Noname manuscript No.
(will be inserted by the editor)

Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition

Dairazalia Sanchez-Cortes[1,2] · **Oya Aran**[1] · **Dinesh Babu Jayagopi**[1,2] · **Marianne Schmid Mast**[3] · **Daniel Gatica-Perez**[1,2]

the date of receipt and acceptance should be inserted later

Abstract In this paper we present a multimodal analysis of emergent leadership in small groups using audio-visual features and discuss our experience in designing and collecting a data corpus for this purpose. The ELEA Audio-Visual Synchronized corpus (ELEA AVS) was collected using a light portable setup and contains recordings of small group meetings. The participants in each group performed the winter survival task and filled in questionnaires related to personality and several social concepts such as leadership and dominance. In addition, the corpus includes annotations on participants' performance in the survival task, and also annotations of social concepts from external viewers. Based on this corpus, we present the feasibility of predicting the emergent leader in small groups using automatically extracted audio and visual features, based on speaking turns and visual attention, and we focus specifically on multimodal features that make use of the looking at participants while speaking and looking at while not speaking measures. Our findings indicate that emergent leadership is related, but not equivalent, to dominance, and while multimodal features bring a moderate degree of effectiveness in inferring the leader, much simpler features extracted from the audio channel are found to give better performance.

[1]Idiap Research Institute
Centre du Parc
Rue Marconi 19
Tel.: +41-277-217-748
Fax: +41-277-217-712
E-mail: (dscortes,oaran,djaya,gatica)@idiap.ch

[2]Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

[3]Institut de Psychologie du Travail et des Organisations
University of Neuchatel (UNINE)
Rue Emile-Argand 11
Switzerland
E-mail: marianne.schmid@unine.ch

Keywords emergent leadership · nonverbal behavior · multimodal cues · small group interactions

1 Introduction

Human interactions are rich on several dimensions. They make use of various communication channels in parallel (verbal and nonverbal, audio and visual, etc) to establish relations, to convey thoughts, and emotions in different social situations, ranging from courtship to family, working in teams and building communities. Psychologists have long studied these interactions of varying scale, to understand behavior, motivation, and emergence of interaction patterns (Poole et al, 2004; Salas et al, 2005). One specific setting of importance is the study of small groups (Bales and Strodtbeck, 1951; Gatica-Perez, 2006).

From the viewpoint of social computing research, a domain rooted on significant developments in data recording, automatic audio-visual analysis, and machine learning (Gatica-Perez, 2009), the aim is to automatically infer human social behavior by observing the interaction among people via multimodal sensing devices that capture the various dimensions of human social interaction. With these developments, results traditionally obtained in psychology can now be revisited with automatic analysis techniques.

Analyzing social interactions requires not only the analysis of single communication modalities such as speech, gesture, facial expression, etc., but it requires also multimodal analysis to infer complex social concepts. To achieve meaningful and reliable results, it is of extreme importance to obtain natural interaction data, which is recorded with appropriate sensors that allow automatic analysis. A number of multimodal corpora depicting group interactions is available in the research community (see summary of corpora available on Table 1). Using these corpora, affect and behavioral cues like facial expressions, prosody, turn-taking patterns, head pose, and gestures have been studied. Furthermore, manual and automatic versions of behavioral cues have been used to infer social constructs like influence, performance, and cohesion. The recording solutions have also varied, from wearable devices, i.e., fully portable sensing, to fixed infrastructure-based sensors.

Despite the availability of these corpora, there are several limitations. One of the most important limitations is related to the naturalness of the interaction. As detailed in Section 2, the scenarios used in these corpora range between experimental setups with scripted meetings to completely natural scenarios. In an experimental setup, the recording environment (consisting of cameras and microphones) is set in such a way that they capture data in the best available way, appropriate for analysis. Although an experimental setup is useful to control the recording environment, the obtained results can not be easily applied to real life scenarios. On the other hand, completely natural recordings recorded without any restrictions are hard to process with the current techniques in terms of audio-visual analysis. Moreover, capturing natural recordings with high quality in their own environment is challenging and the

necessary capturing hardware and software is not easily available. Another limitation is about obtaining necessary annotations. It is difficult to obtain conclusive results without annotations, which can be very subjective depending on the task and require multiple human annotators.

As a first dimension of our work, the ELEA AVS corpus that we use in this study addresses some of these limitations. We use a portable recording setup which allows to record a small group meeting anywhere. Although the scenario we apply is not completely natural, in the sense that the participants are gathered for the purpose of data collection and are given a task, the meeting they perform is natural, without any predefined behaviors. The scenario used in the recordings has been specifically designed to study the possible emergence of leaders. An emergent leader is defined as the person that stands for the group during a face-to-face interaction with no hierarchical roles (predefined) and he/she has the group's sympathy to lead (Stein, 1975). Our corpus also includes a number of annotations on several individual and social concepts collected both from the meeting participants and from multiple external viewers.

As a second dimension of our study, we exploit the use of audio, visual, and multimodal features in small group conversations for the estimation of emergent leadership using unsupervised methods. We present an analysis for the identification of the emergent leader using single as well as multimodal features coded from the audio-video streams. In particular, we focus on the study of features that characterize visual attention and speaking activity of group members. Some of these features are derived from classic studies in psychology (Bales and Strodtbeck, 1951; Efran, 1968) but not yet studied in the context of computational inference of emergent leadership. We first present a correlation analysis between the automatically extracted nonverbal features and the concepts related to emergent leadership. The nonverbal features are extracted from single audio and video streams based on speaking activity and visual attention. Then, we study the performance of the nonverbal features in estimating of the emergent leader in the group. We also explore nonverbal features that are multimodal in nature, such as measures of looking at participants while speaking and the visual dominance ratio. Finally, we present effects of possible misalignments in the multimodal features on the estimation performance. We found that emergent leadership in our study is related, but not equivalent, to dominance, and while multimodal features bring a moderate degree of effectiveness in inferring the leader, much simpler features extracted from the audio channel are found to perform better.

This paper is organized as follows: we first present related work in Section 2. In Section 3, we describe the materials and procedure to collect the corpus. Section 4 explains the annotation encoding scheme. We present the nonverbal features in Section 5. The use of single and multimodal features to infer emergent leadership on the ELEA AVS corpus is presented in Section 6. Finally, we present our conclusions in Section 7.

A preliminary version of this work, covering mainly the discussion on the corpus and the annotations (Sections 2-4), was presented in (Sanchez-Cortes et al, 2011a). In the current paper, we present an analysis on the use of speaking

activity features, visual attention features, and multimodal features that rely on the audio-visual synchrony for estimating the emergent leader.

2 Related work

This section reviews existing corpora presented in the literature to study human behavior in small groups. We also briefly present the features and techniques that are used for the analysis of human interactions.

Most of the corpora that have been collected to study behavior in small groups centered their attention on meeting scenarios where realistic rich interacting patterns can emerge. A detailed look into these corpora reveals a variety of design choices. To promote the interaction between participants, either real or scripted scenarios can be used. The recordings can be done with a wide range of audio-visual sensors. The collected data can be annotated for different aspects, in parallel with the research question in mind. Table 1 summarizes the available corpora focused on small group interactions, described in this section.

The VACE meeting corpus has been recorded using real-world scenarios (war games and military exercises) at the Air Force Institute of Technology (AFIT) (Chen et al, 2005). The aim is to understand the structure in meetings where the objectives are clearly defined, the roles and hierarchy are known, and the planning activity is present.

Natural weekly discussions of a research group, with known roles and hierarchy, has been recorded at ICSI's conference room (Janin et al, 2003). The goal of this corpus is to offer resources to improve automatic speech recognition, transcription, prosody, and dialog modeling.

