Predicting the performance in decision-making tasks: from individual cues to group interaction

Umut Avci, and Oya Aran, *Member, IEEE*

***Abstract*—This paper addresses the problem of predicting the performance of decision making groups. Towards this goal, we evaluate the predictive power of group attributes and discussion dynamics by using automatically extracted features, such as group members' aural and visual cues, interaction between team members, and influence of each team member; as well as self-reported features such as personality- and perception-related cues, hierarchical structure of the group, and individual- and group-level task performances. We tackle the inference problem from two angles depending on the way that features are extracted: (i) holistic approach based on the entire meeting and (ii) sequential approach based on the thin slices of the meeting. In the former, key factors affecting the group performance are identified and the prediction is achieved by Support Vector Machines. As for the latter, we compare and contrast the classification performance of an Influence Model based novel classifier with that of Hidden Markov Model. Experimental results indicate that the group looking cues and the influence cues are major predictors of group performance and the Influence Model outperforms the HMM in almost all experimental conditions. We also show that combining classifiers covering unique aspects of data results in improvement in the classification performance.**

***Index Terms*—social computing, multimodal interaction, group performance analysis.**

I. Introduction

In most of the business-related issues, making correct decisions is more important than making them fast. Companies are willing to devote the time and resources needed to minimize their risks and to protect their financial interests. In many cases, decision-making processes are carried out by groups to leverage the expertise and knowledge of the members despite the fact that decisions made by individuals are faster. Holding effective meetings not only prevents organizations from wasting time and money in vain but also improves productivity. Therefore, identifying factors to run successful meetings and predicting the performance of a group are of great significance for companies.

In the course of decision-making processes, individuals express themselves verbally in a distinctive fashion. Their styles deliver implicit signals that go beyond the literal meaning of the words like sincerity and dominance. Similarly, participants' body language changes in response to the behavior of the interlocutor [49]. Such messages are known as nonverbal behavior and include voice-related variables, gaze, and posture among others [20]. The nonverbal channel is especially useful when genuineness of the communication is an issue, e.g., tone of the voice or the way one looks at another may indicate mockery or sarcasm. Besides, nonverbal behavior is harder to fake as most of the body movements are unconscious forms of expressions, which makes nonverbal cues a reliable source for analyzing group interaction in a discussion [2]. Although nonverbal communication differs depending on the age, gender [25], and culture [12], people tend to follow certain characteristic features. These convey information on one's participation in the discussion (e.g., speaks and interrupts more), role (e.g., leader), and personality (e.g., extrovert). Moreover, a detailed look at the participation cues of group members provides insights into the amount of information shared in the discussion and into the group interaction patterns.

Psychologists laid the foundations of studies on evaluating the group performance in decision-making tasks. They mainly focused on the effects of high-level concepts, e.g., team composition and individual attributes, on the team success rather than speech or vision-based features. Group cohesion and team diversity were shown to affect the group performance in [15] and [22]. It was also found that personality traits [24] and individuals' roles [23] [8] are important factors determining decision quality. Following these findings, computer scientists put their effort on automatically extracting the factors proposed by the psychologists, e.g. cohesion [19], individuals' socio-emotional [45] [3] and functional roles [18]. They further investigated the team effectiveness in terms of the discussion constituents. For this purpose, features characterizing audio and visual patterns of a group performing a decision-making task were used as nonverbal behavioral cues. Performance of the group was observed to be correlated with the audio features that capture speaking turns and lengths [46] and with the visual cues that represent eye contact rate [11]. However, a vast majority of the existing studies have focused solely on determining the factors affecting the team performance by disregarding the inference mechanism. Those few works that take inference into account are far from providing a complete picture of the problem as the approaches were tested for individuals or for single modality.

In this study, we investigate the prediction of the performance of a group performing a decision-making task. To this end, we use a multi-party multimodal dataset of 40 real meetings and extract a large set of features from audio, video, and questionnaires. Extracted features contain information about individual- and group-level nonverbal speaking&looking behavior, individual&group performances, self-reported personalities (Big-Five), and interpersonal perception (of leadership, dominance, competence, and liking). We also consider cues for the internal dynamics of the group, characterizing

Umut Avci is with the Department of Software Engineering, Izmir University of Economics, Izmir, Turkey, e-mail: umut.avci@ieu.edu.tr.

Oya Aran is with the Idiap Research Institue, Martigny, Switzerland, e-mail: oya.aran@idiap.ch.

the hierarchical structure and the interaction between team members. The resulting feature set has been categorized under two headings: features representing the group behavior for the whole meeting, and those account for the temporal information embedded in the discussion. We use the former to identify significant factors affecting the task performance of groups and perform standard machine learning methods to predict group performances in regression and classification tasks. For the latter, inference has been achieved by leveraging a probabilistic sequential approach for classification purposes.

Our paper has several contributions. To our knowledge, this is the most comprehensive quantitative work on the performance prediction in decision-making tasks ever published. While inference on the performance level of groups has been recently addressed for gaze behavior [36], predictive power of different feature sets was unknown. Our study also differs from the works that make prediction for individual performances [6] by focusing on the task-completion rate of teams. As far as extracted features are concerned, we have introduced a novel cue, i.e. *Influence*, that provides information on the interaction between group members. Although Jayagopi et al. proposed an approach to model group interaction patterns, the conversational dynamics were derived only from a single modality, i.e. audio, and the relation between the patterns and the group performance was not addressed [10]. We have further included hierarchical structure in groups with respect to the interpersonal perceptions into the feature set, which had been handled only for the dominance before [9]. Other contributions include the solution of the classification task. First, we have built a new classifier based on the Influence Model (IM). Second, in building the classifier, we have considered the ordering of group members with respect to their influential power to account for the interaction patterns. Finally, the IM has been fit with the capability of processing multivariate binary data.

