Social Signals of Cohesion in Multi-party Interactions

Reshmashree B Kantharaju, Catherine Pelachaud

Social Signals of Cohesion in Multi-party Interactions

Reshmashree B. Kantharaju
ISIR, Sorbonne Université
Paris, France
bangalore_kantharaju@isir.upmc.fr

Catherine Pelachaud
CNRS - ISIR, Sorbonne Université
Paris, France
catherine.pelachaud@upmc.fr

ABSTRACT

Group conversation is a frequently used form of communication for exchanging ideas and making decisions. Cohesion is an emergent phenomenon that describes the members' attraction towards the group and towards working together. In this paper, we present the cohesion labels assigned to segments from [redacted], a multimodal dataset of simulated medical consultations. Then, we present the analysis performed to identify social cues that characterize cohesion and report the accuracy for classifying cohesion. Results show that non-verbal social cues like gaze, facial AUs, laughter etc., indeed convey information regarding the level of cohesion. Finally we present a preliminary evaluation conducted using the prominent cues to simulate a cohesive group of agents.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

Group Cohesion, Multimodal Database, Multi-party, Social Signals

ACM Reference Format:
Reshmashree B. Kantharaju and Catherine Pelachaud. 2021. Social Signals of Cohesion in Multi-party Interactions. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478362

1 INTRODUCTION

Group interaction is a commonly used form of communication among humans. Often the members of a group are involved in discussing, making decisions and exchanging ideas, under different settings (e. g., meeting, conference, party etc.). The interaction between the members is dynamic in nature. Group cohesion is a common phenomena that emerges over time. It generally describes the members' attraction towards the group and the desire to be a part of the group [13]. This shared bond drives the members to stay together and to want to work together. Cohesion can result in improving group performance, satisfaction among members, and increase adherence [4].

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09…$15.00
https://doi.org/10.1145/3472306.3478362

In multi-party interactions, humans communicate and coordinate with each other via a number of verbal and nonverbal behaviours [55]. These behaviours can help us understand the underlying dynamic nature of the interaction. Existing literature on estimating cohesion so far, has mostly focused on the audiovisual features i. e., pause time, speech energy, motion energy and the verbal features i. e., dialog acts. However, they do not provide any insight on the non-verbal social cues that can be observed during interactions e. g., gaze direction, laughter. Therefore, in this research work we aim to recognise the relation between certain social cues and group cohesion. To the best of our knowledge there has been no study so far that has observed cohesion from a audiovisual social behaviour perspective.

Researchers in artificial virtual agents community are integrating social behaviours to develop conversational agents that are context-aware, engaging and able to understand human conversational dynamics [57]. These agents can be potentially effective tools to persuade, motivate and educate the user through interactive group discussions. Considering the surge in the range of healthcare applications using virtual humans, the necessity for studying the interactions between actual healthcare professionals and patients is becoming increasingly important.

The main goal of this research is to develop a multiparty model involving multiple autonomous agents that can motivate and provide information to users. Several multi-party conversations models have been developed to handle turn-taking and display relevant non-verbal behaviour. However, they do not consider the verbal content and the dynamic nature of group conversations completely. For our model, we aim to consider the task and social dimensions of group cohesion to model the shared commitment to group tasks and positive relationship with members respectively. In particular, we are interested in modelling agents as a team capable of displaying cohesive group behaviour. Identifying the most prominent social cues that increase or decrease the perception of cohesion in a group will be used towards building such a model.

In this paper we first present the *Patient Consultation Corpus (PCC)*, a multimodal dataset of simulated health consultations. The annotations will be made available to the research community. Next, we present the analysis of the social signals annotated with respect to cohesion and its dimensions. Since the data from PCC is limited, we include another popularly used dataset i. e., *Augmented Multi-party Interaction Corpus (AMI)* for the analysis. We then present a baseline on the prediction accuracy for the corpora. Finally, a preliminary study using multiple virtual agents displaying the prominent non-verbal cues that characterise group cohesion is presented. The behaviours for this study were scripted. In a future step, we aim to develop a multiparty model that will be able to automatically generate cohesive group behaviours.

2 RESEARCH QUESTIONS

Non-verbal behavioural cues, convey information about individual traits e. g., affect, personality [55] and the attitude of a given individual in an interaction [46]. These behaviours can help us in understanding the type of interaction e. g., conflict or co-operation and the relationship between the members e. g., rapport, interpersonal attitude. At any given point in an interaction, a member, consciously or unconsciously, decides the appropriate social cue to be displayed based the cues observed so far. In order to develop a model for generating non-verbal cues to simulate cohesive behaviour in a group of agents, it is necessary to identify the prominent social cues that characterize cohesiveness. In particular, we want to first identify how fundamental low-level cues i. e., gaze, head movement, laughter, facial expressions, voice-activity can be associated with cohesion (RQ1). These multimodal signals facilitate in managing the interactions and their functionalities are multidimensional. For example, gaze information can be used to signal next speaker as well as to indicate the level of involvement when combined with head movements. High-level social signals e. g., back-channel, turn-taking can be inferred from low-level cues.