Another corpus collected real and scripted meetings on scenarios such as project planning, military exercises, games, chatting and discussion (Burger et al, 2002). The aim of the ISL corpus is to distinguish between different kinds of meetings by characterizing speaking styles.

In the AMI-12 corpus, collected at the Idiap smart meeting room (Jovanovic et al, 2005), the meeting participants have predefined roles and they follow a script. Apart from audio and video resources, a variety of manual annotations that involve verbal, nonverbal and contextual features are available. To study the analysis of dominance, the DOME corpus includes dominance annotations on a subset of the AMI corpus, containing 10 hours of meetings recorded at the Idiap smart meeting room (Aran et al, 2010). To analyze participants' influence in project scenario meetings, a part of the AMI corpus was analyzed, containing 40 meetings recorded at TNO-Soesterberg (Rienks et al, 2006). Several manual annotations are available for this corpus, mostly derived from the audio channel.

Several studies investigated another dimension of social behavior, related to dominance and influence (Chen et al, 2005; Rienks et al, 2006; Kim et al, 2008). Another approach for capturing small group meetings is to use wearable sensors that are able to gather nonverbal signals and proximity data from

short distance transmitters. In (Kim et al, 2008), a corpus was recorded with a wearable sociometer based on two scenarios: brainstorming and problem solving. The aim is to detect social interactions (including dominance) and to promote group collaboration (through real time feedback). For this corpus, nonverbal features and self-reported dominance annotations are available.

Participants' involvement has also been analyzed in small business meetings. In (Campbell et al, 2006), the ATR corpus is presented, which includes recordings of monthly sessions from a real group project meeting. The main goal of this corpus is to identify the type of participation and the flow of the discourse.

The NTT corpus (Otsuka et al, 2005) was collected with the aim of inferring the structure of the meeting and the participants' roles. The corpus contains discussion scenarios in which no roles were assigned. The collected data includes audio, video, and head directions extracted from sensors.

Among the multimodal corpora in the literature, the closest to our work is the Mission Survival Corpus (MSC-1 and MSC-2) (Mana et al, 2007; Pianesi et al, 2007). The data comprises small groups performing the winter survival task. The MSC-1 focuses on the individual behavior during the decision making process; it includes audio and video recordings of four participants and functional roles annotations. The MSC-2 focuses on analyzing performance, group cohesion, and personality, and used the same video recording resources used in MSC-1; in addition they performed an online 3D multi-person tracking during the interaction. For audio recording they reduced the number of sensors to 4 close-talk microphones and one omni-directional microphone placed on the top of the table. The MSC recordings differ from our corpus in terms of participants, given that participants at MSC-1 knew each other. In terms of settings, both corpora (MSC-1 and MSC-2) used a static setup and all the meetings are recorded in a static location in a smart room.

The aim of the multimodal corpora summarized above is to analyze the multimodal human behavior in diverse settings. For the analysis, researchers extracted a variety of features, most of which have their roots from the related research in social psychology. While verbal features from the transcribed texts are used as well, most of these works focus on the nonverbal features. These nonverbal features include audio features such as speaking activity turns and interruptions; visual features such as head/hand gestures, body posture, and gaze. For modeling the social concepts, various techniques have been used including rule based methods, topic models, support vector machines, etc. Detailed information on state-of-the art features and techniques can be found in extensive surveys on the topic (Aran and Gatica-Perez, 2011; Gatica-Perez, 2009).

Although real scenarios have been recorded and several behaviors that emerge in small group interaction have been analyzed in the literature, the *emergent leadership* phenomenon has only been recently explored in (Sanchez-Cortes et al, 2010, 2011b) through audio or visual nonverbal channels.

Table 1 Corpora available for small-group interaction study. The audio sensors/microphones include CTM-close-talk, EWM-earset wireless, TTM-tabletop, LAM-lapel, SBM-sociometer badge, ARM-microphone array, ODM-omnidirectional, FCM-four-channel cardioid, OTM-Other distantly placed microphones. Video sensors include CU-close-up, VC-video camera, WC-webcamera, C360-360 degree camera. Personality annotations correspond to LCB-Craig's Locus of Control of Behavior scale, E-BFMS-Extroversion part of the Big Marker Five Scales, NEO-FFI-NEO-Five Factor Inventory, PRF-Personality Research Form.

Corpus	**Audio/video**	**Questionnaires/annotations**
VACE (Chen et al, 2005)	up to 8 EWM, OTMs, 1 OD and 1 FC 10 VC	conversation transcripts, dominant speaker, language metadata (e.g. floor control), gesture
ICSI (Janin et al, 2003)	4 to 8 CTM	involvement
ISL (Burger et al, 2002)	3 to 9 LAM 3 VCs	word tokens, turns, question/non-question, disfluency
AMI-12 (Jovanovic et al, 2005)	4 CTM, 4 LAM, 1 ARM 4 CU and 3 VC	conversation transcript, addresses, gaze direction, adjacency pairs (question-answer, statement-agreement)
AMI-40 (Rienks et al, 2006)	1 ARM 4 CU and 3 VC	influence ranking (inter-ranking) dominance
AMI (Carletta et al, 2005)	same as AMI-12 and AMI-40 same as AMI-12 and AMI-40	same as AMI-12 and AMI-40, hand and head gestures
DOME (Aran et al, 2010)	same as AMI-12 same as AMI-12	same as AMI-12, dominance annotations
M4 (McCowan et al, 2005)	12 microphones (ARM and LAM) 3 VC	conversation transcript, word segmentation, interest level
NIST (Garofolo et al, 2004)	3 to 9 CT, LAM and OTMs 5 VC	conversation transcript, speaker segmentation
ATR (Campbell et al, 2006)	1 ARM 1 C360, 1-6 VCs	none
MIT (Kim et al, 2008)	4 SBM	dominance, questions and ideas, team performance
NTT (Otsuka et al, 2005)	4 LAM 3 VC	regime estimates (class + directionality) head direction (from magnetic sensors 6-DOF)
MSC-1 (Pianesi et al, 2007)	4 CTM, 6 TTM and 7 ARM 5 VC, 4 WC	functional relational roles (task area and socio-emotional)
MSC-2 (Mana et al, 2007)	4 CTM, 1 ODM same as MSC-1	personality LCB and E-BFMS, group cohesion, individual and group performance
ELEA (Sanchez-Cortes et al, 2011b)	1 ARM same as AMI-12 and 2 WC	personality NEO-FFI, PRF perceived interaction, ranked dominance

The emergent leadership phenomenon arises from group interactions in which participants do not have roles assigned. Since this appears mostly in newly formed groups, the behavior of a participant during this short interaction makes him/her succeed (or fail) as a leader, without considering past information of competence, related task performance or friendship. On the other hand, personality traits might have an impact on leadership skills(Kickul and Neuman, 2000). In (Sanchez-Cortes et al, 2010) we presented our first experiments on a subset of the ELEA corpus. Later on, the full ELEA corpus and a comprehensive study on emergent leadership estimation using communicative nonverbal features was presented in (Sanchez-Cortes et al, 2011b), where performance of audio and visual activity features were described separately as well as aggregated through feature fusion. The details of our experience collecting the ELEA corpus and a brief analysis between emergent leadership and its possible association with personality was presented at the MMC Workshop in 2011 (Sanchez-Cortes et al, 2011a).

In this work, we report the performance in the emergent leader inference using social attention automatically extracted from audio-visual features (e.g. looking at participants while speaking) and using a subset of the ELEA corpus. Additionally, we describe the annotations collected from external observers and we present the perception of the emergent leader from the external observers' point of view.

