

# Detecting Emergent Leaders in Small Groups using Nonverbal Behavior

Dairazalia Sanchez-Cortes<sup>1,3</sup>, Oya Aran<sup>1</sup>, Marianne Schmid Mast<sup>2</sup> and Daniel Gatica-Perez<sup>1,3</sup>

<sup>1</sup>IDIAP Research Institute, Martigny, Switzerland

<sup>2</sup>Institut de psychologie du travail et des organisations, University of Neuchatel, Switzerland

<sup>3</sup>Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

## Abstract

This paper presents first an analysis on how an emergent leader is perceived in newly formed small-groups, and in its second part tackles the emergent leader estimation task, using automatically extracted nonverbal cues. The estimations use rank-based and collective approaches with the combination of acoustic and visual features extracted from a new corpus specifically collected to analyse the emergent leadership phenomena. Our results show that visual information augments acoustic information, and the emergent leader is perceived by his/her peers as an active and dominant person. For the automatic inference of emergent leadership, adding relational information to the nonverbal cues improves the inference of each participant's label in the group.

## Index Terms

Emergent Leadership, Nonverbal behavior

## I. INTRODUCTION

**T**HE team leader in an organization is a role associated with the person having the authority, or a position of power, who directs people towards finishing their jobs, and who has the final say in the fundamentals at work: what, who, where, and when [50]. Since leadership and the interaction between co-workers are critical variables for the success of many of the faced tasks, visionary organizations are nowadays hiring team leaders based on multiple interviews using problem-solving tasks, and through the observation of emergent leaders in assessment centers [20].

In interactions between two or more members of a group, the leader is an agent of change, a person whose acts affect other people more than other people's acts affect him [8]. An emergent leader is defined as the one who naturally arises from an interacting group and has his/her base of power from his peers in the group, rather

than from a higher authority [50]. In face-to-face communication the words represent the verbal information, and everything else is nonverbal communication [35] (voice tone and loudness, eye gaze, head and body gestures, etc.).

The nonverbal channel is especially useful when there is conflict among the verbal and nonverbal channels: when people are engaged in conflicting situations, they often value nonverbal behavior higher, because of the fact that a large part of the internal states and traits are revealed by nonverbal cues, and nonverbal behavior is harder to fake [42].

Nonverbal behavior in group communication has been studied by psychologists for decades, mainly through manual annotations, and ratings from human observers. Nowadays, the automatic extraction of verbal and nonverbal cues from face-to-face interactions in small groups is becoming relevant [18], given its potential to produce large amounts of annotated raw data in an accurate way, saving time and reducing errors compared with manual annotations. Furthermore, as technology advances, the automatic extraction of communicative features from portable sensors (cameras and microphones) is becoming easier and more reliable [34].

The automatic extraction of nonverbal features has been successfully used to infer several social dimensions (dominance, role, status, personality, etc.) that arise in small group conversations [18]. The inferences have been done based on a variety of techniques ranging from rule based estimators [46] to complex machine learning algorithms [26] using audio and visual features. However, most of the existing approaches have not explicitly considered collective inference mechanisms (modeling of the group as a whole, rather than of each individual), which might significantly improve the accuracy of inference when multiple individuals are interacting, as was shown in [30]. Relationships between the different features extracted from the data can be discovered and then used to infer each participant's label simultaneously.

In this work we address the challenging problem of automatically inferring emergent leadership from audio-visual recordings. We present simple methods, as well as more complex machine learning algorithms, to estimate the emergent leaders in small groups using communicative nonverbal cues. The nonverbal features are automatically extracted from a new corpus that has been collected for our study, using portable audio and video sensors. The data consist of approximately 10 hours of audio/video recordings, as well as variables extracted from questionnaires filled by each group member immediately after the recordings.

The contributions of this paper are as follows. First, we present what to our knowledge is the first study on detecting emergent leadership in small face-to-face interaction groups, based on automatic sensing. Second, we showcase the new interaction corpus explicitly collected to study the emergent leadership phenomenon, a corpus which will be publicly available for research purposes. Third, we present a correlation analysis of how the emergent leaders in a group are perceived based on their nonverbal behavior. Forth, we estimate emergent leaders using automatically extracted nonverbal cues. To estimate the emergent leader we used two approaches: a simple, person-

wise, rule-based method, and a collective, group-wise classification approach. Finally, we analyzed the temporal effect of the nonverbal cue extraction process on the accuracy of the emergent leader estimation.

The paper is organized as follows. Section II discusses previous work related to emergent leadership, and the social traits and nonverbal cues associated with it. We summarize our approach in Section III, and then describe the dataset we collected for this work in Section IV. Section V introduces the nonverbal cues used in the experiments, and Section VI describes the leadership estimation methods. We present and discuss experimental results in Section VII and we draw conclusions in Section VIII.

## II. RELATED WORK

In this section we review a few key works closely related to our work, from two distinct fields: social psychology, and social computing.

### A. Social Psychology

Psychologists agree that nonverbal behavior has an important relation with the expression of verticality, which corresponds to relations that suggest position in a low-to-high continuum [22]. The aspects of the vertical dimension include dominance, status, power, and leadership. These concepts are not always clearly distinguished in the literature. In the present paper we focus on the emergent leader, understood as the person who emerges in a group as the one with the most pronounced position on the vertical dimension, thus the individual with the most influence in the group [50]. Given that emergent leadership has been measured using different concepts (dominance, influence, leadership, control), we review the literature concerning all of these aspects of verticality.

The initial studies on emergence of leadership and nonverbal behavior date from the mid-seventies. In 1975, Stein [49] conducted a study on perception of emergent leadership using scenarios in which leaderless groups of eight or nine members worked weekly throughout the semester on a research project. Observers were able to identify emergent leadership in small groups from both verbal and nonverbal information using 20 minute edited recordings from the initial 45 minute meetings. Verbal communication was transcribed from videotapes. Nonverbal communication was tested with a visual-only setup and an audio-visual setup, where the audio is filtered such that it provides only vocal nonverbal information. For emergent leadership, the highest correlation values were obtained between filtered speech and participation, which was defined as the relative amount of time each group member spent talking. In [7], Baird used visual nonverbal cues to predict emergent leadership in a scenario about reaching consensus on a single policy statement in a group of five people, in which volunteers from a introductory course were placed randomly. The videotapes were 20 minutes in length, recorded at different times in the meeting: minutes 10 to 15, 25 to 35, and 45 to 50. At the end of the discussion each participant voted for the emergent leader, defined as the most influential member in the group. Arm and shoulder movements were found to be the main nonverbal

visual cues contributing to participants' perception of leadership. Additionally, gesticulation of shoulders and arms were significantly correlated with eye contact, head agreement, and facial agreement.

The relationship between leadership and several personality traits is also of interest to social psychologists. It has been shown that cognitive ability and two personality traits of the Big-Five model [31] (extroversion and openness to experience) were predictive of emergent leadership behaviors [33]. Groups of four to six participants enrolled in a course, took part in a winter survival simulation, filled in questionnaires of personality, cognitive ability, teamwork effectiveness and emergent leadership. The emergent leader was designated as the one receiving the highest rating scores from the group through measures of interpersonal and self-management behavior, as well as task-related behaviors of a leader. The emergent leaders scored higher on cognitive ability and the personality traits of extroversion and openness to experience. Another study [32] investigated the relationship between leadership style and sociable and aggressive dominance in the context of three unacquainted people trying to decide on the top five candidates out of a group of ten persons who wanted to rent a room. The 20 minute group discussion was recorded, and responses to questionnaires (first glance impression of dominance, socio-emotional and task leadership) were complemented with observations of nonverbal behavior. It was found that although both types of dominance have characteristics that lead to leadership, there was a higher correlation between leadership and social dominance.

It has also been shown that socially dominant people receive more frequent and longer lasting glances from the group, look at others more while speaking, use more gestures, talk more, and take longer turns [38]. On the other hand, aggressively dominant people often attempt to interrupt more, and look at others less while listening [32].

Finally, in another related area, the relationship between dominance and influence in face-to-face groups was analyzed in [2]. Four-person groups of unacquainted people were recorded during 45 minutes while creating an organization and outlining its strategy. A self-dominance report questionnaire was administered, and group members also rated each other on influence, competence and personality. In addition, external observers rated each member along the same dimensions as above. The study concluded that dominant people influence their group more than individuals who are less dominant, by acting competent. In behavioral terms, and in order to attain this influence, they speak the most, and gain more control over the group and the group decisions.

In summary, the literature in psychology has found that human observers can identify emergent leaders in group interactions, and that specific behavioral cues do correlate with emergent leadership. These key findings provide the motivation and basic supporting evidence for our automatic approach.

## *B. Social Computing*

Several recent studies have proposed automated frameworks for analysis of individual and group communicative behavior from nonverbal cues [42], [18]. An extensive and recent summary of state-of-the art and recent approaches

on the analysis of group conversations, in particular modeling social verticality, can be found in [5]. In the context of groups, most existing approaches operate in a two-step process. In the first one, methods extract a number of features from audio signals (related to prosody and turn-taking) [46], [25], [43], video (related to head and body activity or gaze) [36], [23], and wearable sensors (related to body motion or physical proximity) [34], [37].

In the second step, these features are used as input to supervised or unsupervised learning models to infer traits like dominance [47], [27], extroversion and locus of control [43], [36]; relations like roles [57], [16], [17] or status [46], [25]; group attitudes like cooperation and competition [29], tasks like brainstorming [28]; and even fundamental concepts like collective intelligence [56]. Other works use the extracted features to create interactive systems that, through various visualizations of behavioral cues, affect the interaction itself [13], [34], [51], [6]. Our study has some common points with these recent works, in terms of deployed sensors and extracted nonverbal features. However, we address a different aspect of social interaction, namely emergent leadership. Furthermore, we propose a set of visual features that differs from previously investigated cues.

To our knowledge, there are few approaches centered on the computational analysis of emergent leadership. These works have focused on other forms of collaborative environments, such as virtual teams using email and instant messengers [11], social networks using virtual workspaces [53], and music performance [54]. In contrast, our work addresses emergent leadership from a face-to-face, nonverbal perspective, where sensing, feature extraction, and social inference are fully automated. A preliminary version of this work, that involved a subset of the full corpus presented here and only audio cues, was presented in short form at [48].

