corpus neither procedural technicians nor gate-keepers were observed; hence, so we exploited only 4 levels for the task area and only 4 for the socio-emotional one.

In our experiments, the addition of the target subject roles ((ALL, ROLES) condition) caused a performance loss that was more marked for LoC (macro F-score is 0.50 while the baseline is 0.667), when all the acoustic features are used. When the roles are added to the selected features, however, ((SEL, ROLES condition) the results are better both for LoC (0.78 of macro F-score) and for Extraversion (0.81 of macro F-score); in the latter case, the classifier became incapable of finding instances of the less common classes. These results can either be due to a genuine ineffectiveness (or redundancy because functional roles are computed from the same set of features that figure in the ALL condition) of relational role information for the task at hand, or stem from the inaccurate performance of the classifier we used for role assignment.

5.2. Regression

A regression approach was exploited, based on Support Vector Regression (SVR) [Drucker et al., 1997]. Similarly to Support Vector Classification, it produces models that only depend on a subset of the training data, thanks to the cost function that ignores any training data closer to the model prediction than a threshold ε. Moreover, SVR ensures the existence of a global minimum and the optimization of a reliable generalization bound. In ε-SVR the goal is to find a function $f(x)$ that has at most ε deviation from the target for all the training data and at the same time is as flat as possible [Smola and Schölkopf, 2003].

We used an ε-SVR with a Radial Basis Function (RBF) kernel. The cost parameter C, the kernel parameter γ and the threshold ε were estimated through the grid technique by cross-fold validation using a factor of 10.

5.2.1 Feature Selection

We investigated two feature selection procedures: (i) the correlation-based approach on both the acoustic and the visual features, and (ii) the ANOVA-based approach on only the acoustic features (as in the classification task).

The correlation-based feature selection technique [Hall, 1999] selects a subset of features that highly correlate with the target value and have low inter-correlation. This method is used in conjunction with a search strategy, typically Best First that searches the features subset space through a greedy hill-climbing strategy with backtracking. The search may start with an empty set of features and proceed forward (forward search) or with the full set of features and go backward (backward search), or proceed in both directions.

We used the backward and the forward search, applying them both to the features of the target subject and to those of the other members of the group. Table 41 and Table 42 report the results of the two selection procedures for the two personality traits.

LABELS	ACOUSTIC FEATURES	Sel_F		Sel_B	
		Extra	LOC	Extra	LOC
F1 – E	Mean Formant Frequency (Hz)	*		* ▲	* ▲
F2 – E	Mean Confidence in formant frequency	*		* ▲	* ▲
F3 – E	Mean Spectral Entropy			▲	* ▲
F4 – E	Mean of Largest Autocorrelation Peak	* ▲		* ▲	*
F5 – E	Mean of Location of Largest Autocorrelation Peak	*		* ▲	* ▲
F6 – E	Mean Number of Autocorrelation Peaks	▲		▲	▲
F7 –A	Mean Energy in Frame	*	* ▲	* ▲	* ▲
F8 – E	Mean of Time Derivative of Energy in Frame	*	*	* ▲	* ▲
F9 – E	SD of Formant Frequency (Hz)	* ▲		* ▲	
F10 – E	SD of Confidence in formant frequency			* ▲	
F11 – E	SD of Spectral Entropy	* ▲	▲	* ▲	* ▲
F12 – E	SD of Value of Largest Autocorrelation Peak	* ▲	▲	* ▲	▲
F13 – E	SD of Location of Largest Autocorrelation Peak	*		* ▲	* ▲
F14 – E	SD of Number of Autocorrelation Peaks		*	▲	* ▲
F15 – A	SD of Energy in Frame	* ▲		* ▲	* ▲
F16 – E	SD of Time Derivative of Energy in Frame	*		* ▲	▲
F17 – A	Average length of voiced segment (seconds)			▲	* ▲
F18 – A	Average length of speaking segment (seconds)	*		* ▲	▲
F19 – A	Fraction of time speaking	* ▲		* ▲	*
F20 – A	Voicing rate	*		* ▲	▲
F21 – I	Fraction speaking over	*		* ▲	*
F22 – M	Average number of short speaking segments	*		* ▲	* ▲

Table 41. Extracted acoustic features. *= features for the target subject, and ▲= features for the other subjects selected by the two correlation-based selection procedures

LABELS	ACOUSTIC FEATURES	Sel_F		Sel_B	
		Extra	LOC	Extra	LOC
F23	Head fidgeting		*	▲	* ▲
F24	Hands fidgeting			▲	▲
F25	Body fidgeting	*		*	*

Table 42. Extracted visual features, related to Head, Hands, and Body. . *= features for the target subject, and ▲= features for the other subjects selected by the two correlation-based selection procedures

It can be noticed that the forward search (Sel_F) produces a much larger subset of features for Extraversion than for LoC. The backward search (Sel_B), in turn, yields more numerically balanced subsets.

Instead, the ANOVA-based feature selection was performed only on the acoustic features of the target subject, by comparing their means through ANOVA: each feature was treated as a dependent variable in two between-subject analysis of variance, with factor Extraversion (3 levels: Low, score<-1standard deviation, Medium, -1standard deviation≤score≤1standard deviation; High, score>1standard deviation) and LoC (3 levels: Low, Medium, High); significance level was $p<.05$. No adjustment for multiple comparisons was performed, in order to have a more liberal test. Only the features for which the analysis of variance reported significant results were retained, for the each factor, namely: F1, F2, F6, F14, a subset of the Emphasis class, and F21, the Influence feature, for Extraversion, and F1, F6, F14, the same subset of the Emphasis class apart for the mean energy, and F22, the Mimicry feature, for LoC.

5.2.2 Experimental Design

We formulate similar hypotheses to ones formulated in the classification task:

- **Hypothesis 1**. The consideration of the social context improves personality assessment. The social context is encoded through thin slices of the other members of the group.
- **Hypothesis 2**. The selected subsets improve the performance. We investigate if the personality assessment could be made more economical by limiting the analysis to subsets of the features discussed above.

A within-subject design was exploited to address the two hypotheses, with factors ‘Target’ and ‘Others’, each relating to different arrangements of the target subject’s (Target) and of the other participants’ (Others) features.

- ‘Target’ has 3 levels: (i) All features (AllFeat); (ii) the features obtained by means of the correlation-based approach (either Sel_F or Sel_B, see below); (iii) the features provided by the Anova-based procedure (Sel_A).
- ‘Others’ has 4 levels: the same three as for Target, plus a level corresponding to the absence of any features for the other participant (No_Feat). The presence of this level allows to address the contextual hypothesis discussed above.