Cohesion is often associated with bonding, feedback and support. We hypothesize that behaviours corresponding to these are frequent in highly cohesive segments. We look at social attention and back-channeling in particular [1]. Social attention provides the capability to orient attention based on other members' focus of attention to draw inferences regarding their goals [42]. Studying the entire mechanism of social attention is beyond the scope of this paper; we focus on two aspects i. e., joint attention and eye contact. *Joint attention* is the ability to coordinate attention between members in order to share an awareness of the objects or events [39]. *Eye contact* helps in establishing a communicative link [21] and improves the level of involvement [43]. Back-channels are short utterances e. g., yeah, uh-huh, okay to indicate that the listener is actively involved in the conversation and encourages the speaker to continue by providing feedback [32]. The next research question aims to identify how these communicative functions characterise cohesion (RQ2).

Interpersonal synchrony refers to an individuals' temporal co-ordination during social interactions [8]. It is dynamic in nature and transpires across modalities. It has been shown to improve the degree of affiliation between members [29]. Executing tasks in synchrony showed improvement in co-operation and strengthening social attachment among group members [31]. Interaction synchrony manifests in various forms e. g., mirroring where individuals mimic each other's behavior subconsciously, entrainment at the voice prosody level. An exhaustive review on interpersonal synchrony and its analysis can be found in [14]. In our next research question, we want to address whether bodily synchrony is positively correlated to cohesion (RQ3). Finally, we report baseline experiments for audio, visual and audiovisual approaches on the corpora (RQ4).

3 BACKGROUND

Cohesion is one of the most commonly studied phenomenon in group dynamics. Several definitions of cohesion have been presented in specific contexts such as sports team [11] and group psychotherapy [7]. Carron et al. defines cohesion as *"a dynamic process that is reflected in tendency of group to stick together and remain united in pursuit of its goals and objectives"*. A multidimensional model based on two dimensions, group–individual and task–social, was proposed [12]. The former indicates that cohesion results from a member's desire to remain part of the group and the interpersonal attraction toward being a member. The latter reflects the perceived task and social aspects of the group. Cohesion can therefore be defined as the tendency of group members to share a bond, accompanied by feelings of solidarity and the willingness to work together [13]. An observation of the existing models helps us identify two constructs of cohesion i. e., attraction to the group or interpersonal attraction (analogous with social cohesion) and commitment to the task (analogous with task cohesion).

With respect to measuring cohesion, a popular method is the use of self-report questionnaire e. g., *Group Environment Questionnaire (GEQ)*, developed to measure cohesion in sports teams [10]. However, they are time consuming, impractical and prone to a bias from participants choosing socially desirable responses [47]. Potential alternatives include unobtrusive measures e. g., sociometric badges [58] or external observer ratings [30]. In this research work we measure cohesion as perceived by external observers using a questionnaire. Since our main goal is to integrate cohesive behaviours into a group of virtual agents, it is relevant for us to measure perceived level of cohesion.

4 RELATED WORK

In this section we present the datasets available with annotations of cohesion, its estimation and the existing models for multi-party interactions.

4.1 Cohesion

In this section, we first present the corpora and then the literature on cohesion estimation. The Group Analysis of Multimodal Expression of cohesiON (GAME-ON) is one of the few datasets designed to specifically study group cohesion [38]. It consists of 17 groups of three participants in a social setting i. e., escape game recorded in Italian. Self-report questionnaire is used to measure cohesion (GEQ), emotions, leadership, warmth and competence.

The AMI corpus has multimodal recordings of four participants in scenario-driven and real meetings [9]. A portion of the corpus i. e., 120 two-minute segments chosen randomly, was annotated for task and social cohesion [30]. In total, 61 segments with high inter-rater agreement were used for the analysis. They extracted audio, visual and audiovisual features to estimate cohesion using naïve mean value based classifier and support vector machine (SVM). The best performing feature was the total pause time between each individual's turns. Further they found a strong correlation between cohesion level and turn-taking patterns. The Emergent LEader Analysis corpus (ELEA) consists of 40 audiovisual recordings of a task-driven interaction between participants [48]. A portion of the corpus i. e., 115 two-minute segments, was annotated for task and social cohesion [20]. For each of the 19 meetings used, four different scores i. e., mean cohesion, weighted mean cohesion, maximum and minimum cohesion were calculated using all the segments belonging to that particular meeting. They studied the relation

between personality and social behaviours of participants with cohesion. Speech turn, variation of speech energy, and *Agreeableness* (a personality trait) were shown to be related to cohesion.

Inferring cohesion based on content analysis i. e., linguistic and paralinguistic mimicry and convergence, in group discussion was presented in [40]. They found that paralinguistic mimicry was useful in estimating social cohesion which is more openly expressed by nonverbal vocal behaviour than task cohesion. Finally, linguistic features e. g., language use constituents (LUC), discourse markers, disfluencies were used to categorize cohesiveness of a group as cohesive, divisive, or mixed interactions [56]. They found that cohesive interactions are comprised of agreement and alignment with minor disagreements and other forms of rejection. These studies indicate that automatically extracted behavioral cues can be used to estimate perceived levels of cohesion in meetings.