### III. OUR APPROACH

Several observable characteristics are related with emergent leadership, and the nonverbal behavior associated to some of those characteristics can be measured in accurate ways. To analyze the emergence of leadership in small groups, we collected two sets of data per group interaction. The first set includes audio-visual recordings of a group performing a survival task. The second set includes questionnaires filled by each group member, to capture how other participants are perceived by each other. From the questionnaires, we derived several variables as described in Section IV. From the recordings, we automatically extracted a number of nonverbal cues to characterize individual participants. We then analyze the correlation between variables derived from questionnaires and audio and visual features. After this, we develop methods to automatically infer the emergent leader using acoustic and visual nonverbal cues. Finally, we present an analysis of the effect of the temporal support needed to infer leadership from an interaction. Figure 1 summarizes our approach, highlighting the stages of our work:

We analyzed the data, as follows

#### 1) Data collection

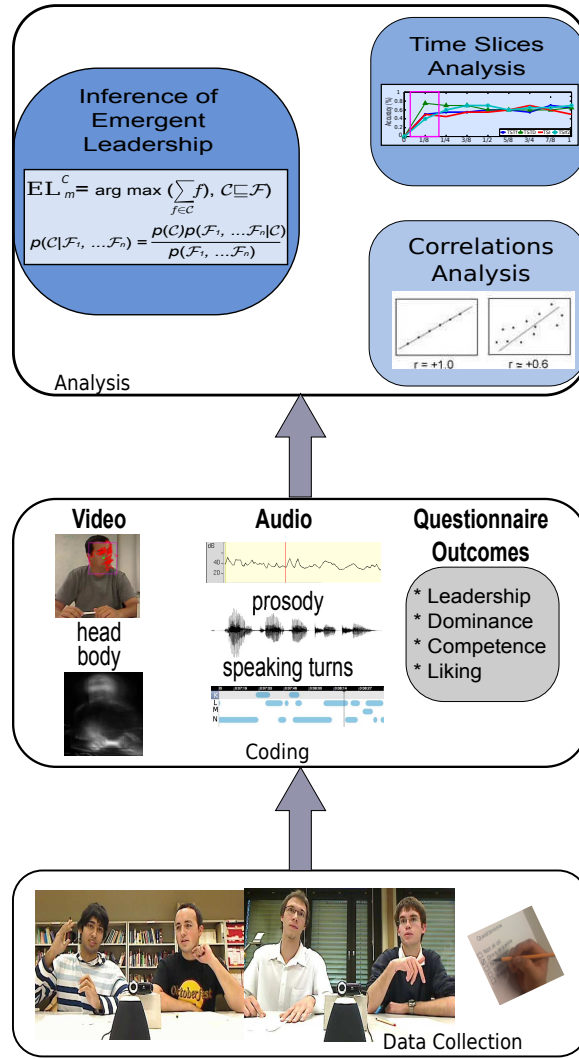


Fig. 1. Visualization of our approach.

- 2) Nonverbal feature extraction
- 3) Preliminary correlation analysis between questionnaire outcomes and nonverbal cues.
- 4) Inference of emergent leadership in small groups using nonverbal cues.
- 5) Analysis of the effect of thin slices in the estimation of emergent leaders.

#### IV. DATASET

The Emergent LEADER corpus (ELEA) consists of 40 meetings, corresponding to approximately 10 hours of recordings. There are 28 four-person meetings and 12 three-person meetings in newly formed groups, i.e. composed of previously unacquainted people. Average age is 25.4 years old (5.5 standard deviation), the gender distribution is 26 females and 46 males. Participants in ELEA meetings are asked to participate in a winter survival task with no roles assigned [33].

**Scenario:** To recruit participants, we posted advertisements in two well known Swiss universities and a business and management school in Switzerland asking for volunteers to participate in a study on casual social interactions. Volunteers were asked to participate in the study for approximately one hour. The recruitment process, questionnaires, and tests were available both in English and French. Volunteers were paid for their participation.

Participants first signed a consent form where it was mentioned that audio and video would be recorded during the group interaction, and data would be used for research purposes. After approval, they chose a letter identifier to preserve their names anonymous in the study. Then they filled questionnaires about themselves, performed the winter survival task, and finally they were asked to fill in questionnaires based on their perceived interaction.

**Sensing infrastructure:** With the aim of recording realistic interactions we chose non invasive audio and video sensors that allow freedom of movement, during the 15-minute face-to-face interaction. People discussed around a rectangular table, with one or two people on either side.

Audio recordings were gathered using the Microcone, a commercial microphone array [1], designed to record small discussion groups (up to 6 individuals) with audio sample rate of 16kHz. As shown in Figure 2, the Microcone (dark object at the center bottom of Figure 2-top) was placed in the center of the discussion table to capture the interaction. The Microcone automatically segments speakers, and provides audio for prosodic cue extraction.

For video recordings, we used two setups, one static setup with six cameras (four close-ups, two side-views, and one center-view), and one portable setup with two webcams (Logitech®Webcam Pro 9000). The video frame rate was 25 fps and 30 fps respectively. Taking advantage of the nowadays video recording devices' portability, in this work we are showing the feasibility to record data in any room, moving from the in lab approach to a more natural in field approach. Figure 2 also shows examples from the ELEA corpus from the portable setup and the static setup.

Among the 40 meetings in the ELEA corpus, 27 were completely recorded with the portable setup, and 10 with the static setup. In three meetings, the portable video recordings were not successfully recorded and thus discarded for experiments.

In the experiments described in Sections VII and VII-B we use the full audio ELEA corpus (40 meetings) and in Sections VII and VII-C we use the portable video corpus (27 meetings - ELEA AV) to control for variability in the video quality.

**Survival task:** For our study, we chose the winter survival task due to the fact that it is the most cited task in studies related to small group performance, decision making and leadership [33]. This scenario has been used as well in the social computing field to automatically estimate functional roles [14], [44], [57] and personality [43]. The task is focused on ranking a list of 12 items in order to survive an airplane crash in winter. The ranking is first performed individually and then as a team, to favor interaction among the participants and allow the emergence of

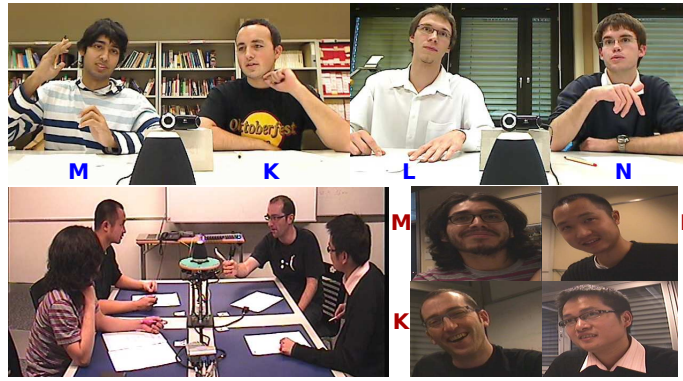


Fig. 2. The recording setups of the ELEA corpus: Top - Portable setup. Bottom - Static setup (Central and closeup views). K, L, M, N are the participants IDs

a leader. The group is aware that they have 15 minutes to discuss and come up with the final ranking list.

The Absolute Individual Scores and the Absolute Group Scores (AIS and AGS respectively) are calculated based on the absolute difference with respect to the survival experts rankings. Further the Absolute Individual Influence in the scores (AII) is calculated based on the absolute difference between the individual and the group ranking. Additionally we considered the influence in the top ranking items. We denote as  $NTII_i^j$  the number of top  $j$  items for participant  $i$  that also appear in the top  $j$  group rankings, and we denote as  $DTII_i^j$  the absolute distance from the top  $j$  individual items with respect to the group ranking (for  $j = 1, 2, 3, \dots, 10$ ). We considered up to the first 10 items in the analysis, given that the last two items are less relevant such that, there was no discussion at all on the ordering. Additionally correlations with respect to number of top individual items (NTII) is close to zero for item 11, and 0 for item 12.

**Questionnaires:** Before the task, participants filled in some tests that measure personality and personality dominance. After the survival test, participants were asked to answer 17 statements that capture how they perceived each participant, including themselves. This instrument was designed adapting existing questionnaires in leadership and dominance. 16 of the statements were evaluated on a five-point scale. The variables included in these statements are: Perceived Leadership (PLead: directs the group, imposes his or her opinion, is involved), Perceived Dominance (PDom: dominates, is in a position of power, asserts him- or herself), and Perceived Competence (PCom: is competent, is intelligent, has a lot of experience). The last statement asked for the Ranking of group Dominance (RDom) for all participants in the group, assigning 1 to the person considered the most dominant during the interaction, and based on the number of participants in the group 3 or 4 respectively to the least dominant. As a result, for each participant, we obtained three or four questionnaire outputs, depending on the number of participants in the group, which reflected the participants' perception. Although they are generally highly correlated we distinguished perceived leadership and perceived dominance to cover different aspects of verticality. Leadership focuses on the influence a person has on the other group members and the task solution during the interaction whereas dominance covers to extent to which a person seeks to stand out and control the others. Finally, participants provided some



demographic information including gender, age, etc.

## V. NONVERBAL FEATURE EXTRACTION

In this section, we present a description of the extracted audio and visual nonverbal features. The audio features include speaking turn and prosodic cues; the visual features include tracking-based features and motion template-based features.

### A. Audio Nonverbal Features

1) *Speaking Turn Features*: The Microcone automatically generates a binary speaker segmentation [1], using as a basic principle a filter-sum beamformer followed by a post-filtering stage, for each of the six spatial segments of the microphone array. The segmentation is stored in a file containing relative time in seconds (start and end), the subject label, and the Microcone sector. Similar techniques (e.g. [39]) have shown that the performance in terms of speech quality is relatively close to the performance using headset microphones, and better than lapels (Signal-to-noise ratio enhancement of 20.63dB for single speakers and 11.27dB in crosstalk). We did not evaluate objectively the quality of the speaker segmentation; but inspected many files and observed that the speaker turns (even if they are short) are detected correctly by the device; furthermore, the device can recover turns' beginning and endings well. Note that as our study aims at aggregating features over longer periods of time, the features tolerate minor errors in the estimation of exact boundaries of speaker turns.

The speaker segmentation results in a binary segmentation for each participant, where status 1 represents speech and status 0 represents non-speech. From the binary segmentation, we compute the following features for each participant:

**Total Speaking Length ( $TSL_i$ )**: The total time that participant  $i$  speaks according to the binary speaking status.

**Total Speaking Turns ( $TST_i$ )**: The number of turns accumulated over the entire meeting for each participant  $i$ , where each turn is a segment defined by a series of active speaking status. We added a variant ( $TSTf_i$ ) which only accumulates turns longer than two seconds.

**Average Speaking Turn Duration ( $AST_i$ )**: The average turn duration per participant  $i$  over the entire meeting.

**Total Successful Interruptions ( $TSI_i$ )**: We use two definitions to calculate this feature:

$TSI_i^1$ : Participant  $i$  interrupts participant  $j$  if  $i$  starts talking when  $j$  is speaking, and  $j$  finishes his/her turn before  $i$  does.

$TSI_i^2$ : Participant  $i$  interrupts participant  $j$  if  $i$  starts talking when  $j$  is speaking; when  $i$  finishes his/her turn  $j$  is not speaking anymore.

For each of the two cases, we added a variant ( $TSIf_i^1$  and  $TSIf_i^2$ ) which accumulates interruptions for turns longer than two seconds.

**Speaking Turn Matrix (STM):** The matrix which counts, as events, who speaks after whom over the entire meeting.

2) *Prosodic nonverbal cues:* With the speaker segmentation, we obtain the speech signal for each participant. We then compute two well known prosodic speech features, energy and pitch (the perceived fundamental frequency ( $F_0$ ) of voice, and it is the rate of vibration of vocal cords). To extract energy, we used Wavesurf, an open source software package. For pitch extraction we used a robust method proposed in [52]. The following variables were computed from energy and pitch:

**Energy Spectral flatness (ESF):** Is a measure often used to discriminate between voiced and unvoiced speech [21] and it is calculated as:

$$ESF = 10 * \log \frac{(\prod_{i=1}^n a_i)^{\frac{1}{n}}}{\frac{1}{n} \sum_{i=1}^n a_i} \quad (1)$$

where  $a_i$  denotes the magnitude of each of the spectral lines  $i$ , and  $n$  is the number of spectral lines.

**Energy variation (EVT):** This feature measures the variation in energy, meaning the loudness perceived by the ear. It is computed dividing the standard deviation by the mean.

We also estimated some statistics from the energy extracted from single speaking turns, like minimum, maximum, median and variance (denoted EMIN, EMAX, EMED, and EVAR).

**Pitch variation (PVT):** This feature measures the pitch variability. It is calculated dividing the standard deviation by the mean.

