Estimation of Cohesion with Feature Categorization on Small Scale Groups

Sheng FANG
Institut des Systèmes Intelligents et de Robotique
UPMC
fang@isir.upmc.fr

Catherine ACHARD
Institut des Systèmes Intelligents et de Robotique
UPMC
catherine.achard@upmc.fr

ABSTRACT

The estimation of cohesion between participants of a small scale meeting is one of the challenging applications of social signal processing. We focus on 2 kinds of relation of cohesion during the interaction, which are: the relation between cohesion of meetings and personality of participants, as well as the relation between cohesion and social behaviours of participants. First of all, we collect the annotations of the cohesion by crowdsourcing. Then, we analyze the correlation between cohesion of meeting and the personality of participants. Finally, we predict the cohesion by using the social behaviours of participants. During the experiments, we find that the Big Five Personalty Trait, Agreeableness, is highly correlated to cohesion compared to other personality traits. Moreover, the social signals, the speech turn and variation of speech energy, are related to the meeting cohesion.

Keywords

SSP, Cohesion, Personality

1. INTRODUCTION

With the development of speech and image processing and the expansion of databases for human to human interactions, there is an increasing interest on computational sociology. In the survey [28], Vinciarelli et al. introduce several applications of Social Signal Processing (SSP), such as role [9, 11] and personality [12] recognition. Both the roles and the personality traits are personal level features, which reveal information on participants. In this paper, we are interesting in group cohesion estimation that is a challenging task as it depends on all participants and their interaction. Moreover, there is today no clear definition of group cohesion that can also be considered as the quality of the interaction. Actually, cohesion between the participants is an important and necessary element for collaboration. One psychological research , one popular definition of cohesion can be found in [4], which indicates that *"Cohesion is now generally considered as the group members' inclinations to forge social bonds, resulting in the group sticking together and remaining united."* Cohesion is widely studied by the psychologists, because it helps to better complete the task, such as achieving better performance of a sport team and making better financial benefits by the strategy decisions of a business meeting. One recent psychological study [22] reveals that *"Cohesion demonstrates more significant relationships with performance when conceptualised using social and task dimensions (but not other) and when analyses are performed at the team level"*. Thus, in this work, the annotation of cohesion consider these two aspects. This paper focuses on using the techniques of SSP to solve the psychological problem about the comprehension of cohesion level in small group interaction. The goal is to understand the elements that leads to a good cohesion during group interactions. We work on two aspects. In the first part, we study the personality traits of the participants to find out if some personality traits contribute positively or negatively to the cohesion during interaction. The second part of this paper is related to cohesion prediction, in which we extract audio and visual social signals and apply the same feature categorization that mentioned in [12].

The rest of this paper is organized as follows. Section 2 presents the state of the art on social signal processing work and psychological research of cohesion, and the database we choose to perform the cohesion study. Then, Section 3 gives information about the online questionnaire we create to annotate the meeting and presents several analysis on the annotations. Section 4 analyzes the relation between Big Five Personality Traits and cohesion. Section 5 explains how we predict cohesion by using social signals and finally, Section 6 draws the conclusions.

2. STATE OF THE ART

Cohesion was firstly studied in psychological research, then it becomes a topic of computational research domain, as SSP. This section gives a brief introduction to the state of the art on psychological research of cohesion and its transposition to the computational field. Then, two databases of small scale group meetings are discussed.

2.1 Psychological Research

For the study of cohesion, there is a long history in psychological research. At the beginning of research on cohesion, most of the social scientists concentrated their effort on the causes of the cohesion, which leads to the difficult task of *a priori* predict how the group interaction will be. For instance, Carron and Brawley [3] define cohesion as *"the individual's motivation to remain in the group"* and *"the individual's perceptions about what the group believes about its closeness"*. These two dimensions about cohesion are difficult to capture. Moreover, they strongly depend on the prior experiences of the individuals. The study of the cause of cohesion results in research focusing on individuals, while the experience of each individuals are hard to define and measure.

Nowadays, psychologists pay more and more attention on the consequences of cohesion and psychologists now treat cohesion as a group phenomenon. For example, the work [4, 21] reveal that group cohesion is correlated to group performance. From the aspect of the consequences of the cohesion, the early theory of cohesion proposed to measure the attraction of the group to the individuals [25]. This measure considers the cohesion as the sum of the attraction of group to every individual. However, there is no concept to treat the group as an entity. Later, Evans and Jarvis [10] proposed a measure that represents the aspect of group level decision about the group goal. Then more and more measures related to the group are proposed to be added to the measures of cohesion. For example, Forsyth D.R. [14] considers that cohesion can be broken down into 4 main components which are Social Relations (SR), Task Relations (TR), Perceived Unity (PU), and EMotions (EM). SR and TR are the measures about social aspect and task aspect. PU measures the perceptions of individuals. Emotion represents if the individuals enjoy the process of the group interactions. These 4 components are used in Section 3 to guide the design of a questionnaire to annotate group cohesion.