4.2 Multi-party Models

There are several models in the literature that are enabled to handle multi-party conversations. The Ymir Turn Taking Model (YTTM), a computational agent-oriented model, implemented up to 12 agents in a virtual world participating in real-time cooperative dialogue [52]. Each agent was configurable for various characteristics e. g., urge-to-speak, yield tolerance. Although these models supported multi-party conversation by handling turn-taking they did not take expression of attitudes and other high-level behaviour into account. Affective real-time turn-taking with expressive interpersonal attitudes by the agents in a group was presented in [45]. It consisted of a turn-taking component, a group behaviour component and a conversational behaviour component. Even though this model handled expression of attitudes, it did not consider speech information. Virtual museum guides with differing opinions and behaviours to provide useful information was presented in [51]. The interaction between the agents was limited to sharing responses. Mission Rehearsal Exercise (MRE) is a model that supports multi-floor dialogue interactions in a 3D virtual environment [54]. Agents were enabled to interact with each other and a human user. The agents were also able to make contact by moving closer (eye or ear shot) and break contact by walking away. A computational model that tracks the conversational dynamics to make turn decisions and render decisions about turns into an appropriate set of low-level behaviours like, coordinated gaze, gesture and speech was developed in [5]. These multi-party models handled turns between the agents and users, and considered the real-time behavioural input from the users. However, none of the existing work considers the high-level group phenomenon i. e., cohesion.

5 PATIENT CONSULTATION CORPUS

Patient Consultation Corpus (PCC), is a multimodal dataset of simulated health consultations between a range of patients with a health condition (diabetes), portrayed by actors, and at least two healthcare professionals with different areas of expertise [50]. It was recorded in collaboration with the University of Dundee. The actors were provided with an overview of relevant medical and social history as well as a brief description of the patient's personality trait, motivations and concerns. The healthcare professional was provided with a brief description of the patient's medical history and health goals. The consultations were spontaneous and had no time restraint to let the interaction take its natural course and to ensure maximum data capture. The corpus consists of nine consultations in total with three to four participants. The first two sessions are from the pilot study and the remaining seven sessions are from the main corpus recordings.

5.1 Cohesion Labels

In this section we present the methodology followed to annotate and assign cohesive labels to PCC corpus. Out of nine sessions, we had to drop three sessions due to incomplete data. The total duration of the remaining data is 126 minutes. The recordings from the individual cameras and overall scene from the six sessions were stitched together to present the complete view of the interaction to the annotators. Thin slice segmentation has been widely used for efficient annotations by external observers [30] since annotating a lengthy video is mentally cumbersome. Moreover, external observers can perceive social interaction by just observing short slice [2]. Therefore, we slice the six sessions into 63 short two-minute segments. We use two-minute slices for two main reasons: (1) to maintain uniformity with the existing annotated data of cohesion; (2) since cohesion is an emergent phenomenon that is conveyed through verbal and non-verbal behaviours, conversational context is necessary to make observations and 30 second slices are too short to convey this information.

5.1.1 Questionnaire. The questionnaire designed aims to measure the collaborative nature of the discussions as well as the social dynamics. As described in Sec. 3, we focus on the two dimensions of group cohesion i.e., task and social. The questions used in psychology research and computational studies [30] on cohesion is used as a basis for the design of our questionnaire. We first grouped questions based on the keywords and discarded redundant questions e. g., questions related to leadership, atmosphere of the group. We selected 10 questions relevant to the scenario in the corpus, where five items were related to the task dimension and the other five were related to the social dimension. We use a 7-point Likert scale for our questionnaire. The list of questionnaire items is presented below.

Task Dimension

1. Every team member seems to have sufficient time to make their contribution.
2. The team seems to share the responsibility for the task.
3. Overall, do the team members appear to be collaborative.*
4. The team members seem to share the same purpose / goal / intentions.*
5. Overall, the members give each other a lot of feedback.

Social Dimension

6. Overall, the group members listen attentively to each other.
7. Overall, the team members seem to be supportive towards each other.*
8. Overall, the work group appear to be in tune/in sync with each other.*
9. Overall, the work group operates spontaneously.
10. Overall, the participants seem to be involved/engaged in the discussion.

5.1.2 Procedure. We recruited annotators from an online crowdsourcing platform named Prolific[1]. Only annotators residing in the United Kingdom with a high proficiency in English and an educational background in psychology or sociology were recruited. We randomly grouped the 63 segments into thirteen groups of five segments. Each group was annotated by five members. To avoid bias and ensure attentiveness, we randomly flipped the scale of the ratings, the order of the questions and the order of the five videos within a group. A questionnaire was presented after each video segment. A PHP web script was used to facilitate the annotation process.

5.1.3 Cohesion Score. In order to calculate the final score for each segment, we first clean up the raw scores. Next, we calculate the inter-rater agreement between the five annotators. Even though previous studies have used a weighted kappa score, we use Interclass Correlation Coefficient (ICC) since we consider the scores to be a continuous variable and ICC is suitable for multi-annotator agreement analysis [26]. Moreover, we are interested in measuring the similarity in ratings instead of the exact level of agreement. We compute ICC between the five annotators belonging to a group using a one-way, average, consistency measure. The agreement level between the five annotators for all the 13 groups were above 75% representing good agreement for all the 63 segments.

To obtain the final score for each segment, we first calculate the average response of each annotator for each segment. Next, the average of the five annotators scores gives us the final score. The obtained values range from 3.56 (lowest) to 6.54 (highest). The mean rating is 5.33 with a standard deviation of 0.67. We categorised the 63 segments using a threshold (mean value) where, 53 segments were labelled as high cohesion and 10 segments as low cohesion. Since the corpus consists of simulated consultations which were usually collaborative, we have a higher number of highly cohesive segments.