We also calculated some statistics from the  $F_0$  from single speech per participant, PMIN, PMAX, PMED, and PVAR.

## B. Visual Nonverbal Features

### 1) Tracking-based features:

a) *Head activity:* Figure 3 summarizes the feature extraction process for the head activity. To measure the head activity of each participant, we first tracked the face with a Particle Filter (PF), using an ellipse face model [24]. The dynamic model of the PF uses a damped velocity model for the position and velocity, and a random walk model for the shape parameters (i.e., the size of the ellipse). As observations, we use the skin color probability image, which has a positive probability for skin color pixels and zero probability for other colors. Skin color models are learned on additional data to calculate the likelihood, we make two measurements based on the ellipse that is defined by the state vector of the particle: The ratio of the skin colored pixels to the total number of pixels (i) inside

the ellipse, and (ii) at the boundary of the ellipse. High likelihood is assigned to cases where the first measurement is high and the latter is low. We additionally apply the mean shift algorithm to move the particle centers to the areas with high skin color probability. This allows to use particles effectively, and requires fewer particles than a standard PF. More details can be found in [3].

Once the face area is estimated by the PF, the optical flow vectors within the face area of two successive frames are calculated to have a fine-grained analysis of head movements. We use the hierarchical Lucas-Kanade optical flow algorithm, using points selected from the face area that indicate strong corners. The OpenCV library is used for the implementation of the optical flow algorithm [10].

Using the optical flow vectors, we calculate the average motion vector to get the average head motion on the  $x$  and  $y$  dimensions. For each participant, we obtain two real-valued vectors,  $hR_x$  and  $hR_y$  with elements  $hR_{x,t}$ ,  $hR_{y,t}$ , one for each dimension, describing the head activity of that participant during the whole meeting.

Furthermore, to identify significant head activity, we first binarized these vectors via automatic thresholding, obtaining the binary vectors  $hB_x$ ,  $hB_y$  with elements  $hB_{x,t}$ ,  $hB_{y,t}$ . The automatic threshold for the  $x$  dimension eliminates small movements, i.e. movements of anxious people, it is calculated as  $\mu_x + \sigma_x$ , where  $\mu_x$  and  $\sigma_x$  are the mean and standard deviation of  $hR_x$  respectively. Computed for each participant in each meeting, the values above the threshold are set to 1, indicating a significant head activity, and rest to 0. This calculation is repeated for the  $y$  dimension as well. The final binary head activity vector,  $hB$ , is then calculated by an OR operation:

$$hB = hB_x \vee hB_y. \quad (2)$$

For each participant, the following features are calculated using  $hR_x$ ,  $hR_y$ , and  $hB$ , which represent the participant's head activity during the meeting.

**Head activity length ( $THL_i$ ):** The total time that participant  $i$  moves his/her head, calculated from  $hB$ .

**Head activity turns ( $THT_i$ ):** Number of turns for each participant  $i$ , where each turn is considered as a continuous head activity, calculated from  $hB$ .

**Head activity average turn duration ( $AHT_i$ ):** The average turn duration for participant  $i$ , calculated from  $hB$ .

**Standard deviation of head activity ( $stdHx_i$ ,  $stdHy_i$ ):** Standard deviation of head activity in  $x$  and  $y$  dimensions, calculated from  $hR_x$  and  $hR_y$ .

*b) Body activity:* Figure 4 summarizes the process for body activity feature extraction. It is measured by simple motion differencing as the background is stationary. Hence, all the moving pixels outside the tracked head area are considered as belonging to the body area. Each frame is converted to a grayscale image,  $F_t$ , and the difference image,  $\Delta_t = F_t - F_{t-1}$  is calculated.

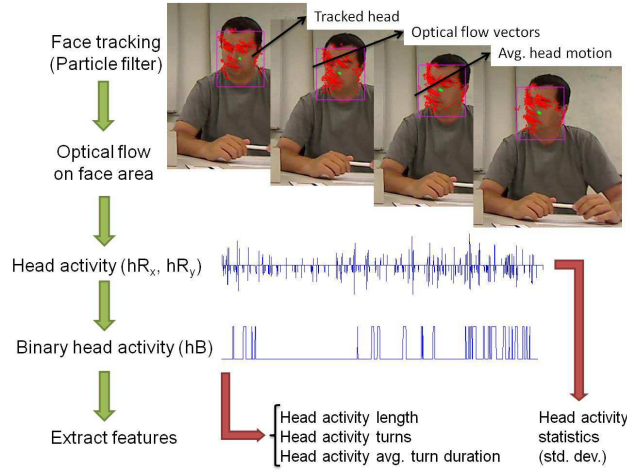


Fig. 3. Head activity feature extraction.

The difference image is thresholded to identify the moving pixels, and then the total number of moving pixels in each frame, normalized by the frame size  $S$ , is recorded. We use a manually selected threshold ( $Th_g = 30$ ) for this purpose, which means that if the difference between the grayscale values of two pixels is greater than this threshold, it is considered as a moving pixel. For each participant, this results in a real-valued vector  $bR$  with elements  $bR_t$  describing the body activity of that participant during the whole meeting:

$$bR_t = \frac{1}{S} \sum (\Delta_t > Th_g). \quad (3)$$

Furthermore, to identify significant body activity, we binarized this vector with a threshold  $Th_f = 0.05$ , (i.e., if 5% of the pixels are moving in that frame, it is considered as a significant body activity), obtaining the binary vector  $bB$ . This threshold value is set such that it captures the global body movements (e.g., leaning), filtering out the local ones.

$$bB_t = \begin{cases} 1, & \text{if } bR_t > Th_f \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

It is important to note that the values of the thresholds are chosen with respect to the video recordings in the ELEA corpus. For different video recordings, different threshold values would be needed.

For each participant, using  $bR$  and  $bB$ , the following features, which represent the participant's body activity during the meeting, are calculated.

**Body activity length ( $TBL_i$ ):** The total time that participant  $i$  moves his/her body, calculated from  $bB$ .

**Body activity turns ( $TBT_i$ ):** The number of turns for each participant  $i$ , where each turn is considered as continuous body activity, calculated from  $bB$ .

**Body activity average turn duration ( $ABT_i$ ):** The average turn duration for participant  $i$ , calculated from  $bB$ .

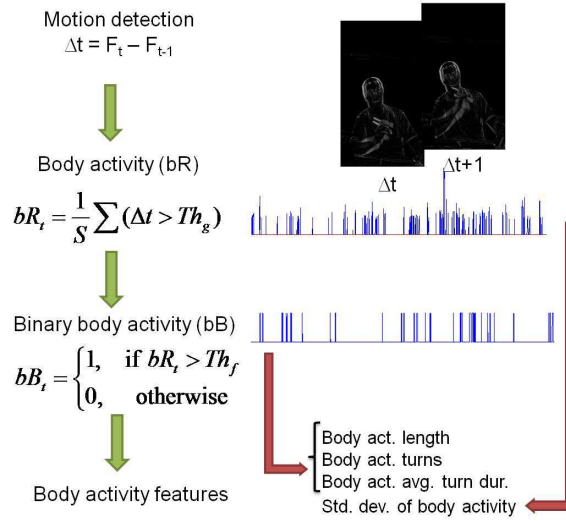


Fig. 4. Body activity feature extraction

**Standard deviation of body activity ( $stdB_i$ ):** Standard deviation of body activity, calculated from  $bR$ .

2) *Motion template based features:* As an alternative approach to characterize visual activity, we use motion templates to extract the full body activity features of each participant throughout the meeting. Bobick and Davis proposed the Motion Energy Image (MEI) and the Motion History Image (MHI) as ways to summarize the spatio-temporal content in a single image [9]. MEI is a binary image showing the location of the motion, whereas MHI is a grayscale image showing both the location and the direction of the motion. Both MEI and MHI are proposed as motion templates to describe short motion, mainly for human action recognition. We propose a modified version of MEI, what we call Weighted Motion Energy Image (wMEI) illustrated in Figure 5. wMEI is proposed to represent the dominant motion regions, and is suitable to be employed as a template for long duration videos. It is a gray scale image describing the location along with the intensity of motion throughout the video in that region.

A wMEI contains the accumulated motion information and is calculated as:

$$wMEI_p(x, y) = \frac{1}{N_p} \sum_{t=1}^T (D_p^t(x, y, t)), \quad (5)$$

where  $D_p^t(x, y, t)$  is a binary image that shows the moving regions for participant  $p$  at time  $t$ ,  $N_p$  is the normalization factor, and  $T$  is the total number of frames. Unlike motion energy images, wMEI is not a binary image. In wMEI, the brighter pixels correspond to regions where there is more motion. wMEI can be normalized by dividing all the pixel values by the maximum pixel value. Alternatively, the length of the video can be used as a normalization factor. Thus, a normalized wMEI describes the motion throughout the video as a gray scale image, where each pixel's intensity indicates the visual activity in that pixel.

For each participant, we calculate the wMEI and extract several statistics as body activity features. These include

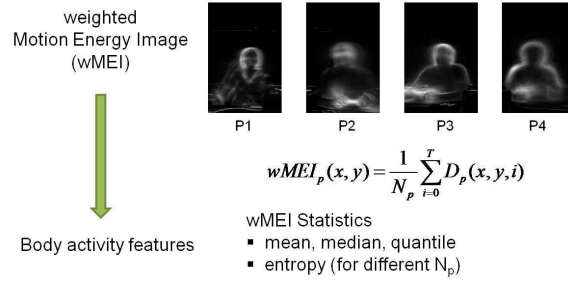


Fig. 5. Weighted motion energy image based body activity feature extraction

the maximum ( $wMEImx_i$ ), mean ( $wMEImn_i$ ), median ( $wMEImd_i$ ), and 75% quantile ( $wMEIqn_i$ ) of the intensity value of wMEI. For mean, median and quantile calculation, we omit zero values in the wMEI and only use the non-zero intensities. In addition to these statistics, we also calculate the entropy. For entropy, we follow three different approaches to obtain the normalized wMEIs on which the entropy is calculated:

- 1)  $wMEIeP$ :  $N_p = \max(\sum_{t=1}^T (D_p^t))$ .
- 2)  $wMEIeA$ :  $N_p = \max(N_{p,1}, N_{p,2}, \dots, N_{p,P})$ .
- 3)  $wMEIeT$ :  $N_p = T$ .

$N_p$  is the normalization factor used in Eq. 5, and  $P$  is the number of participants in a meeting. The first approach,  $wMEIeP$ , uses the maximum value in the wMEI of each participant as the normalization factor. This value is unique for each participant in each meeting. The second and third approaches use a single normalization factor for all participants in the meeting: in  $wMEIeA$  the normalization factor is calculated as the maximum intensity in all the wMEIs of participants in the meeting,  $N_{p,i}$  and  $wMEIeT$  uses the length of the video.

### C. Discussion about selected features

The set of audio features presented in Sections V-A1 and V-A2 can be considered as state-of-the-art in nonverbal behavior analysis, as reported in [18]. In addition, similar features have been used to recognize dominant people [27], [4], roles [45], and to discover leadership styles [26]. Other research works have also used similar features, e.g. [47] to identify concepts like dominance or influence [42]. While a few variants of prosodic measures and several voice quality measures have been proposed [12], the voice quality measures have shown similar performance than prosodic measures, nevertheless, voice quality features are more complex to compute. Taking into account this finding, and considering social vertical analysis works, the features we use in this paper are considered as a good choice.

