Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection

Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro

Fondazione Bruno Kessler
via Sommarive 18
38050 Povo – Trento (Italy)

{mana, lepri, chippendale, cappelle, pianesi, svaizer, zancana}@fbk.eu

ABSTRACT

This paper describes an automatically annotated multimodal corpus of multi-party meetings. The corpus provides for each subject involved in the experimental sessions information on her/his social behavior and personality traits, as well as audio-visual cues (speech rate, pitch and energy, head orientation, head, hand and body fidgeting). The corpus is based on the audio and video recordings of thirteen sessions, which took place in a lab setting equipped with cameras and microphones. Our main concern in collecting this corpus was to investigate the possibility of creating a system capable of automatically analyzing social behaviors and predicting personality traits using audio-visual cues.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces: Computer supported cooperative work Synchronous interaction.

I.2.10. [**Artificial Intelligence**]: Vision and Scene Understanding – Perceptual Reasoning.

General Terms

Your general terms must be any of the following 16 designated terms: Design, Experimentation, Human Factors.

Keywords

Multi-party meetings, multimodal corpus, automatic annotation.

1. INTRODUCTION

In this paper we present a multimodal corpus of multi-party meetings, automatically annotated using audio-visual cues (speech rate, pitch and energy, head orientation, hand fidgeting, and body fidgeting). The multimodal corpus was developed with two goals in mind. Firstly, the collection of multimodal data to support the design and development of systems and tools capable of automatically analyzing social behaviors and predicting personality traits by exploiting audio-visual cues. Secondly, the multimodal data collected is meant to enhance the corpus of empirical data in human interaction, as well to improve our understanding in the multiple social, psychological and emotional aspects involved in multi-party meetings.

Several other multimodal corpora have recently been developed to analyze meetings. Among these, the MM4 corpus [11] and the VACE corpus [5] are relevant to our work since they include low-level cues of human behavior, such as speech, gesture, posture, and gaze in order to interpret high level meeting events. Similar considerations hold also for the AMI corpus described in [14].

In the past, we developed the "Mission Survival Corpus 1" (MSC-1) [13]. This corpus is different from the corpora mentioned above because it was built in a controlled but non-scripted way. In fact, although the groups' composition and the tasks were pre-determined, the behavior of the meetings' participants was spontaneous. In this way, MSC-1 provides a range of natural interactions without the excessive variability that would characterize a corpus of daily meetings.

However, some weaknesses affect MSC-1. First of all, the collected meetings are quite short (their average length is 19' 37"). Secondly, MSC-1 lacks any measure about how good or bad the meetings were. Finally, the audio-visual cues extracted were limited to speech rate, and hands and body fidgeting.

Furthermore, MSC-1 includes functional role annotations [2] but does not include any information about the personality traits of the participants, information that is interesting and useful with respect to our goal to investigate the possibility of developing an automatic personality trait detector.

The paper is organized as follows: Section 2 presents methods and procedure applied for the data collection. Section 3 describes some measures related to personality traits and social behavior. Section 4 presents the results of the automatic annotations. Finally, Section 5 summarizes the present work and draws some possible future works.

2. DATA COLLECTION: METHODS AND PROCEDURES

The goal of the last data collection was to accumulate a large multimodal corpus of multi-party meetings, in which people show some personality traits and social behaviors while discussing and interacting.

ICMI '07 Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, November 15, 2007, Nagoya, Japan

As in [13], in order to promote group discussion we used one of two versions of the Survival Task (ST), often used in experimental and social psychology to elicit decision-making processes in small groups [8].

2.1 The Survival Task

The ST exercise consists of asking participants to reach a consensus on how to survive a disaster scenario. In consensus decision making processes, each participant is asked to express their own opinions. The group is encouraged to discuss each individual proposal through the weighing and evaluation of decision quality.