2.2 SSP Study

As far as we know, the only work about cohesion estimation in task-oriented small scale group was done by Hung et al. [18]. However, there are several other work on similar group level features, for instance, creativity of group [31], rapport of interaction [5, 17], success of learning interaction [30], interest-level of group [16] or quality of Spoken Dialogue Systems [24, 26, 27]. For creativity, Won et al. [31] recorded 52 dyadic interactions by Kinects, who capture 3-Dimension features. Several nonverbal features, such as head angle and right/left arm angle, are extracted. Then they created a measure of synchrony by correlating movements between the members during the dyadic interactions. Finally the authors used synchrony scores as input of 3 machine learning methods, Logistic Regression, Multilayer Perception and Decision Tree, to classify high/low level of creativity. They achieved 86.7% prediction accuracy and discussed implications for methodological approaches to measure nonverbal behaviours and synchrony. Won et al. [30] used a similar approach to predict the success of learning during the interaction between student and teacher. The prediction accuracy achieves 85.7%.

In the research on group interaction, Huang et al. [17] carried out experiments on conversation between a human speaker and a virtual human listener. Two kinds of virtual agents were designed, the responsive agent and the unresponsive one. The responsive agent was designed to perform nonverbal behaviours corresponding to positive feedback and create an impression of active listener. The results show that interaction with a responsive agent creates much stronger feelings of rapport between users and virtual agent. Cassell et al. [5] focused on the differences in the use of dialogue acts and non-verbal behaviours of virtual agent. They found a pattern of verbal and nonverbal behaviours that differentiates the dialogue of friends from that of strangers.

Gatica-Perez et al. [16] detected the interest-level of group. Firstly, they got the annotations of interest-level from external observers. Then the nonverbal audio and visual features were extracted to be integrated into Hidden Markov Model (HMM). Two HMMs were trained for audio and visual features independently to predict the interest-level of group.

Hung et al. [18] estimated the cohesion of group by using nonverbal audio and visual features. Firstly, annotators answered the questionnaire about cohesion for 100 meeting segments. However, only 61 segments, whose annotations are in high agreement, are used in the experiments. Then three types of nonverbal features, audio features, video features and audio-video features, were extracted. All these features were integrated into two classifiers, naïve mean value based classifier and Support Vector Machine (SVM). Each meeting segment was classified into high or low cohesion by using only one social signal. The social signal with best performance is total pause time.

2.3 Database

Cohesion was once studied in computational science in the Augmented Multiparty Interaction (AMI) Corpus [20]. In the scenario of AMI meeting, 4 participants have to design a remote control. Each participant has a different role to play: project manager, marketing expert, user interface designer and industrial designer. The cohesion of meetings have been annotated by external annotators [18]. The main drawback of this database is that roles are played. The roles have strong influence on the way the participants act, while the role players are students who have no real experiences. This drawback results in non-natural performances and reactions during interactions.

So, we prefer using the ELEA [23] corpus. In ELEA, each group composed of 3 or 4 persons, performs a winter survival task. Suppose the participants are survivors after an airplane crash landing, they have to sort the 12 objects from most important to least one. At first, each person thinks alone and gives their own decision. Then, the group members discuss together to generate a final ranking list. The cohesion study on ELEA is based on the discussion process. Although there is a virtual scenario set for ELEA database, the most important interaction part is to discuss the rank of importance of 12 objects. We believe the interactions in ELEA are more natural than the ones of AMI. Moreover, each participant in ELEA is free to interact with other. During the discussion, group members sit at two face-to-face sides of a square table. There is 1 microcone and 2 webcams placed in the middle of the table. Each webcam captures the video of one side of the table. Audios and videos are synchronized. The corpus ELEA contains 40 meetings, 27 of them are well recorded with both audio and video information, while the rest 13 get only audio information. Our experiments are performed in a subset containing the 27 meetings with both audios and videos. ELEA database has a rich set of annotations, including the age and gender of participants, the ranking lists of groups and individuals, etc. Moreover, before the Winter Survival Task, all participants are asked to take a test to measure their personality traits, named self-reported personality traits. For each participant, we have the 5 personality values of OCEAN [19]. Each personality value ranges from 1 to 5.

A major drawback of ELEA is that there is no annotation of cohesion between participants. So we design a questionnaire to create the annotations of cohesion.

3. ANNOTATIONS OF COHESION

Among the 27 meetings with both audios and videos, we only get annotations of cohesion for 19 meetings because of the synchronization issues and bad quality of videos for other 8 meetings.

Ambady et al. show in [1] that external observers can perceive the social interaction by just observing short slices of the interaction. So we divide the 19 synchronized meetings into small slices. Each slice lasts two minutes. The segmentation facilitates the annotations process of ELEA database (annotators have now to watch 2 mn of video instead of 15 mn) and generates more data. The synchronization between audio and video, as well as the division of meeting into slices, are done by using the toolbox FFmpeg [13]. We obtained 115 video slices, where 76 slices are in English and

39 slices are in French. The on-line questionnaire is done based on the 115 videos.

3.1 Design of Online Questionnaire

In this work, we study the cohesion of a small group of participants engaged in a task, which requires the collaboration between all participants. The cohesion between the participants can be reflected by their performance. So the questionnaire is designed to ask the external observers about their perceptions of the meeting as well as the participants' behaviours. The questions used in psychology research [2] and previous computational work about cohesion [18] are pooled together for the design of our questionnaire. First of all, we delete the questions which are already proved to be useless. Then we delete the questions which are not directly related to the 4 components as introduced in [14], e.g., Social Relations (SR), Task Relations (TR), Perceived Unity (PU) and EMotions (EM). Finally, 13 questions are designed in our questionnaire. They are shown in Table 1.