The paper is organized as follows. Section II briefly overviews the literature regarding the performance analysis for the decision-making tasks in terms of both social psychology and social computing. We explain the general structure of our approach and describe the dataset used in the study in Section III. We provide details on the features to be used, feature extraction process, and the correlations between the extracted features and the group performance in Section IV. An introduction to the IM is given in Section V. We then present and discuss experimental results in Section VI. Finally, we conclude our paper in Section VII.

II. Related Work

In this section, we briefly discuss the existing works that are closely related to our study in terms of social psychology and social computing.

A. Social Psychology

Determining the key factors affecting the efficiency of groups is an active topic that attracts the attention of social psychologists for a long time. McGrath et al. stressed that being led by a competent person played a critical role in sustaining a high performance in project teams [23]. To the study, characteristics of a knowledgeable and assertive leader include ensuring a meeting to continue on its track, keeping members involved in discussions, and uncovering the participants' strengths among others. Sundstrom et al. pointed out that success of a group was highly correlated with the team cohesion, i.e. what made a group of people a "team" [15]. Such a property is the spark that lights the flame of team spirit and enables team members to resolve conflicts before it damages the relationship between them.

In [22], Shaw et al. showed that higher decision performances could be achieved by heterogeneously formed teams, i.e. by participants who have different levels of knowledge, skills, and abilities. Later studies revealed that the dissimilarities between team members in experiences and knowledge did not always guarantee to make better decisions. Using heterogeneous teams only for complex tasks and gathering like-minded individuals together were proposed in [28] and [29] respectively. The situation in which the decision problem was solved by the divide and conquer approach, i.e. each of the group member was assigned a unique responsibility, was considered in [5] for different task complexities. It was observed that higher workload led to significant reduction in the team success as well as the individual performance.

In order to highlight the impact of personality types on the performance, Bradley et al. defined four main personality types based on the Myers-Briggs Type Indicator with their subcategories and used trait compositions for the evaluation [24]. The criteria that are used by the majority of the researchers for the personality analysis, however, is based on the Big Five personality traits. There are many studies in the literature that investigate the effect of these factors on the team performance such as [38], [26], and [40] among others.

Haslam et al. and van Dick et al. focused on revealing the relation between the team members' involvement in discussion and the team performance in [1] and [33]. The cases for which the individuals in a group work collectively (in the former) and interact actively (in the latter), i.e. less social loafing, were observed to be positively correlated with the decision performance. Spreitzer et al. analyzed the effectiveness of a team in terms of the clarity of individuals' roles and the level of information provided to the members in addition to the team involvement in [17]. The authors observed a threshold effect such that after a certain level of role clarity and of information on the task was reached group members lost their focus due to micro-managemental duties and information overload, which led to drop in the task performance.

Researchers also evaluated the links between the mental ability of groups and the success in decision-making. The results regarding the cognitive abilities showed inconsistencies from one study to another. Neuman et al. observed a linear relationship between the cognitive capability of a group and the group performance [16]. However, Lepine et al. could not find any evidence to support this finding [21]. An important study on the subject was conducted by Woolley et al.. Their findings showed that the individual intelligence of a group was not the only determinant for the group's collective intelligence [4]. This statement refuted the belief that a group's performance

was bounded by the individual performances of its members, i.e. the performance of a group may be lower (or higher) than the minimum (or maximum) of the individual performances in the group.

B. Social Computing

The problem of analyzing the performance of groups has been addressed only a few times by the social computing community when compared with the related approaches in psychology. In [47], Dong et al. performed a quantitative analysis on a brainstorming task in order to create a model to represent the relationship between the discussion constructs and the group performance. The findings of the study suggested that the number of simultaneous speakers, i.e. generated ideas, is proportional to the group performance. Dong et al. later extended their studies on the performance evaluation of groups by using the mixture of Hidden Markov Processes and Markov jump processes for nonverbal audio features. They observed that groups made better decisions when certain speaking characteristics were followed, i.e. longer clause lengths, faster speaker changes, less standard deviations of pause [42], and more speaking turns, balanced participation rate [46]. Speaking and looking patterns were extracted at the group level by using Latent Dirichlet Allocation (LDA) in [11]. Jayagopi et al. then investigated the connection between the extracted patterns and the meeting descriptors such as group composition and performance. As a result of the work, the group performance was told to be correlated with the unsuccessful interruptions skew and convergent gaze.

In a conversational setting, determining the group dynamics has always been a popular topic. The Influence Model has enabled researchers to further investigate the dynamics, especially in terms of the interaction between participants. One of the first studies on the subject was conducted by Basu et al. [34]. The authors presented an efficient alternative to the interacting Markov Chains, e.g. coupled HMM, with the aim of reducing the parameters to be learned. The experiments with 5 subjects playing a debating game indeed showed that the proposed approach was capable of discovering group interactions in an efficient way. Dong et al. designed two experimental settings in different languages [44]. They found that the type of interaction could be the same regardless the topic of the discussion and the language it was held. A generalized version of the Influence Model was then introduced by Pan et al. [48]. The approach was built upon the idea that the interaction was a dynamic process and could change over time. Hence, the model was updated to account for the fluctuations in the influence between subjects. Varying applications has emerged in years depending on the interpretation of the group interaction. Studies by Dong et al. [43] and Raducanu et al. [7] focused on the determination of roles that group members took. The Influence Model was used to extract functional roles and status with respect to the corporate hierarchy in the former and in the latter respectively. In [37], Escalera et al. developed a framework for extracting the characteristics of a social network. Through the use of IM influences, the authors generated graphs in which the maximum cliques represented people sharing similar opinions. Researchers have also shown interest in classifying conversations with the help of IM. Cristani et al. defined classes based on the participants (adults vs. children) and the context (dispute vs. static) [27]. In [31], class definitions were made based on the emotion in the dialogs as positive, negative, neutral, and undefined. Although the IM model was not used as a standalone classifier in these works, it helped improving the classification accuracy.