5.2 Augmented Multiparty Interaction Corpus

For our analysis, we include the 120 segments from AMI corpus since we have limited number of segments in PCC corpus. Additionally, AMI has been well studied and captures a wider range of behaviours. We obtained the raw scores for the 120 segments from Hung et. al., authors of the study [30]. We calculate the ICC (one-way, average, consistency measure) between the annotators. The agreement level was above 60% representing fair agreement for all the 120 segments. For each segment, we calculate the average of all the 24-questions by each annotator to obtain three annotator-dependent values. Then, we calculate the average of those to obtain the final score. The obtained values range from 2.36 (lowest) to 6.30 (highest). The mean is 4.63 with a standard deviation of 0.89. We assigned a label based on a threshold (mean value) where, 64 segments are labelled as high cohesion and 56 segments as low cohesion.

5.3 Social Cue/Non-verbal Annotation

In this section we describe the annotation of social cues. We follow the same procedure for all the segments from AMI and PCC corpora unless explicitly specified. We mainly make use of the MUMIN annotation scheme [1] to annotate various non-verbal behaviours displayed during interactions. The schema was mainly developed for dyadic interactions and we have adapted it to suit group interactions. A brief description of each tier is provided below. ELAN annotation tool was used for manual annotations of the video segments.

5.3.1 Gaze Direction. We define three types of gaze targets for a given participant i. e., the other members in the group, objects present in the room and unfocused gaze where the target is an arbitrary point. The categorisation of objects is important as AMI meetings are task-oriented, and when a participant was not looking at a member they were usually focusing on the task. We further categorize gaze at other members as *Phead:* gazing at the head and *Pbody:* gazing in the general direction of the member. This helps us in extracting sequences of mutual gaze where two members gaze at each other simultaneously. We partially automated the gaze annotation task to reduce the manual effort. OpenFace [3] is used to automatically extract the gaze angle of the eyes and the head rotation. We discard instances with detection confidence score below 95%. We group the gaze angle data points into four clusters i. e., left, right, straight and down using the k-nearest neighbour algorithm. We extract time-continuous data-points in the same cluster and get the time boundaries i. e., start and end timestamps of a gaze behaviour. Finally, the annotations were manually corrected for any discrepancies.

5.3.2 Facial Action Units. The facial Action Units (AUs) [18] for each participant is extracted using OpenFace. The toolkit outputs the intensity (on a scale of 0 to 5) and the presence of 17 AUs in total. For this study, we use AU intensity values as we are interested in the level of variation of facial expressions displayed by the participants during the interaction. We remove the frames with detection confidence score below 95%. To remove noisy data, intensity values below 0.7 (threshold found empirically by observation) is replaced by zero. We calculate the duration of continuous activation of a given AU. We discard instances that were activated for less than 200ms. For each instance, we calculate the non-zero intensity value. The final set consists of annotations of 17 AUs with the start and end time points and the average intensity of activation for that period. In addition to the AUs, we also automatically extracted the annotations for smile by observing the activation on AU6 (Cheek Raiser) and AU12 (Lip Corner Puller), and corrected them manually.

5.3.3 Head Movement. The presence of head nods in conversation often creates a favorable environment [27] and is commonly associated with attentive listening. As we are interested in understanding the head movement that conveys agreement or comprehension and disagreement in the group, for this work we focus on two event tags i. e., *concord signal:* signals agreement, comprehension, or a positive response, and *discord signal:* signals uncertainty, comprehension failure, or disagreement. A part of the AMI corpus has already been annotated for head movement i. e., 83 of the 120 segments. We manually annotated the remaining 36 segments of AMI corpus and the 63 segments of PCC corpus for concordance and discordance.

[1] https://www.prolific.co

Table 1: Mean intensity values of 17 Action Units (AUs). $*$ indicates that the difference in the values of given AU is statistically significant ($p < 0.05$) when compared with low and high cohesion.

AUs	1*	2*	4	5	6*	7*	9	10*	12*	14	15*	17*	20*	23	25	26*	45
Low	1.010	**1.069**	1.069	0.918	0.991	1.197	0.949	1.193	1.023	1.117	**1.620**	**1.445**	**1.290**	1.126	1.301	**1.358**	0.948
High	**1.046**	1.039	1.065	0.914	**1.045**	**1.282**	0.946	**1.225**	**1.088**	1.108	1.487	1.399	1.248	1.106	1.294	1.307	0.953
p-value	**.0001**	**.014**	.60	.66	**.57E-05**	**.90E-11**	0.78	**.002**	**.15E-09**	.36	**.21E-07**	**.001**	**.029**	.17	.492	**.25E-04**	.470

5.3.4 Laughter. We annotated laughter for both corpora manually using the audiovisual recordings. Audio data was primarily used for annotation of laughter episodes and video data was referred when there was ambiguity. The total of 784 instances from AMI and 112 instances from PCC was extracted.

5.3.5 Social Attention. We consider *joint attention* as instances of shared gaze where more than two participants' look at the same subject of interest. We make the distinction of whether the shared gaze is on a person or object. We consider *eye contact* as instances of mutual gaze between any two participants. We automatically extract the instances from the gaze direction annotation.

5.3.6 Back-channels. For PCC, we extracted the segments with and without speech using automatic voice activity detection algorithm. We used the audio signals from the individual microphones to perform speaker diarization. In automatic speech recognition, the speech signal is segmented into utterance units based on pauses inbetween. Previous studies have used a pause duration ranging from 100ms [37] to 500ms [16]. We consider the most commonly used threshold of 200ms [33]. Therefore, if the silence between two consecutive speech segments was less than 200ms, we merged them into one utterance unit. For AMI, the segmentation of the transcription files were already available. We used these files to annotate back-channels. First we make use of the speech segments to extract instances of overlapping speech. We then extracted the respective audio segments and manually classified whether the overlap was an interruption or back-channel based on the semantic information.