3 The Emergent Leadership Synchronized Corpus

The ELEA AVS corpus is a subset of selected recordings from the ELEA corpus (Sanchez-Cortes et al, 2011b). This subset corresponds to recordings using a fully portable setup, with no video frame dropping. The recordings are audio-video synchronized, allowing multimodal analysis of the emergence of leadership. The corpus consists of 22 meetings (19 meetings with four participants and 3 meetings with three participants).

For the group interactions, three or four people are seated around a table, and the audio and video is recorded, while the participants perform a winter survival task. Before and after the task, the participants fill several questionnaires to be used as ground truth in the analysis of emergent leadership and related concepts. The total duration of the ELEA AVS corpus is approximately five hours. We describe our practical experience with its design and implementation, and discuss results on emergent leaders inference by automatically extracted nonverbal features.

Sensing infrastructure: To collect the audio, we used Dev-Audio's Microcone, a commercial portable microphone array designed to record group focus interactions (McCowan, 2011). The recording device was selected considering its portable nature, high quality voice recording in small group interactions and additionally, it is a noninvasive voice recording device as compared to close talk microphones. This device directly outputs speaker segmentation for each participant (assuming that people do not change seats during the in-

teraction), to our knowledge this is the first multimodal corpus that uses this type of voice recording device.

The video setup uses two wide-angle web cameras (Logitech Webcam PRO 9000), with a frame size of 640x480, at 30 fps. The design of this portable system was chosen such that it is easy to obtain and replicate in diverse settings, and allows adequate resolution and frame rate for our analysis purposes. Although spherical camera systems (either with a single 360 degree lens or with multiple lenses) provide a larger camera view, these cameras are in a higher price range and few of them meet our resolution and lossless frame rate criteria. Given that video recordings could be susceptible to frame dropping, due to reasons not necessarily related to the device, we considered having a device capable to record at least up to 30 fps for a reliable feature extraction. Among the portable video recording systems used in social computing research, in (Campbell et al, 2006), a spherical lens with a frame rate of 12 fps is used. The resolution is low and does not allow the analysis of fine details of participants' movements. In (Otsuka et al, 2008), two omnidirectional cameras with fish eye lenses are used. The system provides high resolution and 30 fps frame rate. In comparison to these video recording systems, our system uses commercial webcams and provides a cheap and easy-to-obtain solution for small group video recordings with sufficient resolution and frame rate.

The setting requires two laptops, one for the microcone and one for the video. Since audio and video were recorded separately, the synchronization was done manually by clapping once in the center of the table and by aligning the streams using the clapping activity. Figures 1 and 2 show a snapshot from the recording scenario and the capture devices respectively.

Fig. 1 A snapshot from the ELEA AVS corpus. The webcam is circled in red (left) and the Microcone is circled in blue (right).

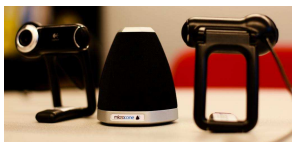


Fig. 2 Capture devices, the webcams and the Microcone, used in the ELEA recordings.

Subjects: Potential volunteers were invited to participate in a study on casual social interactions, the invitations were posted in English and French offering a monetary compensation for their participation. Advertisements were placed in two universities, a research center and a business management school

in French-speaking Switzerland. After participants contacted us by phone or email, they were informed of the process and, if they agreed to participate, cellphone number and email were requested. Since the participants were not supposed to have previous partnership or work relationship, ad-hoc groups were formed and participants were requested to attend the recordings.

85 participants were recruited, of which 31 females and 54 males in mixed teams. 19 teams are four-person and 3 teams are three-person. Average age is 23.1 years, with standard deviation 5.2.

Trust agreement: On arrival, participants signed a trust agreement. The agreement explained the process of the study, and informed them that audio and video recorded will be used only for research purposes and their identity will be anonymized. The agreement emphasizes the participants' right to quit the study at any time. Participants were provided with a copy of the signed agreement, including our complete names and email addresses for their own records.

Survival task: There are several tasks that promote group discussion and decision-making. After reviewing the tasks most often used for training in assessments centers, we chose the winter survival task, given that it promotes interactions among the participants in the group. The participants in the task are supposed to be survivors of an airplane crash. They have 12 items that they have to rank in order of their importance, giving 1 to the item considered the most important to survive as a group, 2 to the second most important, and so on. The task is performed first individually (5 min) and then we asked them to come up with the group ranking (max 15 min). Considering that not all the participants could be familiar with the items, we provided them with slides containing a picture and the definition of the item. The slides were consulted only during the individual ranking, to avoid the occlusion of the cameras during the group discussion.

Questionnaires: Four well structured questionnaires were applied, with the aim of getting ground truth for several variables from the participants in the group. For each participant, we obtained three or four questionnaire outputs, which reflected the participants' perception. The averaged outputs are considered as the ground truth.

First we administered NEO-Five Factor Inventory (NEO-FFI) (Costa and McCrae, 1992), which is a well known measure of the Big Five personality traits: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). We used the self-reported long version of the instrument composed of 60 items, each item has a score from 1 to 5 ('Disagree totally' to 'Total agreement').

This questionnaire was followed by the Personality Research Form (PRF) (Jackson, 1967). This questionnaire yields scores for personality traits relevant to the functioning of individuals in power dominance and leadership. It consists of 16 true-false items. After the personality tests, we recorded the survival task.

After the task, participants filled out a Perceived Interaction Score, that captures perceptions from participants during the interaction, in which they score every participant in the group through four items related to the fol-

lowing concepts: perceived leadership (PLead), perceived dominance (PDom), perceived competence (PCom) and perceived liking (PLike). The 16-item questionnaire can be scored from 1 to 5 ('Not at all' to 'Frequently if not always', respectively). Afterwards they provide a dominance ranking (RDom), i.e., participants were asked to rank the group, given 1 to the most dominant participant, and 3 or 4 for the less dominant, such that they have to include themselves in the ranking, similarly to previous work in dominance annotation (Jayagopi et al, 2009).

Finally, participants were asked to provide additional information including age, and experience in practicing outdoor activities and winter sports in a scale from 1-5 ('Not at all'-'Frequently, if not always'). It was optional to provide additional comments to express their feelings during the interaction and about the process.

4 Questionnaire and annotations

This section describes the coding used to process the collected data and the results of analyzing the questionnaire data.

To keep their identity anonymized, participants chose a letter K, L, M, or N and to link them with their respective questionnaires and audio/video files, the final identifier is defined as: number of group, participant letter, day and month of recording and a letter indicating the gender. Below we describe the computations done from each of the questionnaires.

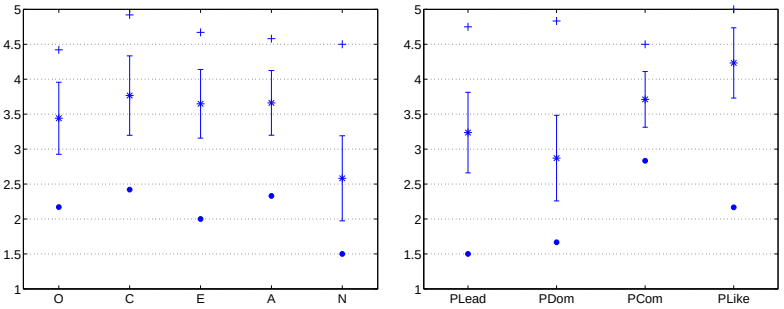
NEO-FFI: From this questionnaire, we compute mean values over the items that correspond to each of the big five traits, taking into account that some items needed to be reversed. For each person we have a vector of five real values between 1.0-5.0. Figure 4 (left) shows the distribution of the self reported personality in the ELEA AVS corpus.