To our knowledge, there are only a few studies that are relatively close to our work. In [6], Lepri et al. developed an SVM-based method to classify task performance of individuals in a group performing Mission Survival Task. The technique takes as input nonverbal behavioral features extracted from one-minute-long sequences, i.e. thin slices, and outputs one of three performance levels as the classification result. Performance prediction in the problem-solving tasks was again investigated at the individual level in [36] and [32]. Unlike Lepri's study, each member was asked to solve an 8-tile puzzle game individually rather than as a group. Performance of the gamers was grouped under three success levels, i.e. low, average, and high, based on their task completion times. The authors performed eye tracking to extract features characterizing participants' ocular behavior. Classification was then achieved by using an SVM model trained with the extracted features and the relative success levels. The approach presented in [39] differs from the previous studies in terms of the level of inference, i.e. the performance prediction is done for groups instead of individuals. The authors extracted nonverbal cues from the audio recordings of discussion groups working on the solution of algebra and geometry problems. Extracted features were used to create vocalization graphs each of which is a Markov chain where transitions represent ordered occurrence of speaking events, for instance a moment of silence is followed by the simultaneous speech of two members. Problem solving performance of groups was predicted by using the k nearest neighbor strategy, i.e. by performing simple graph matching.

In comparison to the abovementioned works in the performance prediction, which either extracted features from a single modality or inferred the task performance for team members separately, we used a large set of features from different sources (nonverbal multimodal cues, personalities, interpersonal perception etc.) and predicted the performance for groups. Consequently, we provide a complete picture of the problem by approaching it from a broader perspective.

III. The Approach and the Dataset

A. Our approach

A graphic summary of our approach is presented in Fig. 1. The dataset used in our work, i.e. the ELEA corpus, contains information about the meeting constructs of 40 groups performing a decision-making task. These include, apart from the audio and video recordings, the task performance of team members as well as the group performance, group members' personalities, and their perception of other participants. The available data have been processed to obtain the behavioral cues. We categorize the features under two headings depending on the meeting segment on which the cues are extracted: (i)

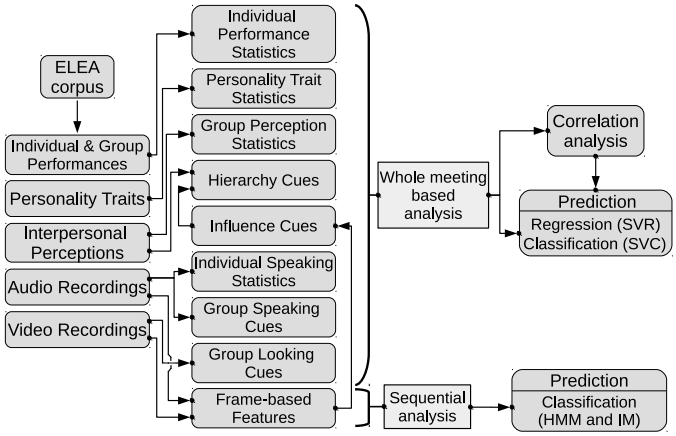

Fig. 1. The flowchart of our approach

whole meeting as a segment (i.e. meeting-based features) and (ii) one-second segments (i.e. frame-based features). We have tackled the prediction problem from two angles based on the feature category. For the meeting-based features, correlation analysis has been performed on the extracted features in order to identify the factors that have significant influence on the group performance. Then, regression and classification problems have been addressed by using the Support Vector Machines for distinct feature sets to account for the differences in the significance of the cues (in terms of their correlation with the group performance) and in the functionality of features (e.g., speaking vs looking cues). For the frame-based features, we have trained classifiers by using the Influence Model [35] and the Hidden Markov Model for audio, video, and visual focus of attention features. We have also presented a fusion procedure to combine classifiers from different modalities.

B. The ELEA corpus

The Emergent LEAder corpus (ELEA) is a multi-party multi-modal dataset that allows gathering information on group performance, dynamics, and structure [14]. To create the data, 148 participants (48 females and 100 males) were recruited from the French-speaking part of Switzerland. The subjects had an average age of 25.4 years with a standard deviation of 5.5 years. The assignment of subjects to groups was randomized in such a way that each group was composed of different participants. As a result, 28 four-person and 12 three-person groups were formed. Each group was asked to perform winter survival task. The procedure was composed of a discussion session as well as a series of questionnaires for behavioral assessment, and took approximately 15 minutes.

Winter survival task is a fictional scenario, the purpose of which is to rank 12 items in order to survive an airplane crash in winter. For each group, the ranking was done individually by participants at the beginning of the task and as a team after the group discussion to elucidate the effects of cooperation and factors determining the group composition, e.g., dominance, leadership. Individual and group rankings were then compared with the rankings of survival experts and Absolute Individual Scores (AIS) and Absolute Group Scores (AGS) were calculated via absolute difference in rank order, as measures of individual and group performance, respectively. Let r_I^i, r_G^i, and r_E^i be ranks of the item $i, i \in 1, \ldots, 12$ in individual, group, and expert rankings respectively. Then, individual and group performances are computed as $AIS = \sum_i \mid r_I^i - r_E^i \mid$ and $AGS = \sum_i \mid r_G^i - r_E^i \mid$. These measures quantify how similar the expert ranking is to individual rankings and group rankings. The smaller the score, the closer the ranked lists to each other, hence the higher the performance. A group G_1 is said to perform better than a group G_2 if the ranking made by G_1 is more similar to the expert rankings than the one made by G_2. In the ELEA corpus, AIS and AGS have ranges of [22,66] and [22,60] respectively.