5.3.7 Interpersonal Synchrony. Interlocutors synchronize in overall body movement in addition to posture and gesture [49]. We conduct a preliminary analysis by observing the level of motion energy during the interaction. Motion energy is defined as the differences in grey-scale pixels between consecutive video-frames [24]. We make use of the Motion Energy Analysis tool (MEA) [44] for automatic extraction. Next, we use Windowed Cross Correlation (WCC) with peak picking algorithm, used widely to study head motion synchrony in dyads [6, 28]. Based on the literature, we define a window size of 4sec with and window overlap of 2sec and maxlag of ±2sec. The obtained matrix is provided as an input to the peak picking algorithm which outputs two vectors corresponding to peak correlations nearest to lag of zero and their respective time lags. In order to avoid bias due to low correlation, only peak correlation values greater than 0.5 are considered. We calculate the pairwise body motion energy synchrony score between every member of the group and then calculate the overall average score for that segment.

6 RESULTS AND DISCUSSION

In this section, we report on the experiments performed on the corpora to address the research questions along with a brief discussion.

6.1 Cohesive Signals

Our first research question was to identify the non-verbal social cues associated with low and high cohesion. We performed the analysis on the combined set of of video segments from both the corpora unless specified explicitly. We performed one-way test and computed Spearmann's correlation coefficient since our data does not follow normal distribution.

Literature has repeatedly shown the prominence of eye gaze in social interactions [41]. It regulates the flow of conversation and understanding in multi-party settings [35]. Overall, we assumed that participants in highly cohesive group gaze at other members frequently. This can be an indication of providing feedback to the speaker, or paying attention [25]. To verify this, we calculate the average of total duration spent by each participant of the group looking at other participants. A one-way Anova indicates that the duration spent by a participant looking at another member is positively correlated to cohesion ($r_s = 0.25$, p= .0006), validating the hypothesis. We also observed the amount of time spent by participants gazing at any task-oriented object i. e., slides that were being presented or notes on the desk for both the groups [2]. Interestingly, we found gaze behaviour related to tasks to be negatively correlated to cohesion ($r_s = -0.18$, p= .04). A possible explanation for this tendency could be that social cohesive behaviour contributes more than task cohesive behaviour for perceived level of cohesion by an external observer. Specifically, longer gaze at slides by the speaker can be perceived as gaze avoidance and disengaging the audience [15].

Facial expressions convey a plethora of information regarding the emotional state of a person. To identify the pertinent AUs associated with group cohesion, we first group the segments based on the cohesion label assigned. Then, we calculate the overall non-zero mean intensity of a given AU for all the participants of all segments belonging to a group. Finally, we perform an independent t-test between these two groups i. e., low cohesion consisting of 66 segments and high cohesion consisting of 117 segments. From the 17 AUs observed we found ten to be significantly discriminatory for low and high cohesion. Out of which, five AUs were associated with low cohesion and five with high cohesion. We observe that cheek raising (AU6), lip corner puller (AU12) that is often associated with happiness and smile [17], had higher intensity values in high

[2]This was calculated only for AMI corpus since PCC did not have any objects in the scene

cohesion segments. Further, we observed that lip corners pulled down (AU15), chin raised (AU17), that is often associated with discontent [53] or sad face [36] occurred more frequently in low cohesion segment. In addition to this, a correlation analysis on the intensities, revealed Brow Lowerer(AU4) and Nose Wrinkler (AU9) to be negatively associated with cohesion levels ($r_s = -0.22$, p= .01) and ($r_s = -0.26$, p= .0004) respectively.

Concordance signals agreement or comprehension, often indicated with a nod, and creates a favorable environment [27]. In our data, we found that the frequency of head nods are higher in high cohesion segments and positively correlated to social cohesion ($r_s = 0.42$, p= .44E-08). Since head movement can signal turn-taking or emphasise a point the speaker is trying to convey, we need to study it further with respect to its various associated functions.

Laughter is a commonly used non-verbal vocalisation that usually indicates a positive affect and cooperative intent. We hypothesize that instances of laughter are frequent in highly cohesive meeting segments. To verify this, we compared the average occurrence of laughter and average duration of laughter for every segment belonging to the two groups. Results indicate that laughter was observed more frequently and lasted longer in high cohesion segments ($M = 2.26$, $SD = 2.31$) than low cohesion segments ($M = .91$, $SD = 0.98$) with ($p = .0008$). We also found that shared laughter instances i. e., where two or more participants laugh simultaneously occurred three times more frequently in high cohesion segments in comparison to low cohesion segments. Literature on laughter in groups state that "laughter establishes a form of bond in social groups and makes people feel more comfortable" [22]. We identify it to be a prominent cue in conveying information about the bond between the members of the group.

Social attention which is studied as eye-contact or shared attention between participants was found to be a good indicator of social cohesion. The pairwise average instances of mutual gaze between any two participants in the group is positively correlated with cohesion ($r_s = 0.24$, p= .001). We can infer that participants in cohesive groups spent longer time holding eye-contact while low cohesive groups spent a shorter amount of time holding the gaze with other participants or avoiding eye-contact. This is in line with Exline et. al., [19], where they state that the duration of eye-contact decreased in non-collaborative conditions. Further, shared attention by three or more members on a task-related object is positively correlated with task cohesion ($r_s = 0.23$, p= .01). This indicates that participants were co-operative during the interactions and perceived to be working together on the task. Finally, we found that backchannels were positively correlated with cohesion ($r_s = 0.44$, p= .45E-9) as found in previous studies [30] which can be interpreted as a sign of support and attentiveness from the listeners.