In our setting, we specifically used the 'plane crashing in Canada' scenario. The group task was to rank 12 items according to their importance for crew member survival. As for the previous data collection, the basic structure of the ST was retained. In particular, the task was a) collaborative within the group (because the group goal was to reach a common solution possibly by global consensus); b) competitive across groups (because all final group lists were ranked and each group was encouraged to find the best solution).

Given the intensive engagement required by the groups to reach a common task solution, a large set of social dynamics and personal attitudes can be observed.

2.2 Experimental Procedure

Before starting each recording session, some general information about the task was given to the participants gathered in separate the room. Each participant was asked to sign a consent-form to preserve privacy and data confidentiality. Then, they filled in a first personality questionnaire (pre-questionnaire – see Section 3). After the questionnaire filling, participants moved into the lab (described in Section 2.3), specially-equipped for audio/video recordings. As each person entered the room, one-at-a-time, they were initiated into the 3D tracking system (see Section 2.3). He/she put on a close-talk microphone and sat around a table, taking up a specific position, marked on the table with a label according to a pre-defined scheme (Subject1-West, Subject2-North, Subject3-East, and Subject4-South). In front of each participant there lay a sheet of paper, presenting the task and the list of objects to be ordered. There was also a further sheet on which to report the final solution and a single pen for all was located at the middle of the table. Unlike the previous data collection, where an experimenter was also seated inside the experiment room, this time only the four participants were present. Instead, the experimenter was sat in a separated room to monitor the acquisition systems and observe the scene through the video-streams. At the end of the task, participants went back to the initial room, where a second questionnaire (post-questionnaire - see Section 3) was to be completed. In the meantime, the experimenter scored the list provided by the groups and finally presented them with their scores, explaining how they were calculated and what was the correct solution.

Globally 52 participants (male: 51.9%; females: 48.1%; average age: 35 years), distributed in thirteen groups, were involved in the data collection.

2.3 Technical Set-up and Recording Procedure

Each session was visually recorded in a specially-equipped room at FBK-irst (see Figure Figure 2.1), by means of four firewire cameras placed in the four corners of the room and four actively driven web cameras (PTZ IP cam) which were installed on the walls surrounding the table.

In order to obtain an optimal detection of acoustic events, four wireless close-talk microphones and one omni-directional microphone placed on tabletop were used to record speech activity during experiments.

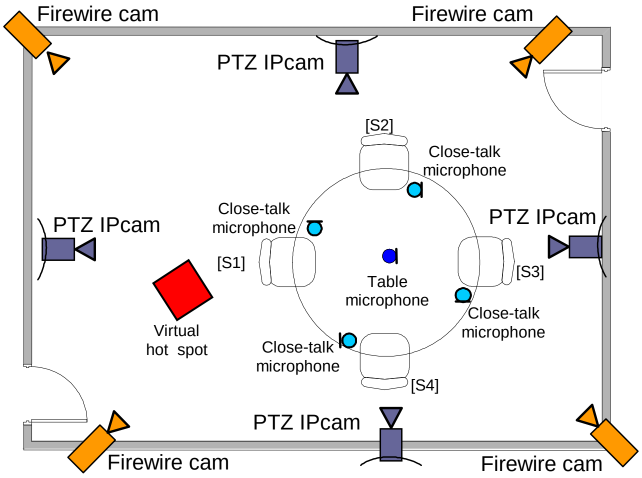

Figure 2.1 The experimental setting

To acquire ideal (i.e. maximized and frontal) images for a fidgeting detector module, the four PTZ cameras had to be instructed to point towards single participants as they moved within the environment. In order to achieve this, it was necessary to run a 3D multi-person tracking module in real-time at the time of recording. Consequently, before a session could be recorded, a visual representation (or model) of each session participant had to be acquired and supplied to the person tracker. The model acquisition procedure is semi-automatic, requiring that each person should stand in a virtual hot spot in the scene for a couple of seconds when they enter the room (See Figure 2.1). Target detection within this area is based upon the matching of extracted image contours with a virtual silhouette of an average adult rendered for different heights and widths.