The sensing infrastructure was composed of audio and video recording devices and was designed to be unobtrusive to provide group members freedom of movement. Audio data was captured by using a commercial microphone array, i.e. Microcone. Video data was recorded solely for 27 of groups with two webcameras.

Group members' perception of other participants and self-perception were obtained by using questionnaires filled after and before the debate session respectively. Questionnaires had 17 statements to measure five perceived variables. The first 16 statements were evaluated by using a five-point scale to account for four variables. Perceived Leadership (PLead) is associated with a person who keeps the meeting in track, and leads the group. Perceived Dominance (PDom) characterizes a person who dominates others, asserts her own will, and is decisive. A person who has sufficient skills, experience, and knowledge is identified with Perceived Competence (PComp). And the one who is well-disposed, considerate, and friendly corresponds to Perceived Liking (PLike). To capture the perception about the participant *x* for instance, other participants needed to score statements like 'I found that the person *x* addresses the group', 'I found that the person *x* imposes his/her views', and 'I found that the person *x* let others do what they choose'. The scores given to the variable-specific statements (PLead for instance) were then averaged over the participants other than *x* to find the perceived variable of *x*. The procedure was repeated for each perceived variable and each participant in the group. The last statement was used to quantify the fifth variable, i.e. Ranking of Dominance (RDom). For this purpose, each group member was asked to rank other participants based on their dominance. A member who is perceived as the most dominant would be given a score of 1 and a member perceived as the least dominant receives a score of 3 or 4 depending on the number of group members. Ranking of Dominance of a person in the group was then computed as:

$$\text{RDom}_i = 1 - \frac{\bar{R}_i}{\sum_{j=1}^{N} \bar{R}_j},\ i \in 1, \ldots, N = 3\,(\text{or}\,4)$$

where $\bar{R}_i$ is the mean value of the ranking scores assigned to *i* by other participants, i.e. $\bar{R}_i = \sum_{j=1, j\neq i}^{N} \frac{rank(j,i)}{N-1}$; $rank(j,i)$: ranking score assigned to *i* by *j*. As a result, for each perceived variable, a three-dimensional (or a four-dimensional depending on the group size) vector was created that showed how a person was perceived or ranked by others.

Personality of the participants were determined based on the dimensions of the Big Five personality traits, i.e. Agree-

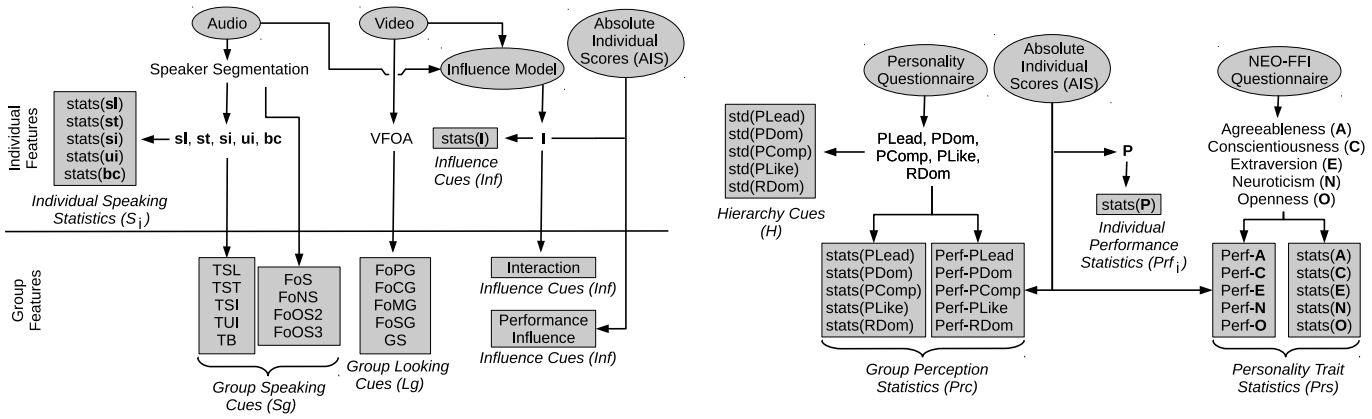

Fig. 2. Schematic representation of meeting-based feature extraction. Ellipses and rectangles show input data and extracted features respectively.

ableness, Conscientiousness, Extraversion, Neuroticism, and Openness. To capture these personality traits, NEO Five-Factor Inventory (NEO-FFI) was used. The NEO-FFI is a questionnaire comprising 60 questions (12 questions for each trait) formatted with a five-point Likert scale, i.e. each question has a score from 1 to 5 ('Disagree totally' to 'Total agreement'). The self-reported questionnaire was asked to be filled by each participant. Afterwards, mean values of 12 questions were calculated for each trait. Eventually, a five-dimensional vector with real values between 1.0-5.0 was generated for each participant, such that the dimensions correspond to the personality traits. Please refer to [14] and [13] in order to find more information on the ELEA dataset. The corpus can be downloaded from https://www.idiap.ch/datasets/elea.