The visual features are as well similar to the ones used in the state-of-the-art on nonverbal communication to characterize the total amount of a person's physical activity [27], [14], [44]; we also propose a novel set of visual features based on motion templates(wMEI). Although there are other visual features (e.g. gaze or facial expression) of potential relevance to characterize leadership, the features presented in Sections V-B1a, V-B1b, and V-B2 are

robust and have been frequently tested in social verticality analysis works [27], [14], [44].

A more detailed literature review on the audio and visual feature extraction and processing for nonverbal communication analysis can be found in [5].

## VI. ESTIMATING THE EMERGENT LEADER

It has been shown in social psychology research that the speaking time has a stronger association with individual dominance than other features, such that people who talk more have more chances to contribute in group interaction between strangers [38]. Similarly to individual dominance, emergent leaders contribute more than nonleaders in a group discussion. If the participation in the group is quantified in single nonverbal behavior variables (like head agreement, postural shift, or rate of verbal participation) each variable alone is a significant predictor of leadership [7], [50]. Considering that there is evidence that the emergent leader can be assessed from single features nonverbal features, we present nonsupervised methods that consider single nonverbal feature methods, as well as supervised methods with combination of features.

We use four approaches to estimate the emergent leader in each group: (i) A rule-based approach, in which the participant with the highest nonverbal feature value in the group is selected as the leader; (ii) rank-level fusion which is an extension of the rule-based estimator to handle fusion of multiple features; (iii) support-vector machine, a supervised learning method. and (iv) a collective classification approach, which uses relational information in addition to the nonverbal feature vector.

### A. Rule-based estimator approach

For the task of estimating the emergent leader, our hypothesis is that the emergent leader in a group is the one who has the highest value of a single nonverbal feature (i.e., the participant with the longest total speaking time). We define a rule-based estimator that selects the participant with the maximum feature value in the group as the emergent leader. Thus, we estimate the leader  $EL_m^f$  for group  $m$  according to feature  $f$  as

$$EL_m^f = \arg \max_p (f_p^m), p \in \{1, 2 \dots P\}, \quad (6)$$

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

### B. Rank-Level Fusion

To investigate whether the combination of features has an advantage over using single features, we fuse rule-based estimators defined on different individual features, and used the ranked feature values of each estimator as recently proposed in [4]. Instead of selecting the participant with the maximum feature value, the participants are

ranked and the rank information is used to fuse different estimators based on different features. For group  $m$ , using feature combination  $\mathcal{C}$ , we sum up the ranks for each participant and select the participant with the highest total rank as the estimated leader:

$$EL_m^{\mathcal{C}} = \arg \max_p \left( \sum_{f \in \mathcal{C}} r_{f_p}^m \right), \quad \mathcal{C} \subseteq \mathcal{F}, \quad (7)$$

where  $r_{f_p}^m$  is the rank of participant  $p$  using feature  $f$  in group  $m$ , and  $\mathcal{F}$  is the set of all features. In case of ties, we select the leader based on the z-normalized scores [4].

### C. Support Vector Machine

As a supervised alternative we used a support vector machine (SVM), a supervised learning method that constructs an hyperplane by mapping the nonverbal input vector in higher dimensions.

$$\sum_j \alpha_j K(x_j, x) = C \quad (8)$$

Where  $K$ , represents the kernel function, in this case a linear kernel,  $\alpha$  parameter that represents a linear combination,  $C$  a constant value, and  $x_j$  the input vector composed of nonverbal features. As implemented in [27], we use the SVM score to rank each participant in the group. The rankings are then used to determine which participant is assigned the Emergent-Leader person label, by considering the point which is furthest from the class boundary. This procedure generates exactly one Emergent-Leader person in the group. For training and testing, we applied the leave-one-meeting-out approach, and the test accuracy is calculated based on the average performance.

### D. Collective Classification

We also investigated a novel approach based on statistical relational learning. Nowadays networked data is ubiquitous, and the relation among instances has been exploited in several ways, ranging from classifying scientific papers with related topics to finding ways to understand centrality in online communities, and the propagation of ideas or opinions [41], [19], [37].

In a network of data, the data instances are related in some ways, and this relation can be learned to infer several instances simultaneously. This is the aim of collective classification [40], [30]. The label estimation for a data point can be influenced by estimations of its neighboring labels.

Taking into account that our data is not independent and possibly not identically distributed, we propose to investigate collective classification in our problem. A collective approach improves probabilistic inference when the data is relational and correlated. In the context of web data analysis, it has been proved that adding relational information when instances are not independent improves inference [30]. As we mentioned in section II-B, there



are nonverbal speaking features highly correlated with dominance, and dominance is also correlated with emergent leadership, as described in section II-A. Our hypothesis is that by considering the relational information and given that the data is correlated, collective inference can improve the leader estimations performed using non-collective approaches.

The data is modeled as follows: We have a graph  $G = (V, E, X, Y, C)$  where  $V$  is the set of participants  $v_i \in V$ ,  $E$  is a set of directed edges, coded from the speaking turn matrix (STM), each  $x_i \in X$  is an attribute vector composed of nonverbal features for participant  $v_i$ , each  $y_i \in Y$  is a label variable for  $v_i$ , and  $C$  is the set of possible labels (i.e. 1 for Emergent Leader or 0 for NonEmergent Leader). Figure 6 shows the model.

To perform collective classification in an efficient way, the Iterative Classification Algorithm (ICA) has been defined [40]. The algorithm makes an initial label estimation  $y_i$  for each  $v_i$ , then iteratively re-estimate the labels based on the estimations for every participant that is interacting with  $v_i$ .

There are two tasks that can be performed using the ICA algorithm, named out-of-sample and in-sample [40]. For the in-sample task, we are given a set of known labels  $Y^K$  for a subset of participants  $V^K \subset V$ , so that  $Y^K = \{y_i | v_i \in V^K\}$ . Then, the task is to infer  $Y^U$ , the values of  $y_i$  for the remaining participants with unknown labels ( $V^U = V - V^K$ ), or a probability distribution over those values. We implemented the three variants for the ICA algorithm described in [40], ICA, ICA<sub>kn</sub> and ICA<sub>c</sub>. All these three algorithms are based on iterations over 5, ICA considers all the estimations from the previous iteration, ICA<sub>kn</sub> uses only known labels  $V^K$  in the first iteration, and from the second to the last iteration it works like ICA. Finally, ICA<sub>c</sub> uses the known and the most confident estimated labels, and increases gradually the number of estimated labels in each iteration.

For the out-of-sample task, no labels are known, thus,  $V^K$  is empty and there are only two variants to the algorithm, namely ICA and ICA<sub>c</sub>.

Several relational features can be used in our problem. The simplest one is coded as a *count*, which represents the number of participants that take turns after participant  $i$  and that belong to a particular class. For instance,  $f_i(0) = 2$  indicates that two participants labeled as non-emergent leaders take turns after participant  $i$ . A second relational feature, called *proportion*, is coded as the proportion of participants taking turns after participant  $i$  and that have a particular label. For instance  $f_i(0) = 2/3$  indicates that three participants take turns after participant  $i$ , from which two are labeled as non-emergent leaders and the label for the third participant is unknown. Finally, the relational feature *multiset* produces a single numerical value for each possible label for the participants that take turns after participant  $i$ . This value can be compared against the mean value from the training set (missing labels are not used). For instance,  $f_i = \{1, 1, 1\}$  means that for participant  $i$  there is one participant labeled as non-leader that takes a turn after him, there is one participant labeled as leader that takes a turn after, and one more participant with an unknown label takes a turn after him.

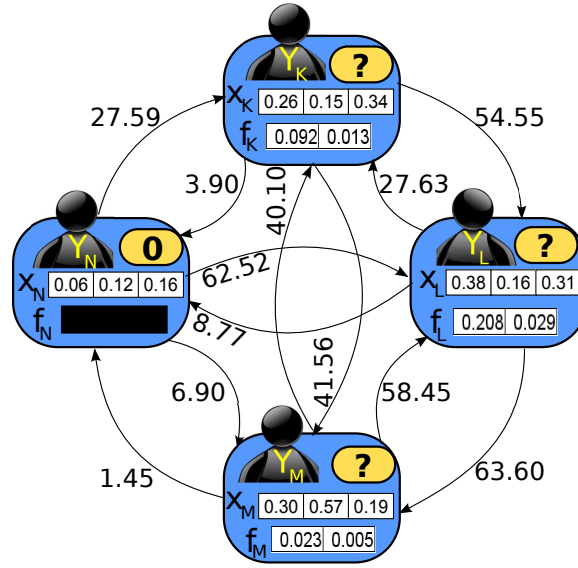


Fig. 6. Data modeled for collective classification algorithm. The weighted links between participants represents percentage of turns taken (the direction indicates who takes the turn).  $x_i$  shows values from three audio features: TSL, AST and TSI<sup>1</sup>. In this case we have a known label, participant N is non-emergent leader ( $y_N = 0$ ). The relational feature *weighted proportion*  $f_i$  is estimated considering the known label  $y_N$ , and the number of participants that have a turn before.

To our knowledge, weighted links have not been explored as a potential relational feature. Given that we have the weights that represent the amount of turns that participants take during the 15-minute interaction, we defined a new relational feature named *weighted proportion*. This relational feature considers weights, direction, and number of participants taking turns after participant  $i$  does. For instance, from figure 6  $f_K(0).IN = (0.2759)/3$  and  $f_K(0).OUT = (0.039)/3$ , where  $f_K(0).IN$  represents that participant  $K$  takes turns 27.59% of the time after participant  $N$  (labeled as class 0) does, the value is then divided by the number of neighbors, i.e. the number of participants that have turns before  $K$  takes a turn.

The ICA algorithm requires a local classifier for training and for the initial labeling. The variant  $ICA_c$  needs as well the confidence values for the labels. For confidence estimation, we use the posterior probability for the most likely label for participant  $v_i$ , calculated with a naive Bayes classifier. The local classification is performed as well using a naive Bayes classifier. For training and testing, we applied the leave-one-meeting-out approach, and the test accuracy is calculated based on the average performance.

#### E. Other possible relational models

Note that other models that consider relational information have been used to predict functional roles in meetings [14]. The influence model takes into account dependencies between pairwise chains with the aim of estimating amount of influence. We have decided to leave the exploration of other relational models as part of future work.

TABLE I

PEARSON CORRELATION VALUES BETWEEN VARIABLES FROM QUESTIONNAIRES OUTCOMES. SIGNIFICANCE VALUES \* :  $p < 0.005$ ,  
+ :  $p < 0.05$ .

	PLead	PDom	PCom	PLike	RDom
PLead		0.81*	0.31*	-0.29+	0.75 *
PDom			0.28*	-0.32*	0.73*
PCom				0.20	0.29*
PLike					-0.33*
RDom					

## VII. EXPERIMENTS AND RESULTS

In this section, we first present a correlation analysis between self-reported questionnaires and nonverbal features, we then present results on leadership estimation. Finally, we report results on the effect of the observation window.

### A. Correlation Analysis

We used the full ELEA corpus for this analysis, variables and speech features, and for the visual analysis we used the subset ELEA AV. We validate correlations, calculating the Pearson correlations per group, then applying a Fisher transformation, and finally we test if the correlations are statistically significant with a t-test, at 5% significance level (i.e.,  $p < 0.05$ ).

**Questionnaire output analysis.** First, we analyse the correlation of the questionnaire outputs filled by people after the interaction. Each perceived variable is averaged over all participants per group, and the group ranking is normalized according to the number of participants per group. Table I shows the Pearson correlation values. PLead shows significant correlation with PDom and RDom (0.81 and 0.75, respectively). These results suggest that the emergent leader is perceived as a dominant person by the other participants. Interestingly, the correlation between perceived leadership and competence is significant but less strong, and lower between perceived or ranked dominance and competence, which suggests that participants might not have used often the latter construct as part of their judgments.