We analysed the bodily motion synchrony, a simplified version of interpersonal synchrony, between the participants of the group. We extracted the change in body motion energy for the participants of the group and plotted it, ref. Figure 1. The plot shows the change in motion energy between four participants from a segments belonging to high cohesion and low cohesion group. We can easily observe that for high cohesion segment, along with high energy values, the overall variation is frequent than in low cohesion segment. Further, we observe a synchronous variation in the motion energy between participants, while its almost non-existent

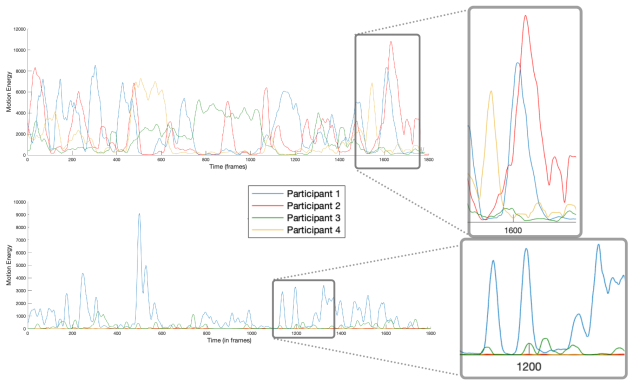

Figure 1: Plot of body motion energy level computed over time for all the participants belonging to low (below) and high (above) cohesion group.

for low cohesion segment. This is line with studies in literature that state that synchronous movement contributes to social cohesion by increasing social closeness, and reducing stress [23].

6.2 Cohesion Estimation

Next, we wanted to verify whether the non-verbal features identified can effectively predict the level of cohesion in groups. We report both classification and regression results on both corpora using (10-fold stratified cross-validation. We perform a grid-search approach to find the optimum value for cost and gamma and report the results with this setting. The most significantly discriminatory ($p < 0.05$) features found in our data analysis for low and high cohesion was used to form our feature set for the prediction task. The best performing feature set i. e., average duration spent gazing at other members, average mutual gaze instances between all the participant pairs, shared attention between three members gazing at the fourth member, average laughter duration and average instances of backchannels, achieved an accuracy of 78.1% in classification task (SVM) and RMSE rate of 0.63 for the regression task (Gaussian PMR). From these results, we can interpret that the information conveyed by the listener is important for estimating cohesion as this is an indication of feedback and group-level features are highly important for predicting cohesion.

7 PRELIMINARY EVALUATION

In this section we present a preliminary study conducted incorporating the prominent non-verbal cues recognised in our earlier analyses. The objective of this evaluation is to measure the perceived level of cohesiveness of a group of virtual agents. To simulate a cohesive group of agents we incorporate head nods along with gaze and smile. A basic LSTM network implemented in Keras with optimised hyper-parameters is trained on cohesive video segments separately to predict speaker and listener behaviours for the agents. The input to the model is binary encoding of gaze direction (speaker/listener), nod (yes/no) and smile (yes/no). The predicted output was then manually translated into a BML script to be executed by the virtual

agents. Additionally, we changed the gaze pattern of the agents when speaking to the user to gaze at the user[3].

Figure 2: Screenshot of the interaction

The designed scenario has four virtual agents interacting with each other and addressing the user about stress management as shown in Figure 2. Based on the results from a previous study on persuasiveness [34], the role of providing advice was assigned to an older authoritative agent while a younger peer coach acted as a supportive coach. We also made use of vicarious persuasion techniques where one agent presents an argument to persuade another agent while indirectly persuading the user. We use a between-group study with one group watching agents displaying cohesive behaviours generated by the model and the other group watching agents displaying behaviours from a random generator. The agents and the dialogue content remained the same for both the conditions. Four questions were shortlisted from the 10-item questionnaire, based on their relevancy to the study, to measure the cohesiveness of the group (see Section 5.1.1). Each user was presented with a series of recordings of the interaction between the agents. The user was prompted to provide their input using on-screen buttons with four pre-defined responses during the interaction. The interaction on stress management lasted about four minutes. The user then completed a questionnaire at the end of the interaction. The pre-study questionnaires measures the Negative Attitude towards Robots Scale (NARS) adapted to virtual agents (4-items) and persuadability of the user (5-items). The post-questionnaire measures the cohesiveness (4-items), the credibility (3-items) and the persuasiveness (3-items) of the group.

The participants were recruited from an online survey platform, Prolific. We had set three specifications to recruit participants, i.e., aged above 50, proficient in English and has been diagnosed with chronic disease [4]. In total we had 32 participants taking part in our evaluation study where 10 participants were in the age group of 51-60 and 22 were in the age group above 60. 36% of the participants were male while 64% were female. A one-way Anova for used for the analysis of the responses. The perceived level of cohesion was slightly higher for the condition using our model (m=4.03) in comparison to random behaviour model (m=3.53) for participants measured to be persuadable (n=16) and the difference was slightly significant (p = 0.1). We further grouped the participants based on NARS questionnaire, and did not find any significant results.