For each experimental condition, the training instances included the average values of the relevant acoustic and visual feature, computed over a 1-minute window. The analysis was conducted through a leave-one-out procedure. At each of the 48 folds, training was conducted on the data of all but one subject, who was used for testing.

5.2.3 Results and Discussion

Our figure of merit is the squared regression error, SSEER=$(y_{obs}-y_{pred})^2$. Results are compared to those obtained by the base model that always returns the average (27 for LoC and 47 for Extraversion. Its mean SSERR are 59.70 (standard deviation=60.14) for LoC and 63.63 (standard deviation=93.35) for Extraversion.
T-tests ($p<.05$ with Bonferroni corrections) were first conducted comparing the performance of the features obtained by means of the forward (Sel_F) and backward (Sel_B) search for the correlation-based method in the following conditions: (SEL_F, No_Feat) vs. (Sel_B, No_Feat); (SEL_F, All_Feat) vs. (Sel_B, All_Feat); (Sel_F, Sel_F) vs. (Sel_B, Sel_B); (All_Feat, Sel_F) vs. (All_Feat, Sel_B). The two sets of features never produced significant differences for Extraversion, while Sel_B was consistently superior to Sel_F for LoC. Hence, in the following we will consider only Sel_F for Extraversion and Sel_B for LoC.
A repeated measure analysis of variance for Extraversion revealed only a Target main effect ($F_{1.435,\ 47}=6.802$, $p=.004$, with Greenhouse-Geisser correction). According to pairwise comparisons on Target's marginals, Target=All_Feat is significantly lower than the other two levels ($p<.0001$). Finally, all the conditions with Target=All_Feat have SSERR values that are not pairwise statistically different (t-tests, $p<0.05$, Bonferroni correction). Hence, no condition is better than (All_Feat, No_Feat) and there is no evidence that the exploitation of the context (as encoded by the Others' features) improves the results. In other words, both Hypothesis 1 and Hypothesis 2 cannot be maintained. Finally, (All_Feat, No_Feat) is better than the baseline.

Target	Others:	No_Feat	All_Feat	Sel-B	Sel_A	
	All_Feat	19.45 (58.38) *	25.04 (69.98) *	24.13 (61.41) *	26.20 (72.45) *	23.78 (65.69)
	Sel-B	34.09 (68.65)	44.64 (80.93) *	26.63 (69.45) *	45.92 (80.23)	37.05 (75.21)
	Sel_A	35.02 (76.09) *	39.63 (115.06)	49.48 (84.57)	40.57 (102.43) *	41.27 (95.89)
		29.53 (67.99)	36.44 (90.46)	33.41 (72.84)	37.56 (85.79)	

Table 43. Average SSERR and standard deviations for Extraversion. * = conditions that are significantly better than the baseline.

Target	Others:	No_Feat	All_Feat	Sel_F	Sel_A	
	All_Feat	17.78 (45.11) *	11.87 (30.23)	12.58 (32.17)	15.85 (30.03)	14.52 (36.38)
	Sel_F	33.82 (56.42)	27.35 (60.58) *	13.07 (34.91)	39.65 (54.27)	28.47 (53.00)
	Sel_A	33.23 (50.94)	29.73 (94.92)	53.32 (59.90)	26.39 (61.33)	35.69 (69.09)
		28.31 (51.22)	22.98 (67.31)	26.32 (47.82)	27.30 (52.44)	

Table 44. Average SSERR and standard deviations for LoC. * = conditions that are significantly better than the baseline.

Another repeated measure ANOVA for LoC produced both Target ($F_{1.546, 47}$=12.362, p<.0001) and Target*Others ($F_{1.815, 47}$=4.838, p<0.05) effects. Concerning marginals, Target=All_Feat is better than the others (pairwise t-tests, p<0.05, Bonferroni correction). The interaction is due to Others=Sel_B that produces very low SSERR values in two cases out of three (see Table 3). Conditions (All_Feat, All-Feat), (All_Feat, Sel_B) and (Sel_B, Sel_B) do not pairwise statistically differ, provide the best results and are all better than the baseline. Hence, for LoC both Hypothesis 1 and Hypothesis 2 are verified, the latter limited to a few cases.

Hence, the data analysis shows that the two traits we have considered behave differently concerning our hypotheses. In the case of Extraversion, no feature selection procedure provided results that were no worse than those obtained by means of All_Feat for the target subject, and there was no evidence that the consideration of the interaction context improve performance. LoC, in turn, seems more capable of taking advantage of one of the feature selection procedure (Sel_B) and, what is more, there are clear signs that LOC’s manifestation (and/or understanding by an external observer) improves if the social context is considered.

Chapter 6

6. Automatic Prediction of Individual Performance

Endowing machines with the capability of predicting specific behavioural outcomes, such as task performance, is both of the highest theoretical interest and needed to provide important practical results applicable to a host of different goals: group facilitation and coaching; group management, personal skill improvement, etc.
In this chapter we investigate the prediction of individual performance (the task to solve is the Mission Survival one) by means of short sequences of nonverbal behaviour.

6.1. Classification

For our task, we exploited as dataset the 12 meetings contained in the MS-II corpus. As features, we used the 4 classes of acoustic honest signals (Conversational Activity, Emphasis, Mimicry, and Influence) and a class of visual features, Fidgeting. So, by exploiting 1-minute-long thin slices of non-verbal behaviour classification experiments were conducted whereby the system had to assign the thin slices of a subject the right class of individual performance (Low, Medium, and High). Groups of four people were asked to sit around a table and engage in the Mission Survival Task. As said in Chapter 3, the task consisted in imaging a plane crash and in ranking a list of 12 proposed items, according to their importance for survival. During the decision making process, each participant was asked to express her/his own opinions and preferences, and the group was encouraged to discuss each individual proposal before producing the final ranking.
Our classification experiment addressed the following aspects:

- whether subsets of the original audio-visual features could do any better than the full set;
- whether consideration of the context of interaction, encoded by means of the audio-visual features of the other members of the group, provided any advantage;
- whether the predictive power of our thin slices differs according to their temporal position (beginning of the meeting vs. central part vs. final part).

In general, the assumption underlying this study is that non-verbal social behaviour can be used to predict individual performance in social tasks. One might therefore wonder whether the prediction task is better accomplished when the context of social interaction is taken into consideration. So, our two experimental hypotheses are the following:

- **Hypothesis 1**. The consideration of the social context improves performance assessment. For our purposes, the social context is modelled by means of the other participants' thin slices.
- **Hypothesis 2.** Our feature selection approaches improve the learning performance.