IV. Features

The ELEA corpus provides four types of data with different characteristics. These include (1) audio and video recordings of groups performing winter survival task, (2) individual and group task performances, (3) participant's perception of other group members, and (4) participant's personalities. Based on this information, we extracted 9 sets of features reflecting distinct properties of the dataset. Absolute Individual Scores (AIS) and Absolute Group Scores (AGS) were directly used for performance-related features. For data acquired from questionnaires, personality-related and perception-related features were extracted from averaged and normalized questionnaire variables. Unimodal and multi-modal aural and visual cues were obtained after processing audio and video recordings as explained in the following paragraphs.

Audio processing is straightforward as the Microcone is capable of segmenting speakers automatically. The device stores the segmentation information in triplets as the subject label, speaking times of the relative subject in seconds (start and end), and the Microcone sector. This input is then used to create a binary segmentation for each participant, where status 0 and 1 correspond to silence and speech, respectively.

Video recordings were processed for two types of motion, i.e. head activity and body activity. *Head activity* detection starts with tracking the face of each participant. To this aim, face area is estimated by using a particle filter with an elliptic face model. Then, the optical flow vectors within the face area of two successive frames are computed to detect the changes in the head movement. The average head motion on the x and y dimensions is computed as the average motion vector given the optical flow vectors. As a result, two real-valued vectors are found for each dimension and binarized via automatic thresholding. Resulting binary head motion is calculated by an OR operation between two dimensions. Simple motion differencing is used to extract *Body activity* given that the background is stationary. All moving pixels except those belonging to the tracked head region are considered as the body area. Each frame is converted to a grayscale image and the difference between two successive frames are computed (Δt). It is assumed that moving pixels exist if Δt is greater than a threshold. Then, total number of moving pixels in each frame is normalized by the frame size. Finally, binary body motion is found by only considering frames that have a moving pixel rate of 5% minimum (Please refer to [14] for more details). For multi-modal feature extraction, we used binary segmentations created for each motion type, where status 0 and 1 represents stillness and movement of head (body) respectively.

A. Extraction

In the following, we present the details of the feature extraction process for distinct groups. We selected the meeting segments to be processed in two ways by considering one-second segments or the whole meeting. The former was used for the influence cues and frame-based features, and the latter was employed for the rest of the features. The features extracted by using one-second segments served as input to the sequential probabilistic methods in Sections V and VI-B. Those extracted from the whole meeting were used to report the results in Sections IV-B and VI-A.

Meeting-based feature extraction process is shown in Fig. 2. We represented input data as ellipses and the extracted features as rectangles. To keep the graph as simple as possible, feature names were given in their abbreviated forms and different function names were defined. Five types of statistical measures, i.e. *Mean*, *Max*, *Range*, *Top Range*, and *Hellinger*, were represented as 'stats'. Standard deviation of a variable

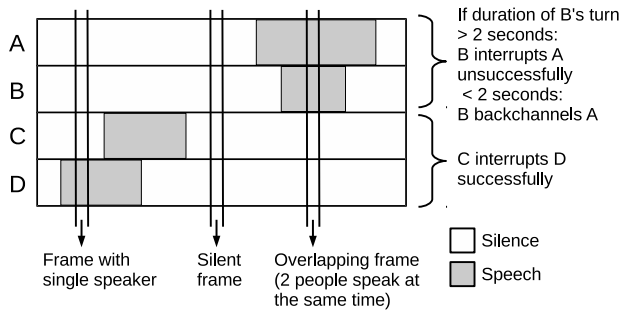

Fig. 3. Graphical explanation of Group Speaking Cues

was coded as 'std' while 'Performance' feature was shortened as 'Perf'.

1) Group Speaking Cues: Group-based speaking cues give insights about the group's turn taking behavior for the whole meeting. We extracted 9 features from the speech segmentation to capture group members' participation in the discussion, the rate of overlapped speech, and the amount of silence.

Initially, for each individual; speaking length (**sl**), speaking turns (**st**), successful interruptions (**si**), unsuccessful interruptions (**ui**), and backchannels (**bc**) were computed. Speaking length is defined as the total time that a participant speaks based on her/his binary speaking status. Speaking turn shows how many times a member takes the ground, i.e. total number of speech segments for the member. Interruptions represent how many times a person successfully and unsuccessfully interrupts other participants. A person (P_1) successfully interrupts another person (P_2) if (P_1) starts speaking during the speech of (P_2) and finishes speaking after the turn of (P_2). P_1 unsuccessfully interrupts P_2 if P_1 speaks more than 2 seconds and, starts and finishes speaking during the speech of P_2. If P_1 speaks less than 2 seconds under the same conditions, P_1 backchannels P_2. Backchannel, in this case, is the total number of times that a person backchannels other participants. Fig. 3 presents examples of group speaking cues for the turn taking behavior of four participants. Summation of these measures for all participants in a group then generates group participation features that include *Total Speaking Length* (TSL), *Total Speaking Turns* (TST), *Total Successful Interruptions* (TSI), *Total Unsuccessful Interruptions* (TUI), and *Total Backchannels* (TB).

Features for overlapped speech and silence are composed of *Fraction of Silence* (FoS), *Fraction of Non-overlapped Speech* (FoNS), *Fraction of Overlapped Speech for 2 People* (FoOS2), and *Fraction of Overlapped Speech for 3 People* (FoOS3). These show in order the fraction of meeting duration for which no participant speaks, a single participant speaks, more than two participants speak simultaneously, and more than three participants speak simultaneously. We refer the readers to [11] for the details of the group-based speaking cue extraction process.

