IDIAP RESEARCH REPORT

ESTIMATING COHESION IN SMALL GROUPS USING AUDIO-VISUAL NONVERBAL BEHAVIOR

Hayley Hung Daniel Gatica-Perez

Idiap-RR-12-2010

JUNE 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Estimating Cohesion in Small Groups using Audio-Visual Nonverbal Behavior

Hayley Hung, *Member, IEEE,* Daniel Gatica-Perez, *Member, IEEE,*

***Abstract*—Cohesiveness in teams is an essential part of ensuring the smooth running of task-oriented groups. Research in social psychology and management has shown that good cohesion in groups can be correlated with team effectiveness or productivity so automatically estimating group cohesion for team training can be a useful tool. This paper addresses the problem of analyzing group behavior within the context of cohesion. 4 hours of audio-visual group meeting data was used for collecting annotations on the cohesiveness of 4-participant teams. We propose a series of audio and video features, which are inspired by findings in the social sciences literature. Our study is validated on as set of 61 2-minute meeting segments which showed high agreement amongst human annotators who were asked to identify meetings which have high or low cohesion.**

I. Introduction

COHESION in teams is a necessary part of collaboration either for social or professional motives. Though definitions of cohesion by social psychologists have varied considerably, depending on the domain in which it is studied, a good definition can be found in Casey-Campbell and Martens' recent critical assessment of the group cohesion-performance literature: "Cohesion is now generally considered as the group members' inclinations to forge social bonds, resulting in the group sticking together and remaining united." [11] (p 223). However cohesion is defined, it is undeniable that there has been considerable interest in this concept in the organizational management world, due to its relation to group performance. However, the link between a group or team and its performance is not limited to tasks carried out in business organizations where financial gain, and perhaps power and influence, can be seen as the principal motivation for success. In practice, the vast body of psychology literature on cohesion in groups relates to contexts ranging from team sports [10], to group psychotherapy [5]back or military training [23]. Studying cohesion in each of these domains has led to a plethora of theories about what cohesion is.

Despite the challenging nature of cohesion as a group behavioral phenomenon, we show that human annotations can have strong agreement under certain circumstances, and that simple nonverbal audio and visual cues can represent reasonably well, the perceptions of levels of group cohesion in a task-based scenario. Inspired by findings from social psychology, we investigate effective automatically extracted audio, visual, and audio-visual cues that can be used to estimate cohesion levels in groups.

Our aim here is to investigate systematically, automatic features that can be used to measure cohesion levels in groups rather than to develop sophisticated classifiers for this task. To our knowledge, this is perhaps the first study that attempts to automatically estimate cohesion in task-based meetings. Specifically, our contributions are :

1) Investigating methods for estimating the cohesion in task-based group meetings using automatically extracted audio, visual, and audio-visual cues. Simple and more sophisticated classification methods are also investigated to highlight the discriminative power of the features.
2) Collecting and studying human annotations of this behavioral construct as a means of understanding how cohesion is perceived by external observers, and also to establish a reference for evaluating automated methods.

The remainder of this paper is organized as follows: Section II describes related work in both social sciences and computing; Section III gives an overview of our cohesion estimation approach; Section IV describes the data and annotation process that was used to gather perceptions of group cohesion. Section V describes the audio, video, and audio-visual nonverbal cues that were extracted for the classifiers; Section VI describes the experiments that were carried out; Section VII shows and discusses the results and we conclude in Section VIII.

II. Related work

A. Cohesion in the Social Sciences

There has been considerable interest in cohesion in groups. The term 'group' in itself refers to any collection of people that can range from a size of 2 to hundreds or thousands, depending on the context. Here, we concentrate on analyzing cohesion in small groups in face-to-face encounters. Despite this focus, there is still a considerable amount of literature concerning groups of this size. We do not provide an exhaustive review of group cohesion here but refer the reader to a comprehensive review by Casey-Campbell and Martens [11].

Part of the relevance of studying cohesion in groups is because of its benefits in terms of an individual's need to feel a sense of belonging (whether to their work place or social life). In addition, group cohesion has been suggested to be well correlated with performance in some studies [11], [40]. Many definitions of cohesion have been approached through specific contexts such as team sports, group psychotherapy etc. Psychologists initially approached cohesion from the perspective of what causes cohesion, rather than what its consequences are. Causes or antecedents of cohesion are

H. Hung is a research fellow at the University of Amsterdam, Netherlands

D. Gatica-Perez is with Idiap Research Institute, Switzerland and Ecole Polytechnique de Lausanne (EPFL), Switzerland

difficult to measure since they have to be measured before any interaction between group members has occurred. In addition, it may be difficult for investigators to distinguish the actual causes from the antecedents. For example, one of Carron and Brawley's definitions of cohesion is "the individual's personal motivations to remain in the group" as well as "the individual's perceptions about what the group believes about its closeness" (p 90) [9], we see that these two aspects can be greatly influenced by a person's prior experience with some or all of the group members, or indeed their predisposition to need to belong or be affiliated with a particular group [11]. On the other hand, if we consider the outcomes of cohesion, these can be altogether more readily measurable, such as the duration of membership within a group, the influence that members can have on each other [20], increased organizational citizenship [25], or reduced absenteeism [30].

When measuring group cohesion, a popular early theory by Seashore suggests that cohesion is a construct that defines "the resultant of all forces of members acting to remain in the group, including both driving forces toward the group and restraining forces against leaving the group" (p. 11) [34]. However, the problem with this definition is that it considers all the individual perceptions in a group to be the summation of all opinions, which fail to consider the group as an entity in itself. Later Evans and Jarvis [19] suggested a two-dimensional construct which argued that in addition to attraction to a group, the degree to which the group decides what goals are important for its members was also key to cohesion. This moved the definition of cohesion away from just a social phenomenon into a behavioral construct that could be both *task* and *socially* oriented. That is, as suggested by Mullen and Cooper [32], the cohesiveness-performance effect was more attributed to a group's commitment towards the task rather than each other.

Bollen and Hoyle [4] also proposed a two-dimensional model of cohesion relating to belongingness and morale, where the latter represented a more affective element of cohesion. In this case, both aspects of cohesion could be considered to be social. Zaccaro and Lowe [39] suggested that a multidimensional approach to cohesion is "supported if each type of cohesion has different consequences" (p.556). Their approach suggests that task cohesiveness leads to better performance while social cohesion can limit maximum performance. However, Zaccaro and McCoy [40] also found that both types of cohesion are required to succeed on a group task. It has also been argued that good team performance can be viewed as an antecedent of high cohesion, making the cohesion-performance relation cyclic [11].

When social scientists started to look at the consequences rather than the causes of group cohesion, it became easier to treat cohesion more as a group phenomenon rather than an aggregation of individual perceptions. This led to theories on cohesion as a multi-dimensional construct such as that of Carron, who defined cohesion as "a dynamic process that is reflected in the tendency for a group to stick together and remain united in the pursuit of its instrumental objectives and /or for the satisfaction of member affective needs" (p213) [10]. Carron and Brawley [9] went on to define cohesion as a four-dimensional construct combining the perceptions of the group as a whole, as well as the capability of the group to address each individual's needs, as well as how this would relate to task and social cohesion. Other psychologists have also considered cohesion as a construct with vertical and horizontal components; Siebold [35] suggested that group leaders hold a group together and encourage a sense of pride in the group. He defined two axes to team cohesion, namely horizontal cohesion (related to peer bonding) and vertical cohesion (related to having a caring leader and also pride and shared values, needs, and goals within the group).

So far we have described cohesion theories based on general terms, and often in terms of inward states that may not be so easy to observe automatically. However, for automatic analysis, we must look more closely at elements of the behavioral construct that could be correlated with measurable interactive behavior. Braaten [5] suggested 5 factors that affect group cohesion in group psychotherapy: attraction and bonding, support and caring, listening and empathy, support and caring, self-disclosure and feedback, process performance and goal attainment. We will use some of the findings in the cohesion literature to inspire the design of our features later.