A full factorial within-subject design was exploited to test the two hypotheses, with factors 'Target' and 'Others', each relating to different arrangements of the target subject's (Target) and of the other participants' (Others) features.

- 'Target' had 3 levels: (i) All features (All_Feat); (ii) the features obtained by means of the SMO selection approach (Smo_Feat); (iii) the features provided by the Simple Logistic algorithm (Sl_Feat).
- 'Others' had 4 levels: the same three as for Target, plus a level corresponding to the absence of any features for the other participants (No_Feat).

For each experimental condition, the training instances included the average values of the relevant acoustic and visual feature, computed over a 1-minute window; this way, the total number of generated instances corresponded to the total meetings' duration in minutes (i.e. 366 minutes). The analyses were conducted using the leave-one meeting out method, whereby at each fold the classifier is trained on 11 meetings and then tested on the subjects of the left-out one.
From a machine learning point of view, we modeled our task as a three classes-classification task: Low (13 subjects, 0-24 score range), Medium (21 subjects, 25-30) and High (14 subjects, scores higher than 30) performance. Then, classification was conducted by means of a SVM. As in the previous tasks, the "one-against-one" method [Hsu and Lin, 2002] was used to deal with a multi-class classification task. We used The bound-constrained SVM classification algorithm with a RBF kernel was used. The cost parameter C and the kernel parameter γ were estimated through the grid technique by cross-fold validation using a factor of 10. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size.

6.1.1 Feature Selection

As we stated in our Hypothesis 2, we also wanted to investigate the possibility that subsets of the considered features yielded better results than the full set. To this end, we resorted to the classifier subset evaluator, which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. In particular, we exploited the Sequential Minimal Optimization (SMO) algorithm jointly with a linear kernel and the Simple Logistic (SL) algorithm.
The SMO algorithm is the implementation of the John Platt's sequential minimal optimization algorithm for training a Support Vector Machine [Platt, 1998]. We decided to use this technique as classification schema because this algorithm is extremely close to the algorithm we used in our classification task.
The Simple Logistic algorithm is a way for combining linear logistic regression and tree induction, two learning techniques with complementary advantages and disadvantages: the linear logistic regression is a simple linear model of the data and the process of model fitting is quite stable with low variance but potential high biases; on the opposite, the tree induction has high variance but low potential biases. The approach proposed by [Landwehr et al., 2005] is based on the combination of a tree structure and a logistic regression model in order to deal with classification tasks, overcoming the drawbacks of the single techniques.
Both these classification schemas have been used jointly with a searching algorithm, the Best-First search, that searches the space of attribute subsets by means of a greedy hill-climbing strategy, augmented with a backtracking facility.

We applied these techniques separately on the two feature vectors of the "target subject" and of the "other participants". The results are reported in Table 44: in this table the set of all acoustic and visual features extracted from the MS-2 corpus is reported and the selected features related to the target subject are marked with ▲, while those related to the other participants are marked with □.

ACOUSTIST FEATURES			**SELECTED FEATURES**	
CLASSES	**LABELS**	**DESCRIPTION**	**SMO**	**SL**
EMPHASIS	F1	Mean of Formant Frequency (Hz)		
	F2	Mean of Confidence in formant frequency	▲ □	▲
	F3	Mean of Spectral Entropy	□	
	F4	Mean of Largest Autocorrelation Peak	▲	▲ □
	F5	Mean of Location of Largest Autocorrelation Peak	▲ □	▲ □
	F6	Mean of Number of Autocorrelation Peaks	▲ □	
	F8	Mean of Time Derivative of Energy in Frame	□	
	F9	Standard deviation of Formant Frequency (Hz)	▲ □	▲
	F10	Standard deviation of Confidence in formant frequency		▲
	F11	Standard deviation of Spectral Entropy		
	F12	Standard deviation of Value of Largest Autocorrelation Peak	□	□
	F13	Standard deviation of Location of Largest Autocorrelation Peak	□	
	F14	Standard deviation of Number of Autocorrelation Peaks		
	F16	Standard deviation of Time Derivative of Energy in Frame		□
ACTIVITY	F7	Mean of Energy in Frame		
	F15	Standard deviation of Energy in Frame		
	F17	Average length of voiced segment (seconds)	□	
	F18	Average length of speaking segment (seconds)	□	□
	F19	Fraction of time speaking		
	F20	Voicing rate	□	
INFLUENCE	F21	Fraction speaking over	□	
MIMICRY	F22	Average number of short speaking segments		

VISUAL FEATURES		**SELECTED FEATURES**	
LABELS	**DESCRIPTION**	**SMO**	**SL**
F23	Head fidgeting	▲ □	□
F24	Hands fidgeting	▲ □	▲ □
F25	Body fidgeting	▲	▲

Table 45. Acoustic and visual features. ▲= features for the target subject, and □ = features for the other subjects. Selected by the SMO and SL algorithms

As can be seen, the SL algorithm selects only Emphasis and visual features, both for the target and other people; in other words, SL focuses on people motivation and general activation. The SMO algorithm picks up features in all the classes, with the exception of Mimicry, hence leaving open the possibility that a broader view on social behaviour is necessary to attain better prediction of performance.

6.2. Results and Discussion

Table 45 reports the Macro_F values, computed as the harmonic means of the macro averaged one-class Precision and Recall scores, for the various experimental conditions.

		Others			
		All_Feat	No_Feat	Smo_Feat	Sl_Feat
Target	All_Feat	0.227	0.321	0.264	0.268
	Smo_Feat	0.266	**0.407**	0.329	0.348
	Sl_Feat	0.260	**0.415**	0.341	0.330

Table 46. Macro F values

A repeated measure analysis of variance was conducted on the accuracy data from each fold and Table 46 shows these results. Observing the table, we see that main effects were not significant (at p<.05, with Greenhouse-Geisser correction), while Target*Others interaction was significant (F=2.535, p<.05; Greenhouse-Geisser correction). The interaction effect is due to the (significantly) higher performance of (Smo_Feat, No_Feat) and (Sl_Feat, No_Feat).