After this model acquisition, both audio recording and vision tracking systems are ready to be run.

3. PERSONALITY TRAITS AND SOCIAL BEHAVIOR MEASURES

The personality questionnaires, filled in by the participants before and immediately after the meetings, were the Italian version of Craig's Locus of Control of Behavior scale (LCB) [7], and part of the Big Marker Five Scales (BFMS) related to the Extraversion dimension [12]. In addition to this information relating to personality traits, we also collected information about group cohesion, asking them to indicate to what extent they felt themselves and the other participants as group members.

In the post-questionnaire we also asked them to provide us with a personal solution to the task. Therefore, in addition to scoring the Group's Performance, we also measured Individual Performance by scoring the individual solutions provided by each subject.

Finally, for each measure 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 [10] and the Shapiro-Wilk statistic [18]. For both tests, if the p-values of the two tests are greater than 0.05, the distribution is normal.

3.1 Locus of Control

A Locus of Control (LoC) orientation is a belief about whether the outcomes of our actions are dependant upon what we do (internal orientation) or on events outside of our control (external orientation) [16]. In our experiments we measured the LoC by means of a questionnaire, composed of 17 questions/items with a rating scale from 0 to 5 points [7].

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 [7]. As depicted in Table 1, the distribution is normal.

Table 1. LoC: Tests of Normality

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

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

(a) Lilliefors Significance Correction

3.2 Extraversion (Big Five Scale)

To measure Extraversion, we used the extraversion sub-scale of the Big Five Marker Scale.

The Big Five is a "comprehensive system, a framework for virtually organizing all personality traits" [23]. This framework includes five factors of personality: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Creativity. In particular, Extraversion is the quantity and the intensity of a subject's interpersonal reactions, emotional expressiveness, and sociability.

Table 2. Extraversion: Tests of Normality

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

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

(a) Lilliefors Significance Correction

The mean and the standard deviation of the Extraversion raw scores are 43.6 and 10.3. The results of the tests of normality, reported in the Table 2, show us that the distribution is normal.

3.3 Group Cohesion

The interest in group cohesion pervades several research areas, including social psychology, group dynamics, and computer-mediated communication ([21] and [15]). Perceived cohesiveness encompasses an individual's sense of *belonging* to a particular group and his/her feelings of *morale* associated with membership in the group [2]. Perceived cohesiveness reflects an individual's appraisal of her relationship to the group, so that higher group cohesion corresponds to positive feelings of membership.
The mean and the standard deviation of the participants' scores are 32.3 and 6.2 respectively. Also in this case the distribution is normal (see Table 3).

Table 3. Group Cohesion: Test of Normality

Kolmogorov-Smirnov (a)			Shapiro-Wilk		
Statistic	df	Sig.	Statistic	df	Sig.
0.105	52	0.200(*)	0.973	52	0.294

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

(a) Lilliefors Significance Correction

3.4 Individual and Group Performance

The Individual and Group performances have been measured by scoring the individual and group solutions provided.

Scoring consists of considering the first five items of the final list, scoring each one with their position in the correct task solution and summing the scores. In this way, the range of possible scores goes from 15 (best score) to 79 (worst score).

The mean and the standard deviation for the raw scores of Individual Performance are 80.1 and 7.1, while for the Group Performance are 80.6 and 7.4.

By comparing the measures of individual and group performances, we can note that they are very similar.

On one side, this might be due to the fact that the individual ranking was following the group one. In this case the group discussion, done to reach a consensus decision, might have influenced the individual positions.

On the other side, the two performances might be similar simply because most of the participants had the same opinion and agreed on the group solution given.

To better understand the reasons for this similarity, a deeper analysis of the solutions should be conducted. In this analysis we should consider the ranking of the specific items, their order, and in particular the discussions between each subject and the others, trying to detect possible impositions or negotiations.