Table 1: Full view of questionnaire with relation between questions and Social Relations (SR), Task Relations (TR), Perceived Unity (PU) and EMotions (EM)

Questions	SR	TR	PU	EM
Do you think all members contribute to the task? (Yes/No)		√		
Overall, does the group enjoy the task? (Not enjoy at all/enjoy)		√		√
Overall, the collaborations between all the participants to perform the task are (Bad/Good)		√	√	
Overall, does the atmosphere of the group seem more jovial or serious?(Jovial/Serious)	√			√
Does all the team members have sufficient time to defend their opinions? (Yes/No)		√	√	
Do all group members enjoy each other's company? (Yes/No)	√		√	√
How involved/engaged in the discussion do the participants seem? (Not evolved/completely involved)	√		√	
Overall, does the work group operate spontaneously. (Yes/No)	√		√	
Overall, do all group members listen attentively to each other. (Yes/No)	√		√	
Overall, are there good interactions between all the participants? (Good/Not)	√		√	√
Does the meeting progresses efficiently. (Yes/No)		√		
Overall, does each participant of the group have an equivalent participation rate. (Yes/No)	√		√	
Do the participants of the group appear to have a strong bond/relationship? (Yes/No)	√		√	

We apply online-questionnaire on the website *www.crowdflower.com*. Although the answers of some questions are binary, the answers are set in 7-point scales to give the annotators more choices. To avoid the bias during the selection of the answers of annotators, the order of answers are flipped for some of the questions (from the answer representing good cohesion to the one with bad or in the opposite order).

As some annotators may respond arbitrarily to go faster, the design of questionnaire in platform *CrowdFlower* requires the annotators to finish watching the entire video slices before making the annotation. Moreover, the platform proposes to add test questions to check the responses quality. Thus, to detect arbitrary responses, we add 2 questions which are respectively similar to 2 initial questions :

Original designed questions:
1) *Overall, the atmosphere of the meeting is tense or relaxed?*
2) *Do the participants appear comfortable or uncomfortable with each other*?

Added questions:
1) *Overall, does the atmosphere of the group seem more jovial or serious?*
2) *Do all group members enjoy each other's company?*

Let us recall that the answer to each question is seven-point scales. We impose the answers to paired questions to be consistent: they should be either both positive or both negative. If it is not the case, we consider the answers as not pertinent and remove all the answers of the concerned annotator.

3.2 Measure of Cohesion for Video Slice

The cohesion of each video slice is derived from the answers of the 5 annotators. Each annotator gives answers to the 15 questions presented above. To generate the general cohesion from the answers of the 5 annotators, there are 2 problems to solve.

1) How to generate a cohesion value for one annotator? We name this value as **annotator-dependent cohesion**. It results from the responses to the 15 questions for the concerned slice and annotator.

2) How to generate a cohesion value for one video slice? We name this value as **cohesion of each video slice**. It results from the 5 annotator-dependent cohesion values.

There are 2 solutions to calculate the annotator-dependent cohesion value. The first solution, which is intuitive, consists in computing the average value of answers of the 15 questions. In the second solution, we consider that some questions may be correlated and that the correlation biases the mean value. So we apply a Principal Component Analysis [29] and keep only values on the first dimension (the first dimension keeps 34% of the inertia).

Once the annotator-dependent cohesion values estimated, the cohesion of each video slice is computed as the average of the 5 annotator-dependent cohesion values.

Figure 1 shows the cohesion of the 115 video slices when annotator-dependent cohesion values are estimated using the average value or the PCA. The 2 solutions are very close each other, as confirmed by their correlation coefficient:

$$R(Cohesion_{PCA}, Cohesion_{Mean}) = 0.997$$

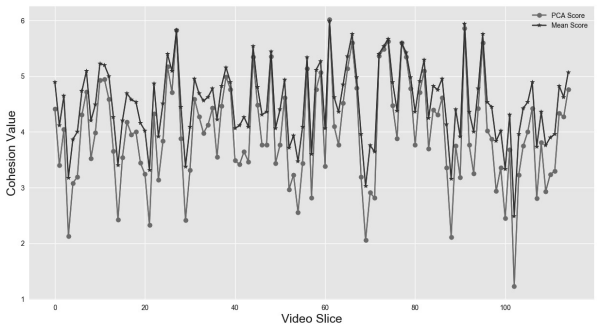

Figure 1: Mean cohesion values and PCA cohesion values for the 115 video slices

In the following, we choose to compute the annotator-dependent values as the average of the 15 answers for each annotator.

3.3 Agreement Analysis

To recall the setup of the online questionnaire, there are 5 annotators for each 2-min video slice. Each annotator gives answers to the 15 questions. The cohesion of the video slice is defined as the average of the 15 questions of the 5 annotators.