		Others				
		All_Feat	No_Feat	Smo_Feat	Sl_Feat	
Target	All_Feat	0.341 (0.376)	0.352 (0.279)	0.322 (0.334)	0.339 (0.330)	0.338
	Smo_Feat	0.332 (0.316)	**0.433** (0.330)	0.358 (0.300)	0.404 (0.300)	0.382
	Sl_Feat	0.322 (0.293)	**0.459** (0.359)	0.361 (0.290)	0.381 (0.283)	0.381
		0.332	0.415	0.347	0.375	

Table 47. Acccuracy

In conclusion, our Hypothesis 2 can be at least partially accepted, in that the best conditions exploit the subsets produced by our feature selection procedures. Hypotheses 1, however, is not supported: there is no evidence that the consideration of the social context improves the prediction of performance. These conclusions are supported also by the macro-F measure scores in Table 45.
We also compared the performance of our best classifier with the trivial classifier, which exploits only the prior probability of each performance class: Table 47 compares the precision and recall figures of (Sl-Feat, No_Feat) with those of the trivial classifier.
The former is generally superior, both in terms of performance and of recall, with the exception of recall on H (χ^2=194.87, df=4, p<.0001).

	Precision		**Recall**	
	Sl_Feat, No_Feat	**TC**	**Sl_Feat, No_Feat**	**TC**
Low	0.373	0.283	0.318	0.270
Medium	0.497	0.436	0.666	0.438
High	0.434	0.281	0.270	0.292

Table 48. P(recision) and R(ecall) values for (Sl_Feat, No_Feat) and the Trivial Classifier (TC)

Finally, we analyzed whether the capability of predicting the performance is affected by the temporal position of the thin slice: initial position (first 9 minutes), central position (the following 10 minutes) and final (the last 9 minutes). We restricted our attention to condition (Sl_Feat, No_Feat), summed up the counts of correct classifications for each temporal class across all the subjects, and then ran a χ^2 test comparing the resulting distribution with the uniform one. The non-significance of the comparison (χ^2=1.264, df=2, p>0.05) shows that the classifier performance does not vary according to the temporal position of the thin slices.
In general, our experiments showed that few social signals (a subset of Emphasis class and visual features related to the subjects' head, hands and body energy) have turned out to be enough to take accuracy close to 0.5 and provide a statistically significant improvement over the trivial classifier. However, it doesn't seem that the consideration of the social context improves the results in predicting individual performance. Finally, an interesting result is that the effectiveness of the prediction does not seem to be affected by the temporal position of thin slices. Taken together, these results suggest that measurements of the level of personal motivation/involvement in the interaction (as provided by Emphasis and visual activity features) taken early on in the meeting provide a good indication of what the final performance in the considered task is going to be like.
On a less positive tone, even the best conditions have yielded relatively low values for both accuracy and macro-F score; in particular, class High has a recall value that is lower than that of the criterion. A number of factors can be responsible for this state of affairs: a) the limited size of the considered sample (we used only 12 meetings and 48 subjects for our experiments); b) the rather artificial and not so competitive nature of the task might have limited participants' involvement,

restricting the range of relevant social behaviours; c) the usage of a classification approach rather than of ordinal regression (or regression tout-court). Conceivably, by taking ordinal information in full account ordinal regression might provide better results than classification, which neglects it; d) limitations due to the use of the sole social signals. It might well be that even with 'social' task, as the one considered here, other cues are needed to improve prediction; e.g., those relating to the level of people attention and involvement in the task (as distinguished by attention and involvement in the interaction).

Needless to say, some more investigation is needed to understand the relative importance of these factors and provide better prediction of people performance in social settings. Still, it seems to us that this study opens interesting prospects and indicates relevant avenues for a new and promising research area: the automatic prediction of individual performance in social settings.

Chapter 7

7. Conclusions

In this thesis, we discussed the possibility of exploiting simple acoustic and visual non-verbal features from social interaction to automatically extract information about the participants' behavior. The usage of simple non-verbal features (*social signals*), and the focus on short behavioral sequences (*thin slices*) makes the minimalist approach to human behavior understanding, discussed here, particularly suited to automatic systems. Our results suggest that dynamic and transitory properties, such as the functional role an individual plays in a meeting, and persistent properties, such as personality factors, and behavioural outcomes, such as task performance, can be extracted with a reasonably high accuracy, showing that machines can be made capable to exploit thin slices of behavior in a fashion that closely resembles their usage by humans [Ambady and Rosenthal, 1992].
In particular, our experimental results on three different tasks (automatic recognition of functional roles, automatic prediction of personality traits, and automatic prediction of individual performance) provided the following evidence:

- Regarding the recognition of functional roles, the class of social signals called Conversational Activity seems to provide for as much (if not more) classificatory power than the whole set of acoustic and visual features.
- Again for the recognition of functional roles, the Influence Model seems to be a good learning technique to model these dynamic social behaviours: the performance obtained using this algorithm is comparable to the inter-coder reliability on the same corpus of data. Moreover, an interesting observation is that the Influence Model seems to be generalizable to different numbers of participants in the group. The ability to automatically adapt to different sized groups without retraining would allow a great increase in the flexibility and applicability of automatic role classification technology.
- As concerns the prediction of the personality traits, our studies (classification and regression) not only show the feasibility of automatically assessing personality traits based on thin slices of behaviour (our figures of merit are all much higher than the baseline, and higher than those reported in the few studies on the topic published so far, see for example [Mairesse et al., 2007).; but they also indicate which sets or subsets of features are more appropriate. Moreover, our experimental results (in particular for the regression case) seem to show a different contextual sensitivity of Extraversion and LoC. Probably a reflection of deep differences between these two traits: Extraversion is more directly linked to (certain) behavioural manifestations than LoC, for which the social context acts a moderating factor.

- For the prediction of individual performance, few social signals (a subset of the Emphasis class and the set of fidgeting features) have turned out to be enough to provide a statistically significant improvement over the trivial classifier. At the same time, it doesn't seem that the consideration of the social context improves the results. Finally, the effectiveness of the prediction does not seem to be affected by the temporal position of thin slices. Taken together, these results suggest that measurements of the level of personal motivation/involvement in the interaction taken early on in the meeting provide a good indication of what the final performance in the considered task is going to be like.

Of course, our results are still based on lab data and more evidence is needed from real case studies in more ecological setting (workplaces, schools, and houses). Yet, if definitely proven feasible, the automatic extraction of information about human behavior and about human characteristics from thin slices of behavior opens important scenarios both for the field of human-computer interaction and for the field of human-human, computer mediated interaction.

7.1 Future work

Our work opens some important areas for future experiments and research. First of all, a further step is adding more features, in particular on the visual side. To this end, we are planning to add information related to the amount of attention received (focus of attention), 3D postures, features extracted by facial expression recognition algorithms (e.g., appearance features such as the texture of the facial skin in some specific facial areas including wrinkles, furrows and bulges; geometric features such as the shapes of facial components and the locations of facial fiducial points) [Tian et al., 2005; Pantic and Bartlett, 2007], emotional facial expressions like happiness and anger [Pantic and Rothkrantz, 2003; Sebe et al., 2006; Zeng et al., 2009].