**Survival task top ranking analysis.** Given that the task in the groups is to come up with a group rank list (composed of twelve items), we review the correlations with the aim of discovering the individual influence in the group. We analyze the correlation between the number of individual items in the top group list against the perceived variables from questionnaires. We use two approaches: In the first one, we count the number of items in the top group rank (see table II); In the second approach, we consider the absolute difference of the individual items with respect to the top group rank, and normalize with respect to the number of items in the top rank (Table III). If one item is not in the top rank, it is assigned with the maximum distance + 1. From Table II we can see that the emergent leader did not necessarily convince the group to select his/her two top individual items in the group rank, in contrast with the participants that were ranked as the most dominant (RDom). On the other hand,

TABLE II

CORRELATION VALUES BETWEEN VARIABLES FROM QUESTIONNAIRES AND THE NUMBER OF INDIVIDUAL ITEMS IN THE TOP RANK LIST FROM THE WINTER SURVIVAL TASK. SIGNIFICANCE VALUES \* :  $p < 0.005$ , + :  $p < 0.05$ .

	PLead	PDom	PCom	PLike	RDom
TOP1	0.18	0.15	0.03	-0.01	0.248 <sup>+</sup>
TOP2	0.19	0.18	0.04	-0.003	0.17
TOP3	0.32 <sup>+</sup>	0.36*	0.15	0.003	0.28 <sup>+</sup>
TOP4	0.32 <sup>+</sup>	0.35 <sup>+</sup>	0.16	-0.03	0.308 <sup>+</sup>
TOP5	0.23 <sup>+</sup>	0.14	0.15 <sup>+</sup>	0.06	0.158 <sup>+</sup>
TOP6	0.28 <sup>+</sup>	0.22 <sup>+</sup>	0.21	-0.05	0.24 <sup>+</sup>
TOP7	0.27*	0.28 <sup>+</sup>	0.20 <sup>+</sup>	-0.08	0.22 <sup>+</sup>
TOP8	0.28 <sup>+</sup>	0.24 <sup>+</sup>	0.39*	0.16	0.17
TOP9	0.13	0.20 <sup>+</sup>	0.21 <sup>+</sup>	0.21	0.19
TOP10	0.01	0.05	-0.03	0.09	0.18

TABLE III

CORRELATION VALUES BETWEEN VARIABLES FROM QUESTIONNAIRES AND ABSOLUTE DISTANCE IN TOP RANK ITEMS FROM THE WINTER SURVIVAL TASK. SIGNIFICANCE VALUES \* :  $p < 0.005$ , + :  $p < 0.05$ .

	PLead	PDom	PCom	PLike	RDom
TOP1	-0.17	-0.17 <sup>+</sup>	-0.05	0.01	-0.23 <sup>+</sup>
TOP2	-0.18	-0.18 <sup>+</sup>	-0.03	-0.01	-0.15
TOP3	-0.28 <sup>+</sup>	-0.32 <sup>+</sup>	-0.11	-0.003	-0.27 <sup>+</sup>
TOP4	-0.36 <sup>+</sup>	-0.37*	-0.15	0.02	-0.33*
TOP5	-0.32*	-0.30 <sup>+</sup>	-0.16	-0.06	-0.23*
TOP6	-0.37*	-0.35*	-0.22 <sup>+</sup>	-0.02	-0.28*
TOP7	-0.31*	-0.32*	-0.21 <sup>+</sup>	0.02	-0.24 <sup>+</sup>
TOP8	-0.32*	-0.31*	-0.40*	-0.17	-0.24*
TOP9	-0.25*	-0.28 <sup>+</sup>	-0.24 <sup>+</sup>	-0.18	-0.28 <sup>+</sup>
TOP10	-0.25 <sup>+</sup>	-0.31*	-0.08	-0.05	-0.32*

stronger effects are observed both for leadership and dominance when one allows more items in the top group rank (top 3 - top 8)

From Table III we can see another facet of the influence that the emergent leader has with respect to the final group ranking. In particular, the most dominant people (PDom and RDom) might try to make the final group rank as similar as possible to their individual ranking list (TOP 1 to TOP 3). In this case, negative correlations are due to the absolute distance: the closest the individual list with respect to the group list, the smallest the difference. As shown in [32], dominant people tend to get their way in small group tasks related to ranking preferences.

Finally, we explored as well the individual performance in the survival task (AIS), the only significant finding was a correlation of value  $-0.21$  between AIS and PCom, with  $p=0.01$ . This might suggest that the individual performance in the ranking task, has a slight effect in the perception of competence from the group.

**Nonverbal speaking behavior and perception from participants.** Table IV shows Pearson correlation values between questionnaire outputs and individual audio nonverbal features. As we can see, there is a correlation between several features and PLead, suggesting that emergent leadership perception has a connection to the person who talks the most, has more turns, and interrupts the most. Furthermore, several nonverbal cues have also correlation (although with lower values) with perceived or ranked dominance. This confirms previous work showing that these features are reasonably correlated with dominance in groups [18] [38]. Finally, the interruptions (TSIf<sup>2</sup>) have a medium correlation with judgment of competence [2]. As shown in [55], emergent leaders do not necessarily have

TABLE IV

CORRELATION VALUES BETWEEN VARIABLES FROM QUESTIONNAIRES AND NONVERBAL ACOUSTIC FEATURES ON THE FULL ELEA CORPUS. SIGNIFICANCE VALUES \* :  $p < 0.005$ , + :  $p < 0.05$ . FOR ENERGY AND PITCH FEATURES, ONLY SIGNIFICANT CORRELATIONS WITH AT LEAST ONE OF THE CONCEPTS ARE SHOWN.

	PLead	PDom	PCom	PLike	RDom
TSL	0.52*	0.44*	0.14	-0.12*	0.34*
TST	0.35 <sup>+</sup>	0.28	0.19	0.05	0.07*
TSTf	0.47*	0.40*	0.2	-0.06 <sup>+</sup>	0.22*
AST	0.47*	0.39*	0.16	-0.10 <sup>+</sup>	0.28*
TSI <sup>1</sup>	0.40*	0.33*	0.17	-0.01 <sup>+</sup>	0.19*
TSIf <sup>1</sup>	0.45*	0.40*	0.17	-0.07 <sup>+</sup>	0.27*
TSI <sup>2</sup>	0.28 <sup>+</sup>	0.27*	0.15	0.01	0.12*
TSIf <sup>2</sup>	0.46*	0.41*	0.20 <sup>+</sup>	-0.03 <sup>+</sup>	0.30*
EMIN	-0.36*	-0.24 <sup>+</sup>	-0.03	0.25	-0.29 <sup>+</sup>
EMED	-0.19 <sup>+</sup>	-0.15	0.01	-0.16	-0.21
PVAR	-0.13	-0.19	-0.12	0.05	-0.27 <sup>+</sup>
PVT	-0.12	-0.20 <sup>+</sup>	-0.01	0.03	-0.22 <sup>+</sup>

TABLE V

CORRELATION VALUES BETWEEN VARIABLES FROM QUESTIONNAIRES AND NONVERBAL VISUAL FEATURES ON ELEA AV CORPUS. SIGNIFICANCE VALUES \* :  $p < 0.005$ , + :  $p < 0.05$

	PLead	PDom	PCom	PLike	RDom
THL	0.12	0.11	0.27 <sup>+</sup>	0.11	0.15
THT	0.25 <sup>+</sup>	0.25 <sup>+</sup>	0.31*	-0.04	0.12
AHT	-0.17	-0.18	-0.30 <sup>+</sup>	0.08	-0.06
stdHx	-0.04	-0.02	-0.12	0.06	-0.02
stdHy	-0.13	-0.18	-0.28 <sup>+</sup>	0.08	-0.15
TBL	0.37 <sup>+</sup>	0.34 <sup>+</sup>	0.001	-0.19	0.27 <sup>+</sup>
TBT	0.34 <sup>+</sup>	0.29 <sup>+</sup>	-0.03	-0.18	0.23
ABT	0.27 <sup>+</sup>	0.27 <sup>+</sup>	0.001	-0.23	0.23 <sup>+</sup>
stdB	0.32 <sup>+</sup>	0.33 <sup>+</sup>	-0.06	-0.16	0.17
wMEIeP	0.26	0.19	-0.04	-0.24 <sup>+</sup>	0.18
wMEIeT	-0.05	-0.10	0.17	0.19	0.01
wMEIeA	0.42 <sup>+</sup>	0.37 <sup>+</sup>	-0.03	-0.26 <sup>+</sup>	0.25 <sup>+</sup>
wMEImx	0.27 <sup>+</sup>	0.25 <sup>+</sup>	0.05	-0.08	0.20
wMEImn	0.31 <sup>+</sup>	0.30 <sup>+</sup>	-0.01	-0.08	0.14
wMEImd	0.37 <sup>+</sup>	0.27	-0.07	-0.26 <sup>+</sup>	0.25 <sup>+</sup>
wMEIqn	0.36*	0.32 <sup>+</sup>	0.04	-0.13	0.23 <sup>+</sup>

to be the highest participators when they are perceived as competent in a task which could be interpreted from no significant correlations between PCom and the speaking turn features, although moderate correlation was found between competence and leadership and dominance in Table I.

**Nonverbal visual behavior and perception from participants.** We use the 27 meetings recorded with the portable setup from the ELEA corpus that include both audio and video recordings, which we call ELEA Audio-Visual (AV) corpus. Pearson correlation values between individual visual nonverbal features and questionnaire outputs are shown in Table V. Significant correlations can be observed between PLead and body activity (TBL, TBT, ABT, and stdB), and PLead and motion statistics (wMEImx, wMEImn, wMEImd, and wMEIqn). Which supports Baird's affirmation [7], that gesticulation of arms and shoulders is an important contributor in the perception of emergent leadership. PDom and RDom have as well significant correlations with body activity (TBL and ABT) and motion statistics, as exposed in [15], dominant individuals are highly noticeable by their body movements and gestures, in association with their vocal cues.

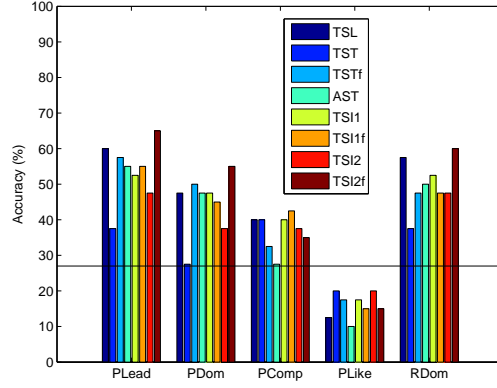


Fig. 7. The accuracy of speaking turn features on the ELEA full corpus and rule-based estimation. The black horizontal line shows the random baseline.

TABLE VI

RESULTS OF RANK-LEVEL FUSION ON THE FULL ELEA CORPUS. THE FEATURES COMBINED ARE LISTED IN THE LAST COLUMN.

	Acc(%)	Fused variables
PLead	72.5	AST, TSI <sup>2</sup> , TSI <sup>2f</sup> , EMED, EVAR
PDom	65	AST, TSI <sup>1</sup> , TSI <sup>2f</sup> , EVAR, PMED
PCom	55	TST, TSI <sup>1</sup> , TSI <sup>1f</sup> , TSI <sup>2f</sup> , EMIN, EVAR, PMIN, PMED
PLike	40	EMIN
RDom	72.5	TSL, AST, TSI <sup>2f</sup> , EMED

### B. Leadership Estimation using Audio Nonverbal Cues

In this section, we present the results for each of the estimation methods and the audio nonverbal cues. For the evaluation of our approach, we use the variables from the questionnaires as ground truth.