Concerning the tests of Normality, as shown in Table 4, in this case the p-values obtained do not lead us to assume that the distributions are normal.

Table 4. Individual and Group Performance: Tests of Normality

	Kolmogorov-Smirnov (a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Individual Performance	0.131	52	0.026	0.940	52	0.011
Group Performance	0.196	52	0.000	0.889	52	0.000

(a) Lilliefors Significance Correction

4. DATA ANNOTATION

The collected multimodal corpus consists of audio and video recordings of thirteen meetings, having durations between 28’.14” – 34’.09” and an average length of 31’.10” The total length of the collected audio-video is 6h 47’ 8”.

As in MSC-1, the collected corpus has been enriched by adding some automatic annotations concerning speech activity and 3D tracking of body activity. However, in this case the extracted audio-visual cues are not only limited to speech rate and hands/ body fidgeting but also include pitch and energy values from the audio side and head orientation and head fidgeting on the body tracking side[1].

4.1 Speech activity

Speech activity refers to the detection of the presence/absence of human speech. For a good detection, human speech must be distinguished from non-speech (silence, breaths, yawns, coughing, and noises – e.g. noises caused by the subjects when touching the microphones). At the same time, speech can be analyzed to extract information about waveform, spectrogram, pitch, energy, formants, etc.

4.1.1 Voice Activity Detection

The speech activity of each participant in our Mission Survival Task was recorded by the close-talk microphones, and it has been automatically segmented at a 500ms frame rate and labelled by means of a VAD (Voice Activity Detector) [4]. For each session, the VAD detects participant’s speech activity and produces 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).

Given the VAD output, we calculated the speech rate of each participant by measuring the individual contribution with respect to the global speech of the session. For each participant, this was achieved by considering all the temporal frames with value '1' (without any distinction of "overlapped speech", i.e. when two or more people are speaking simultaneously).

[1] Contrary to MSC-1, at the moment this corpus does not include any annotation of functional roles. In the future we will consider the possibility of adding this labeling.

The mean speech rate is 0.32 and its standard deviation is 0.13, while from the tests of normality we can assume that the distribution is normal (see Table 5).

Table 5. Speech Rate: Tests of Normality

Kolmogorov-Smirnov (a)			Shapiro-Wilk		
Statistic	df	Sig.	Statistic	df	Sig.
0.072	52	0.200(*)	0.987	52	0.825

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

(a) Lilliefors Significance Correction

4.1.2 Pitch and Energy

A speech analysis of the recorded audio was conducted using the Wavesurfer tool [17] to extract information about pitch and energy.

The pitch analysis tries to capture the fundamental frequency of the sound source by analyzing the final speech utterance. The fundamental frequency is the dominating frequency of the sound produced by the vocal chords. It gives the strongest correlation as to how the listener perceives the speakers' intonation and stress. The energy expresses the speech power during the vocal emission.

For each participant we extracted pitch and energy values of his/her audio. Then, we normalized these values and calculated their mean value per subject to be used as a possible acoustic feature. Globally, the means of pitch and energy were 0.14 and 0.33 respectively, while the standard deviations were 0.04 and 0.06.

Table 6. Pitch and Energy: Tests of Normality

	Kolmogorov-Smirnov (a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pitch	0.124	52	0.044	0.969	52	0.187
Energy	0.147	52	0.007	0.954	52	0.044

(a) Lilliefors Significance Correction

Concerning the results of the Normality tests, depicted in Table 6, only the Shapiro-Wilks test does not reject the assumption of normality for the pitch distribution, while both tests imply that the energy distribution is not normal.

4.2 3D Tracking of Body Activity

4.2.1 Fidgeting

Fidgeting is defined as “a condition of restlessness as manifested by nervous movements” [22], and it can reveal important clues about the emotional state and activity of an individual [24]. Using visual means, fidgeting signatures can be detected by employing techniques such as optical flow or Memory History Images (MHIs) [9].