We apply the agreement analysis between annotators by using the metric of Weighted Cohen's Kappa (WCK) [7] that is an extension of Cohen's Kappa [6] for non binary responses. WCK agreement is a metric designed to measure the inter-rater agreement between 2 raters. It varies between -1 and 1, which means no agreement with negative values and a perfect agreement with a value close to 1. The agreement between the 5 annotators is calculated as the average value of agreement for all possible pairs of annotators. Figure 2 shows the distribution of WCK agreement and cohesion values. Each point of the Figure 2 represents a video slice.

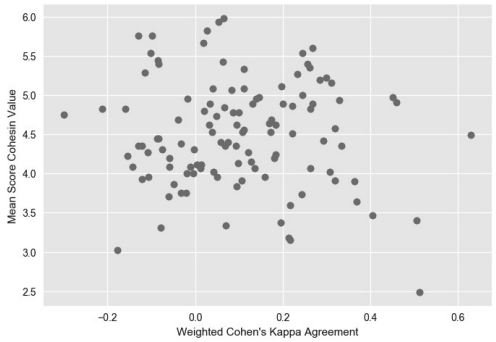

Figure 2: Distribution between Agreement and Cohesion Value

The points of the Figure 2 have not a particular shape and thus, no correlation exist between agreement and cohesion values. Moreover, there are more video slices with good cohesion than the ones with bad cohesion (neutral cohesion value is 4 as responses varies between 1 and 7). This seems logical as all the participants are volunteers to participate to the experiments.

3.4 Measure of Cohesion for Whole Meeting

As detailed before, the 15 minutes meeting videos are cut into 2 minutes video slices. This division improves the quality of online questionnaire, since the annotators need great patience to finish a 15-minute long video and answer the questions. Short video makes the work of annotators easier. Meanwhile, we do not collect the cohesion measure of the whole meeting. As far as we know, there is no research on the relation between the cohesion of meeting and the cohesion of video slices, which compose the meeting. Thus, we propose to define several meeting-level cohesion measures using the cohesion values of video slices as below:

1. Mean cohesion (MeanCo) is the average of the cohesion values of all video slices belonging to a same meeting. For instance, for a meeting divided into n video slices, whose cohesion values are $\{c_1, c_2, \ldots, c_n\}$. The MeanCo value is:

$$MeanCo = \frac{\sum_{i=1}^{n} c_i}{n} \quad (1)$$

2. Weighted mean cohesion value (WMeanCO) is estimated using the agreement values to weight video slice cohesion. Let us note $\{k_1, k_1, \ldots, k_n\}$, the agreement level of video slices. As negative agreement values correspond to no agreement, only positive values are used to compute the weighted mean value WMeanCO:

$$WMeanCO = \frac{\sum_{i|k_i>0} k_i * c_i}{\sum_{i|k_i>0} k_i} \quad (2)$$

3. Minimum cohesion (MinCO) value that measures the worst cohesion of all the video slices of a same meeting.

4. Maximum cohesion (MaxCO) value that measures the best cohesion of all the video slices of a same meeting.

We analyze these 4 measures in the following sections.

4. RELATION BETWEEN COHESION AND PERSONALITY TRAITS

In this section, we analyze the relation between cohesion and personality traits. As personality traits are defined for each participant and as the number of participant varies according to the meeting, it is difficult to keep all personality traits of all participants. So we calculate 4 statistics, e.g., minimum (min), maximum (max), standard deviation (std), and average value (mean), for each personality trait between all the participants in a same meeting. Moreover, we use the Big Five Personality traits annotations *OCEAN* [19] to describe the personality. This 5 personality traits are Openness to experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N).

Table 2 shows correlation coefficients between the 19 meeting level cohesion measure and the statistics of personality traits as previously defined. Correlation coefficients whose absolute value are larger than 0.3 are marked in bold and red.

Table 2: Correlation coefficient between personality and cohesion

	MeanCo	WMeanCo	MinCo	MaxCo
O_{std}	-0.05	-0.16	0.02	-0.27
O_{mean}	-0.17	**-0.36**	-0.20	-0.17
O_{min}	-0.25	**-0.30**	-0.22	-0.16
O_{max}	**-0.31**	**-0.45***	-0.23	**-0.41***
C_{std}	0.16	0.16	0.08	**0.32**
C_{mean}	0.09	-0.008	-0.08	-0.001
C_{min}	-0.06	-0.11	-0.20	-0.17
C_{max}	0.15	0.089	-0.11	0.17
E_{std}	-0.15	-0.27	-0.28	-0.08
E_{mean}	**0.37**	**0.30**	**0.39***	0.18
E_{min}	**0.31**	**0.33**	**0.37**	0.12
E_{max}	0.19	0.07	0.10	0.03
A_{std}	0.25	0.29	**0.35**	0.08
A_{mean}	**0.41***	0.17	0.24	**0.48****
A_{min}	0.11	-0.07	-0.06	0.28
A_{max}	**0.55****	**0.36**	**0.47****	**0.50****
N_{std}	0.11	0.17	**0.35**	-0.03
N_{mean}	0.02	-0.13	0.22	0.09
N_{min}	0.02	-0.27	0.04	0.07
N_{max}	0.08	0.03	0.33	0.03

O, C, E, A and N are abbreviations for Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. 'std', 'mean', 'min', 'max' refers to standard deviation, mean, minimum and maximum. MeanCo, WMeanCo, MinCo and MaxCo are the Mean, Weighted mean, minimum and maximum cohesion values among each meeting.
$**$ means the correlation coefficient has a 95% confidence level and $*$ refers to the correlation coefficient has a 90% confidence level

Several conclusions can be drawn from values in Table 2:

(1) Conscientiousness and Neuroticism personality traits have a small influence on meeting level cohesion, since most of the correlation coefficients are small.