2) Group Looking Cues: Group-based looking cues aims at revealing the gaze behavior of the group. As in the case of speaking cues, group-based cues were built from the participants' individual features for the whole meeting. The visual target of each participant was estimated by using the head pose angle. For this purpose, visual data acquired from the webcameras were processed [11]. The head location and pose were jointly estimated based on a standard state-space formulation within a dynamic, probabilistic framework. Through the use of particle filters, the location, scale, and the discretized pose of the head were matched with different states. The visual focus of attention (VFOA) was then predicted by Maximum a Posteriori (MAP) rule by using the pose of the head. The MAP rule assumes a Gaussian distribution with mean and standard deviation prespecified manually in the state space for visual targets each corresponding to a distinct participant and table area. This VFOA feature indicates whether a participant looks at another participant, or at a point on the meeting table, or stays unfocused. By processing the VFOA features, we defined five group-level looking cues. *Fraction of People Gaze* (FoPG) measures the ratio of interpersonal looking interaction and tells if participants look more at each other or more at the table, or stay unfocused most of the time. *Fraction of Convergent Gaze* (FoCG), *Fraction of Mutual Gaze* (FoMG), and *Fraction of Shared Gaze* (FoSG) correspond to the fraction of meeting duration for which a participant is being looked at by all the other participants, two participants look at each other, and two participants look at the third participant respectively. *Gaze Skew* (GS) sheds light on the uniformity of the participant's gaze interaction, i.e. participants equally look at each other, or a person is being looked at more than others. We will cover this aspect in detail in the following sections. Further information on group-based looking cue extraction can be found in [11].

3) Personality Trait Statistics: The effect of the group composition was evaluated via personality trait features. We used two sources in the feature extraction, the NEO-FFI questionnaire variables and the individual performances of group members (AIS). The former assigns each group member a score for each personality variable. Then for each trait, scores of the group members were used to calculate five statistics. These are the average of individual scores in the group (*Mean*), the maximum score in the group (*Max*), the difference between the maximum and the minimum scores in the group (*Range*), the difference between the maximum and the second maximum scores in the group (*Top Range*), and the skew of the scores in the group (*Hellinger*). Skew is the measure of distribution and is computed by the Hellinger distance. Assume that for a group of 4 people, we have a vector $\mathbf{E}$ whose elements $ex(i)$, $i \in 1,2,3,4$ correspond to the individual scores for Extraversion. The basic statistics are computed as *Mean*($\mathbf{E}$), *Max*($\mathbf{E}$), *Range*($\mathbf{E}$), and *Top Range*($\mathbf{E}$). For the skew, a ranked vector $\mathbf{p}$ is formed by ordering the ratio of Extraversion for each participant, i.e. $ex(i)/\sum_i ex(i)$. This vector is compared with a vector of uniform distribution $\mathbf{q}$, whose elements are $1/|\mathbf{p}|$, $|\mathbf{p}|$ being the cardinality of the group. The comparison is done by the Hellinger distance with the help of Bhattacharya coefficient as $HD(\mathbf{p},\mathbf{q}) = \sqrt{1 - BCoeff(\mathbf{p},\mathbf{q})}$, where $BCoeff(\mathbf{p},\mathbf{q}) = \sum_i \sqrt{p(i) * q(i)}$. The Hellinger distance takes values between 0 and 1, where 0 indicates a uniform meeting. Computation of these statistics for five personality traits results in 25 features such as *Mean-Agreeableness*, *Max-Agreeableness*,

Range-Agreeableness, *Top Range-Agreeableness*, *Hellinger-Agreeableness* and their counterparts for other traits.

We introduced additional features based on the individual performances and the personality variables. For each trait, a person having the highest trait score in the group was selected and her/his individual performance was used as a feature. Hence, five cues were formed from AIS with respect to performance scores as *Performance-Agreeableness*, *Performance-Conscientiousness*, *Performance-Extraversion*, *Performance-Neuroticism*, and *Performance-Openness*.

4) Individual Speaking Statistics: Group speaking cues reflect the speaking behavior of the group as a whole. In this case, individual effects of the members may be missed. To fill this gap, we processed individual speaking features presented in Section IV-A1, i.e. **sl**, **st**, **si**, **ui**, and **bc**. We first normalized each individual feature by the summation operation, e.g., $sl(i)/\sum_i sl(i)$, where $sl(i)$ corresponds to the speaking length of the participant i for $i \in 1, 2, 3, 4$. Afterwards, by employing an analogous notation for **st**, **si**, **ui**, and **bc**, five statistics were extracted as explained before. As a result, we obtained 25 features as *Mean-Speaking Length*, *Max-Speaking Length*, *Range-Speaking Length*, *Top Range-Speaking Length*, *Hellinger-Speaking Length* and their counterparts.

5) Individual Performance Statistics: The basic intuition suggests that there would be a relation between the participants' individual performances and the group performance. In order to judge the correctness of this hypothesis, we extracted AIS statistics. Let **P** denote the performance vector of a group whose elements $perf(i)$ consist of the individual performances for participants $i; i \in 1, 2, 3, 4$. **P** being the input, AIS statistics were obtained for *Mean-Performance*, *Max-Performance*, *Range-Performance*, *Top Range-Performance*, and *Hellinger-Performance*.

6) Group Perception Statistics: Group Perception Statistics were designed to observe the impact of the people with specific roles like leaders and dominants based on the perceived variables, i.e. PLead, PDom, PComp, PLike. We followed the same steps taken for the personality trait statistics. Therefore, two sources were used in the feature extraction, the perception questionnaire variables and the individual performances of group members. The former assigns each group member a score for each perceived variable. Scores of the group members were used to calculate the statistics for PLead, PDom, PComp, PLike, and RDom. Eventually, we had 25 features as *Mean-PLead*, *Max-PLead*, *Range-PLead*, *Top Range-PLead*, *Hellinger-PLead* and their counterparts.