Regarding the automatic recognition of functional roles, some important areas for future work are the following ones:

- Investigate the mutual dependencies between functional roles and personality traits: for example verifying the predictive power of these roles for the detection of personality traits.
- Apply the Functional Role Coding Scheme (FRCS) to more ecological interactions (e.g., real meetings at workplace) or to other small group interactions played in different settings (e.g., political debates, classroom interactions, interactions among friends, TV reality shows, etc).
- Apply and devise machine learning techniques able to deal with very unbalanced dataset (e.g. SMOTE [Chawla et al., 2002]). Our current algorithms does not perform very well at classifying the low-frequency classes (Orienteer/Seeker for Task area roles, and Attacker/Supporter for Socio-Emotional area roles).

As concerns the prediction of personality traits, given the significant results obtained in this thesis the following research directions are open for future works:

- Provide for more comprehensive personality assessments that can be actually used in realistic settings—e.g., by considering the full set of Big Five's scales. Conceivably, this move might require enlarging the scope of the context explored beyond the social ones. It is well known, in fact, that traits such as Extraversions are more deeply involved in social behaviour than others, such as Conscientiousness. Another direction for a move towards practical impact is towards addressing traits that, much as the Locus of Control considered here, have been argued to be important for the relationship and the interaction between humans and machines (e.g., Computer Anxiety).

- Comparing the personality attribution performances of humans and machines in same settings.
- Improve the modelling of the social context (e.g., how the behaviour of the other participants influence the target subject's reactions). Theoretically, the verbal and non-verbal behavior of a given target subject is a manifestation of/ caused by his/her personality At the same time, the verbal and non-verbal behavior is modulated by two different factors: (i) the social context = the behavior of the other members of the group, and (ii) one's beliefs/attitudes concerning the other members.
- Use and devise machine learning models able to represent and explain intermediate states (goals, emotions, or simply some significant and interesting combinations of our low-level features) among the personality traits and the behavioural (visual and acoustic) observations. Having this goal, the usage of generative graphical models seems a better choice than using only discriminative and strongly data-driven approaches, such as SVMs.
- Last, but not least, there comes the important task to connect personality traits to behaviours, attitudes and beliefs of interest in a given scenario for the purposes of personalization and adaptation. One might, therefore, inquiry which interaction style and/or specific product choice are more appropriate to people exhibiting a given level personality profile, and then use this information to adapt the system behaviour.

Instead, concerning the automatic prediction of individual performance we are planning to pursue the following investigations:

- the use of different task (e.g., solving simple problems, playing games such as chess, crosswords, puzzles, etc.). In fact, the rather artificial nature of the Mission Survival task might have limited participants' involvement, restricting the range of relevant social behaviours.
- the usage of ordinal regression (or regression tout-court). Conceivably, by taking ordinal information in full account ordinal regression might provide better results than classification, which neglects it.

Bibliography

[1] Albright, L., Kenny, D.A., and Malloy, T.E. Consensus in personality judgments at zero acquaintance. Journal of Personality and Social Psychology, 55(3), pp. 387-395. 1988

[2] Ambady, N., Bernieri, F. J., and Richeson, J. A. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), Advances in Experimental Social Psychology, 32, pp. 201-271. 2000

[3] Ambady, N., and Rosenthal, R. Thin slices of expressive behaviors as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin, 111, pp. 256-274, 1992

[4] André, E., Klense, M., Gebhard, P., Allen, S., Rist, T. Integrating models of personality and emotions into lifelike characters. In Proceedings of the Workshop on Affect in Interaction - Towards a New Generation of Interfaces, pp 136-149, 1999.

[5] Andrew, K. It's in Your Nature: A Pluralistic Folk Psychology. Synthese, 165 (1) 13-29. 2008

[6] Argamon, S., Dhawle, S., Koppel, M., and Pennbaker, J. Lexical predictors of personality type. In Proceedings of Interface and the Classification Society of North America, 2005.

[7] Asavathiratham, C., Roy, S., Lesieutre, B., and Verghese, G. The influence model. In IEEE Control Systems Magazine, Special Issue on Complex Systems, (12) 2001.

[8] Ba, S.O., and Odobez, J.M. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Mar. 2008.

[9] Bales, R.F. Personality and interpersonal behavior. New York: Holt, Rinehart and Winston, 1970.

[10] Banerjee, S., and Rudnicky, A.I. 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 Language Processing (Interspeech 2004-ICSLP), Jeju Island, Korea, October 2004.

[11] Barzilay, R., Collins, M., Hirschberg, J., and Whittaker, S. The rules behind roles: identifying speaker role in radio broadcasts. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 679-684, July 30-August 03, 2000

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

[13] Basu, S., Choudhury, T., Clarkson, B., and Pentland, A. Towards measuring human interactions in conversational settings. In Proceedings of the IEEE CVPR Int. Workshop on Cues in Communication (CVPR-CUES), Kauai, Dec. 2001.

[14] Benne, K.D., Sheats, P. Functional Roles of Group Members, Journal of Social Issues 4, 41-49. (1948)

[15] Biddle, B.J. Role theory: Expectations, identities, and behaviors New York: Academic Press. 1979

[16] Blanck, P.D., Rosenthal, R., Vannicelli, M., and Lee, T.D. Therapists' tone of voice: descriptive, psychometric, interactional and competence analyses. Journal of Social and Clinical Psychology, 4, pp. 154-178. 1986.