(2) Openness to experience is negatively correlated with meeting level cohesion. As the definition of Openness is *"Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty. High openness can be perceived as unpredictability or lack of focus"*, it is natural that group cohesion is negatively influenced by openness in these task oriented meetings.

(3) The Extraversion is positively correlated with group cohesion. The definition of Extraversion is *"Extraversion frequently associates with sociable, assertive, talkative, active, energy, positive emotions, and surgency. A person with high Extraversion is often considered to be attention-seeking, and domineering."* So the extraversive person can easily contributes to the meeting level cohesion.

(4) Agreeableness is positively correlated with cohesion as expected. Actually, agreeableness is defined as friendliness and social conformity.

To conclude, group cohesion mainly depends on 3 personality traits of people involved the group for a task oriented meeting. Ideally, the participants should present high Agreeableness and Extraversion, but low Openness to experience. The results have been obtained from 19 meetings. This number is sufficient to estimate correlation values but does not allow to learn a prediction model from all personality traits.

5. RELATION BETWEEN COHESION AND SOCIAL SIGNALS

In this section, we first introduce the social signals we extracted. Then, we analyze the correlation between social signals and cohesion. Finally, we perform the experiments of cohesion estimation and analyze the experimental results.

5.1 Social Signals

Social signals extracted from audios and videos are group in 2 categories: Intra-Personal and One_VS_All features as proposed in [12]. Intra-Personal features contain information about only one participant, even if interacting with others. One_VS_All features encode the interactions between a participant and the rest of group. The goal is to identify if one of this set of features is more relevant to predict cohesion values.

5.1.1 Intra-Personal Social Signals

The social signals include features based on speech activity, prosody, and visual activity. We extract the following features from speech activity:

- Speech Length (SL_t): number of frames during which the target participant p_t speaks.
- Speech Turn (ST_t): number of times the speech state of p_t changes from 'not speaking' to 'speaking' (we do not consider natural stops when a participant is talking; in other words, we ignore short silences whose duration is less than 1 second).
- Mean Length of Speech Turns ($MLST_t$): $MLST_t$ is calculated as SL_t/ST_t.

We extract one prosodic social signal:

- Standard deviation of energy ($Estd_t$): the standard deviation of energy of participant p_t.

The visual activity indicates if the target participant is moving or not. We extract visual activity features of head/hand areas and get the energy of motion of each frame by using Motion Energy Image (MEI) [8]. A simple definition of motion energy is the proportion of moving pixels. Because the distance from participants to the camera are not the same, we normalize the motion energy by the area of the face. The visual activity features are defined as follows:

- Head Motion Energy (HME_t): motion energy of head for participant p_t divided by the area of the face.
- Body Motion Energy (BME_t): motion energy of body for participant p_t divided by the area of the face.

5.1.2 One_VS_All Social Signals

One_VS_All features encode the general relationships between target participant p_t and the rest of the group. They are defined as:

- Successful Speech Interrupt (SSI_t): number of times p_t interrupts others.
- unSuccessful Speech Interrupt ($uSSI_t$): number of times p_t unsuccessfully interrupts others.
- Speech Back Channel (SBC_t): back channel measures the turn taking pattern like p_t-p_i-p_t. When participant p_t stops talking, participant p_i takes the talking turn. Then participant p_t takes back the talking turn when p_i stops. We interpret this pattern as an interaction between p_t and p_i.
- Silence (SLC_t): It measures the silence time after the speech turn. When participant p_t stops his speech turn and there is no active speech after the stop.
- Total Speech Conflict (TSC): number of frames during which there are more than 1 participant speaking at the same time.

All the social signals are summarized in the Table 3. Between these 11 features, the first 10 are extracted for a particular participant p_t. So, we use 4 statistics (minimum, maximum, mean and standard deviation) to represent group-level features. The total number of features is 41, including 24 Intra-Personal social signal based features and 17 One_VS_All social signal based features as follows.

Table 3: Social signals used to predict cohesion

	Name	Description
INTRA-PERSONAL	SL_t	Speech length
	ST_t	Speech turn
	$MLST_t$	Mean length of speech turns
	$Estd_t$	Standard deviation of energy
	HME_t	Head motion energy
	BME_t	Body motion energy
ONE_VS_ALL	SSI_t	Total successful speech interruptions of p_t
	$uSSI_t$	Total unsuccessful speech interruptions of p_t
	$TSBC_t$	Total speech back channel of p_t
	SLC_t	Silence time of p_t
	TSC	Total speech conflict during interaction

5.2 Analysis of the Relation between Social Signals and Cohesion

Table 4 shows Pearson correlation coefficients between statistics of social signals and cohesion values of video slices. Correlation coefficients with absolute values are larger than 0.3, are marked in red and bold and those with absolute values between 0.2 and 0.3 are marked in blue and bold.