B. Aspects of Cohesion and Computing

To our knowledge no work has been done to approach the problem of computational analysis of cohesion in teams. The most similar work to that proposed here concerns aspects of cohesion, which are more related to observations of interactive phenomena such as interest levels in groups [21], rapport [12], [22], attractiveness [28], mimicry [2], [26], or synchrony [7]. Such related work generally falls into three categories of interaction; dyadic human-human, dyadic human-computer, or multi-party human-human interaction.

In terms of human-human dyadic interactions, Madan et al. [28] tried to predict the interest of pairs during a speed-dating event. They extracted features from vocal signals that represented engagement, stress, and mirroring behavior and trained a support vector machine to see if romantic attraction could be discriminated from attraction for friendship or business reasons. Campbell attempted to measure the degree of synchrony and rapport between dyads, using speaking activity features [6]. From these features, he showed that synchrony at the speech activity level could also be identified, and suggested a measure of conversational flow that could be used to observe the change in developing relationships between previously unacquainted dyads. He has also carried out experiments to show that body motion and speaking activity were correlated between individuals [7] in a four-person conversation. However, while the analysis of the data was based on automatically extracted cues, an evaluation of the success of the measures experimentally for the target behavioral constructs such as synchrony and mimicry was not carried out. In addition, it would be difficult to draw strong conclusions from the work since the data set consists of few pairs of dyads or groups.

Moving beyond dyadic human-human interactions interactions, much work has been carried out on finding ways to make the human-computer interaction experience more pleasant. To this end, some work has concentrated on trying to create virtual agents that exhibit natural affective interactive behavior

such as rapport [12], [22] and mimicry [2], [26]. Gratch et al. [22] conducted experiments using a virtual agent that could be controlled by a person listening to the speaker. Pairs of speakers and listeners were assigned either 'responsive' or 'unresponsive' virtual agents during a recount of a previously observed incident. In the 'unresponsive' condition, the avatar's movements were created randomly, and not based on the speaker's or listener's behavior. Results showed that mimicry in a virtual agent (the 'responsive' mode) led to an increase in speaker fluency, duration of the interaction, and feelings of rapport. Cassell et al. [12] moved this further by studying if the dynamic nature of rapport, as a relationship develops, could also be synthesized in an embodied conversational agent.

In terms of automated studies of groups of people, Gatica-Perez et al. [21] addressed the problem of estimating group interest levels in meetings. Audio and video features were extracted to measure vocal pitch, energy, speaking rate, and visual information such as coarse head and body motion, and body pose information. These features were then integrated into a Hidden Markov Model (HMM) framework by fusing single modality features together before training the model or using a Multi-stream HMM that trains an audio and visual model independently before merging likelihoods by multiplication at each time step.

While all the work presented above tries to identify ways of measuring how well people are involved or getting along during a conversation, this only addresses the social aspect of cohesion. As already mentioned, cohesion can be divided into social and task aspects. In this work, we address both these issues to see if they could be identified through facets of cohesion.

III. Our Approach

In our experiments, we extracted audio, video, and audio-visual cues related to aspects of group cohesion, to see which would represent meetings with high or low cohesion. We tried both a simple supervised method and also using a more powerful supervised classifier summarized in Figure 1.

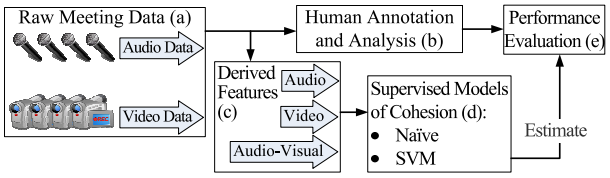


Fig. 1. Summary of our approach.

(a): The raw data from which our experiments were conducted is described in Section IV.

(b): The annotation of cohesion in the meeting data is described in Section IV. These annotations were studied and a subset from this was selected for experiments

(c): Audio, video and audio-visual nonverbal cues were extracted by taking inspiration from findings in social psychology, as described in Section V.

(d): Two different supervised methods were used for classifying the cohesion level in the meeting segments. These are described in Section VI.

(e): The two classification methods and feature modalities were evaluated using the data set selected in (b) and the results are discussed in VII.

We organized our experiments based on the following goals:

1) whether external observers can perceive differing levels of group cohesion.
2) whether some automatically extracted audio, video and audio-visual nonverbal cues can be used to infer or explain differing levels of group cohesion.
3) whether automatically extracted audio-visual cues are more effective for estimating levels of group cohesion than any cues extracted from a single modality.

IV. Data

The Augmented Multiparty Interaction (AMI) Corpus contains small group meetings recorded using audio-visual sensors [8]. It contains both meetings created for volunteers to take part in a scenario where each was assigned a role, placed into a team of four, and asked to design a remote control, and also a small set of meetings which are taken from real meetings where colleagues or acquaintances come together to discuss a topic related to their real life. The meeting room and the examples of the typical video data are shown in Figure 2. In addition to cameras capturing all the meeting participants at varying degrees of granularity, headset microphones were also used to record each person's voice with high quality. All meetings involve four participants.

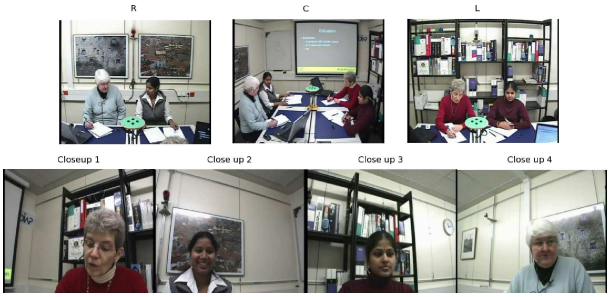


Fig. 2. Top: the layout of the meeting room and the spacing of the participants around the table. Bottom: view from each participant's close-view camera.

A. Annotation Procedure

To our knowledge, most previous work in the social sciences that has investigated aspects of cohesion in groups has tended to use the participants themselves to gather perceptions of a group or individual's behavior [11]. Since we frame the problem of estimating cohesion for aiding meeting browsing or data mining, it follows that external observations may be more appropriate for our task.

A pool of 21 annotators were used to annotate 120 2-minute non-overlapping meetings segments from the Idiap AMI meetings. 100 meeting segments were taken equally from each of the 10 teams in the corpus who were asked to design remote controls. 20 meeting segments were taken from two other groups who were involved in real (rather than scenario-based) meetings. One involved discussing movies that could be shown at a film club, and the second was a meeting to discuss the new allocation of offices to members of staff. Meeting segments in our data set were purposefully chosen so that the participants remained seated through the duration of the slice. The 120 meeting segments were divided into 10 sets of 12 (1 from each team) where each set was assigned to a group of 3 annotators. Most annotators labeled one set of 12

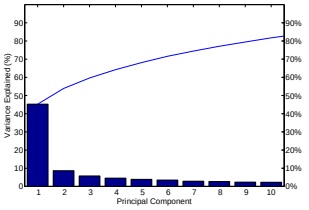


Fig. 3. The eigenvalue for each component (bar chart) and the cumulative percentage of the explained variability (line).

meetings. Some volunteered to label more sets and were placed in new groups of 3 annotators. Note that these annotators never labeled two different sets of meetings segments in parallel.

The length of each segment may appear short but Ambady et al. [1] have suggested that behavioral constructs can be perceived sufficiently from observing short segments or thin slices of interactive behavior. To our knowledge while other similar interactive behaviors (e.g. rapport) have been investigated using the thin slice concept, the exact relationship between group cohesion and thin slices of behavior has not been documented. For the sake of simplicity, since all of our data is organized into meeting segments, we will refer to them as meetings for the remainder of this paper.