PRF: Since this questionnaire is of the form true-false, we mapped the values to 1-0, such that we accumulated the number of items corresponding to power or dominance. In the data set we have two values, one corresponding to the number of items related to leadership and dominance, and a second value that represents the mean value.

Perceived interaction scores: For this questionnaire we calculated mean values for each of the perceived variables PLead, PDom, PCom, PLike, using the judgment from the other participants (i.e., not herself/himself). We consider as ground truth the annotations from the perceived interactions, such that the emergent leader in the group is the participant with the highest mean value of perceived leadership, and similarly for the related concepts. Figure 4 (right) shows the distribution of the values for the perceived variables in the ELEA AVS corpus.

Ranked Dominance: We calculated the value per participant as the mean value of the rank assigned from the other participants.

Survival task performance: Although there is no unique solution for the winter survival task, there is a ranking provided by experts, that justify the



(a) (b)

Fig. 3 Averaged values from the ELEA AVS corpus. The plot shows: minimum (•), mean (*), standard deviation (I), and maximum (+) averaged value. Of (a) personality traits (O-Openness to Experience, C-Conscientiousness, E-Extraversion, A-Agreeableness and N-Neuroticism) and (b) perceived variables (PLead-Leadership, PDom-Dominance, PCom-Competence, PLike-Liking).

item rank order with more chances to survive. We used the survival experts' ranking list to code some variables related to performance and influence, the description can be found in (Sanchez-Cortes et al, 2011b).

Perception of Leadership and Dominance from external observers: Using the questionnaires that the participants filled based on their interaction, we extracted the views of the participants themselves on the perceived interaction. However, research shows that the perception of the participants themselves and external observers differ (Dunbar and Burgoon, 2005). To be able to evaluate these differences, we also collected judgments from external observers for two of the variables, leadership and dominance.

We use the same questionnaire as filled by the participants, focusing only on leadership and dominance and excluding the questions related to other concepts. For each meeting, we assigned two external observers, one male and one female, who watched the first five minutes of the meeting video and answered eight questions for each of the participants in the meeting. The mean values are then calculated for the variables of external observers: ELead and EDom.

5 Automatic Nonverbal Features

In addition to manual coding, our corpus includes a number of automatically extracted features. Table 2 summarizes the list of features extracted from the corpus, described in this section. We first describe speaking activity features, then the visual attention features, and finally audio-visual features that combine speaking activity and attention.

Table 2 Feature groups: SA-Speaking Activity, AT-Visual Attention, AV-Audio-visual features.

Feature type	Acronym	Definition
Speaking Activity (SA)	SPL	Speaking Length
	SPT	Speaking Turns
	SPI	Speaking Interruptions
	ASP	Average Speaking Turn Duration
Visual Attention (AT)	ATR	Attention Received
	ATG	Attention Given
	ATQ	Attention Quotient (ATR/ATG)
	ATC	Attention Center
Audio-Visual (AV)	LWS	Looking While Speaking
	LWL	Looking While Listening
	BLWS	Being Looked While Speaking
	CAWS	Center of Attention While Speaking
	VDR	Visual Dominance Ratio (LWS/LWL)

5.1 Speaking Activity Features

The Microcone automatically generates a speaker segmentation (McCowan, 2011), which is easily converted to a binary segmentation in which the speaking status is represented as 1, and 0 represents a non-speaking status. From the segmentation we coded the following speaking turn features:

Speaking Length (SPL): Contains the total speaking time of each participant i during the meeting.

Speaking Turns (SPT): Accumulates total turns over the entire meeting for each participant i, the turn is defined by a series of active speaking status.

Speaking Interruptions (SPI): Accumulates total interruptions over the entire meetings. Participant i interrupts participant j if i starts talking when j is speaking; when i finishes his/her turn j is not speaking anymore.

Average Speaking Turn Duration (ASP): Represents the averaged turn duration for each participant i during the meeting.

This set of features have been used by other researchers in previous computational works to characterize individual behavior in group interactions, specifically to recognize dominant behavior and status (Rienks and Heylen, 2005; Jayagopi et al, 2009).

Furthermore, speaking time has been identified in social psychology literature as a strong indicator of dominance (Mast, 2002).

5.2 Visual Attention Features

The extracted visual features are based on attention (denoted VFOA for Visual Focus of Attention), specifically 'who is looking at whom or what'. First, we extract the VFOA and then construct features that could characterize an individual's behavior in group interactions. Gaze cues, along with conversational cues are known to be informative to characterize small group interactions (Knapp and Hall, 2008). Apart from facilitating the turn-taking patterns, they also signal socially relevant information, for example dominance or status (Hall et al, 2005; Harrigan, 2005).

As tracking eye gaze requires high-resolution videos, and head direction sufficiently captures eye gaze direction in conversational settings (Stiefelhagen and Zhu, 2002), we first estimate the head pose automatically. The head pose is characterized by three angles: pan, tilt, and roll. Then, we assign the head pose to a discrete VFOA label in every frame. We use the method proposed in (Ricci and Odobez, 2009), that employs a dynamic, probabilistic framework to estimate the head location and pose jointly based on a standard state-space formulation. The states correspond to location and scale of the head as well as discretized head pose. The observation model uses both color features and texture features (based on Histograms of Oriented Gradients (HOG)). The inference is done using particle filters which represents the distribution of states at each frame by a finite set of samples (or particles). The left image in Fig. 4 shows the tracker output location which is computed as the mean (in green color) and median (in red color) of the states distribution. The top right part of Fig. 4 shows the estimated pan and tilt head pose angles represented by the green line over a semi-circumference spanning $\pm 90°$.

Fig. 4 Tracking, head-pose estimation, and VFOA estimation for an individual in a group interaction in the ELEA AVS corpus. See main text for details.

Considering pan and tilt only, the VFOA is later estimated by Maximum a Posteriori (MAP) rule. The MAP rule assumes a Gaussian distribution with mean and standard deviation pre-specified manually (in the pan and tilt space), for each of the five visual targets T1 to T5. Fig. 5 shows the position of these visual targets with respect to the configuration of the room. T1, T2 are the participants sitting opposite to the participants shown in Figure 4. T3 is the participant sitting next to the tracked participant. T4 and T5 represent the table area close to the tracked participant and participant T3, respectively, UN stands for unfocused (i.e. any other possible VFOA). The bottom right part of Fig. 4 shows the estimated VFOA target (T1 for this particular frame).

In order to assess the VFOA recognition accuracy, we carried out manual annotations of the VFOA of every participant, for one randomly chosen

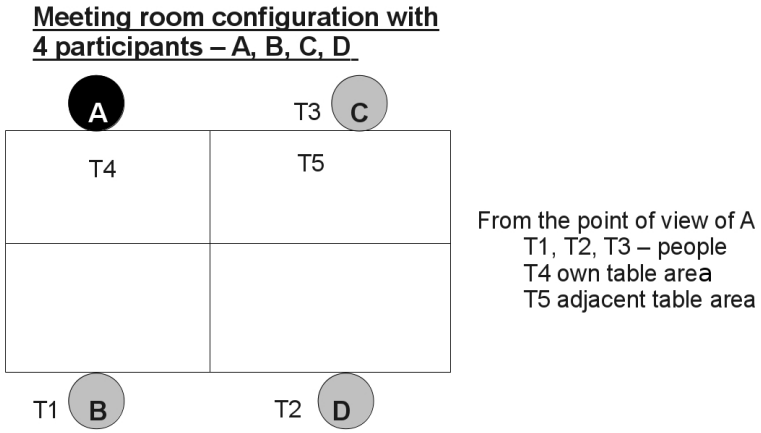


Fig. 5 The configuration of the meeting room (where the group interaction took place)

discussion in the ELEA AVS corpus. Every 15 seconds, the VFOA of every participant was annotated using one annotator. The automatic method had an accuracy of 42% (frame-level) when compared to the manual annotation. The cases where the method failed belonged to two categories. The first one was due to tracking failures, which were typically due to background color effects or illumination issues. The second source of error are inaccuracies in head-pose estimation. Errors in tilt estimation sometimes resulted in the wrong assignment of automatic VFOA targets. Our method used a fixed mapping from head-pose angles to VFOA. As mentioned in the previous paragraph, this mapping was pre-specified for every participant. Importantly, typical VFOA accuracies obtained with similar methods in other group interaction data (e.g. the AMI corpus) are roughly in this order [see for instance Ba and Odobez (2009)]. Also, note that more sophisticated methods, which for instance model the joint VFOA of multiple people (Ba and Odobez, 2011), could probably result in higher recognition performance but have not been explored here.