Table 4: Correlation coefficients between statistics of social cues and cohesion

	min	max	mean	std
SL_t	0.12	0.11	**0.30****	0.01
ST_t	0.19	**0.31****	**0.27****	0.09
$MLST_t$	0.07	**-0.21***	**-0.20***	**-0.22****
$Estd_t$	**0.35****	**0.33****	**0.35****	-0.04
HME_t	**0.30****	0.10	-0.05	0.18
BME_t	-0.06	-0.06	-0.06	-0.004
SSI_t	**0.34****	**0.32****	**0.33****	0.09
$uSSI_t$	0.16	0.13	**0.25***	0.08
$TSBC_t$	-0.007	0.08	0.03	0.11
SLC_t	-0.15	**-0.21****	**-0.21****	**-0.21****
TSC	**0.37****			

The first column lists all the social signals defined in Table 3. The row represent the 4 statistics among participants: standard deviation, mean, minimum and maximum. Values in the table are the correlation coefficients between the statistics of social signals and the cohesion of each video slice. $**$ means the correlation coefficient has a 95% confidence level and $*$ means the correlation coefficient has a 90% confidence level

The statistics of $Estd$, SSI, ST and TSC are more correlated to cohesion than other social signals. According to the correlation analysis, we can observe that:

- Good cohesion happens in the meeting, where the speech energy of each participant is variable ($Estd$).
- Good cohesion happens in the meeting with lots of successful interrupts (SSI).
- Good cohesion happens in the meeting with lots of speaking turn (ST).
- Good cohesion happens in the meeting with good level of speech conflict (TSC).

On the opposite, some social signals are negatively correlated with cohesion:

- Good cohesion happens in the meeting, where the mean length of speaking turn is small ($MLST$).
- Good cohesion happens in the meeting with little silence (SLC).

5.3 Prediction of Cohesion by Social Signals

5.3.1 Classification Model

We treat the problem of cohesion prediction as a classification problem and divide all video slices into the ones with higher cohesion and the ones with lower cohesion. The high/low categories are clustered by the median of cohesion values of all video slices. We use Ridge Regression (RR) [15] to model cohesion values by minimizing:

$$E = \sum_{j=1}^{N} (c_j - \mathbf{x}_j^T \boldsymbol{\beta})^2 + k \parallel \boldsymbol{\beta} \parallel^2 \quad (3)$$

where c_j is the cohesion value predicted using the $n = 41$ features of example $\mathbf{x}_j = [f_{j,1}, \ldots, f_{j,n}]^T$, N is the number of examples and k is the ridge parameter, which improves the conditioning of the problem and reduces the variance of the estimates. The solution of Equation 3 is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad (4)$$

where $\mathbf{X}$ is the matrix obtained by concatenating feature vectors.

In this article, and as in [18], we transfer Ridge Regression (RR) [15] to a classification model: all the original cohesion values are first clustered by their median value to get new binary labels. Values larger than the median value are labelled 1 and the smaller ones are marked as 0. The RR is trained with these binary values. During the prediction process, the predicted values are compared to 0.5 to generate the final binary label. This adjusted model is named *RR-0/1*.

As 41 features are used to train the predictive model, the 115 meeting video slices are not sufficient to obtain good capacity in generalization. Thus, we propose to apply a feature selection algorithm and more particularly, a wrapper-based method, which uses a heuristic search to find the feature subset $S_{selected}$ generating the best prediction accuracy.

Let us note $S_{all} = \{f_i\}$, $i = 1, \ldots, n$ the whole feature set, where $n = 41$ is the total number of features. At the beginning, $S_{selected}$ is empty and every single feature is used independently to make the prediction. The one with the best prediction accuracy is selected and added to $S_{selected}$. Other features are added one by one to $S_{selected}$. At each iteration, we first rank the remaining features according to their performance in prediction (by adding them to $S_{selected}$) and then select the feature with the best accuracy. The process stops when there is no more remaining feature. At the end of feature selection algorithm, the subset associated to the highest accuracy is selected.

The process of feature selection is summarized in Algorithm 1.

Data: Whole feature set $S_{all} = \{f_i | i = 1, 2, \ldots, n\}$
Result: Subset of features
Select the first feature who has the best prediction performance, add this feature to $S^1_{selected}$
$n_f = 1$
while $(n_f < n)$ **do**
  Construct $(n - n_f)$ subsets composed of $S^{n_f}_{selected}$ and one remaining feature, then calculate the accuracy of the corresponding subsets
  Sort the subsets according to their accuracies in a descending order and select the first subset as $S^{n_f+1}_{selected}$
  $n_f = n_f + 1$
end
Select the subsets $S^{n_f}_{selected}$ with highest prediction accuracy.