Performance-related features were extracted based on the AIS and the perception variables. For each role, a person having the highest perceived variable score in the group was selected and her/his individual performance was used as a feature. Hence, five cues were built from AIS with respect to performance scores as *Performance-PLead*, *Performance-PDom*, *Performance-PComp*, *Performance-PLike*, and *Performance-RDom*.

7) Influence Cues: Influence is a dominance-like measure, computed automatically by using the Influence Model (see Section V). The method creates scores quantifying the interaction between group members and presents it as a matrix. These scores are used to assign each member an influence value that reflects the effect of a participant on other group members. We produced our cues by importing the influence values obtained as a result of the multimodal analysis, i.e. joint processing of audio and video features, from those reported in [41].

Initially, the influence statistics were extracted. Let **I** show the influence vector of a group whose elements $inf(i)$ are composed of the influence values of members $i; i \in 1, 2, 3, 4$. **I** being the input, influence statistics were obtained for *Mean-Influence*, *Max-Influence*, *Range-Influence*, *Top Range-Influence*, and *Hellinger-Influence*. Secondly, the most influential person in the group was selected, i.e. $infPerson = \arg max_i(inf(i))$, and her/his AIS score was used as the performance feature, i.e. *Performance-Influence*. Finally, the form of group interaction was used as the *Interaction* feature. The feature takes a value among 1, 2, and 3 that corresponds to one-to-one, many-to-many, and one-to-many interaction. In one-to-one interaction, a participant has influence on only another participant. A participant has influence on at least two participants for one-to-many interaction. In the last case, many participants have influence on many others. To find such relation, we first applied a threshold (0.8 in our case) on the influence matrix and assumed that a person has influence on another if the threshold exceeds the influence score. For each participant, resulting interaction type was automatically retrieved by calculating the number of people that the participant affects and that the participant is affected by.

8) Hierarchy Cues: In a recent study, Frauendorfer et al. state that hierarchically structured groups perform better in decision tasks than the unstructured ones [9]. They assume that a hierarchical structure exists if a person in the group is in a position of power. As a measure of power, perceived dominance was selected and its standard deviation was used as an indicator of hierarchy. Following this idea, we extracted power hierarchies for all perceived variables and the influence. Statistically, hierarchy cues and the Range in Group Perception Statistics can be thought of as similar measures quantifying the variability in the data. However, they provide different information especially when the data is skewed and the ranges are the same. In this case, hierarchy cues can be used to make predictions on the group performances. Another point to be considered is the type of structure that is conveyed in the hierarchy. In terms of the influence, hierarchy corresponds to an interaction structure that is understood as the extent to which a person can interact with other group members. Let **PL** denote the vector of a group whose elements $PLead(i)$ include the PLead scores for participants $i; i \in 1, 2, 3, 4$. Standard deviation of **PL** leads to the hierarchy feature with respect to perceived leadership, i.e. *Hierarchy-PLead*. An analogous notation applies for the other perceived variables and the influence. Hence, we also have *Hierarchy-PDom*, *Hierarchy-PComp*, *Hierarchy-PLike*, *Hierarchy-RDom*, and *Hierarchy-Influence*.

9) Frame-based Features: The features extracted before provide a single value for each group by using the whole meeting as a segment to be processed. In this section, we processed the meeting with thin slices, in one-second segments.

Consequently, extracted features present sequence of values for each group rather than a unique measure. We used the binary speech and motion segmentations as well as the visual focus of attention information.

Frame-based audio features were acquired by utilizing the binary speech segmentation. We extracted the following cues for each participant.

Speaking Status: The binary speech segmentation for participant i.

Successful Interruptions-Audio: A binary feature where status 1 represents time instants with Successful Interruptions and status 0 otherwise, given that *Participant i interrupts participant j if i starts talking while j is speaking, and i finishes her/his turn after j does*.

Unsuccessful Interruptions-Audio: A binary feature where status 1 represents time instants with Unsuccessful Interruptions and status 0 otherwise, given that *Participant i interrupts participant j if i starts talking while j is speaking, and i finishes her/his turn before j does*.

Frame-based video features were obtained by using the binary head and body motion segmentations. The following cues were extracted for each participant.

Head Motion Status: The binary segmentation for head motion for participant i.

Body Motion Status: The binary segmentation for body motion for participant i.

Successful Interruptions-Head and Successful Interruptions-Body: The binary features where status 1 represents time instants with Successful Interruptions and status 0 otherwise, given that *Participant i interrupts participant j if i starts acting while j is moving, and i finishes her/his turn after j does*, respectively for head and body motion.

Unsuccessful Interruptions-Head and Unsuccessful Interruptions-Body: The binary features where status 1 represents time instants with Unsuccessful Interruptions and status 0 otherwise, given that *Participant i interrupts participant j if i starts acting while j is moving, and i finishes her/his turn before j does*, respectively for head and body motion.

The visual focus of attention provides information about participants' visual target, i.e. who looks where at a specific time. For frame-based VFOA features, we only considered inter-personal gaze by ignoring targets for the table area and unfocused. Such a choice aims at determining the effect of the participant who attracts more attention than others. For this purpose, we defined a multivariate binary feature for each participant $i; i \in 1,2,3,4$ such that each dimension of the feature corresponds to a distinct participant $j; j \in l : l \neq i$ in the group and feature is assigned 1 at the time instant when the participant i is being looked at by the participant j.