[17] Bloom, G., Castagna, C., and Warren, W. More than mentors: Principal coaching. Leadership. May/June. 2003.
[18] Borkenau, P., and Liebler, A. Trait inferences: Sources of validity at zero acquaintance. Journal of Personality and Social Psychology, 62, pp. 645-657. 1992
[19] Bormann, E. Small Group Communication: Theory and Practice, 3rd ed., Harper & Row, New York, 1990.
[20] Bradbury, J.W., and Vehrenkamp, S. *Principles of Animal Communication*. Sunderland, MA: Sinauer. 1998.
[21] Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition. In Data Mining and Knowledge Discovery, Vol. 2, No. 2, pp. 121-167. (1998)
[22] Carli, G., Gretter, G. A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End based on DSP32C VME Board. In Proceedings of International Conference on Signal Processing, Applications and Technology, ICSPAT, (1992).
[23] Carrere, S., and Gottman, J.M. Predicting divorce among newlyweds from the first three minutes of a marital conflict discussion. Family Process, 38, pp. 293-301
[24] Cassell, J., & Bickmore, T. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. User Modeling and User-Adapted Interaction, 13, 89–132, (2003)
[25] Castelfranchi, F. D. Rosis. How can personality factors contribute to make agents more 'believable', Proceedings of the Workshop on Behavior Planning for Lifelike Characters and Avatars, Sitges, Spain, 25-35, 1999.
[26] Cattell, R.B. *Personality and motivation: Structure and measurement.* New York: Harcourt, Brace & World. 1957.
[27] Cattell, H.E.P, and Mead, A.D. The 16 Personality Factor Questionnaire (16PF). In G.J. Boyle, G. Matthews, and D.H. Saklofske (Eds.), *Handbook of personality theory and testing: Vol. 2: Personality measurement and assessment.* London: Sage. 2007.
[28] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 2002, 16: 341-378.
[29] Chippendale, P. Towards Automatic Body Language Annotation. In Proceedings of 7th International Conference on Automatic Face and Gesture Recognition - FG2006 (IEEE) Southampton, UK, April 2006.
[30] Choudhury, T., Pentland, A., 2003. Sensing and modeling human networks using the sociometer. In Proceedings of the 7th IEEE International Symposium on Wearable Computers, 216-222.
[31] Cohen, I., Sebe, N., Garg, A., Len, M.S., and Huang, T.S. Facial expression recognition from video sequences. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02), vol. II, pp. 121-124, Lausanne, Switzerland. 2002
[32] Cohen, J. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* Vol.20, No.1, pp. 37–46. 1960
[33] Costa, P.T., and McCrae, R. R. NEO PI-R Professional Manual. Psychological Assessment Resources, Odessa, FL., 1992
[34] Costanzo, M., and Archer, D. Interpreting the espressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior, 13,* pp. 225–245. 1989

[35] Cristani, M., Pesarin, A., Drioli, C., Perina, A., Tavano A., and Murino, V. Auditory Dialog Analysis and Understanding by Generative Modelling of Interactional Dynamics, Second IEEE Workshop on CVPR for Human Communicative Behavior Analysis, pp. 103-109, 2009

[36] Cristianini, N., and Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambridge University Press; 2000.

[37] Curhan J. and Pentland A. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology, 92(3), pp. 802–811, 2007.

[38] Dabbs, J.M., Bernieri, F.J., Strong, R.K., Campo, R., and Milun, R. Going on stage: Testosterone in greetings and meetings. *Journal of Research in Personality*, 35, pp. 27-40. 2001

[39] DePaulo, B.M. Nonverbal behavior and self-presentation. Psychological Bulletin, 111, pp. 203-243. 1992

[40] Dietterich T. G. Machine Learning for Sequential Data: A Review. In T. Caelli (ed.) Lectures Notes in Computer Science. Springer-Verlag, 2002.

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

[42] Dong, W., and Pentland, A. Modeling Influence Between Experts. Artificial Intelligence for Human Computing: pp. 170-189. 2007

[43] Donnellan, M. B., Conger, R. D., & Bryant, C. M. The Big Five and enduring marriages. Journal of Research in Personality, 38, 481–504, (2004)

[44] Doyle M. and Straus D. How To Make Meetings Work. The Berkley Publishing Group, New York, NY. 1993.

[45] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. Support Vector Regression Machines. In Advances in Neural Information Processing Systems 9 NIPS, pp. 155-161, MIT Press (1997).

[46] Ekman, P., and Friesen, W. V. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, pp. 49- 98. 1969

[47] Exline, R.V., Ellyson, S.L., and Long, B., Visual behavior as an aspect of power role relationships. In Advances in the Study of Communication and Affect, Plenum Press, New York.

[48] Eysenck, H.J. Dimensions of personality: 16, 5 or 3? criteria for a taxonomic paradigm. Personality and Individual Differences, 12 (8), 773–790, (1991)

[49] Eysenck, H.J., and Eysenck, S.B.G. Manual of the Eysenck Personality Inventory. London: University of London Press, 1964.

[50] Falcon V., Leonardi C., Pianesi F., Zancanaro M., Annotation of Group Behaviour: a Proposal for a Coding Scheme. Proc. of Workshop on Multimodal Multiparty Multimodal Processing at ACM-ICMI 2005, 39-46. (2005).

[51] Farma, T., and Cortivonis, I. Un Questionario sul "Locus of Control": Suo Utilizzo nel Contesto Italiano (A Questionnaire on the Locus of Control: Its Use in the Italian Context). Ricerca in Psicoterapia. Vol. 2, 2000.

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

[53] Favre, S., Salamin, H., Dines, J., Vinciarelli, A. Role recognition in multiparty recordings using social affiliation networks and discrete distributions, in: Proceedings of the International Conference on Multimodal Interfaces (ICMI), Chania, Oct. 2008.

[54] Funder, D.C. Personality. Annual Review of Psychology, 52, pp. 197-221. 2001

[55] Funder, D.C., and Sneed, C. D. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality and Social Psychology, 64 (3), 479–490, (1993)

[56] Furnham, D. Language and Personality. In Giles, H., & Robinson, W. (Eds.), Handbook of Language and Social Psychology. Winley, 1990.

[57] Gada, N.M., Bernieri, F., Grahe, J.E., Zuroff, D., and Koestner, R. Love: How do observers perceive it. Paper presented at the Midwestern Psychological Association Convention. Chicago. 1997.

[58] Goren-Bar, D., Graziola, I., Pianesi, F., Zancanaro, M. Influence of Personality Factors on Visitors' Attitudes towards Adaptivity Dimensions for Mobile Museum Guides'. In User Modeling and User Adapted Interaction: The Journal of Personalization Research, 16 (1), 31-62. 2006

[59] Graziola I., Pianesi P., Zancanaro M., Goren-Bar D. Dimensions of Adaptivity in Mobile Systems: Personality and People's Attitudes. In Proceedings of Intelligent User Interfaces IUI05, San Diego, CA (2005).

[60] Grucza, R.A., and Goldberg, L.R. The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89*, 167-187. 2007

[61] Guyon, I., and Elisseeff, A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003), pp. 1157-1182, 2003.

[62] Hall, J.A., Coats, E.J., and LeBeau, L.S., Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychological Bulletin. v131 i6. pp. 898-924. 2005.

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

[64] Hall, M.A. Correlation-based Feature Selection for Machine Learning. PhD dissertation, Department of Computer Science, University of Waikato (1999).