Algorithm 1: Feature selection algorithm

5.3.2 Experimental Setup

The quality of data plays an important role in the experiments. So we propose to remove data with low agreement level between annotators. Meanwhile, the quantity of data also influences the performance of machine learning model. In order to keep balance between quantity and quality of data, we set a threshold as *0* to filter the video slices by agreement level. As mentioned before, negative agreement level means no agreement, so we only keep the video slices with positive agreement values, i.e., 81 video slices of 115 are chosen. Moreover, the leave-one-out cross validation is applied to optimize the hyper parameter k (equation 3) of *RR-0/1*. The average accuracy during leave-one-out cross validation is considered as the performance of cohesion prediction. The hyper parameter k of ridge regression is selected in $\{2^i | i = -8, -7, \ldots, 7, 8\}$.

5.3.3 Prediction without feature selection

In this section, we perform the cohesion estimation with the Intra-Personal based group features, One_VS_All based group features and all features, without the feature selection algorithm. Results, presented in Table 5, show that both Intra-Personal and One_

VS_All features lead to poor accuracies. Moreover, the prediction accuracy by using the combination of all features is lower than the one by using any simple feature category.

Table 5: Prediction performance with all features

Features set used	Prediction Accuracy
Intra-personal based features	61.73%
One_VS_All based features	60.49%
all features	58.02%

The bad performance and the fall of accuracy with all features (Intra-Personal and One_VS_All features) may be caused by over-fitting or under-fitting. To have a clear understanding of the 2 phenomena, we study the learning curves and plot the performances (accuracy) according to the number of training data on training and testing data sets.

As mentioned in Section 5.3.2, the *RR-0/1* model contains a hyper parameter k that is optimized in $\{2^i | i = -8, -7, \ldots, 7, 8\}$. We calculate the accuracy several times with different values for parameter k and compute the average value and standard deviation of these accuracies.

Figure 3: Learning curves by using Intra-personal based features

Figure 3 shows the plot of learning curves by using Intra-Personal based features. Training and testing curves show low convergent tendency. Moreover, the prediction accuracy on training data is quite good with values higher than 80%; while the one on testing data is always low. There is a great gap between testing and training curve. These evidences prove that classification with Intra-personal features suffer from over-fitting. The same phenomenon appears using all features.

There are several solutions to improve the prediction accuracy, e.g. get more data or apply feature selection algorithm. Because the ELEA database was collected in a room well designed, it's hard to reconstruct the room for getting more data. In this work, we solve the over-fitting problem by feature selection.

5.3.4 *Prediction with Feature Selection*

In this section, we estimate the cohesion using Intra-Personal and One_VS_All features by applying the feature selection process introduced in algorithm 1. Figure 4 shows the evolution of prediction accuracy according to the number of selected features.

The best prediction is achieved by using only 3 features and it is interesting to note that all the 3 selected features are Intra-Personal based group features : ST_{std} (standard deviation of speaking turn), BME_{mean} (mean of body motion energy) and $Estd_{max}$ (maximal value of the standard deviation of energy). This reveals that One_VS_All group based features are less efficient to

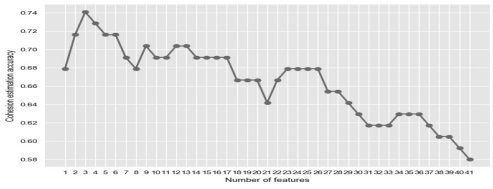

Figure 4: Accuracy of cohesion prediction vs the number of features with All based group features

estimate cohesion in our experiments. The same conclusion was drawn in [12] for a personality traits recognition application.

Table 6 shows the results of prediction accuracy by using Intra-Personal and One_VS_All features. It's easy to notice that the prediction accuracy is greatly improved with the feature selection process.

Table 6: Prediction with and without feature selection

	Prediction Accuracy
without features selection	58.02%
with features selection	74.41%

Figure 5 shows the learning curves by using the selected features. We perform the same study on the learning curves about the convergence and performance level. Both curves converge towards the same point corresponding to the accuracy of 75%. Thus, the over-fitting problem does not appear using feature selection.

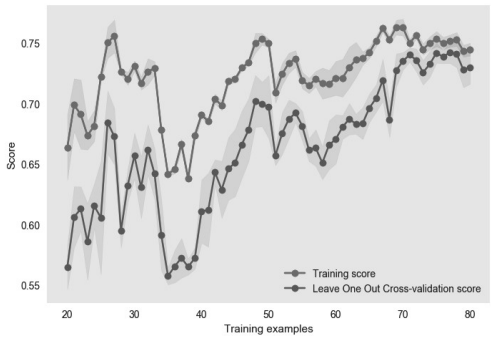

Figure 5: Learning curve by using the selected features

6. CONCLUSION

In this work, we study group cohesion on the ELEA database by using personality traits and social signals.

By analyzing the relation between personality traits and the cohesion, we find that the Big Five Personalty Trait, Agreeableness, is highly correlated to cohesion compared to other personality traits. According to the definition of Agreeableness and our correlation analysis, we believe a participant with high Agreeableness can help to improve the level of cohesion during interactions. Extraversion and Openness to experience also show a strong correlation with group cohesion.

We also apply the estimation of cohesion by using social signals extracted from audios and videos. These signals are clustered into Intra-Personal features and One_VS_All features categories. During the experiments, we find speech turn and variation of speech

energy are related to the meeting cohesion. Moreover, we improve prediction accuracy by applying a feature selection algorithm.