1) *Rule-based estimator*: We calculate the accuracy of the rule-based estimator by comparing the ground truth emergent leader with the participant who has the highest value for each of the nonverbal cues (Equation 6). Figure 7 shows the accuracy using single speaking turn features, where the best accuracy for variable PLead is achieved using TSI<sup>2f</sup> with 63.5%, followed by TSL with 60%. Random performance in this case is 27.5% given that the full corpus has 40 meetings, from which 28 meetings have 4 participants, and 12 meetings have 3 participants.

We also explored the performance of the prosodic features using the rule-based estimator. Figure 8 shows accuracy for energy and pitch, from which we can observe that all prosodic features performed better than speaking turn features for the variable PLike (e.g. EMIN, with 40.0%) and PCom (EMED, with 32.5%). Although the accuracy does not improve estimations performed with the top speaking turn features for variables PLead, PDom, PCom, and RDom, they do provide some discriminatory information.

2) *Rank-level fusion*: For the rank-level fusion the highest accuracy for PLead is 72.5% combining AST, TSI<sup>1</sup>, TSI<sup>2f</sup>, EMED, and EVAR. Table VI shows the combinations of relevant nonverbal audio features to estimate emergent leadership and other related concepts.

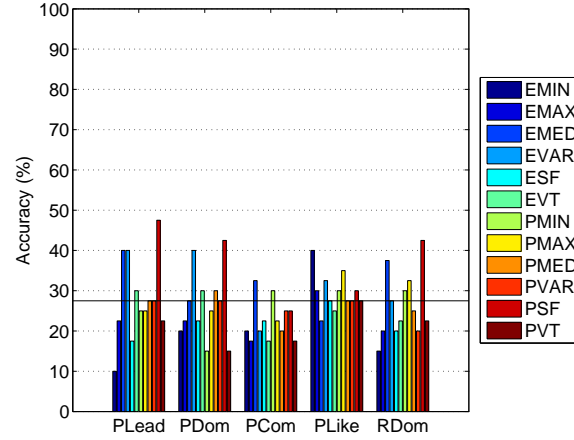


Fig. 8. The accuracy of energy and pitch on the ELEA full corpus and rule-based estimation. The black horizontal line shows the random baseline.

TABLE VII  
BEST RESULTS OF SVM ON THE FULL ELEA CORPUS, USING ONLY AUDIO FEATURES. RANDOM PERFORMANCE 27.5%

Acc(%)	features
PLead 67.9	AST, TSIf <sup>2</sup> , EVAR, EVT
PDom 64.3	AST, TSIf <sup>1</sup> , TSIf <sup>2</sup> , EMIN, EMED, EVAR, EVT
PCom 48.8	AST, TSIf <sup>2</sup> , EVAR, EVT
PLike 55.4	TSIf <sup>1</sup> , TSIf <sup>2</sup>
RDom 66.7	AST, TSL, TSTf, TSIf <sup>2</sup>

3) *Support vector machine*: Using only audio nonverbal features, the results improved random performance. Table VII shows accuracy results on SVM. The accuracy is calculated on the label assigned to the emergent leader (one per group, i.e. the participant with highest probability) or the related concepts compared with the ground truth. For this task the baseline accuracy is 27.5% considering that we have 28 meetings with four participants, and 12 meetings with three participants.

The use of SVM improves slightly the accuracy obtained for PLead using Rank-level Fusion from 67.5 to 67.9%. For the variable RDom improvement with respect to Rank-level fusion is 1.7%. Although our SVM-results for RDom are lower than the ones presented in [47] with up to 75% accuracy, and in [27] with up to 91.2% when there is Full agreement for the most dominant person from annotators, and up to 75.4% when there is majority agreement, is worth to mention that the scenarios differ, such that in our scenario no roles are assigned in the recordings.

4) *Collective classification*: The collective classification discriminates the emergent leader and related concepts in the group using posterior probabilities. In each group the emergent leader is the participant with the highest posterior probability.

Using the relational information with collective classification improved the accuracy to estimate the emergent leader and related concepts. The nonverbal features are selected based on the highest correlation values mentioned in section VII-A. We applied both the out-of-sample (two variants) and in-sample (the three variants) approaches

TABLE VIII

BEST RESULTS OF COLLECTIVE CLASSIFICATION ON THE FULL ELEA CORPUS. OUT-OF-SAMPLE TASK USING ONLY AUDIO FEATURES.

	Acc(%)	features	ICA variant
PLead	72.0	AST, TSIf <sup>2</sup> , EVAR, EVT	ICA <sub>c</sub>
PDom	60.1	AST, TSI, TSIf <sup>2</sup> , EMIN, EMED, EVAR, EVT	ICA <sub>c</sub>
PCom	46.4	TSL, TSTf, AST, TSIf <sup>2</sup> , EMIN, EMED, EVAR	ICA <sub>c</sub>
PLike	55.4	All Speaking Turn Features	ICA <sub>c</sub>
RDom	61.9	PVAR, PSF, PVT	ICA <sub>c</sub>

TABLE IX

BEST RESULTS OF COLLECTIVE CLASSIFICATION ON THE FULL ELEA CORPUS. IN-SAMPLE TASK USING ONLY AUDIO FEATURES.

	Acc(%)	features	ICA variant
PLead	70.2	TSL, TSTf, TSIf <sup>2</sup> , EMED	ICA <sub>c</sub>
PDom	58.3	AST, TSIf <sup>2</sup> , PMIN	ICA
PCom	57.7	AST, TSIf <sup>2</sup> , EVAR, EVT	ICA <sub>c</sub>
PLike	53.6	PVAR, PSF, PVT	ICA <sub>c</sub>
RDom	76.2	AST, TSIf <sup>2</sup>	ICA <sub>c</sub>

described in section VI-D.

For the out-of-sample task, the accuracy is calculated on the label assigned to the emergent leader or the related concepts compared with the ground truth. For this task the baseline accuracy is 27.5% considering that we have 28 meetings with four participants, and 12 meetings with three participants. Table VIII shows accuracy with the out-of-sample variant from the ICA algorithm using only audio features.

For the in-sample variant, we provide a known label per group. Since we notice from the rule based-estimator and the rank-level fusion method that participants with the lowest feature values are often perceived neither as leaders nor as most dominant, we labeled these participants as Non-Emergent Leader/Non-Most Dominant. The test is then performed using this known label and estimating the leader or related concepts out of two or three participants respectively per group. For this task, the baseline accuracy is 38.3%. Table IX shows the accuracy results for audio features on the full ELEA corpus using the In-sample variant from the ICA algorithm.

From the ICA variant to estimate the concepts related to emergent leadership, we can observe that the variant ICA<sub>c</sub> (which uses the known and most confident estimated labels in each iteration) has the best performance for most of the cases. The best accuracy for emergent leadership estimation is 70.2% when we provide a well-known label (i.e., when we have a participant that is non-leader in this task, see Table IX), and 72% when the group does not have any known label (Table VIII).

Table X shows best accuracy results from the four methods on the full ELEA corpus using audio nonverbal features. In all cases the Rank-level Fusion outperformed accuracies obtained with the Rule-based Estimator, and almost all accuracies obtained with SVM (except from PLike). Similarly, CC outperformed accuracies from Rule-based Estimator and SVM in almost all cases (except PDom). Although CC did not outperformed Rank-level Fusion in all cases, the difference in accuracy for PLead is only 0.5%. As we can observe, best accuracy for PLead is 72.5% and for PDom is 65% using Rank-level Fusion. For RDom and PCom the best accuracy is 76.2



TABLE X

BEST ACCURACY PERFORMANCE FOR EMERGENT LEADERSHIP AND RELATED CONCEPTS FROM THE FOUR METHODS ON THE FULL ELEA CORPUS WITH ONLY AUDIO NONVERBAL FEATURES

	<b>PLead</b>	<b>PDom</b>	<b>PCom</b>	<b>PLike</b>	<b>RDom</b>
Baseline	27.5	27.5	27.5	27.5	27.5
Rule-based Estimator	63.7	53.7	42.5	40	61.3
Rank-level Fusion	<b>72.5</b>	<b>65</b>	55	40	72.5
SVM	67.9	64.3	48.8	<b>55.4</b>	66.7
CC-Out-of-Sample	72	60.1	46.4	<b>55.4</b>	61.9
CC-In-Sample*	70.2	58.3	<b>57.7</b>	53.6	<b>76.2</b>

and 57.7 respectively, using CC-In-Sample. For PLike the best accuracy (55.4%) was performed with SVM and CC-Out-of-Sample.

### C. Leadership Estimation using Audio-Visual Nonverbal Cues - ELEA AV Corpus

In this section, we present the results for each of the three estimation methods and the audio, visual and audio-visual cases. As described in Section IV to control for variability in the video quality, in this section we used a part of the full ELEA corpus recorded with the portable setup (i.e., 27 meetings) referred as ELEA AV.

1) *Rule-Based Estimator*: Among the 27 meetings from the ELEA AV corpus, there are six meetings with three participants and 21 meetings with four participants. This gives a random baseline performance of 26.8% for the estimation of the emergent leader (or the other social constructs) among the meeting participants. Figures 9 and 10 shows the accuracy of the audio features and the visual features for the five tasks on the ELEA AV corpus. Speaking turn features perform better than prosodic features for all the tasks, except for PLike. The results on the visual features show that for PLead, PDom, and RDom the body activity features and wMEI based features perform better than the head activity ones. On the contrary, for PComp and PLike, head activity features perform better. Some visual nonverbal features perform quite poorly in some tasks, giving accuracies below the baseline.

Figure 10 shows the accuracy of the visual features for the five tasks on the ELEA AV corpus. The results show that for PLead, PDom, and RDom, the body activity features and wMEI based features perform better than the head activity ones. The highest performance for emergent leadership is 55.6% and is achieved by TBL, TBT, stdB and wMEIqn features. On the contrary, for PComp and PLike, head activity features perform better (THL and stdHy, with 51.85% and 44.4% respectively). Some visual nonverbal features perform quite poorly in some tasks, giving accuracies below the baseline, as the rule-based estimator was applied on single features, no fusion experiments are reported here.

2) *Rank-Level fusion*: We performed an exhaustive search for all feature combinations up to six features on the ELEA AV corpus. Figure 11 shows the accuracies of best single audio nonverbal feature, best single video nonverbal feature, audio-visual fusion, audio-only fusion, and video-only fusion on the ELEA AV corpus. We also show the confidence intervals of the best accuracy, with 95% confidence, with respect to the number of examples in

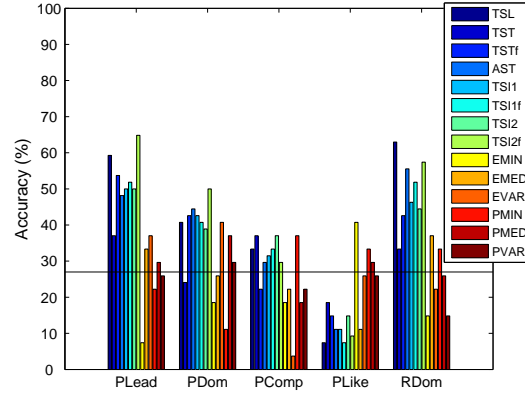


Fig. 9. The accuracy of audio nonverbal features on the ELEA AV corpus. The black horizontal line shows the random baseline.