[65] Hare, A.P. Types of roles in small groups: a bit of history and a current perspective. Small Group Research. v25, pp. 433-448. 1994.

[66] Hasson, O. Toward a general theory of biological signalling. Journal of Theoretical Biology 185: 139-156. 1997.

[67] Heider, F. The psychology of interpersonal relations. Wiley. New York. 1957

[68] Hogan, R., Curphy, G. J., & Hogan, J. What we know about leadership: Effectiveness and personality. American Psychologist, 49 (6), 493–504, (1994)

[69] Hsu, C.W. A, and Lin C.-J. A Comparison of Methods for Multi-Class Support Vector Machines. IEEE Transactions on Neural Networks, 13, pp. 415-425, 2002.

[70] Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. 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, 2007.

[71] Hung, H., Hunag, Y., Yeo, C., and Gatica-Perez, D. Associating Audio-Visual Activity Cues in a Dominance Estimation Framework. In Proceedings of the First IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2008

[72] Hung, H., Jayagopy, D.B., Ba, S.O., Odobez, J.M., and Gatica-Perez, D. Investigating Automatic Dominance Estimation in Groups From Visual Attention and Speaking Activity. In Proceedings of International Conference on Multimodal Interfaces, 2008

[73] Jayagopi, D.B., Ba. S., Odobez, J.M., and Gatica-Perez, D. Predicting Two Facets of Social Verticality in Meetings from Five-Minute Time Slices and Nonverbal Cues. In Proceedings of the International Conference on Multimodal Interfaces (ICMI), Chania, October 2008

[74] Jayagopi, D.B:, Hung, H., Yeo, C., and Gatica-Perez, D. Predicting the Dominant Clique in Meetings through Fusion of Nonverbal Cues. In Proceedings of ACM MM 2008

[75] Jayagopi, D.B., Hung, H., Yeo, C., and Gatica-Perez, D. Modeling Dominance in Group Conversations from Nonverbal Activity Cues. IEEE Trans. on Audio, Speech, and Language Processing, Special Issue on Multimodal Processing for Speech-based Interactions, Vol. 17, No. 3, pp. 501-513. 2009

[76] John, O. P., Donahue, E. M., & Kentle, R. L. The "Big Five" Inventory: Versions 4a and 5b. Tech. rep., Berkeley: University of California, Institute of Personality and Social Research, (1991)

[77] John, O. P., Srivastava, S. The Big five trait taxonomy: History, measurement and theoretical perspectives. In Pervian, L. A. & John, O. P., (Eds.) Handbook of personality theory and research. Guilford Press. New York, 1999.

[78] Johnson, R. D., Marakas, G., Plamer, J. W. Individual Perceptions Regarding the Capabilities and Roles of Computing Technology: Development of The Computing Technology Continuum of Perspective. Ms. 2002.

[79] Katz, D., and Kahn, R.L. The social psychology of organizations (2nd ed.). (JohnWiley, New York). 1978

[80] Keltner, D., and Haidt, J. The social functions of emotions at four levels of analysis. *Cognition and Emotion, 13*, pp. 505-522. 1999

[81] Kenny, D. A. Interpersonal perception: A social relations analysis. New York: Guilford Press. 1994.

[82] Kressel U. Pairwise Classification and Support Vector Machines. In B. Scholkopf, C.J.C. Burges, and A.J. Smola (eds.) Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA, 1999.

[83] Komarraju, M., & Karau, S. J. The relationship between the Big Five personality traits and academic motivation. Personality and Individual Differences, 39, 557–567, (2005)

[84] Landis, J.R, and Koch, G. The measurement of observer agreement for categorical data. Biometrics; 33: pp. 159-174. 1977.

[85] Landwehr N., Hall M., and Frank E. Logistic Model Tree. In Machine Learning 59 (1-2) 161-205, 2005

[86] Lilliefors, H. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, Vol. 62. pp. 399-402. 1967

[87] Mairesse, F., & Walker, M. Automatic recognition of personality in conversation. In Proceedings of HLT-NAACL, (2006a)

[88] Mairesse, F., Walker, M. Words marks the nerds: computational models of personality recognition through language. In Proceedings of the 28th Annual Conference of the Cognitive Science Society, pp 543-548, 2006.

[89] Mairesse F., Walker M.A., Mehl M.R., and Moore R.K. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In Journal of Artificial Intelligence Research 30, pp.457-500. 2007.

[90] Maynard Smith, J., and Harper D.G.C. Animal Signals. Oxford University Press. 2003.

[91] Mallory P., and Miller V. A possible basis for the association of voice characteristics and personality traits. Speech Monograph, 25, pp. 255-260, (1958).

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

[93] McCrae, R.R., and John, O.P. An introduction to the five-factor model and its applications. Journal of Personality, 60, pp. 175-215. 1992

[94] McGrath, J.E. Groups: Interaction and performance. Englewood Cliffs, N.J.: Prentice Hall. 1984

[95] McLarney-Vesotski, A. R., Bernieri, F., & Rempala, D. Personality perception: A developmental study. Journal of Research in Personality, 40 (5), 652–674, (2006)

[96] Oberlander, J. and Nowson, S. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In Proceedings of the Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 627-634, 2006.

[97] Olguín, D., O., Gloor, P. A., Pentland, A., 2009 Capturing Individual and Group Behavior with Wearable Sensors. In Proceeding of AAAI Spring Symposium on Human Behavior Modeling.

[98] Oliver, N., Rosario, B., and Pentland, A., A Bayesian computer vision system for modeling human interactions. In IEEE transactions on pattern analysis and machine learning, 22(8) 831-843, 2000.

[99] Oviatt, S. Multimodal Interfaces. In Handbook of Human-Computer Interaction, (ed. by J. Jacko & A. Sears), Lawrence Erlbaum: New Jersey, 2002.

[100] Pantic, M. ,and Bartlett, M.S. Machine analysis of facial expressions. In K. Delac and M. Grgic (Eds), *Handbook of Face Recognition* (pp. 377-416). Vienna, Austria: I-Tech Education and Publishing. 2007

[101] Pantic, M., and Rothkrantz, L.J.M. Toward an affective-sensitive multimodal human-computer interaction. Proceedings of the IEEE, 91(9): pp. 1370-1390. 2003

[102] Paunonen S.V., On the accuracy of ratings of personality by strangers. Journal of Personality and Social Psychology 61, pp. 471–477. 1991

[103] Paunonen, S.V., and Jackson, D. N. What is beyond the Big Five? plenty!. Journal of Personality, 68 (5), 821–836, (2000)

[104] Peabody, D. and Goldberg, L.R. Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology, 57*, 552-567. 1989.