For each subject, the output of the fidgeting analysis consists of a data structure containing: an absolute timestamp, followed by three parameters relating to the fidgeting energy for the head, body and hands. At the end of a session, these values were then normalized to the maximum observed activity of each participant during the entire meeting.

Globally, the statistical mean of hand and body fidgeting are 0.05 and 0.03 respectively, while their standard deviations are 0.03 and 0.02. Concerning the tests of normality, Table 7 shows that both distributions are not normal.

Table 7. Hands and Body Fidgeting: Tests of Normality

	Kolmogorov-Smirnov (a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hands	0.157	52	0.003	0.922	52	0,002
Body	0.183	52	0.000	0.893	52	0.000

(a) Lilliefors Significance Correction

4.2.2 Head orientation

Stiefelhagen and colleagues [20] stated that the potential of head orientation detection, i.e. "whom is looking at whom" in an around-the-table setting could provide valuable human interaction information. Using an extension of the same multi-person tracker used to provide positional information, we were able to annotate the head-orientation of each participant in real-time. From this, we were able to make an estimate of each person's focus of attention.

The data output from the 3D head-tracking module consisted of a timestamp together with pan and tilt information. Head orientation was subsequently categorised into four cardinal values North, South, East or West together with a flag for 'looking down' (i.e. when a subject's head was oriented towards the table). As for body and head fidgeting, we extracted the fidgeting values for each participant and normalized them. Globally, the statistical mean of head orientation and fidgeting is 0.04 and its standard deviation is 0.03. Also in this case we cannot assume the distribution normality (see Table 8).

Table 8. Head Orientation: Tests of Normality

Kolmogorov-Smirnov (a)			Shapiro-Wilk		
Statistic	df	Sig.	Statistic	df	Sig.
0.161	52	0.002	0.906	52	0.01

(a) Lilliefors Significance Correction

5. SOME PRELIMINARY ANALYSES

One of the goals of the data collection was to amass a large corpus of meetings in which people show some personality traits and social behaviors while discussing and interacting. This data is then to be used to investigate the possibility of developing an automatic system able to detect personality traits and social behaviors based on audio-visual features.

Given this goal, we have started to perform some correlation analyses between our measures (personality traits, group and individual performance, group cohesion score) and our set of audio-visual features (speech rate, pitch, energy, head fidgeting, body fidgeting and hands fidgeting) by using Z-standard scores (mean=0, standard deviation=1).

From this preliminary analysis it becomes evident that in our corpus there is a significant positive relationship between extraversion and the two audio features, pitch and energy. Furthermore, the group performance is inversely related to fidgeting activity and head orientation fluctuations. That means that groups with a better performance in the Survival Task tend to be composed by subjects with low values of fidgeting energy and stable head orientation. On the contrary, both for LoC, Group Cohesion and Individual Performance it seems that there is no significant correlation with the audio-visual features.

The next steps of our work will be to perform more analyses to investigate the relationships between our measures and the audio-visual features. In particular, we plan to conduct linear and non-linear regression analyses using personality traits and social behaviors measures as dependent variables. To do that, we are going to use Support Vector Machines (SVMs) because they offer some computational and performance advantages. In particular, SVMs (i) represent a solution by means of only a small subset of training points; (ii) ensure the existence of a global minimum; (iii) and the optimization of a reliable generalization bound [19].

6. CONCLUSION AND FUTURE WORKS

In this paper we have presented a multimodal corpus of multi-party meetings, automatically annotated with audio-visual cues.

The corpus consists of audio and video recordings of thirteen meetings, for a total length of almost 7 hours. The collected data has been enriched by adding some automatic annotations concerning speech activity and the 3D tracking of body activity.

One of the goals of the data collection was to collect a large corpus of meetings, in which people show some personality traits and social behaviors while discussing and interacting.