In this evaluation study we tried to measure the perceived level of cohesion and how this in turn affects the trust in the agents and their persuasiveness. The study had to be performed online with pre-recorded videos which hindered the quality of interaction. Even though we tried our best to record high-quality videos, we are not sure whether the participants were able to watch them in the same setting. Since the differences in a listener executing a smile or nod is very subtle the participants might have missed it. Also, the environmental conditions could affect the results which was beyond our control. Regarding the perceived persuasiveness, we found there was no significant difference. This could be attributed to the fact that we used the same dialogue content and agents for both the conditions and only the non-verbal behaviours were different. Some participants found the automatic text-to-speech generated audio to be very artificial and their expectations of agents were probably a bit technically unrealistic, which could have affected their rating. Overall, the participants found the study to be quite interesting and an enjoyable experience.

8 CONCLUSION AND FUTURE WORK

In this paper, we provide an insight into the non-verbal social cues pertinent to group cohesion. In particular, we focused on gaze direction, head movement, laughter, social attention, back-channel and bodily motion energy synchrony. From the results obtained we can conclude that non-verbal social cues indeed convey information regarding the level of cohesion and can be used for automatic prediction tasks. The results from this work will contribute towards developing a computational model to simulate a cohesive group of virtual agents. We aim to have a distributed implementation where each agent decides independently the social cue to be displayed based on a set of observations rather than a common model assigning behaviours to the agents. This sort of implementation handles dynamic variation in number of agents involved in the interaction. Future work includes implementing a neural network model with temporal consideration of the interaction and representation of group features.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant Agreement Number 769553. We would like to thank our collaborators at University of Dundee, Roessingh Research and Development and University of Twente for the design and recording of the PCC corpus. We are also grateful to Dr. Hayley Hung of TUDelft for sharing the AMI cohesion annotation dataset with us.

REFERENCES

[1] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41, 3-4 (2007), 273–287.

[2] N. Ambady and R. Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111, 2 (1992), 256.

[3] T. Baltrušaitis, P. Robinson, and L. Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[3] We included this based on a previous evaluation with just the agents discussing among themselves without gazing at the user. The users found this to be unnatural. We did not encode user in our training data, but we plan to incorporate this factor in our future model.

[4] The study was performed in conjunction with a project evaluating the perception of virtual agents as coaches for ageing adults

[4] D. Beal, R. Cohen, M. Burke, and C. McLendon. 2003. Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology* 88, 6 (2003), 989.

[5] D. Bohus and E. Horvitz. 2010. *Computational Models for Multiparty Turn Taking*. Technical Report. Microsoft Research Technical Report MSR-TR 2010-115.

[6] S. Boker, J. Rotondo, M. Xu, and K. King. 2002. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods* 7, 3 (2002), 338.

[7] L. Braaten. 1991. Group cohesion: A new multidimensional model. *Group* 15, 1 (1991), 39–55.

[8] J. Burgoon, L. Stern, and L. Dillman. 2007. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*. Springer, 28–39.

[10] A. Carron. 1982. Cohesiveness in Sport Groups: Interpretations and Considerations. *Journal of Sport Psychology* 4, 2 (1982), 123–138.

[11] A. Carron and P. Chelladurai. 1981. The dynamics of group cohesion in sport. *Journal of Sport Psychology* 3, 2 (1981), 123–139.

[12] A. Carron, N. Widmeyer, and L. Brawley. 1985. The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. *Journal of Sport Psychology* 7, 3 (1985), 244–266.

[13] M. Casey-Campbell and M. Martens. 2009. Sticking it all together: A critical assessment of the group cohesion–performance literature. *International Journal of Management Reviews* 11, 2 (2009), 223–246.

[14] M. Chetouani, E. Delaherche, G. Dumas, and D. Cohen. 2017. 15 Interpersonal Synchrony: From Social Perception to Social Interaction. *Social signal processing* (2017), 202.

[15] J. Collins. 2004. Education techniques for lifelong learning: giving a PowerPoint presentation: the art of communicating effectively. *Radiographics* 24, 4 (2004), 1185–1192.

[16] J. Edlund, M. Heldner, and J. Gustafson. 2005. Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen* (2005), 576–587.

[17] P. Ekman, R. Davidson, and W. Friesen. 1990. The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology* 58, 2 (1990), 342.

[18] P. Ekman, W.V. Friesen, and J.C. Hager. 2002. *Facial action coding system (FACS). A human face*. Research Nexus, Salt Lake City.

[19] R. Exline. 1963. Explorations in the process of person perception: Visual interaction in relation to competition, sex, and need for affiliation. *Journal of Personality* (1963).

[20] S. Fang and C. Achard. 2018. Estimation of Cohesion with Feature Categorization on Small Scale Groups. (2018). WACAI.

[21] T. Farroni, G. Csibra, F. Simion, and M. Johnson. 2002. Eye contact detection in humans from birth. *Proc. of the National academy of sciences* 99, 14 (2002), 9602–9605.

[22] Phillip Glenn. 2003. *Laughter in interaction*. Vol. 18. Cambridge University Press.

[23] A. Göritz and M. Rennung. 2019. Interpersonal synchrony increases social cohesion, reduces work-related stress and prevents sickdays: a longitudinal field experiment. *(GIO)* 50, 1 (2019), 83–94.

[24] K. Grammer, M. Honda, A. Juette, and A. Schmitt. 1999. Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *Journal of personality and social psychology* 77, 3 (1999), 487.

[25] E. Gu and N. Badler. 2006. Visual attention and eye gaze during multiparty conversations with distractions. In *International workshop on intelligent virtual agents*. Springer, 193–204.

[26] Kevin A H. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8 (2012).