To annotate the meetings for cohesion, terms used in the psychology literature [4], [5], [9], [23], [35] were pooled together to create a questionnaire containing 27 questions. These included scoring the group interactions based on how comfortable participants were, how integrated the team appeared, how well they knew each other, how engaged or involved they were, whether they shared the same goal, etc. The complete list of questions is listed in the Appendix. There were also a few questions that were inspired by terms used in the literature, that had either been measured before using automated methods (e.g. rapport and dominance) or which we felt could be directly related to measurable nonverbal cues. An initial set of questions were chosen for a pilot annotation study. From this study, the questionnaire was modified to remove confusion and ambiguity that was found in the pilot study.

For each question, annotators were asked to score their response on a 7-point scale. To ensure that annotators thought carefully about each of the questions, the valences of each answer were randomly flipped. If participants were unsure of the answer to any of the questions, they could skip it. The annotators viewed their corresponding meeting segments through a web interface so that all the groups and meeting times were made anonymous.

B. Analyzing the Annotations

To analyze the distribution of the annotations, we first carried out principal component analysis on the data. This analysis showed that 45.3% of the variance in the annotation data could be explained by the first principal component, with the first 6 components containing just over 70% of the data variance. A graph of the eigenvalues is shown in Figure 3.

We examined the annotation data further by plotting the data using the first two components of the data, shown in Figure 4. Here we see that the questions are arranged mostly into two clusters along the first principal component. On further inspection, we found that these corresponded to the orientation

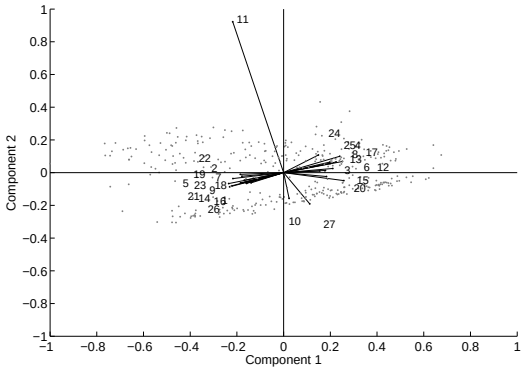


Fig. 4. Plot showing the loadings (black lines) and scores (grey dots) for the data when observed in relation to the first and second principal components of the annotation data. Each line shows the loading of each question in the principal component space such that a longer line indicates the variability of the vector in the two components and is labeled with a number which corresponds to the question number (see appendix for the mapping of questions to numbers).

of the scale for each question. So questions which we expected to have high scores correlating with high cohesion appearing in one cluster while those we expected to have low scores corresponding to high cohesion, appeared in the other cluster. This observation suggests that very similar scoring patterns were occurring for a significant number of different meetings. We also observe that questions 10, 11 and 27 are more strongly loaded by the second principal component. Question 10 corresponds to a 'yes'/'no' question in the questionnaire, which asks whether there is a strong leader in the group. If the answer to this question is yes, annotators were asked to answer question 11 (on a 7-point scale) to indicate the extent to which the leader brought the rest of the group together. The other question relates to whether each team member has sufficient time to make a contribution to the discussion.

To analyze the agreement amongst annotators and across meetings and questions, we used the kappa agreement measure. Furthermore, since the orientation for some of the answers were flipped, these were modified to ensure that all the valences for each question were aligned. Since the annotation scores were on a scale, we used the weighted kappa measure [14] with a linear decay from the confusion matrix diagonal. For each meeting, the weighted kappa was computed for each pair-wise combination of the 3 annotators. Then, the average of all 6 kappa values was taken to be the mean weighted kappa agreement for that meeting.

From these kappa scores, we were able to observe the variation in kappa across different meetings or questions. From the principal component analysis of the data, we found that the data points formed a continuous manifold, from meetings which exhibited very high cohesion to those with low cohesion. By plotting the kappa agreement for each meeting against the mean score for all the questions in each category, we found that the relationship showed a very distinct characteristic, as shown in Figure 5. We chose to take the average for each score for every question since previous work on cohesion [34] tends to take all attributes of cohesion to be of an additive form.

From analyzing this distribution, we see that agreements about the cohesion levels for each meeting was higher at the

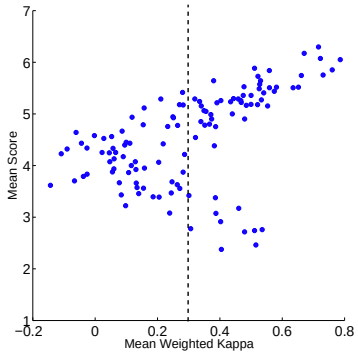


Fig. 5. Distribution of the mean weighted kappa for each meeting and the corresponding mean annotator scores. Lower mean scores correspond to meetings with lower cohesion.

two extremes of the scale. We also observe that meetings exhibiting scores related to higher cohesion tended to be more numerous than those exhibiting lower scores. Finally, we selected 61 points where the kappa agreement was above 0.3 [1], shown by the dotted line in Figure 5. This consisted of 50 points which exhibited high cohesion traits and 11 with low cohesion where all 12 teams were represented. These 61 points were used for our experiments on automatically estimating whether a group had high or low cohesion.

1) Social vs Task Cohesion Analysis: The annotations were analyzed to see if there were any trends between the questionnaire scores for high and low cohesion meetings after partitioning the questions according to whether they represented task or social cohesion better [9], [35]. Given the findings from the psychology literature, we hypothesized that meetings that were not very cohesive could exhibit behavior that was highly socially cohesive but could score lower in task cohesion [39]. This was because the data we used was made of groups of volunteers, most of whom knew each other well as friends and colleagues. For some teams the atmosphere felt quite informal, with laughter and joking. On the other hand, in terms of meetings that exhibited highly cohesive behavior, we hypothesized that those scoring highest could potentially score higher on task cohesion and slightly lower on social cohesion as the participants would be more concentrated on the task rather than on each other [39].

The questions from the annotated questionnaire were divided depending on whether they were more indicative of task or social cohesion (see Appendix). Questions that were assigned to the social cohesion category were related to aspects such as whether the teammates appeared to be involved/engaged in the discussion, have good rapport, or whether participants were in tune with each other. For the task cohesion case, questions such as whether the group appeared to share the responsibility/purpose/goal or intentions for the task, whether the morale was high, or whether teammates were collaborative [35]. In all, 17 questions were used for the overall social cohesion score while 8 were used for the task cohesion condition. Two questions about leadership were removed since one requested a yes/no answer and the other was only answered if it was considered that there was a strong leader in the group. In addition there was a question directly asking about the cohesion levels, that was also not used because this could have been interpreted both as a question about social and/or task cohesion.

The human annotations were processed as follows. Firstly, the mean score per question for each meeting was taken. Then, the lowest scores over all social or task cohesion questions per meeting segment were calculated. These are shown for both low and high cohesion meetings in Figure 6 and constitutes, on average, the annotator's most pessimistic scores for a given meeting. From Figure 6, we observe that for the low cohesion meetings, the scores for both social and task cohesion questions tend to be lower (1.39 and 1.72 on average, respectively) than for those meetings labeled as high cohesion (3.81 and 4.57 on average, respectively). Also, task cohesion questions more often had higher scores than social cohesion questions. If we take the difference between the mean lowest social and task cohesion scores across the high and low cohesion data, we see that the difference between task and social cohesion is higher for the high cohesion meetings (0.76) than for the low cohesion case (0.34). This suggests that the meeting segments we used were generally more task-cohesive in the high cohesion case, which could be considered to be in line with the nature of the AMI meetings, which were all task-based. In addition, it is likely that behavioral attributes that correlate to task cohesion would be more difficult to automatically estimate than social cohesion since the former is less readily measurable from the interaction data.

V. Nonverbal Cue Extraction

The discussion in Section II-A shows the variety of definitions of cohesion. Often, the measures used are based around questions asked about the affective state of the members of the group. The main difference between the works in psychology and that presented here is that we use annotations taken from external observers, so much more emphasis is placed on third-party observations of instantaneous interactions within the closed world of each meeting segment that was used in our data set. Therefore the inspiration for the features described in this section represent the elements of cohesion related to the observable aspects of the construct such as rapport, involvement, mimicry etc, which have been found to be correlated to aspects of cohesion.