From the recognized VFOA labels, i.e. the visual target of each participant, the following features that capture socially relevant information are extracted:

Attention Received (ATR): ATR is the number of frames in which the participant i is looked by the other participants.

Attention Given (ATG): ATG is the number of frames in which a participant i looks at other participants.

Attention Quotient (ATQ): is the ratio between the amount of attention that participant i received from the other participants (ATR) and the amount of attention that participant i gives to the other participants in the group (ATG).

Attention Center (ATC): ATC is the total number of frames in which participant i received attention from all the participants in the group at the same time.

Similar features were originally used by Hung et al., (Hung et al, 2008) to characterize dominance in small groups in the AMI corpus. Furthermore, other related features have been used to capture connections between attention and personality (Subramanian et al, 2010), and to investigate interpersonal influence (Otsuka et al, 2006). Furthermore, attention features have been discussed in some of the classic works in social psychology on dominance and nonverbal behavior (Efran, 1968; Cook and Smith, 1975).

5.3 Multimodal Features

The fusion of features obtained from different channels can provide a better understanding of the group interactions (Otsuka et al, 2005). As described by Dovidio, the proportions of look-speak and look-listen in a conversation provide information about dominance and power (Dovidio and Ellyson, 1982). This finding has been verified with automatic features by Hung et al. (Hung et al, 2008). Considering that we have extracted features related to speaking turns from the audio channel, and attention from the visual channel, we extracted the following variables.

Looking while Speaking (LWS): Amount of attention (in frames) that participant i gives to the participants in the group while i is speaking.

Looking while Listening (LWL): Amount of attention (in frames) that participant i gives to the participants in the group while i is not speaking. Note that we cannot infer that a person is listening, so we simply approximate this by non-speaking.

Being Looked at while Speaking (BLWS): Amount of attention that participant i receives from the other participants while i is speaking.

Center of Attention while Speaking (CAWS): Number of frames that participant i is the center of attention (i.e. all the participants are looking at her/him at the same time) while i is speaking.

Visual Dominance Ratio (VDR): Ratio of Looking while Speaking and Looking while Listening (LWS/LWL).

Some of the described multimodal features were used in (Hung et al, 2008) and showed to be useful to analyze dominance in the AMI corpus.

To compute these features, audio-visual synchronization is needed. To achieve this, the audio and visual channels were aligned, by manually localizing the synchronization point for each audio-visual sequence (i.e. using the clapping event that indicates the beginning of the group interaction).

6 Using Nonverbal Behavior to Identify Emergent Leaders

We now describe the use of nonverbal behavioral cues to identify the emergent leader in the group. To have a clear understanding on how the various features perform, we define an unsupervised rule-based inference that selects

Table 3 Pearson correlations between speaking activity features and perceived variables, significance values $+ : p < 0.05$, $* : p < 0.01$. SPL-Speaking Length, SPT-Speaking Turns, SPI-Speaking Interruptions and ASP-Average Speaking Turn Duration.

	SPL	SPT	SPI	ASP
PLead	0.506^{*}	0.492^{*}	0.548^{*}	0.46^{*}
PDom	0.290^{+}	0.373^{*}	0.403^{+}	0.235
RDom	0.492^{*}	0.408^{*}	0.538^{*}	0.474^{*}

the participant with the maximum feature value in the group as the emergent leader.

$$EL_m^f = \arg\max_p(f_p^m), p \in \{1, 2 \dots P\}, \quad (1)$$

where p is the participant number, f is a nonverbal feature, f_p^m is the value of feature f for participant p in group m, and P is the number of participants in the group (3 or 4 in our case).

The selection of the inference method is based on the research done by Baird, which states that a simple predictor of leadership can be constructed using single nonverbal behavioral features like head nodding, body shift, or verbal participation (Baird, 1977; Stein and Heller, 1979).

We use the perceived variables (PLead, PDom and RDom) from the questionnaires defined in Section 4 as ground truth. Random performance in this case is 26.1%, given that the synchronized corpus has 22 meetings, of which 19 meetings have four participants, and 3 meetings have three participants.

6.1 Speaking Activity Features

In this section we present correlations and inferences of the emergent leader in the group, using only speaking activity features.

Table 3 shows correlations between the speaking activity features and the perceived variables. The Pearson correlations are calculated per group, followed by a Fisher transformation and a t-test at 5% significance level. As we can observe the amount of interruptions (SPI) are significantly correlated with PLead, PDom and RDom. Similarly, significant correlations between interruptions and concepts of dominance and leadership have been reported in (Hung et al, 2008; Jayagopi et al, 2009; Sanchez-Cortes et al, 2011b).

In Table 4 we can observe accuracy performance of single nonverbal speaking cues extracted and the inference method in Equation 1. SPI has the best accuracy (72.7%), which is significantly higher than random performance. It has been shown that the single acoustic channel can provide good accuracy performance in the prediction on dominance in small groups (up to 85%) using SPL (Jayagopi et al, 2009), although in our case, the survival task and the scenario with unacquainted people resulted on a more challenging case (up to 54.5%).

Table 4 Accuracy (%) performance from speaking activity cues on the ELEA AVS corpus. Random performance is 26.1%

	SPL	SPT	SPI	ASP
PLead	54.5	45.5	**72.7**	45.5
PDom	31.8	40.9	**45.5**	40.9
RDom	54.5	36.4	**63.6**	50.0

Table 5 Pearson correlation and features from attention, significance values $^{+}: p < 0.05$, $*: p < 0.01$. ATR-Attention Received, ATG-Attention Given, ATQ-Attention Quotient and ATC-Attention Center.

	ATR	ATG	ATQ	ATC
PLead	0.330*	0.013	0.233^{+}	0.286*
PDom	0.374*	-0.060	0.306^{+}	0.355*
RDom	0.306*	0.075	0.134	0.173

Table 6 Accuracy (%) performance from visual attention features on the ELEA AVS corpus. Random performance is 26.1%

	ATR	ATG	ATQ	ATC
PLead	**59.1**	27.3	40.9	40.9
PDom	**68.2**	40.9	59.1	54.6
RDom	**45.5**	22.7	22.7	27.3

6.2 Visual Attention Features

In this section we present correlations between the visual attention and the perceived variables, followed by the results of the emergent leader inference (and related concepts) using the estimator defined in Equation 1. Table 5 shows Pearson correlations between the features extracted from attention and the perceived variables. The Pearson correlations are calculated per group, followed by Fisher transformation and a t-test at 5% significance level. As we can observe, there are significant correlations between ATR, and the variables PLead, PDom and RDom.

Single features obtained from visual attention help to identify the emergent leader up to 59.1%. The amount of attention received (ATR) from participants is the most informative cue, followed by the amount of attention received from the group (ATC) with 40.9%. For the case of PDom, the best performance is 68.2% as well with the feature ATR, this reflects that the most dominant participant receives the largest amount of visual attention in the group. The results are shown in Table 6.