[105] Pentland, A. A Computational Model of Social Signaling. Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06). Volume 1, Issue 2006 , Page(s):1080 – 1083 (2006)

[106] Pentland, A. Honest Signals: how they shape our world. (In press). MIT Press, September 2008.

[107] Pentland, A. Socially Aware Computation and Communication. IEEE Computer, May 2005

[108] Pentland, A., Social Signal Processing. In IEEE Signal Processing Magazine, 108-111, July 2007

[109] Perugini, M. and Di Blas L. Analyzing Personality-Related Adjectives from an Eticemic Perspective: the Big Five Marker Scale (BFMS) and the Italian AB5C Taxonomy. In De Raad, B., & Perugini, M. (Eds.), Big Five Assessment, Hogrefe und Huber Publishers. Göttingen, 281-304 (2002)

[110] Pianesi, F., Zancanaro, M., Falcon, V., and Not, E. Toward supporting group dynamics. In Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, Athens, Greece, 2006

[111] Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press, 1998

[112] Psathas, G. Conversation Analysis: The Study of Talk in Interaction, Sage Publications. 1995

[113] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE,* **77** (2), 257-286.

[114] Raducanu, B., Vitria, J., and Gatica-Perez, D. You are Fired! Nonverbal Role Analysis in Competitive Meetings. In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taiwan, April, 2009

[115] Reeves, B., & Nass, C. The Media Equation. University of Chicago Press, (1996)

[116] Richmond, V. P., and McCroskey, J. C. *Nonverbal Behavior in Interpersonal Relations,* 3rd edition. Needham Heights, MA: Allyn & Bacon. 1995.

[117] Rienks R., and Heylen D., Dominance Detection in Meetings Using Easily Obtainable Features. In Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms. Edinburgh, Scotland, October 2006.

[118] Rienks R., Zhang D., Gatica Perez D., and Post W. Detection and Application of Influence Rankings in Small Group Meetings. In Proceedings of International Conference of Multimodal Interfaces ICMI-06, 2006.

[119] Rienks, R., Nijholt, A., and Heylen, D. Verbal behavior of the more and the less influential meeting participant. In Proceedings of the International Conference on Multimodal

Interfaces (ICMI), Workshop on Tagging, Mining and Retrieval of Human Related Activity Information, Nagoya, Oct. 2007.

[120] Rosenthal, R., Hall, J.A., Archer, D., DiMatteo, M.R., and Rogers, P.L. Measuring sensitivity to non-verbal communication: The PONS test. In A. Wolfgang (Ed.), Nonverbal behavior : Applications and cross-cultural implications. New York: Academic Press. 1979

[121] Rotter J.B. Generalized Expectancies for Internal versus External Control of Reinforcment. In Psychological Monographs, 80 (1, Whole N. 609), (1965)

[122] Salazar, A. An Analysis of the Development and Evolution of Roles in the Small Group. Small Group Research, 27, 4, pp. 475-503, 1996.

[123] Shapiro, S.S., and Wilk, M.B. An analysis of variance test for normality (complete samples). Biometrika, 52, 3 and 4, pages 591-611. 1965.

[124] Scheflen, A.E. The significance of posture in communication systems. Psychiatry, 27: pp. 316-331. 1964

[125] Scherer K.R. Personality Inference from Voice Quality: the Loud Voice of Extraversion. European Journal of Social Psychology, vol. 8, pp. 467-487, 1978.

[126] Scherer K.R. Personality markers in speech. In Scherer K.R. and Giles H. (eds.) Social Markers in Speech, pp. 147-209 Cambridge University Press, (1979)

[127] Scott-Phillips, T. C. On the Correct Application of Animal Signalling Theory to Human Communication, in A. D. M. Smith, K. Smith and R. Ferrer i Cancho (eds.), *The Evolution of Language: Proceedings of the 7th International Conference on the Evolution of Language*, Singapore: World Scientific, p.275-282. 2008.

[128] Sebe, N., Cohen, I., Gevers, T., and Huang, T. Emotion Recognition Based on Joint Visual and Audio Cues. In Proceedings of International Conference on Pattern Recognition (ICPR 2006), pp. 1136-1139, Hong Kong, August 2006.

[129] Shrout, P.E., and Fiske, D.W. Nonverbal behaviors and social evaluation. Journal of Personality, 49(2): pp. 115-128, 1981

[130] Siegman A., and Pope B. Personality variable associated with productivity and verbal fluency in the initial interview. Proceedings of the 73rd Annual Conference of the American Psychological Association, 1965.

[131] Sigurdsson, J. F. Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. Personality and Individual Differences, 12 (6), 617–624, (1991)

[132] Smith-Lovin L., and Brody, C., Interruptions in group discussions: the effects of gender and group composition. American Sociological Review 54, pp. 424–435, 1989

[133] Smola , A. J., and Schölkopf, B. A Tutorial on Support Vector Regression. Statistics and Computing, (2003)

[134] Stoltzman, W. Toward a Social Signaling Framework: Activity and Emphasis in Speech. MEng Thesis, MIT. 2006

[135] Tian, Y., Kanade, T., and Cohn, J.F. Facial Expression Analysis In *Handbook of Face Recognition*, Edited by Stan Li and A.K. Jain, Springer-Verlag January, 2004.

[136] Vapnik, V.N. *The Nature of Statistical Learning Theory*. Springer, 1995.

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

[138] Vinciarelli A., Pantic M. and Herve B., Social Signal Processing: Survey of an Emerging Domain. Journal of Image and Vision Computing. 2009.

[139] Waxer, P. Nonverbal cues for depth in depression: set versus no ser. Journal of Consulting and Clinical Psychology. 44. 493. 1976

[140] Waxer, P. Nonverbal cues for anxiety: an examination of emotional leakage. Journal of Abnormal Psychology, 86, pp. 306-314

[141] Weiner, B. *Achievement motivation and attribution theory.* Morristown, N.J.: General Learning Press. 1974

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

[143] Yule, G. Pragmatics. Oxford University Press. 1996

[144] Zahavi, A. and Zahavi, A. The handicap principle: a missing piece of Darwin's puzzle. Oxford University Press. Oxford. 1997.

[145] Zeng, Z., Pantic, M., Roisman, G.L., and Huang, T.H. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1): pp. 39-58. 2009

[146] Zhou, X. and Conati, C. Inferring user goals from personality and behavior in a causal model of user affect. In Proceedings of the 8th international Conference on intelligent User interfaces (Miami, Florida, USA, January 12 - 15, 2003), 211-218, ACM, New York, NY. 2003.