For all the following cues that are described below, the features were extracted either on a participant/individual, or group level. For individual-based features, the name labeling follows a protocol for which the first term corresponds to the method of representation for the group through some function of each individual's feature value in the meeting while the second term indicates how the feature values for each person is represented. For features extracted on an individual level, each person's mean feature value was calculated first before the mean (**MeanMean**), minimum (**MinMean**), and maximum (**MaxMean**) of the 4 values (one for each team member) was calculated. If group-level features were extracted, the sum (**Total**), variance (**Var**), minimum (**Min**), median (**Med**), and maximum (**Max**) were calculated. If only one term is shown, this means that the representation was taken from

[1]Note that a kappa agreement of 0.3 is typically considered low.

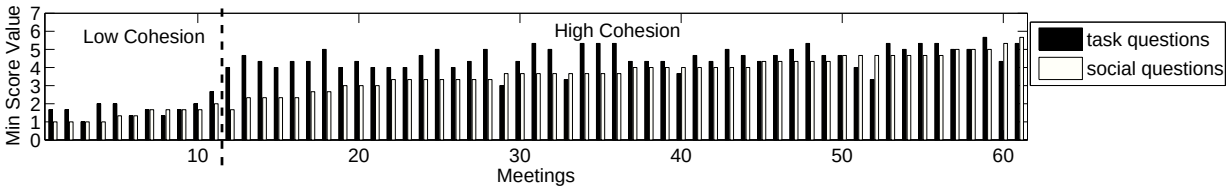


Fig. 6. Separation of the lowest mean annotator scores across the relevant questions for each of the meetings, depending on whether the question was more task or socially oriented. See the Appendix for the social and task cohesion questions.

pooling together all values for every individual in the group into a group-based representation. Due to the large number of features that were used for our experiments, there was not room to show the performance of each feature type. We include the names here to indicate that different variants were tried out systematically. For space reasons, and for narrative clarity, only those with a significant effect on the cohesion estimation performance are finally presented and discussed.

A. Audio Cues

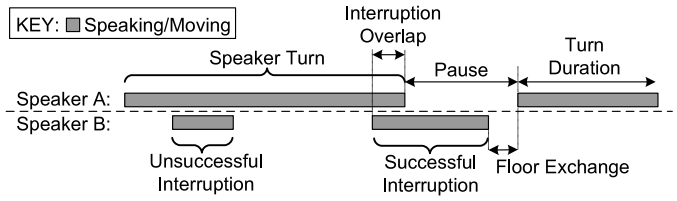


Fig. 7. Summary of different turn-based features that were extracted. They can be used analogously for visual activity.

Audio cues were extracted by firstly automatically segmenting the audio signal from each headset microphone using the voice activity detection method of [17]. From this, various cues related to the speaking activity in the group was generated. We also extracted cues based on the prosody of the original vocal signal. We designed the cues based roughly on the questions that we used in the questionnaire (see Appendix), hypothesizing that each cue would be able to distinguish between perceptions of high and low cohesion.

Inspired by previous work [6], [7], [24], [29] on the power of using turn-taking and other audio features to measure interactive behavior, we have devised audio features which are based on this. They can be roughly summarized into 5 categories: periods between each individual's turns, times between floor exchanges, turn durations, overlapping speech, and prosody. Figure 7 illustrates some of the turn-based features that will be extracted. For space reasons we do not provide equations for each of them but present them conceptually.

Pauses Between Individual Turns

To quantify the degree of participation of each individual in the group, we use features related to the pause time between each person's turns.

- The pause time (**PauseTime**) between a person's turn and that same person's next turn. We hypothesize that during meetings perceived as having high cohesion, there tends to be more equal participation among the participants so everyone will take a lot of turns but will also need to allow time for their fellow teammates to talk.
- The turn to pause ratio (**TurnPauseRatio**) represents the ratio of the amount of time spent speaking and the time between a person's speaking turns. It is hypothesized that those who are more involved in the discussion may tend to have more equal amounts of both speaking and listening; the active listening time is coarsely approximated by the time between a person's turns.

Pauses between floor exchanges

To measure the flow of the conversation, we measure the time between exchanges of the floor to see how quickly turns are passed between participants in the meeting.

- The silence time (**Silence**): It is likely that there will be more periods of silence when teammates are uncomfortable, such as if unacquainted participants are meeting for the first time. These periods of silence can also include times when someone pauses and carries on speaking and no one else tries to grab the floor.
- The time between all floor exchanges (**FloorExch**) approximates whether the pace of the conversation is fast or slow. Conversations at a fast pace will tend to have less time between floor exchanges. Floor exchanges that occur if the person taking over the floor begins before the other has stopped are not considered.

Turn lengths

A turn is a continuous time interval when someone is speaking or their binary speaking activity is 1.

- The turn durations (**TurnDuration**) are hypothesized to be approximately equal for all participants in highly involved conversations.
- The speaking time (**SpeakingTime**) over all teammates may be higher for highly cohesive groups since there would be more activity in a meeting.
- The ratio of short to long turns (**ShortLongTurnRatio**) represents the number of times that someone talks for a long time compared to the number of times they speak for a duration less than a thresholded time (which coarsely approximates the length of a backchannel). This represents the group's ability to provide ideas to the discussion, compared to just giving feedback to other participants by using shorter turns.
- The total number of short turns, (**BackChannel**) or what what could be considered backchannels. We would expect that highly cohesive meetings would contain more team members giving each other positive feedback. Therefore, there will likely be a higher number of back-channels.

Overlapping speech

A period of overlapping speech occurs when one or more people speak at the same time. Overlapping speech can be symptomatic of conflict [37], engagement between participants [36], or is often used for providing backchannels for partici-

pants who do not want to take over the floor.

- The total overlap time (**Overlap**) measures the amount of time that at least two people are speaking at the same time. We hypothesize that more overlapping speech might be due to conflict so the cohesion will tend to be lower.
- The successful interruption overlap time (**InterruptionOverlap**) represents the total amount of time in the meeting that accounts for overlapping speech where one person successfully interrupts someone else. (**Interruption**) represents the same feature when a successful interruption is treated as a event rather than as time.
- The unsuccessful interruption overlap time (**FailedInterruptionOverlap**) represent the total amount of time in the meeting that someone talks but does not take over the floor from the speaker. Such overlaps can be caused both by back-channels but also failed challenges for the floor. (**FailedInterruption**) represents the same feature treated as an event rather than as time.

Prosody

Aside from turn-based features, prosodic cues can measure aspects of a person's vocal signal [33].

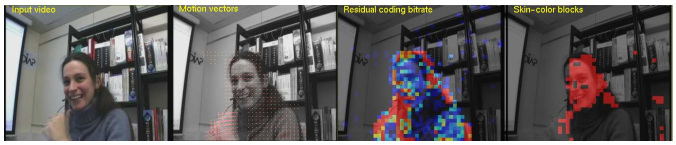
- The speaking energy (**Energy**) coarsely approximates levels of vocal excitement or effort. It is expected that a group with high cohesion will tend to have higher levels of speaking energy on the whole. The energy was computed by taking the sum of the absolute speech signal values over a 32ms sliding frame.
- The speaker overlap energy (**OverlapEnergy**) represents the average energy that is observed for any participant when they are speaking at the same time as at least one other person. Speech overlaps can occur for reasons such as agreement, disagreement, or backchannels. It is expected that during periods of conflict, the speech energy of one or all the people speaking during the overlap period to be relatively high, compared to those who speak over each other for collaborative narrative reasons.
- The speaking rate (**SpeakingRate**) for any person represents the pace of the conversation. It is computed using the mrate estimator by Morgan [31]. For meetings with high cohesion or feelings of being in sync, it is likely that the speaking rate will be relatively high compared to meetings with low cohesion.
- The speaking rate during overlapping speech (**OverlapSpeakingRate**) extracts features related to the nature of the speaking rate during periods when more than one person is speaking. It is expected that the speaking rate during periods of overlapping speech would be higher if the conversation flow or rapport is high.