Further, we reviewed the correlations between the visual attention and the acoustic nonverbal features. In Table 7 we can observe significant correlations between the attention received ATR and SPL, SPT and SPI. Also the correlations between ATQ and, SPL, SPT and SPI are significant. Finally a low but significant correlations can be observed between ATC and, SPL and SPT.

Table 7 Pearson correlation between acoustic nonverbal features and attention, significance values $^{+}: p < 0.05$, $*: p < 0.01$

	ATR	ATG	ATQ	ATC
SPL	0.214^{+}	-0.007	0.183^{+}	0.117^{*}
SPT	0.218^{+}	-0.063	0.224^{+}	0.147^{*}
SPI	0.327^{*}	-0.154	0.381^{*}	0.230
ASP	0.143	0.004	0.138	0.063

Table 8 Pearson correlation between multimodal features and attention, significance values $^{+}: p < 0.05$, $*: p < 0.01$. LWS-Looking while Speaking, LWL-Looking while Listening, BLWS-Being Looked at while Speaking, CAWS-Center of Attention while Speaking and VDR-Visual Dominance Ratio.

	ATR	ATG	ATQ	ATC
LWS	0.131	0.449^{*}	-0.189	-0.037
LWL	-0.351^{*}	0.560^{*}	-0.654^{*}	-0.368^{*}
BLWS	0.751^{*}	-0.118	0.562^{*}	0.581^{*}
CAWS	0.808^{*}	-0.218	0.667^{*}	0.852^{*}
VDR	0.328^{*}	0.011	0.244^{+}	0.185

The correlations between SPL and ATR, although lower compared with the ones reported in Subramanian et al (2010) using a winter survival task scenario, show that the attention received in small groups is correlated to the total amount of speaking activity and, in our case it also correlates with the successful interruptions to grab the floor.

6.3 Multimodal features

In this section we first present correlations between multimodal (i.e. audio-visual) and single features, followed by the results of identification of the emergent leader and related concepts using multimodal features. Table 8 shows correlations between multimodal features and visual attention features, as we can observe there are significant correlations between CAWS and ATR, CAWS and ATC, and, CAWS and ATQ. The strong correlations suggest that being the center of group attention while speaking is connected to the amount of attention received as much as being the visual attention center during the meeting. Similarly, significant correlations can be observed between BLWS and ATR, BLWS and ATQ, and BLWS and ATC. Finally, there are significant negative correlations between LWL and ATR, ATQ and ATC.

Considering that nonverbal behavior extracted from audio and visual single channel can be used to identify the emergent leaders (Sanchez-Cortes et al, 2011b), multimodal features extracted from synchronized audio and video might provide a better understanding of the nonverbal behavior of the emergent leader. Figure 9 shows performance using the unsupervised method and the multimodal features, where the best performance to identify the leader is using either BLWS or CAWS with up to 63.6%.

Table 9 Accuracy (%) performance from frame based multimodal features on the ELEA AVS corpus. Random performance is 26.14%.

	LWS	LWL	BLWS	CAWS	VDR
PLead	50.0	40.9	**63.6**	**63.6**	50.0
PDom	31.8	40.9	59.1	**63.6**	36.4
RDom	50.0	36.4	45.5	45.5	**54.5**

Table 10 Accuracy (%) performance from event based multimodal features on the ELEA AVS corpus. Random performance is 26.14%.

	LWS	LWL	BLWS	CAWS	VDR
PLead	50.0	54.5	54.5	**68.2**	50.0
PDom	40.9	45.5	45.5	**59.1**	36.4
RDom	50.0	45.5	**54.5**	50.0	**54.5**

Table 11 Best accuracy performance (%) from the single and multimodal features on the ELEA AVS corpus. Random performance is 26.1%. SPI-Speaking Interruptions, ATR-Attention Received, CAWS-Center of Attention while Speaking, VDR-Visual Dominance Ratio.

	Variable	Accuracy (%)	**feature**
SA	PLead	**72.7**	SPI
	PDom	45.5	SPI
	RDom	**63.6**	SPI
AT	PLead	59.1	ATR
	PDom	**68.2**	ATR
	RDom	45.5	ATR
AV	PLead	63.6	CAWS
	PDom	63.6	CAWS
	RDom	54.5	VDR

With the aim of having a better understanding on how multimodal features can perform for PLead, PDom and RDom, we also considered an event-based evaluation strategy. To do this, we count only the times that an event (i.e. segment of consecutive frames with the same multimodal feature) occurs during the meeting instead of counting the exact number of frames in which this event occurs. Considering this option, we can observe in Figure 10 that the event-based accuracy to infer the emergent leader in the group increases up to 68.2%, on the other hand the inference of the perceived dominant participant in the group decreases from 63.6% to 59.1% for the best multimodal feature (CAWS).

Finally the performance of single and multimodal features is summarized in Table 11, as we can observe the best single predictor of emergent leader is SPI, followed by CAWS.

Our findings in PLead and RDom, show that the speaking nonverbal cues perform better than the visual and multimodal cues, as similarly reported in dominance estimation using the AMI corpus in (Hung et al, 2008; Jayagopi et al, 2009). Additionally, the visual attention performance reported in (Hung et al, 2008) was estimated considering manual annotations, which perhaps re-

flect a better performance in both the visual and the multimodal features. In contrast in the ELEA AVS corpus the visual attention cues performed better for the perceived dominance (PDom) than only audio cues and slightly better than the multimodal cues. Overall the variable PLead achieved the highest accuracy performance using single speaking activity features. In addition, the correlations performed suggest a connection between leadership and dominance but they are not exactly the same. Finally, the results showed in Table 11 suggest that the recorded scenario in the ELEA AVS corpus is more challenging than the existing small group corpora used to estimate vertical dimensions, although the data is limited to 22 recordings, and the numbers need to be considered carefully.

6.4 Time Delay in Multimodal features

As discussed is Section 3, frame dropping can occur during video recordings, given to several reasons including applications running in background. To test the effects of possible misalignment between the audio and the video channels, we define a alignment-match from the video frame i with a window from i to $i + \delta$ with the respective audio stream, where δ denotes the width of the temporal window in frames. Figure 6 shows the time delay synchronization window applied.

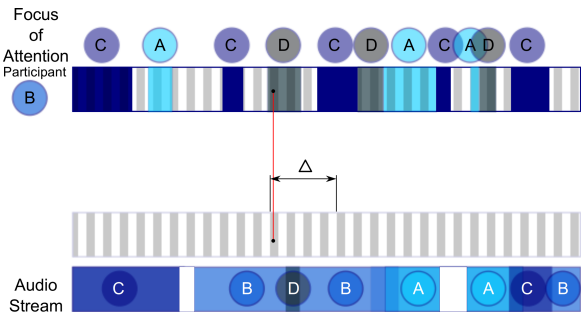


Fig. 6 Frame alignment window between visual attention and speaking activity streams. The frame i in the attention stream, is aligned with a slide window from the frame i to the frame $i + \delta$ in the speaking activity stream.

A video generated from different audio and video channels could be susceptible to frame dropping, while playing the merged video, if it is not well synchronized, we could notice a delay between the visual activity (while speaking) and the audio sound. Considering that our corpus was collected using separated audio and video recording devices, we explored the impact of the delay in the multimodal extracted features. In our experience, as it is most likely that the frame dropping occurs in the video stream, we considered the effect of slight dropping frame in the video channel on the multimodal features.

Figure 7 shows the accuracy considering the variables PLead, PDom and RDom). The X axis represents the amount of frames considered (δ from 1 to 60). The Y axis represents the accuracy performance, using the Equation 1 defined in the beginning of Section 6.