[27] Uri Hadar, Timothy J Steiner, Ewan C Grant, and F Clifford Rose. 1984. The timing of shifts of head postures during conservation. *Human Movement Science* 3, 3 (1984), 237–245.

[28] Z. Hammal, J. Cohn, and D. George. 2014. Interpersonal coordination of headmotion in distressed couples. *IEEE transactions on affective computing* 5, 2 (2014), 155–167.

[29] M. Hove and J. Risen. 2009. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition* 27, 6 (2009), 949–960.

[30] H. Hung and D. Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[31] J. Jackson, J. Jong, D. Bilkey, H. Whitehouse, S. Zollmann, C. McNaughton, and J. Halberstadt. 2018. Synchrony and physiological arousal increase cohesion and cooperation in large naturalistic groups. *Scientific reports* 8, 1 (2018), 1–8.

[32] K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto. 2013. Gaze and Turn-taking Behavior in Casual Conversational Interactions. *ACM Transactions on Interactive Intelligent Systems*, Article 12 (2013), 30 pages.

[33] K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–30.

[34] R. Kantharaju, D. De Franco, A. Pease, and C. Pelachaud. 2018. Is Two Better than One?: Effects of Multiple Agents on User Persuasion. In *Proc. of the 18th International Conference on Intelligent Virtual Agents*. ACM, 255–262.

[35] A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63.

[36] C. Kohler, T. Turner, N. Stolar, W. Bilker, C. Brensinger, R. Gur, and R. Gur. 2004. Differences in facial expressions of four universal emotions. *Psychiatry research* 128, 3 (2004), 235–244.

[37] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech* 41, 3-4 (1998), 295–321.

[38] L. Maman, E. Ceccaldi, N. Lehmann-Willenbrock, L. Likforman-Sulem, M. Chetouani, G. Volpe, and G. Varni. 2020. GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis. *IEEE Access* 8 (2020), 124185–124203. https://doi.org/10.1109/ACCESS.2020.3005719

[39] P. Mundy, M. Sigman, J. Ungerer, and T. Sherman. 1986. Defining the social deficits of autism: The contribution of non-verbal communication measures. *Journal of child psychology and psychiatry* 27, 5 (1986), 657–669.

[40] M. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung. 2017. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 206–215.

[41] D. Novick, B. Hansen, and K. Ward. 1996. Coordinating turn-taking with gaze. In *Proc. of 4th International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1888–1891.

[42] L. Nummenmaa and A. Calder. 2009. Neural mechanisms of social attention. *Trends in cognitive sciences* 13, 3 (2009), 135–143.

[43] C. Oertel. 2013. Towards developing a model for group involvement and individual engagement. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 349–352.

[44] F. Ramseyer and W. Tschacher. 2011. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology* 79, 3 (2011), 284.

[45] B. Ravenet, A. Cafaro, B. Biancardi, M. Ochs, and C. Pelachaud. 2015. Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. In *Proc. International Conference on Intelligent Virtual Agents*. Springer, 375–388.

[46] V. Richmond, J. McCroskey, and S. Payne. 1991. *Nonverbal Behavior in Interpersonal Relations*. Prentice Hall Englewood Cliffs, NJ.

[47] E. Salas, R. Grossman, A. Hughes, and C. Coultas. 2015. Measuring team cohesion: Observations from the science. *Human factors* 57, 3 (2015), 365–374.

[48] D. Sanchez-Cortes, O. Aran, M. Mast, and D. Gatica-Perez. 2010. Identifying emergent leadership in small groups using nonverbal communicative cues. In *Workshop on machine learning for multimodal interaction*. 1–4.

[49] K. Shockley, M. Santana, and C. Fowler. 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* 29, 2 (2003), 326.

[50] Mark Snaith, Nicholas Conway, Tessa Beinema, Dominic De Franco, Alison Pease, Reshmashree Kantharaju, Mathilde Janier, Gerwin Huizing, Catherine Pelachaud, and Harm op den Akker. 2021. A multimodal corpus of simulated consultations between a patient and multiple healthcare professionals. *Language resources and evaluation* (2021), 1–16.

[51] W. Swartout, D. Traum, R. Artstein, D. Noren, P. Debevec, K. Bronnenkant, J. Williams, A. Leuski, S. Narayanan, D. Piepol, et al. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In *Proc. International Conference on Intelligent Virtual Agents*. Springer, 286–300.

[52] K. Thórisson, O. Gislason, G. Jonsdottir, and H. Thorisson. 2010. A Multiparty Multimodal Architecture for Realtime Turntaking. In *Proc. Intelligent Virtual Agents*. Springer, 350–356.

[53] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 2 (2001), 97–115.

[54] D. Traum and J. Rickel. 2002. Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In *Proc. of the First International Joint Conference on Autonomous agents and Multi-agent systems*. ACM, 766–773.

[55] A. Vinciarelli, M. Pantic, and H. Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.

[56] W. Wang, K. Precoda, R. Hadsell, Z. Kira, C. Richey, and G. Jiva. 2012. Detecting leadership and cohesion in spoken interactions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5105–5108.

[57] Q. Xu, L. Li, and G. Wang. 2013. Designing engagement-aware agents for multiparty conversations. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2233–2242.

[58] Y. Zhang, J. Olenick, C. Chang, S. Kozlowski, and H. Hung. 2018. The I in team: Mining personal social interaction routine with topic models from long-term team data. In *23rd International Conference on Intelligent User Interfaces*. 421–426.