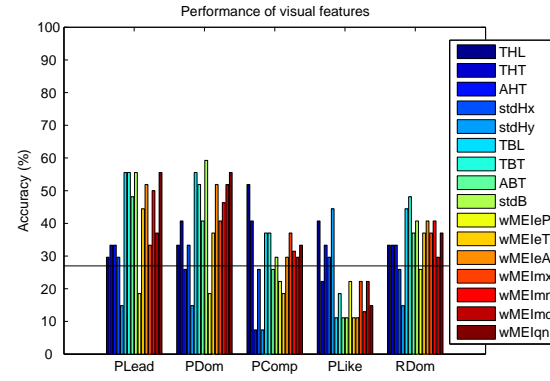


Fig. 10. The accuracy of visual nonverbal features on the ELEA AV corpus. The black horizontal line shows the random baseline.

the dataset. The results show that for PLead and RDom best audio feature provides higher accuracies than the best visual features. This fact is reversed for PDom, PComp, and PLike. For each of the variables, audio-visual fusion provides the highest accuracy, better than audio-only or visual-only fusion. Table XI shows the fused variables, giving the highest accuracy for each of the tasks. The highest achieved accuracy for leadership is 85.2% and corresponds to a variety of the extracted features. The best achievable performance for dominance is slightly lower than the one reported in [27], [4] which investigated a subset of the AMI meeting corpus that is based on a different group task.

For a more detailed look into fused variables, we analyzed the pairwise feature selection frequency in the best

TABLE XI

RESULTS OF RANK-LEVEL FUSION ON THE ELEA AV CORPUS. THE LAST COLUMN OF THE TABLE SUMMARIZES THE FUSED FEATURES WITH RESPECT TO THE FEATURE GROUPS (ST: SPEAKING TURN, HA: HEAD ACTIVITY, BA: BODY ACTIVITY, MT: WMEI BASED FEATURES, EN: ENERGY, PI: PITCH)

	Acc(%)	Fused variables	Feature Groups
PLead	85.2	TSL, TSI <sup>1</sup> , TSI <sup>2</sup> , THT, TBT, EMED	ST, HA, BA, EN
PDom	74.1	TSI <sup>1</sup> , THT, wMEIqn, EVAR	ST, HA, MT, EN
PCom	59.3	THL, PMIN	HA, PI
PLike	59.3	THL, AHT, EMIN, PMIN	HA, PI
RDom	77.8	TSL, AST, TSI <sup>2</sup> , wMEImx, EMED, EMIN	ST, MT, EN

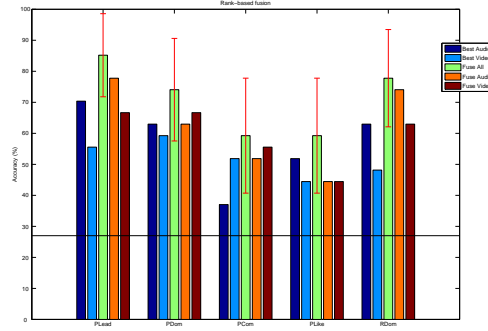


Fig. 11. Audio-visual, audio-only, and visual-only score-level fusion results on the ELEA AV corpus. The accuracies of best single audio nonverbal feature and best single video nonverbal feature are also shown. The black horizontal line shows the random baseline.

combinations of rank-level fusion, as there are multiple combinations giving the same best accuracy. For simplicity, instead of reporting the actual frequencies of features, we grouped the features into six feature groups, and report the pairwise frequencies of the feature groups (see Figure 12). The feature groups are defined with respect to the type of features.

- ST: Speaking turn features (Section V-A1).
- HA: Head activity features (Section V-B1a).
- BA: Body activity features (Section V-B1b).
- MT: wMEI based features (Section V-B2).
- EN: Energy features (Section V-A2).
- PI: Pitch features (Section V-A2).

In Figure 12, the diagonal corresponds to the selection frequency of that feature group, whereas off-diagonals indicate the pairwise frequencies: the brighter the pixel, the higher the frequency. Several interesting conclusions can be made from these results:

- For all the variables, audio-visual fusion is essential.
- Head activity is more important for PLead, whereas it is not used in PDom or in RDom. Instead, PDom and RDom use body activity or wMEI based features as visual information.
- Pitch information is not used in PLead, PDom, and RDom. However it is informative for PComp and PLike.
- For PComp and PLike, head activity, energy and pitch are the most informative features. Speaking turn features have a very little effect for these two variables.

3) *Collective classification*: Table XII shows the accuracies for emergent leader and related concepts for the out-of-sample task on the ELEA AV corpus. The accuracy is calculated based on the correct estimation of the emergent leader and related concepts, for this task the random baseline performance is 26.8 given that the ELEA AV corpus has 21 meetings with four participants and 6 meetings with three participants. For the out-of-sample task,

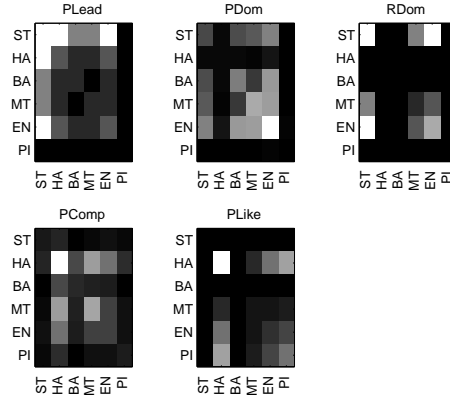


Fig. 12. Pairwise frequency of feature groups in best combinations

adding visual information increases accuracy estimation using ICA algorithm, PLead, PDom and Pcom increased accuracy with respect to only audio accuracy performance. The best accuracy obtained for PLead is 81.0%.

TABLE XII

BEST RESULTS OF COLLECTIVE CLASSIFICATION, OUT-OF-SAMPLE TASK USING AUDIO AND VISUAL FEATURES ON THE ELEA AV CORPUS. FEATURE GROUPS: ST-SPEAKING TURN, HA-HEAD ACTIVITY, BA-BODY ACTIVITY, MT-MOTION (WMEI BASED), EN-ENERGY, PI-PITCH.

		Acc(%)	feature_group
Audio	PLead	59.5	ST
	PDom	58.3	ST, EN
	PCom	41.7	EN
	Like	63.1	ST, EN
	RDom	67.9	ST
Visual	PLead	70.2	HA, BA
	PDom	67.9	BA
	PCom	35.7	HA
	Like	50.0	MT, BA
	RDom	53.6	BA
AV	PLead	81.0	ST, BA
	PDom	70.2	ST, BA
	PCom	46.4	HA, EN, PI
	Like	53.6	ST, EN
	RDom	65.5	ST, EN

Table XIII shows the averaged accuracy results for the in-sample task (i.e. a known label per group). Again, since participants with the lowest feature values are not perceived as leaders nor most dominant, we labeled these participants as Non-EmergentLeader/Non-MostDominant. The test is performed using this known label and estimating the emergent leader and related concepts, for this task the baseline accuracy is 37.0%.

From the accuracy estimations with the in-sample task, in general terms, less features are needed to discriminate between emergent leaders and non-emergent leaders. Additionally the performance for the emergent leader with respect to PLead-85.7% and RDom-82.1% are better than the baseline performance (37.0%). This confirms the statement of McDowell et al. [40], which affirms that having known labels for the test phase provides better accuracy in realistic scenarios. From Tables XII and XIII we can observe that for the variables PLead and PDom having audio and visual information performed better, in contrast with PLike and RDom for which only audio

TABLE XIII

BEST RESULTS OF COLLECTIVE CLASSIFICATION, IN-SAMPLE TASK USING AUDIO AND VISUAL FEATURES ON THE ELEA AV CORPUS.  
FEATURE GROUPS: ST-SPEAKING TURN, HA-HEAD ACTIVITY, BA-BODY ACTIVITY, MT-MOTION (WMEI BASED), EN-ENERGY, PI-PITCH.

	Concept	Acc(%)	feature group
Audio	PLead	63.7	ST
	PDom	61.9	ST
	PCom	57.1	ST, EN
	Like	75.0	ST
	RDom	82.1	ST
Visual	PLead	78.6	HA, BA
	PDom	67.9	BA
	PCom	51.2	HA
	Like	42.9	BA
	RDom	72.6	MT, BA
AV	PLead	85.7	ST, EN, BA
	PDom	70.2	ST, EN, BA
	PCom	48.8	HA, EN, PI
	Like	65.5	ST, EN
	RDom	81.0	ST, EN

TABLE XIV

BEST ACCURACY PERFORMANCE (%) FOR EMERGENT LEADERSHIP ON THE ELEA-AV CORPUS

	PLead	PDom	PCom	PLike	RDom
Baseline	26.8	26.8	26.8	26.8	26.8
Rule-based Estimator	70.4	51.9	37	40.7	63.0
Rank-level Fusion	85.2	<b>74.1</b>	<b>59.3</b>	59.3	77.8
CC-Out-of-Sample	81.0	70.0	46.4	63.1	67.9
CC-In-Sample*	<b>85.7</b>	70.2	57.1	<b>75.0</b>	<b>82.1</b>

features performed better than the combination of features. Finally for PCom, on one hand the combination of audio and visual features, performed better than only audio or only visual information; on the other hand if we provide an example, only audio features performed better than the combination.

Table XIV shows the best performance from the non-collective and collective approaches, from which we can observe that for PLead, PLike and RDom, CC has the best accuracies 85.7, 75.0 and 82.1 respectively. For the variables PDom and PCom the Rank-level Fusion method performed better than collective classification, and is worth to mention that training for this method is not needed.

In summary the collective classification approach overall improved the estimation of the emergent leader up to 85.7% using audio and visual information.

#### D. Observation Window Analysis

We performed an analysis to explore the temporal support needed by our approach. For brevity we focus our analysis on the audio features and their relationship with leadership and dominance. We computed the same audio features described in section V-A1, originally computed for the whole interaction, for smaller observation windows (or thin slices), and then estimated the emergent leader and related concepts with the rule-based estimator per slice.

We explored three type of slices:

- Accumulated Slices: We defined the duration of the slices as multiples of 1/8 of the original duration, where each slice starts from the beginning of the interaction.

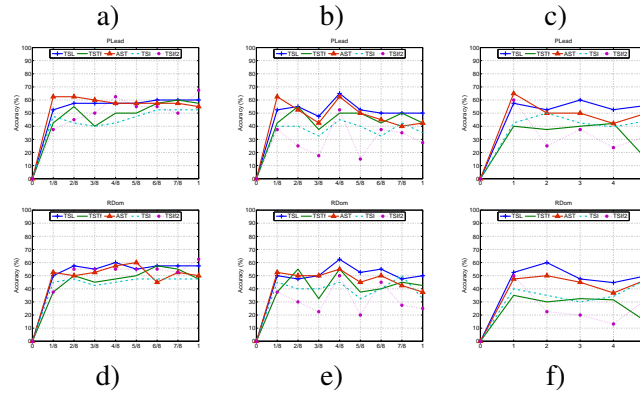


Fig. 13. Observation window analysis for speaking turn features on the full ELEA corpus. (a-c): Results for PLead: a) accumulated slices; b) non-accumulated slices; and c) non-accumulated slices with overlaps. (d-f): Results for RDom: d) accumulated slices; e) non-accumulated slices; and f) slices with overlaps.