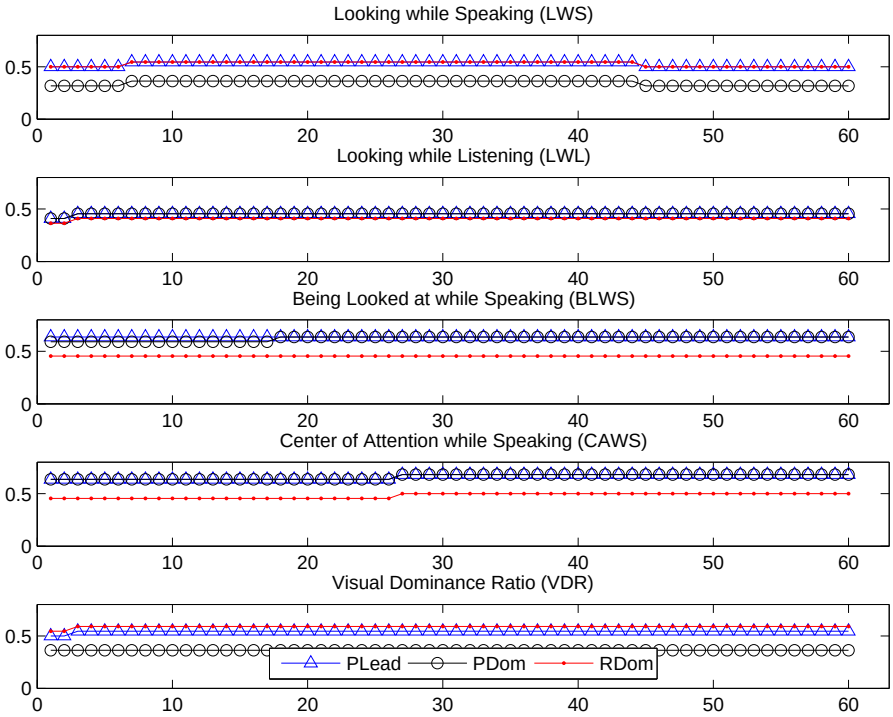


Fig. 7 Accuracy performance (%) from multimodal features using a time delay alignment window with the audio stream.

As we can observe, in Figure 7 the multimodal features can be robust in frame dropping situations, if an alignment window is considered with respect to the audio stream. Additionally, for the case of the frame by frame synchronized features ($\delta = 0$), it is most likely that very short turns might be missed, due to the misalignment, but on the other hand longer turns will be captured more accurately.

6.5 Results with External Annotations

We finally present emergent leader inference results using the external perceptions of dominance and leadership presented in Section 4. Considering the judgment from the external observers on the perception of the leader in the group (ELead), our framework obtains up to 50% accuracy, with the features

ASP, ATR and CAWS; for the case of dominance (EDom) the highest accuracy is 59.1% using ATR and CAWS. These results are lower than the ones obtained with the judgment of the participants in the group. It is worth to mention that the external observers gave their highest scores to the same person in the ELead and EDom measures. In 95% of the cases, one participant in the group is perceived by external observers as both leader and dominant. On the other hand, considering the perception from the interacting participants, in only 63.6% of the cases the same participant is perceived as the leader and perceived as dominant. Figure 8 shows averaged values from ELead, EDom and PLead, PDom from the ELEA AVS corpus. Calculating Pearson correlation between leadership and dominance, for ELead and EDom we have a significant correlation of 0.96 ($p = 3.31E-12$), for the case of PLead and PDom we have 0.75 ($p = 1.39E-04$). This suggests that there is a connection between the leadership and dominance in both the perception of the participants in the group and perception from external observers.

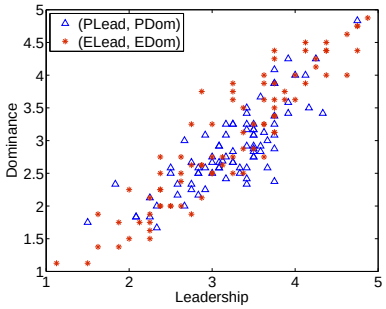


Fig. 8 Comparison of External annotations and perception from participants, for all individuals in the ELEA AVS corpus (N = 85). Each data point shows the averaged perceived values of leadership and dominance for each participant, either (PLead, PDom) or (ELead, EDom), for perceived and external annotations respectively.

6.6 Discussion

Our findings in the ELEA AVS corpus reveal that speaking activity is a better estimator of emergent leadership than visual attention. On the other hand, the amount of visual attention received is more informative for the perception of dominance between the participants in the group. Although the multimodal features are not the best descriptor of leadership, nor of dominance, they provide some information about the perceived leadership during the interaction, such that being the center of attention while speaking correlates with being perceived as the leader.

We also observed strong correlations between perceived leadership and perceived dominance, for both the participants in the group and the external observers. However, the external observers perceived the leader in the group as

a dominant person most of the time, in contrast with the perception of the participants in the group where leadership and dominance are less correlated.

Although we employed similar automatically extracted features and similar methods to identify leadership and dominance as in (Hung et al, 2008; Jayagopi et al, 2009), our results suggest that the scenario recorded in the ELEA corpus is more challenging, in comparison with corpora recorded with small group scenarios that follow a script and have pre-assigned roles.

There are some limitations to be aware of. First, the corpus is relatively small, despite our efforts to collect data. This has to do with the requirement of having to engage only people who do not know each other, and shows the difficulty of collecting data even with portable sensors. The size of the corpus puts limits on the statistical confidence of the results. Second, the VFOA features that are automatically extracted are known to have a performance that is not very high (42% frame-level accuracy on a subsample of the data). We did not conduct studies using clean manual VFOA labels. This is clearly an important thing to do but involves a significant amount of manual work for the five hours of the corpus, and could be part of future work. Third, clearly other better inference methods could have been used, but in this work we made the explicit decision of using something relatively simple. Future work could extend this part using other, more complex, machine learning methods.

7 Conclusions

We presented in this paper a new data corpus, collected with the aim of analyzing emergent leadership in small groups. The novelty of our synchronized corpus is that it is collected with a portable recording solution, and it contains a detailed set of questionnaires related to perceived leadership, personality, and performance collected from the participants in each group.

The annotations available for every group include the big five personality scores, scores on dominance and leadership, scores from perceived and self reported leadership, dominance, competence, and likability, as well as external observer annotations for the same characteristics. The corpus also includes individual and group outcomes from the performed survival task, coded as individual performance, group performance, and individual influence. Finally, the corpus includes automatically extracted features from speaking activity, visual focus of attention, and multimodal features.

As an illustration of research questions that can be addressed with this corpus, we presented a brief analysis on inference of emergent leadership using audio features based on speaking activity, video features based on visual attention, as well as multimodal features. We also compared leadership and dominance perception between external observers and participants in the group.

As future work, the effect of other interesting automatically extracted features, including floor patterns and emotional states on estimating the leader in the group can be investigated. The floor patterns and the emotional states of the participants can be extracted based on their nonverbal behavior to

explore the impact on the perception and estimation of leadership and dominance. Similarly, emerging social interactions, such as involvement or control, which are known to be informative for leadership, can also be studied. Another dimension of future work would be to study the personality of the participants as an influence factor during the interaction and its influence on the perception of the leader in the group.

Acknowledgements We thank Iain McCowan (dev-audio) for technical support; Denise Frauendorfer and Pilar Lorente (University of Neuchatel), Radu-Andrei Negoescu (Idiap) for help during the data collection and for data processing, Jean-Marc Odobez (Idiap) for sharing code for VFOA extraction, and all the participants in the recordings. This research was supported by CONACYT (Mexico) through a doctoral scholarship, and supported by the projects SONVB (Swiss NSF), NOVICOM (EU-FP7-IEF) and SOBE (Swiss NSF Ambizione grant no: PZ00P2-136811).