From Single Features to Ranked Vectors

Some of the features described above can be used in a different framework, where rather than extracting a single scalar value for each meeting, we take each the feature value for each participant in the meeting and concatenate them in descending order to create a feature vector of ranked values.

Overall Group Distributions and Relational Features

Group-based distributional features were computed by taking all the values for that cue, over all frames for each person. They were amalgamated into a histogram to approximate the distribution of the feature for the group.



(a) (b) (c) (d)

Fig. 8. Compressed domain video feature extraction. (a) Original image, (b) Motion vectors, (c) Residual coding bit-rate, (d) skin-colored regions.

As well as accumulating group-based distributions of all the features, the turn-taking patterns were also coded in terms of a matrix, representing how often each person speaks after everyone else in the group (**WhoSpeaksNextMatrix**). This feature is a coarse measure of interpersonal influence since someone is more likely to speak after someone who has influenced them during the conversation. The idea is similar to the influence model by Basu et al. [3].

B. Video Cues

Visual activity for each person in the meeting was extracted from the close view cameras (see Figure 2 and 8) using the compressed domain processing devised by Yeo and Ramchandran [38]. Video streams that have been compressed using MPEG4 encoding contains a collection of group-of-picture (GOP) which is structured with an Intra-coded frame or I-frame while the rest are predicted frames or P-frames. Motion vectors, illustrated in Figure 8(b), are generated from motion compensation during video encoding; for each source block that is encoded in a predictive fashion, its motion vectors indicate which predictor block from the reference frame (in this case the previous frame for our compressed video data) is to be used. After motion compensation, the DCT-transform coefficients of the residual signal (the difference between the block to be encoded and its prediction from the reference frame) are quantized and entropy coded. The *residual coding bitrate*, illustrated in Figure 8(c), is the number of bits used to encode this transformed residual signal. While the motion vector captures gross block translation, it fails to fully account for non-rigid motion such as lips moving. On the other hand, the residual coding bitrate is able to capture the level of such motion, since a temporal change that is not well-modeled by the block translational model will result in a residual with higher energy, and hence require more bits to entropy encode.

Since we hope to capture subtle changes in visual behavior, we used the average residual coding bitrate averaged over the skin-colored regions (see Figure 8(d)) of each close-view camera to create a frame-based representation of personal visual activity. This extracted feature vector has similar properties to the speaking energy and could therefore be manipulated analogously to the audio cues by simply replacing the speaking activity vector by the corresponding visual activity.

We found in previous work [24], that using analogous cues for the visual features allowed for a systematic way of comparing both audio and video features. We consider the raw visual activity values generated from the residual coding bitrate represents a form of motion energy, which can be analogous to speaking energy from the vocal signal. The only

audio cue where an analogous video cue was not generated was for the speaking rate. For reasons of simplicity and clarity, the same feature names will be used when describing the video features but indications will be made, where appropriate, to differentiate video and audio cues.

C. Audio-Visual Cues

Different audio-visual cues were extracted by combining the audio and video activity of each participant. Features were designed based on the idea of trying to capture patterns related to the audio-visual interactions between each person.

Motion During Overlapping Speech

Overlapping speech can be treated as periods of dominance assertion [37] or collaboration [36]. Measuring the amount of visual activity of each person during periods of overlapping speech (**OverlapMotionEnergy**) is likely to indicate the degree of involvement in the meeting.

Motion When not Speaking

The amount of visual activity when a person is not speaking (**SilentMotion**) indicates the level of correlation between a person actively listening while others speak. It is likely that when someone speaks, active listeners may be more visually active (e.g. nodding or shaking their head) than disinterested participants.

Audiovisual Synchrony

Audiovisual synchrony either for the same person or between people is a good indicator of rapport and comfort. It is said that those who have high self-synchrony tend to be more at ease [15] and that those that get along well tend to be well synchronized together [7], [27]. Self synchrony is defined as the synchrony between vocal and/or gestural behavior of the same person. The mutual information [16] for a sliding window of (4s with a shift of 2s) between speaking and visual activity was accumulated from distributions for each participant to form a measure of self-synchrony (**SelfSync**). In addition, the average mutual information for every pair-wise combinations of audio and visual activity features between participants were accumulated. This formed the basis for cues that measured the degree of inter-personal synchrony (**InterPersonalSync**) when the mutual information was computed across differing modalities. In keeping with the findings of Campbell [7], we found that the mutual information for self-synchrony tended to be higher than inter-personal synchrony of any combination of people and modality.

VI. Estimating High and Low Cohesion Meetings

We started initially by using a simple algorithm to estimate whether a meeting had high or low cohesion. For each feature, the mean value for each class was calculated and then a threshold was generated using the mean of the two values.

To minimize problems with over-fitting the data of the high cohesion class, which had many more data points, the high cohesion data was randomly sampled so that there was an equal number of data points in each class. The experiments were carried out using a leave-one-out approach to separate the test and training data. Finally, for each feature and each test data point, the experiments were carried out 100 times to account for variations in the sampling process.

The final performance is given as an average of these trials. In addition, to study the improvement that could be obtained from more powerful supervised methods, we performed the same experiments with the same train-test data partitioning using support vector machines (SVMs) and a linear kernel. We hoped that using both classification methods would allow us to better analyze which features, if any would represent the data better. In addition, the simple method uses much fewer computations to train the system. For features that used cues related to a coarse approximation of backchannels, we found empirically that 4s was a good threshold to use [24].

VII. Results

The results tables presented in this section have been ordered for easy comparison between tables. Each feature type has been clustered according to the descriptions in Section V. Analogous groups of video features have been renamed accordingly. Since a large number of features were tested the tables in this section provide a selection of the best performing and also those that are interesting to compare with results from other modalities or methods.

A. Estimating Cohesion using Audio Cues

1) Naïve Classifier Results: The results using the simple approach is shown in Table I. The second and third columns of the table show the average number of times that the meeting is correctly classified as having low or high cohesion, respectively. The total number of data points in the low and high cohesion classes was 11 and 50 respectively. The fourth column shows the average classification accuracy across all 100 trials and the final column shows the standard error. On the whole, our features performed well, achieving performance significantly above the baseline when one class is chosen randomly (50%). The best performing feature (90%) was **TotalPauseTime**, which always had a high value for highly cohesive meetings. This feature is particularly interesting because it represents how actively attentive each team member is to the others in the group. The attentiveness can be shown through taking and discussing further a team member's ideas or providing many back-channels. This feature will have low values if one person tends to talk a lot while the others don't say anything. The other feature that also performed very well was **MaxOverlapSpeakingRate** (89%). Interestingly, the former feature captures the total time that all participants spend not talking between taking a turn, while the latter captures the times when more than one person is talking at the same time, representing both active as well as passive participation. The third best performing feature was the **MinMeanTurnDuration** (87%) feature. We would expect that in high cohesion meetings, everyone is participating a lot so the minimum average turn length will tend to be higher.