This corpus could be helpful to improve our understanding in the multiple social, psychological and emotional aspects involved in multi-party meetings. At the same time the corpus could be useful to develop systems that, starting from audio-visual cues, automatically analyze social behaviors and predict personality traits of participating in multi-party meetings.

According to this aim, we plan to conduct further analyses to investigate more deeply relationships between our measures and the rich audio-visual features.

7. ACKNOWLEDGMENTS

The data collection described in this paper has been carried out within the CHIL project (http://chil.server.de/). We wish to thank all participants in the data collection. A special thanks goes to Alberto Battocchi and Franca Rossi for their cooperation.

8. REFERENCES

[1] AMI, Augmented Multi-party Interaction, (http://www.amiproject.org/).

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

[3] Bollen, K. A. and Hoyle, R. H. Perceived Cohesion: a Conceptual and Empirical Examination," *Social Forces*,69,2, (1990), pp. 479—504.

[4] Carli, G., Gretter, G. A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End based on DSP32C VME Board. In *Proc. ICSPAT*, (1992).

[5] Chen, L., Rose, R.T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D.,Tuttle, R., Huang, T.: VACE multimodal meeting corpus. *Proc. of Multimodal Interaction and Related Machine Learning Algorithms.* (2005).

[6] CHIL, Computers In the Human Interaction Loop, (http://chil.server.de/).

[7] Farma, T. and Cortivonis I. 2000. '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.

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

[9] James W. Davis: 'Hierarchical Motion History Images for Recognizing Human Motion', IEEE Workshop on Detection and Recognition of Events in Video, pp. 39, 2001

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

[11] McCowan, D. Gatica-Perez, S. Bengio, D. Moore and H. Bourlard. Towards Computer Understanding of Human Interactions. In: Ambient Intelligence, E. Aarts, R. Collier, E. van Loenen & B. de Ruyter (eds.), *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, pp. 235-251. (2004)

[12] Perugini, M. and Di Blas, L. 2002 Analyzing Personality-related Adjectives from an Eticemic Perspective: The Big Five marker Scales (BFMS) and the Italian AB5C Taxonomy. In B. De Raad and M. Perugini (eds.) Big Five Assessment, Hogrefe und Huber Publishers, Göttingen, 281-304.

[13] Pianesi F., Zancanaro M., Lepri B., Cappelletti A. *in press* Multimodal Annotated Corpora of Consensus Decision Making Meetings. To appear in *The Journal of Language Resources and Evaluation*

[14] Rienks, R., Zhang D., Gatica-Perez D., Post W. Detection and Application of Influence Rankings in Small Group Meetings. In Proceedings of ICMI'06. Banff, CA. (2006)

[15] Rosenfeld, L. B., & Gilbert,J. R.. The measurement of cohesion and its relationship to dimensions of self-disclosure in classroom settings. *Small Group Behavior*, 20, (1989), 291-301

[16] Rotter, J.B. 1966. Generalized Expectancies for Internal versus External Control of Reinforcement. Psychological Monographs, 80 (1, Whole N.609).

[17] Sjölander, K., and Beskow, J. 2000. WaveSurfer- an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proc of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 282-285). Beijing.

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

[19] Smola, A. and Schölkopf, B. A Tutorial on Support Vector Regression. Statistics and Computing, 2003.

[20] Stiefelhagen, R., Zhan, J., Waibel, A. Modeling focus of attention for meeting indexing. In *CHI '02 extended abstracts on Human factors in computing systems* (2002).

[21] Stokes, J. Towards an understanding of cohesion in personal change groups. International. *Journal of Group Psychotherapy*, 33, (1983), 449-467.

[22] The American Heritage Dictionary of English Language

[23] Wiggins, J. S. 1996 *The Five-Factor model of Personality*. New York, NY: Guilford Press.

[24] Zancanaro M., Lepri B. and Pianesi F., 'Automatic detection of group functional roles in face to face interactions'. In proceedings ICMI 2006, pp. 28-34, 2006.