- Non-Accumulated Slices: Each slice is exactly  $1/8$  of the total duration, with no overlaps.
- Non-Accumulated Slices with Overlaps: The slice size is 5 minutes with two-minute overlaps, the first slice starts from the beginning of the interaction.

Figure 13 shows the accuracy of the three types of slices with respect to the variables PLead and RDom. We can observe that for accumulated slices (Figures 13 (a) and (d)), after the first half of the recording (7.3 minutes on average), the estimations follow a trend and change only slightly.

Figures 13 (b) and (e) show the accuracy for the non-accumulated slice rule-based estimation with respect to PLead and RDom. We can often capture the score performance by just looking at the slice in the middle (slice from  $3/8$  to  $4/8$ ), in which the person that speaks the most, takes more turns, and interrupts more is perceived as the emergent leader and as well the most dominant.

Finally, considering bigger slices with overlaps, in this case a five minutes slice, we can observe from Figure 13 (f) that dominance is more likely to be clearly perceived approaching the middle of the meeting (minutes 3-8), and the emergent leadership is more highly noticeable in the middle (Figure 13 (c) window 3: minutes 6-11). From Figure 13 (c), we can also observe that the emergent leader talks the most during the first five minutes, then the other participants take turns, and during the middle of the meeting the leader again has the highest speaking time and turns. Based on the task performed, we can probably interpret these results as follows: the leader organizes the group (first five minutes), listens to opinions from the group (minutes 3-8), and then leads the interaction.

## VIII. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a computational framework to estimate emergent leadership in newly formed groups from nonverbal behavior, by combining speaking turns, prosodic features, as well as visual activity and motion. To our knowledge, our work is the first attempt that addresses automatic emergent leadership analysis in groups from audio-visual sensed data. We first designed and collected a new audio-visual group interaction corpus for our study.

We then evaluated the effectiveness of individual and combined audio and visual features in detecting the emergent leader and related constructs using non-collective and collective approaches. Based on the results of a correlation analysis we noticed that the emergent leader was perceived by his/her peers as an active and dominant person, who talks the most, has more turns and interruptions and has a variation in the tone of voice and energy. To estimate the emergent leader, the combination of acoustic and visual information performed better than single modalities for both non-collective and collective approaches. Additionally, estimating collectively the emergent leader in the group, when we have a well known non-leader participant, increases the accuracy up to 85.7%.

Regarding concepts related to leadership in the literature, we found that for the perception of competence the most informative nonverbal cues came from head activity and pitch (59.3% accuracy estimation). For the case of perceived liking (75.0% accuracy estimation if we provide a sample), the most informative features were extracted from speaking turn features. Note also, that the results for perceived leadership and perceived dominance (and rank dominance) were sometimes similar and sometimes different. Given that the variables were highly correlated, it is not surprising that the results are similar. However, we opted not to combine the three measures because they capture somewhat different aspects of verticality. In the present paper we focused on leadership, the skill to influence other group members and the task solution during an interaction.

Through an analysis of observation windows, we found that although the entire interaction is needed to perform the task, to estimate the emergent leader only the first half (approximately 7 minutes) or a slice of the interaction around the middle was required. This finding could be explored in more detail given the potential value for applications that could provide reasonably accurate estimations with less data.

For future work, other nonverbal features like gaze and smiling can be extracted and coded to complement the visual features presented in this paper. Personality traits could be explored as complements to the construction of the concepts related to emergent leadership. We plan to investigate these two lines of work.

In relation to the annotations, we would like to address the issue of comparing judgments of external observers and judgments from participants in the group. Given that the judgment comparison is a research issue in psychology, we plan to collect annotations and perform a study on them.

Finally, we could explore the performance of other graphical models that have successfully recognized functional roles (e.g. the Influence model).

#### ACKNOWLEDGMENT

The authors would like to thank Iain McCowan (dev-audio) for technical support with the Microcone device, Denise Frauendorfer, Pilar Lorente and Radu-Andrei Negoescu for valuable help during the collection and data processing, and the participants in the ELEA corpus for their time and enthusiasm. This research was supported by

CONACYT (Mexico) through a doctoral studies scholarship, and partially supported by the EU project NOVICOM and SNF project SONVB.

## REFERENCES

- [1] The Microcone website, 2011.
- [2] C. Anderson and G. J. Kilduff. Why do dominant personalities attain influence in face-to-face groups? the competence- signaling effects of trait dominance. *Journal of Personality and Social Psychology*, 96(2):491–503, 2009.
- [3] O. Aran. *Vision Based Sign Language Recognition: Modeling and Recognizing Isolated Signs With Manual and Non-manual Components*. PhD thesis, Bogazici University, Istanbul, Turkey, 2008.
- [4] O. Aran and D. Gatica-Perez. Fusing audio visual nonverbal cues to detect dominance in small group conversations. In *ICPR*, Aug 2010.
- [5] O. Aran and D. Gatica-Perez. Analysis of group conversations: Modeling social verticality. In A. A. Salah and T. Gevers, editors, *Computer Analysis of Human Behavior*, Lecture Notes in Computer Science. Springer, 2011.
- [6] K. Bachour, F. Kaplan, and P. Dillenbourg. An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Transactions on Learning Technologies*, 3(3):203–213, 2010.
- [7] J. E. Baird. Some non-verbal elements of leadership emergence. *Southern Speech Communication Journal*, 42(4):352–361, 1977.
- [8] B. M. Bass. *Bass and Stogdill's handbook of leadership. Theory, research, and managerial applications*. Free Press, 1990.
- [9] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.
- [10] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [11] T. A. Carte, L. Chidambaram, and A. Becker. Emergent leadership in self-managed virtual teams. *Group decision and negotiation*, 15:323–343, 2006.
- [12] M. Charfuelan, M. Schrder, and I. Steiner. Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In *Interspeech 2010*, pages 2558–2561, 2010.
- [13] J. M. DiMicco, A. Pandolfo, and W. Bender. Influencing group participation with a shared display. In *CSCW*, pages 614–623, New York, NY, USA, 2004. ACM.
- [14] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *International Conference on Multimodal Interfaces*, ICMI '07, pages 271–278, 2007.
- [15] 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.
- [16] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *International Conference on Mutlimodal Interfaces (ICMI)*, pages 29–36, 2008.
- [17] N. P. Garg, S. Favre, H. Salamin, D. H. Tur, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and social network analysis. In *International Conference on Mutlimedia ACM*, pages 693–696, 2008.
- [18] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing*, 1(12), Dec 2009.
- [19] P. A. Gloor and Y. Zhao. Analyzing actors and their discussion topics by semantic social network analysis. *Information Visualisation, International Conference on*, 0:130–135, 2006.
- [20] L. D. Goodstein and R. I. Lanyon. Applications of personality assessment to the workplace: a review. *Journal of Business and Psychology*, 13(3), 1999.



- [21] J. Gray, A. and J. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 22(3):207 – 217, June 1974.
- [22] J. A. Hall, E. J. Coats, and L. Smith. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological bulletin*, 131(6):898–924, 2005.
- [23] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *IACM Int.Conf. on Multimedia (ACM MM)*, Sep 2007.
- [24] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In B. Buxton and R. Cipolla, editors, *Computer Vision ECCV '96*, volume 1064 of *Lecture Notes in Computer Science*, pages 343–356. Springer Berlin / Heidelberg, 1996.
- [25] D. Jayagopi and D. Gatica-Perez. Discovering group nonverbal conversational patterns with topics. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2009.
- [26] D. Jayagopi and D. Gatica-Perez. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, 12(8):790–802, Dec 2010.
- [27] D. 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*, 17(3), Mar 2009.
- [28] D. Jayagopi, T. Kim, A. Pentland, and D. Gatica-Perez. Recognizing conversational context in group interaction using privacy-sensitive mobile sensors. In *Mobile and Ubiquitous Multimedia*, Dec 2010.
- [29] D. Jayagopi, B. Raducanu, and D. Gatica-Perez. Characterizing conversational group dynamics using nonverbal behavior. In *ICME*, Jun 2009.
- [30] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2004.
- [31] O. P. John and L. A. Pervin. The big five factor taxonomy: Dimensions of personality in the natural language and in questionnaires. pages 66–100, 1990.
- [32] A. K. Kalma, L. Visser, and A. Peeters. Sociable and aggressive dominance: Personality differences in leadership style? *Leadership Quarterly*, 4(1):45–64, 1993.
- [33] J. Kickul and G. Neuman. Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and ksas. *Journal of Business and Psychology*, 15(1), 2000.
- [34] T. Kim, A. Chang, L. Holland, and A. Pentland. Meeting mediator: enhancing group collaboration with sociometric feedback. In *Conference on CSCW*, pages 457–466, 2008.
- [35] M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning, 2008.
- [36] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Employing social gaze and speaking activity for automatic determination of the extraversion trait. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2010.
- [37] A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. Pervasive sensing to model political opinions in face-to-face networks. In *Pervasive*, 2011.
- [38] M. S. Mast. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication research*, 28(3):420–450, 2002.
- [39] I. McCowan, M. Krishna, D. Gatica-Perez, D. Moore, and S. Ba. Speech acquisition in meetings with an audio-visual sensor array. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1382 –1385, july 2005.
- [40] L. K. McDowell, K. M. Gupta, and D. W. Aha. Cautious collective classification. *Journal of Machine Learning Research*, 10:2777–2836, 2009.
- [41] J. Neville, D. Jensen, and B. Gallagher. Simple estimators for relational bayesian classifiers. *Data Mining, IEEE International Conference on*, 0:609, 2003.

- [42] A. Pentland. *Honest signals: how they shape our world*. MIT Press, 2008.
- [43] F. Pianesi, N. Mana, and A. Cappelletti. Multimodal recognition of personality traits in social interactions. In *International Conference on Multimodal Interfaces (ICMI)*, Oct 2008.
- [44] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41:409–429, 2007.
- [45] B. Raducanu and D. Gatica-Perez. Inferring competitive role patterns in reality tv show through nonverbal analysis. *Multimedia Tools and Applications*, pages 1–20.
- [46] B. Raducanu and D. Gatica-Perez. You are fired! nonverbal role analysis in competitive meetings. In *ICASSP*, Apr 2009.
- [47] R. Rienks and D. Heylen. Automatic dominance detection in meetings using easily detectable features. In *Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.
- [48] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez. Identifying emergent leadership in small groups using nonverbal communicative cues. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2010.
- [49] R. T. Stein. Identifying emergent leaders from verbal and nonverbal communications. *Personality and Social Psychology*, 32(1):125–135, 1975.
- [50] R. T. Stein and T. Heller. 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, 1979.
- [51] J. Sturm, O. Herwijnen, A. Eyck, and J. Terken. Influencing social dynamics in meetings through a peripheral display. In *International Conference on Multimodal Interfaces*, 2007.
- [52] D. Talkin. A robust algorithm for pitch tracking (rapt). In *In Speech Coding and Synthesis*, pages 495–518. Elsevier Science, 1995.
- [53] P. Temdee, B. Thipakorn, B. Sirinaovakul, and H. Schelhowe. Of collaborative learning team: An approach for emergent leadership roles identification by using social network analysis. In *Technologies for E-Learning and Digital Entertainment*, volume 3942 of *Lecture Notes in Computer Science*, pages 745–754. Springer, 2006.
- [54] G. Varni, G. Volpe, and A. Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, 12(6):576–590.
- [55] D. K. Wentworth and L. R. Anderson. Emergent leadership as a function of sex and task type. *Sex Roles*, 11(5/6):513–524, 1984.
- [56] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010.
- [57] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2006.