We kept a record for each feature type of how consistent it was in terms of its orientation relative to the classes as a consequence of the random subsampling of the high cohesion data during training. We would expect stable features to exhibit the same trends consistently regardless of the training data so that for example, features that should have a high value to indicate high cohesion always did so. We found that for the top three performing features, the estimation of whether the

Features	Low	High	Class. Acc.(%)	Std Err. (%)
Pauses Between Individual Turns				
TotalPauseTime	82	92	**90**	**0**
MinPauseTime	83	79	80	3
TurnPauseRatio	43	80	73	2
MaxTurnSilenceRatio	35	94	83	0
Pauses Between Floor Exchanges				
MedFloorExch	36	86	77	4
TotalSilence	48	51	51	11
Turn Lengths				
ShortLongTurnRatio	54	69	66	6
MinMeanTurnDuration	82	88	**87**	**1**
TotalSpeakingTime	72	56	59	6
BackChannels	85	67	70	4
Overlapping Speech				
TotalOverlap	87	77	79	3
Interruption	89	75	77	2
InterruptionOverlap	91	78	80	4
FailedInterruption	85	67	70	4
FailedInterruptionOverlap	81	63	66	3
Prosodic Cues				
TotalEnergy	70	54	57	3
MinMeanEnergy	81	66	68	5
VarEnergy	76	49	54	3
MaxOverlapEnergy	97	75	79	6
MinMeanSpeakingRate	91	73	76	1
MaxOverlapSpeakingRate	81	91	**89**	**1**

TABLE I
RESULTS USING SINGLE AUDIO FEATURES AND THE SIMPLE BINARY CLASSIFIER. THE SECOND AND THIRD COLUMN SHOWS THE PERCENTAGE OF CORRECTLY CLASSIFIED MEETINGS AS LOW COHESION (11 SEGMENTS) OR HIGH COHESION (50 SEGMENTS) RESPECTIVELY. THE FOURTH AND FIFTH COLUMNS SHOW THE OVERALL MEAN CLASSIFICATION ACCURACY AND THE CORRESPONDING STANDARD ERROR.

meeting segment showed high cohesion was consistently based on whether the feature value was above the threshold. This is also consistent with the corresponding standard errors.

An interesting result was also obtained with the **TotalOverlap** feature, which we expected would be negatively correlated with cohesion. However, it appears that more overlap is a reliable sign of high cohesion in our data. This aligns with findings in social psychology that interruptions are indicative of good rapport such as when people are able to finish each other's sentences [36].

2) SVM Results: Results when using the SVM are shown in Table II. Here we also show results using the feature vectors and the overall distribution of some features for the whole group. In general the ranked participant features tended to perform at least as well as their corresponding scalar counterparts. However, the best ranked participant feature, **TotalPauseTime**, with a performance of 86% classification accuracy did not outperform the scalar version of **TotalPauseTime**, which had a classification accuracy of 90%.

Group-based distributional features also performed well with the best performance at 79% achieved by the **OverlapEnergy** distribution and the **OverlapSpeakingRate** distribution. Surprisingly, the distribution of the speaking energy on its own resulted only in a classification accuracy of 49% while generating a distribution just based on the energy during overlapped speech led to a significant performance improvement (79%).

Features	Low	High	Class. Acc.(%)	Std. Err. (%)
Pauses Between Individual Turns				
TotalPauseTime	82	91	**90**	**1**
MinPauseTime	82	82	82	2
TurnPauseRatio	51	73	71	3
MaxTurnSilenceRatio	51	88	77	2
Pauses Between Floor Exchanges				
MedFloorExch	41	79	73	7
TotalSilence	69	40	52	10
Turn Lengths				
ShortLongTurnRatio	54	65	66	6
MinMeanTurnDuration	82	88	**87**	**1**
TotalSpeakingTime	80	54	60	6
BackChannels	75	63	72	2
Overlapping Speech				
TotalOverlap	74	71	81	2
Interruption	79	68	78	1
InterruptionOverlap	89	70	82	3
FailedInterruption	55	66	72	2
FailedInterruptionOverlap	70	58	67	2
Prosodic Cues				
TotalEnergy	72	54	57	3
MinMeanEnergy	81	67	69	5
VarEnergy	78	49	55	3
MaxOverlapEnergy	86	78	84	4
MinMeanSpeakingRate	90	75	77	1
MaxOverlapSpeakingRate	81	91	**89**	**1**
Ranked Participant Features				
MeanDuration	82	85	84	5
MeanTurnMeanPauseRatio	84	83	83	6
TotalPause	80	88	86	**4**
TotalOverlapEnergy	92	79	82	4
TotalEnergy	85	74	76	3
Interruption	94	79	82	2
TotalOverlapSpeakingRate	100	87	**89**	4
Group Distribution Features				
OverlapEnergy	85	78	79	7
PauseDuration	98	71	76	6
OverlapSpeakingRate	96	75	79	5
Relational Features				
WhoSpeaksNextMatrix	100	83	86	4

TABLE II
SUMMARY OF THE AUDIO RESULTS OBTAINED USING AN SVM CLASSIFIER. SEE TABLE I FOR COLUMN CONTENTS DESCRIPTIONS.

The matrix that captures who speaks after who (**WhoSpeaksNextMatrix**) performed better than all of the group distribution features (86%) but did not outperform **TotalPauseTime**. The performance is still comparable and highlights that the way that people exchange turns in a meeting segment is significantly correlated with the cohesion levels. In addition, it was the only case where all low-cohesion meetings were always correctly identified, regardless of the corresponding high cohesion meetings that were used for training the data. Overall, despite the use of a more powerful classifier, using SVMs did not significantly improve the performance of the various features that were tested. This may be due to factors such as the small training data size. It also suggests that the best features may indeed be discriminative for the target task.

B. Estimating Cohesion using Video Cues

1) Naïve Classifier Results: We applied the same features extracted from video to the same classification task, and the

Features	Low	High	Class. Acc. (%)	Std Err. (%)
Pauses Between Individual Turns				
TotalPauseTime	59	58	58	11
MinPauseTime	70	58	60	2
Pauses Between Floor Exchanges				
MedFloorExch	47	45	45	7
TotalSilence	72	53	57	3
Turn Lengths				
ShortLongTurnRatio	47	47	47	6
Overlapping Speech				
TotalOverlap	68	57	59	8
FailedInterruptionOverlap	52	66	64	6
Prosodic Cues				
TotalEnergy	42	83	**76**	**3**
MinMeanEnergy	42	80	73	**8**
VarEnergy	57	86	**81**	9
MaxOverlapEnergy	39	65	60	6

TABLE III
RESULTS OBTAINED USING NAÏVE CLASSIFIER AND VIDEO FEATURES. SEE TABLE I FOR DESCRIPTIONS OF THE CONTENTS OF EACH COLUMN.

Features	Low	High	Class. Acc. (%)	Std Err. (%)
Pauses Between Individual Turns				
TotalPauseTime	64	58	59	11
MinPauseTime	71	58	60	2
Pauses Between Floor Exchanges				
MedFloorExch	56	43	46	9
TotalSilence	73	53	57	3
Motion Turn Lengths				
ShortLongTurnRatio	55	47	49	4
Overlapping Visual Activity				
TotalOverlap	70	57	60	8
FailedInterruptionOverlap	55	66	64	6
Visual Energy Cues				
TotalEnergy	51	80	75	3
MinMeanEnergy	45	78	72	8
VarEnergy	40	85	77	17
MaxOverlapEnergy	45	64	61	19
Ranked Participant Features				
TotalEnergy	85	82	**83**	**4**
FailedInterruptionOverlap	74	69	70	15
TotalPause	68	66	66	4
Group Distribution Features				
TurnDuration	81	50	55	6
PauseDuration	86	52	58	5
Relational Features				
WhoSpeaksNextMatrix	88	63	68	8

TABLE IV
RESULTS OBTAINED USING AN SVM CLASSIFIER AND VIDEO FEATURES. SEE TABLE I FOR DESCRIPTIONS OF THE CONTENTS OF EACH COLUMN.