Our work also shows several limitations. The first limitation is the lack of data. With the data we have, we give some interesting conclusions, but they need to be confirmed with more meetings. The second limitation lies on the techniques of feature extraction. Social Signal Processing is a high level application where more meaningful features, such as emotion of speech or face, gaze direction, engagement, etc., can help improve the performance. Some progresses have to be done to extract these features in a wild environment.

7. ACKNOWLEDGMENTS

This work was performed within the Labex SMART (ANR- 11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02

References

[1] N. Ambady and R. Rosenthal. "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." In: *Psychological Bulletin* 111.2 (1992), pp. 256–274.

[2] L. B. Buchanan. "The impact of big five personality characteristics on group cohesion and creative task performance". PhD thesis. 1998.

[3] A. V. Carron and L. R. Brawley. "Cohesion: Conceptual and measurement issues". In: *Small Group Research* 31.1 (2000), pp. 89–106.

[4] M. Casey-Campbell and M. L. Martens. "Sticking it all together: A critical assessment of the group cohesion–performance literature". In: *International Journal of Management Reviews* 11.2 (2009), pp. 223–246.

[5] J. Cassell, A. J. Gill, and P. A. Tepper. "Coordination in conversation and rapport". In: *Workshop on Embodied Language Processing*. Association for Computational Linguistics. 2007, pp. 41–50.

[6] J. Cohen. "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[7] J. Cohen. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." In: *Psychological Bulletin* 70.4 (1968), p. 213.

[8] J. W. Davis and A. F. Bobick. "The representation and recognition of human movement using temporal templates". In: *Computer Vision and Pattern Recognition*. IEEE. 1997, pp. 928–934.

[9] W. Dong et al. "Using the influence model to recognize functional roles in meetings". In: *International Conference on Multimodal Interfaces*. ACM. 2007, pp. 271–278.

[10] N. J. Evans and P. A. Jarvis. "Group cohesion: A review and reevaluation". In: *Small Group Behavior* 11.4 (1980), pp. 359–370.

[11] S. Fang, C. Achard, and S. Dubuisson. "Modeling the synchrony between interacting people: application to role recognition". In: *Multimedia Tools and Applications* (2016), pp. 1–16.

[12] S. Fang, C. Achard, and S. Dubuisson. "Personality classification and behaviour interpretation: an approach based on feature categories". In: *International Conference on Multimodal Interaction*. ACM. 2016, pp. 225–232.

[13] FFmpeg. Available at https://ffmpeg.org.

[14] D. R. Forsyth. *Group dynamics*. Cengage Learning, 2009.

[15] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, 2001.

[16] D. Gatica-Perez et al. "Detecting group interest-level in meetings". In: *International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 2005, pp. I–489.

[17] L. Huang, L.-P. Morency, and J. Gratch. "Virtual Rapport 2.0". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2011, pp. 68–79.

[18] H. Hung and D. Gatica-Perez. "Estimating cohesion in small groups using audio-visual nonverbal behavior". In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 563–575.

[19] O. P. John and L. A. Pervin. "The Big Five factor taxonomy: Dimensions of personality in the natural language and in questionnaires". In: *Handbook of personality: Theory and research*. 1990, pp. 66–100.

[20] I. McCowan et al. "The AMI meeting corpus". In: *International Conference on Methods and Techniques in Behavioral Research*. Vol. 88. 2005.

[21] B. Mullen and C. Copper. *The relation between group cohesiveness and performance: An integration.* 1994.

[22] E. Salas et al. "Measuring team cohesion: Observations from the science". In: *Human Factors* 57.3 (2015), pp. 365–374.

[23] D. Sanchez-Cortes et al. "A nonverbal behavior approach to identify emergent leaders in small groups". In: *IEEE Transactions on Multimedia* 14.3 (2012), pp. 816–832.

[24] A. Schmitt, B. Schatz, and W. Minker. "Modeling and predicting quality in spoken human-computer interaction". In: *Conference of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics. 2011, pp. 173–184.

[25] S. E. Seashore et al. *Group cohesiveness in the industrial work group*. University of Michigan Ann Arbor, 1954.

[26] S. Ultes and W. Minker. "Interaction quality estimation in spoken dialogue systems using hybrid-hmms". In: *Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2014, pp. 208–217.

[27] S. Ultes, A. Schmitt, and W. Minker. "Analysis of Temporal Features for Interaction Quality Estimation". In: *Dialogues with Social Robots*. Springer, 2017, pp. 367–379.

[28] A. Vinciarelli, M. Pantic, and H. Bourlard. "Social signal processing: Survey of an emerging domain". In: *Image and Vision Computing* 27.12 (2009), pp. 1743–1759.

[29] S. Wold, K. Esbensen, and P. Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (1987), pp. 37–52.

[30] A. S. Won, J. N. Bailenson, and J. H. Janssen. "Automatic detection of nonverbal behavior predicts learning in dyadic interactions". In: *IEEE Transactions on Affective Computing* 5.2 (2014), pp. 112–125.

[31] A. S. Won et al. "Automatically detected nonverbal behavior predicts creativity in collaborating dyads". In: *Journal of Nonverbal Behavior* 38.3 (2014), pp. 389–408.