References

Aran O, Gatica-Perez D (2011) Analysis of group conversations: Modeling social verticality. In: Salah AA, Gevers T (eds) Computer Analysis of Human Behavior, Springer London, pp 293–322

Aran O, Hung H, Gatica-Perez D (2010) A multimodal corpus for studying dominance in small group conversations. In: Workshop International Conference on Language Resources and Evaluation, LREC

Ba S, Odobez J (2009) Recognizing visual focus of attention from head pose in natural meetings. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39(1):16–33

Ba S, Odobez J (2011) Multi-person visual focus of attention from head pose and meeting contextual cues. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(1):101–116

Baird JE (1977) Some non-verbal elements of leadership emergence. Southern Speech Communication Journal 42(4):352–361

Bales R, Strodtbeck F (1951) Phases in group problem-solving. Journal of Abnormal and Social Psychology 46:485–495

Burger S, MacLaren V, Yu H (2002) The isl meeting corpus: the impact of meeting type on speech style. In: International Conference on Spoken Language Processing, Interspeech-ICSLP

Campbell N, Sadanobu T, Imura M, Iwahashi N, Noriko S, Douxchamps D (2006) A multimedia database of meeting and informal interactions for tracking participant involvement and discourse flow. In: Workshop International Conference on Language Resources and Evaluation, LREC

Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, J Kadlec VK, Kraaij W, Kronenthal M, Lathoud G, Lincoln M, Lisowska A, McCowan I, Post W, Reidsma D, Wellner P (2005) The ami meeting corpus: A pre-announcement. In: Workshop on Machine Learning and Multimodal Interaction, ICMI-MLMI

Chen L, Rose TR, Parrill F, Han X, Tu J, Huang Z, Harper M, Quek F, McNeill D, Tuttle R, Huang T (2005) Vace multimodal meeting corpus. In: Workshop on Machine Learning and Multimodal Interaction, ICMI-MLMI

Cook M, Smith JMC (1975) The role of gaze in impression formation. British Journal of Social and Clinical Psychology 14(1):19–25

Costa P, McCrae R (1992) NEO PI-R profesional manual

Dovidio J, Ellyson S (1982) Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. Social Psychology Quarterly 45(2):106–113

Dunbar NE, Burgoon JK (2005) Perceptions of power and interactional dominance in interpersonal relationships. Journal of Social and Personal Relationships 22(2):207–233

Efran J (1968) Looking for approval: effects of visual behavior of approbation from persons differing in importance. Journal of Personality and Social Psychology 10(1):21–25

Garofolo I, Michel M, Laprun C, Stanford V, Tabassi E (2004) The nist meeting room pilot. In: International Conference on Language Resources and Evaluation, LREC

Gatica-Perez D (2006) Analyzing group interactions in conversations: a review. In: International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp 41–46

Gatica-Perez D (2009) Automatic nonverbal analysis of social interaction in small groups: a review. Image and Vision Computing 1(12)

Hall JA, Coats EJ, Smith L (2005) Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. Psychological bulletin 131(6):898–924

Harrigan J (2005) Proxemics, kinesics, and gaze. The new handbook of methods in nonverbal behavior research pp 137–198

Hung H, Jayagopi DB, Ba S, Odobez JM, Gatica-Perez D (2008) Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: International Conference on Multimodal Interfaces, ICMI, pp 233–236

Jackson DN (1967) Personality research form manual. Research Psychologists Press

Janin A, Baron D, Edwards J, Ellis D, Gelbart D, Morgan N, Peskin B, Pfau T, Shriberg E, Stolcke A, Wooters C (2003) The icsi meeting corpus. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP

Jayagopi D, Hung H, Yeo C, Gatica-Perez D (2009) Modeling dominance in group conversations using nonverbal activity cues. IEEE Transactions on Audio, Speech and Language Processing 17(3)

Jovanovic N, op den Akke R, Nijholt A (2005) A corpus for studying addressing behavior in multi-party dialogues. In: The sixth SigDial conference on Discourse and Dialogue

Kickul J, Neuman G (2000) Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and ksas. Journal of Business and Psychology 15(1)

Kim T, Chang A, Holland L, Pentland A (2008) Meeting mediator: enhancing group collaboration with sociometric feedback. In: Conference on Computer Supported Cooperative Work, CSCW

Knapp ML, Hall JA (2008) Nonverbal Communication in Human Interaction. Wadsworth, Cengage Learning

Mana N, Lepri B, Chippendale P, Cappelletti A, Pianesi F, Svaizer P, Zancanaro M (2007) Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In: Workshop on Tagging, mining and retrieval of human related activity information, TMR

Mast MS (2002) Dominance as expressed and inferred through speaking time: A meta-analysis. Human Communication research 28(3):420–450

McCowan I (2011) `http://www.dev-audio.com/`

McCowan I, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D (2005) Automatic analysis of multimodal group actions in meetings. PAMI 27(3):305–317

Otsuka K, Takemae Y, Yamato J, Murase H (2005) Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns and head directions and utterances. In: International Conference on Multimodal Interfaces, ICMI

Otsuka K, Yamato J, Takemae Y, Murase H (2006) Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In: CHI '06 extended abstracts on Human factors in computing systems, ACM, New York, NY, USA, CHI EA '06, pp 1175–1180

Otsuka K, Araki S, Ishizuka K, Fujimoto M, Heinrich M, Yamato J (2008) A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In: International Conference on Multimodal Interfaces, ICMI

Pianesi F, Zancanaro M, Lepri B, Cappelletti A (2007) A multimodal annotated corpus of consensus decision making meetings. Language Resources and Evaluation 41:409–429

Poole MS, Holligshead AB, McGrath JE, Moreland RL, Rohrbaugh J (2004) Interdisciplinary perspectives on small groups. Small Group Research 35(1):3–16

Ricci E, Odobez J (2009) Learning large margin likelihoods for realtime head pose tracking. In: International Conference on Image Processing, ICIP

Rienks R, Heylen D (2005) Dominance detection in meetings using easily obtainable features. In: In Bourlard, H., and Renals, S. (Eds.), Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Springer Verlag, pp 76–86

Rienks R, Zhang D, Gatica-Perez D, Post W (2006) Detection and application of influence rankings in small group meetings. In: International Conference on Multimodal Interfaces, ICMI

Salas E, Sims DE, Burke CS (2005) Is there a big five in teamwork. Small group research 36(5):555–599

Sanchez-Cortes D, Aran O, Mast MS, Gatica-Perez D (2010) Identifying emergent leadership in small groups using nonverbal communicative cues. In:

International Conference on Multimodal Interfaces, ICMI
Sanchez-Cortes D, Aran O, Gatica-Perez D (2011a) An audio visual corpus for emergent leader analysis. In: Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, ICM-MLMI
Sanchez-Cortes D, Aran O, Mast MS, Gatica-Perez D (2012b) A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Transactions on Multimedia 14(3):816–832
Stein RT (1975) Identifying emergent leaders from verbal and nonverbal communications. Personality and Social Psychology 32(1):125–135
Stein RT, Heller T (1979) An empirical analysis of the correlations between leadership status and participation rates reported in the literature. Journal of Personality and Social Psychology 37(11):1993–2002
Stiefelhagen R, Zhu J (2002) Head orientation and gaze direction in meetings. In: CHI'02 Extended abstracts on Human factors in computing systems, CHI EA '02
Subramanian R, Staiano J, Kalimeri K, Sebe N, Pianesi F (2010) Putting the pieces together: Multimodal analysis in social attention in meetings. In: ACM Multimedia, MM