Feature	Low	High	Class. Acc. (%)	Std Err. (%)
Synchrony				
MaxInterPersonalSync	72	65	66	3
MeanSelfSyncJoint	58	64	63	3
Motion When not Speaking				
MeanSilentMotion	43	84	76	2
MaxMeanSilentMotion	57	86	81	1
Motion During Overlapping Speech				
MaxOverlapMotionEnergy	39	65	60	20

TABLE V
RESULTS USING AUDIO-VISUAL FEATURES WITH THE NAÏVE BINARY CLASSIFIER. SEE TABLE I FOR COLUMN CONTENTS DESCRIPTIONS.

results using the simple classifier are shown in Table III. Here we see that the best performing features are different, (81%) when using **VarEnergy**. This is likely as the people who are actively involved in both listening and speaking often have a large range of visual activity or motion energy [18]. Other features which are based on visual activity generally perform well however, if we observe the performance for each class, we see that the performance on low cohesion meetings is considerably worse. Compared to the audio features, fewer video features exhibited consistent trends for the same cohesion level. For the **VarEnergy** feature, high cohesion meeting segments that were classified correctly were consistently of low value. This aligns with the idea that those more cohesive interactions will tend to have convergent behavior as a manifestation of mimicry [27]. If we compare the performance of the audio features with analogous video cues, we see that in general, the video cues do not outperform the audio cues. However, visual energy tends to produce significantly better performance than speaking energy.

2) SVM Results: Table IV summarizes some of the performance of the same visual features when trained with an SVM. The best performing features in this case was one of the ranked participant features using the **TotalEnergy** for each participant (83%). While it outperforms its audio counterpart, overall, it does not beat the performance of the audio feature **TotalPauseTime** and its corresponding standard errors also shows that it less stable than the audio feature. **VarEnergy** also performed well but was also less stable than the corresponding audio feature. All other ranked participant features which are related to discrete visual activity tended to perform significantly worse than those using the visual activity as a raw signal. Finally, features based on the overall group distribution do not perform as well as the other features.

C. Estimating Cohesion using Audio-Visual Cues

1) Naïve Classifier Results: Finally, we performed the same comparative experiments between our simple and the SVM classifier for audio-visual cues. Table V shows the results using the simple binary classifier. The best performing features **MaxMeanSilentMotion** with a classification accuracy of 81%. Overall however, the best performing feature still does not outperform the audio-only cue **TotalPauseTime**. It is also interesting to see that features relating to the inter-personal synchrony perform worst. The self-synchrony feature also performed comparably to the other audio-visual interpersonal synchrony measures, indicating that it seems to have some discriminative power for measuring cohesion. Note also that for classifying the low cohesion meetings, the synchrony features tended to perform slightly better than the rest.

2) SVM Results: If we compare the performance of the audio-visual cue performance when SVMs are used, as shown in Table VI, we see that the performance is comparable to using the simple method. The ranked participant feature, **TotalOverlapEnergy** also performed well, but did not outperform the audio-only feature **TotalOverlap**.

D. Results Summary

Overall, the results show the discriminative power of some features derived from single and joint modalities. Figure 9

Feature	Low	High	Class.Acc. (%)	Std Err. (%)
Synchrony				
MaxInterPersonalSync	73	65	67	3
MeanSelfSyncJoint	59	64	63	3
Motion When not Speaking				
MeanSilentMotion	52	80	75	3
MaxMeanSilentMotion	70	84	81	2
Motion During Overlapping Speech				
MaxOverlapMotionEnergy	45	64	61	19
Ranked Participant Features				
TotalOverlapEnergy	80	74	75	7

TABLE VI
RESULTS USING AUDIO-VISUAL FEATURES AND THE SVM CLASSIFIER. SEE TABLE I FOR DESCRIPTIONS OF THE CONTENTS OF EACH COLUMN.

summarizes some of the differences between the classification accuracy when considering single modalities. In particular, the SVM and simple binary classifiers both work comparably well for most scalar features, which suggests that the features themselves are discriminative. Also, it is interesting to see that when raw visual activity features are used to represent the visual activity of the participants, the performance is close to the top-performing features using audio cues.

We performed significance testing using a two-sided t-test on all feature performance comparisons that have been mentioned in this section and found that all results were significant ($p<0.001$). The tests were carried out with each pair of features by taking the performances that were generated from all 100 runs of the leave-one-out resampling process.

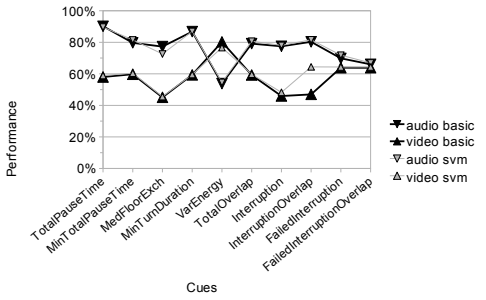


Fig. 9. Comparison of the performance of selected audio and visual features. The performance using both the naïve and SVM classifiers are also shown.

If we refer back to our original goals listed in the introduction, we have shown that external observers can perceive and agree on differing levels of cohesion. In addition, we were able to use nonverbal cues to identify two levels of perceived cohesion. However, we were unable to find audio-visual features that would out-perform single-modality features.

VIII. DISCUSSION AND CONCLUSION

In this paper, we have demonstrated promising results on automatically estimating high and low levels of group cohesion using automatically extracted audio, video, and audio-visual cues. 240 minutes of data were used to collect third-party human perceptions of cohesion, and this data was analyzed to define a data set for the evaluation of automated measures of cohesion. The best performing feature was a scalar audio cue which accumulated the total pause time between each individual's turns during a meeting segment. Using this method, meetings were classified correctly 90% of the time. Video cues also performed quite well, with a top performance of 83% when using the ranked participant feature that used the total visual activity for each person in the meeting. In terms of audio-visual cues, the best performing accumulated the visual activity during periods of overlapped speech, achieving a classification accuracy of 82%.

Our results have shown that automatically extracted behavioral cues can be used to estimate perceived levels of cohesion in meetings based on questionnaire terms that were not directly labeling the cues themselves. To our knowledge this is the first time that an attempt has been made to use automatically extracted nonverbal cues to estimate group cohesion levels and our results indicate a strong correlation between cohesion levels and turn-taking patterns. Our work also attempts to correlate systematically, nonverbal cues with perceived group cohesion which, to our knowledge has not yet been studied by social scientists.

Furthermore, this study constitutes the first computational exploration into a behavioral construct that while important in practice, remains to be fully understood by social scientists. Future work in this area should investigate the design and annotation of group meetings both from an internal and external perspective, to mirror work in the social sciences. The use of either internal or external observations of a group in terms of cohesion remains an open debate [11] and suggests that there may be interesting models that can be formed based on how the use of automatically extracted cues may be used to differentiate internal and external perceptions of a group. For example, highly socially cohesive groups may tend to have the perception that they are very task-cohesive while external observers may see that the group spends more energy enhancing or reinforcing their social/emotional bonds to the detriment of the task. This may help to train teams to align internal perceptions of how their team is performing, more objectively so that improvements to task cohesion can be made. In addition, the data that we have used captures behavior that was carried out by groups of volunteers so the motivations really for joining a group and remaining loyal to it is yet to be explored. Furthermore, future analysis of automatically estimated cues for estimating social and task cohesion levels may contribute to explaining the cohesion-performance relation [39], [40]. Finally, the models that were used here were secondary to the investigation of cues. It would be beneficial in the future to investigate further, how more sophisticated models could be used to capture the cohesive interactive behavior of teammates such as their interpersonal synchrony and how the social and task cohesion behavioral elements of teams can be estimated separately.

ACKNOWLEDGMENTS

This work was done while H. Hung was working at Idiap Research Institute. The authors acknowledge the support of the European Project AMIDA (Augmented Multi-party Interaction with Distance Access) and the Swiss National Science Foundation through the NCCR IM2 (Interactive MultiModal Information Management).

IX. APPENDIX

Below are the questions that were used for the human annotations of our data. They have been organized in terms

of task and social cohesion, followed by all other questions. The numbers before each question indicate the ordering of the questions in the original questionnaire. The source for each term is provided at the end of each question. Questions without a citation were chosen from our own interpretation of cohesion and how they could be related to nonverbal cues.

Task cohesion

2. Does the team seem to share the responsibility for the task? [5]

3. Do you feel that team members share the same purpose/goal/intentions? [23]

4. Overall, how enthusiastic is the group? [5]

7. How is the morale of the team? [5], [13]

8. Overall, do the members give each other a lot of feedback? [5]

19. Overall, do the team members appear to be collaborative? [5]

27. Does every team member seem to have sufficient time to make their contribution? [5], [9]

Social cohesion

1. Overall, do you feel that the work group operates spontaneously? [5]

5. Overall, how involved/engaged in the discussion do the participants seem? [5]

6. Do the team members seem to enjoy each other's company? [5], [9]

9. Does the team seem to have a good rapport? [5]

12. Overall, does the atmosphere of the group seem more jovial or serious? [9]

13. Overall, does the work group appear to be in tune/in sync with each other? [5]

15. Overall, does there appear to be equal participation from the group? [5]

16. Overall, do the group members listen attentively to each other? [5]

17. Overall, does the team appear to be integrated? [13]

18. Do the team members appear to be receptive to each other? [5]

19. Do the participants appear comfortable or uncomfortable with each other? [5]

21. Is there a strong sense of belonging in the work group? [9], [13]

22. Overall, does the atmosphere seem tense or relaxed?

23. Does the work group appear to have a strong bond? [9], [13]

24. How is the pace of the conversation?

25. Overall do the team members seem to be supportive towards each other? [5]

26. How well do you think the participants know each other?

Miscellaneous

Is there a leader in the group? If you answered YES , does the leader bring the rest of the group together?

Overall, how cohesive does the group appear? [35]

REFERENCES

[1] N. Ambady and H. M. Gray. On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality & Social Psychology*, 83(4):947–961, 2002.

[2] J. N. Bailenson, N. Yee, K. Patel, and A. C. Beall. Detecting digital chameleons. *Computer Human Behavior*, 24(1):66–87, 2008.

[3] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, and Alex (Sandy) Pentland. Learning human interactions with the influence model. Technical report, MIT MEDIA LABORATORY TECHNICAL NOTE, 2001.

[4] K. A. Bollen and R. H. Hoyle. Perceived cohesion: a conceptual and empircal examination. *Social Forces*, 69:479–504, 1990.

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

[6] N. Campbell. Individual traits of speaking style and speech rhythm in a spoken discourse. In Anna Esposito, Nikolaos G. Bourbakis, Nikolaos M. Avouris, and Ioannis Hatzilygeroudis, editors, *COST 2102 Workshop (Patras)*, volume 5042 of *Lecture Notes in Computer Science*, pages 107–120. Springer, 2007.

[7] N. Campbell. Multimodal processing of discourse information; the effect of synchrony. *International Symposium on Universal Communication*, 0:12–15, 2008.

[8] J.C. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2005.

[9] A.V. Carron and L.R. Brawley. Cohesion: Conceptual measurement issues. *Small Group Research*, 31(1):89–105, February 2000.

[10] A.V. Carron, L.R. Brawley, and W.N. Widmeyer. *Advances in sport and exercise psychology measurement*, chapter The measurement of cohesiveness in sport groups, pages 213–226. Fitness Information Technology. Morgantown, 1998.

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

[12] J. Cassell, A. J. Gill, and P. A. Tepper. Coordination in conversation and rapport. In *EmbodiedNLP '07: Proceedings of the Workshop on Embodied Language Processing*, pages 41–50, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[13] Wynne W Chin, Wm David Salisbury, Allison W Pearson, and Matthew J Stollak. Perceived Cohesion in Small Groups - Adapting and Testing the Perceived Cohesion Scale in a Small-Group Setting. *Small Group Research*, 30(6):751–766, 2009.

[14] J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, (70):213–220, 1968.

[15] W.S. Condon and W.D. Ogston. Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, 143(4):338, 1966.

[16] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, 1991.

[17] J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *INTERSPEECH*, 2006.

[18] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.

[19] C.R. Evans and P. A. Jarvis. Group cohesion: a review and reevaluation. *Small Group Research*, 11:359–370, 1980.

[20] N. Friedkin. Social cohesion. *Annual Review of Sociology*, 30:409–425, 2004.

[21] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[22] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. Virtual rapport. In *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 14–27. Springer, 2006.

[23] J. Griffith. Further considerations concerning the cohesion-performance relation in military settings. *Armed Forces & Society*, 34(1):138–147, October 2007.

[24] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[25] R. E. Kidwell, K. W. Mossholder, and N. Bennett. Cohesiveness and organizational citizenship behavior: A multilevel analysis using work groups and individuals. *Journal of Management*, 23:775–793, 1997.

[26] N. C. Krämer, N. Simons, and S. Kopp. The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In *Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science*, pages 238–251. Springer, 2007.

[27] J. Lakin and T.L. Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14:334–339, 2003.

[28] Caneel R. Madan A. and Pentland A. Voices of attraction. In *proceedings of Augmented Cognition, (AugCog) HCI*, 2005.

[29] Marianne Schmid Mast. Dominance as expressed and inferred through speaking time. *Human Communication Research*, (3):420–450, July 2002.

[30] A. Mikalachki. *Group Cohesion Reconsidered*. University of Western Ontario, School of Business Administration., London, Ontario, 1969.

[31] N. Morgon. mrate estimater, Accessed July 25, 2009.

[32] B. Mullen and C. Cooper. The relation between group cohesiveness and performance: An integration. 115:210–222, 1994.

[33] Alex (Sandy) Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2008.
[34] S. E. Seashore. *Group Cohesiveness in the Industrial Work Group*. University of Michigan Press, 311 Maynard Street, Ann Arbor, Michigan 48108, 1954.
[35] G. L. Siebold. The evolution of the measurement of cohesion. *Military Psychology*, 11(1):5–26, 1999.
[36] D. Tannen. *Gender and Discourse*, chapter Interpreting Interruption in Conversation, pages 53–83. Oxford Univesrity Press, 1993.
[37] C. West and D. H. Zimmerman. *Language, Gender, and Society*, chapter Small Insults: A study of interruptions in cross-sex conversations between unaquainted persons, pages 103–117. Newbury House, 1983.
[38] C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.
[39] S. J. Zaccaro and C. A. Lowe. Cohesiveness and performance on an additive task: evidence for multidimensionality. *Journal of Social Psychology*, 128:547–558, 1988.
[40] S. J. Zaccaro and C. McCoy. The effects of task and interpersonal cohesiveness on performance of a disjunctive group task. *Journal of Applied Psychology*, 18:837–851, 1988.

Hayley Hung Hayley Hung is a Marie Curie post-doctoral research fellow at the University of Amsterdam in The Netherlands. Between 2007-2010, she was a postdoctoral researcher at Idiap Research Institute, Switzerland. She graduated with an MEng degree in Electronic and Electrical Engineering from Imperial College, London in 2002 and PhD in computer vision from Queen Mary University of London in 2007, which was funded by the EPSRC, UK and QinetiQ Ltd. In 2009 she won the Institute of Engineering and Technology (IET) written premium competition. She received the West Midlands Runner Up Young Engineer for Britain Award for designing an automatic aligning antenna in 1997. She was awarded the Lucent Technologies Global Science Scholarship in 1998. She is a member of both the IEEE and IET.

Daniel Gatica-Perez Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001, receiving the Yang Research Award for his doctoral work. He is now a senior researcher at Idiap Research Institute, Martigny, Switzerland, where he directs the Social Computing group. His recent work has developed statistical methods to analyze small groups at work in multisensor spaces, populations using cell phones in urban environments, and on-line communities in social media. He has published over 100 refereed papers in journals, books, and conferences in his research areas. He currently serves as Associate Editor of the IEEE Transactions on Multimedia, Image and Vision Computing, Machine Vision and Applications, and the Journal of Ambient Intelligence and Smart Environments. He is a member of the IEEE.