B. Correlation Analysis

In this part, we analyze the correlation between the meeting-based features (Sections IV-A1 to IV-A8) and the group performance, i.e. AGS. For this purpose, Pearson correlation coefficients and the p-values were calculated. Table I depicts significant correlations, i.e. those with p-values less than $0.1(^*)$ and $0.05(^{**})$, for features extracted by using all groups and four-person groups for which video recordings are available, resulting in 40 groups and 21 groups, respectively. The latter set enabled us to evaluate the effect of the group looking cues. Although there were 27 groups with video recordings, we disregarded 6 three-person groups to provide a comparable ground for the experimental evaluation. We will detail the reasons of such selection in Section VI-B1.

TABLE I
CORRELATIONS BETWEEN MEETING-BASED FEATURES AND THE GROUP PERFORMANCES ($* : p < 0.1$, $** : p < 0.05$).

Feature Set	**Feature**	R for all ELEA groups	R for 4-person groups with video
Group Speaking Cues (Sg)	Total Speaking Length	-0.327^{**}	-
	Fraction of Silence	0.431^{**}	-
Group Looking Cues (Lg)	Fraction of Convergent Gaze	-	0.562^{**}
Personality Trait Statistics (Prs)	Range-Openness	-	0.431^{*}
	Hellinger-Openness	-	0.445^{**}
	Performance-Openness	0.288^{*}	-
	Range-Conscientiousness	-	0.525^{**}
	Top Range-Conscientiousness	-	0.510^{**}
	Hellinger-Conscientiousness	-	0.511^{**}
	Range-Extraversion	-0.333^{**}	-
	Hellinger-Extraversion	-0.375^{**}	-
Individual Speaking Statistics (S_i)	Top Range-Speaking Length	0.272^{*}	-
	Top Range-Speaking Turns	0.269^{*}	-
	Max-Unsuccessful Interruptions	0.311^{*}	0.543^{**}
	Range-Unsuccessful Interruptions	0.306^{*}	0.431^{*}
	Top Range-Unsuccessful Interruptions	0.309^{*}	0.499^{**}
	Hellinger-Unsuccessful Interruptions	0.333^{**}	-
Individual Performance Statistics (Prf_i)	Mean-Performance	0.395^{**}	-
	Max-Performance	0.354^{**}	-
	Hellinger-Performance	0.264^{*}	-
Group Perception Statistics (Prc)	Range-PDom	0.305^{**}	-
	Hellinger-PDom	0.362^{**}	-
	Max-PCom	0.312^{**}	-
	Range-PCom	0.299^{*}	-
	Top Range-PCom	0.320^{**}	-
	Hellinger-PCom	0.306^{*}	-
	Performance-PLead	0.604^{**}	0.445^{**}
	Performance-PDom	0.318^{**}	-
	Performance-PCom	0.541^{**}	-
	Performance-RDom	0.468^{**}	0.585^{**}
Influence Cues (Inf)	Max-Influence	0.317^{**}	-
	Range-Influence	0.307^{*}	-
	Hellinger-Influence	0.294^{*}	-
	Performance-Influence	0.595^{**}	0.439^{**}
	Interaction	0.365^{**}	-
Hierarchy Cues (H)	Hierarchy-PDom	0.352^{**}	-
	Hierarchy-PCom	0.342^{**}	-
	Hierarchy-Influence	0.318^{*}	-

As far as all groups are concerned, it is seen from the third column of Table I that significant correlations exist for all feature sets except Group looking cues. In group speaking cues, *Total Speaking Length* and *Fraction of Silence* have correlations of -0.327 and 0.431. Even if they are different in sign, they characterize the same aspect of the groups due to their complementary nature. Groups that speak less in total, i.e. that stay silent more, perform better than those having higher talking time. Personality trait statistics capture two aspects regarding *Openness* and *Extraversion*. The higher the performance of the most open person to experience in a group the higher the group performance. Range and Hellinger distance with respect to Extraversion have negative correlation (-0.333 and -0.375) . This implies that groups formed by similarly extrovert people, i.e. the more uniform group wrt. extraversion,

have higher group performance. Information on *Speaking Length*, *Speaking Turns*, and *Unsuccessful Interruptions* is provided by the Individual Speaking Statistics. Top Range for *Speaking Length* and *Speaking Turns* indicates that the groups with higher difference between the top two members wrt. hierarchy of *Speaking Length* and *Speaking Turns* perform better. The higher performances are achieved also for groups in which some participants have much more unsuccessful interruptions than others, i.e. higher diversity wrt. unsuccessful interruption count. Positive correlations observed for the Individual Performance Statistics indicate that average and maximum performance of the group members are determining factors for the success of the decision task. The correlation for the Hellinger distance suggests that the higher the amount of divergence in participants' performances in a group, the better the task performance. Statistical features computed for the Group Perception correlate with the group success positively. They all mean that groups in which some members are perceived much more dominant and competent than others make better decisions. In addition, it is observed that the performance of the leader, dominant and competent person in a group is a strong indicator of the performance of the group. Correlations for the Influence Cues show similarities to those for the Group Perception. Groups whose members have considerably different levels of influence (*Range* and *Hellinger*) reach better success rates, which is also parallel to the influential member's performance. *Interaction* feature reveals that one-to-one and one-to-many interaction types correspond to lower and higher performance levels respectively. Finally, *Hierarchy Cues* point out that groups in which a hierarchical structure exists with respect to Dominance, Competence, and Influence perform better than groups with a flat hierarchical structure.

The last column of Table I presents correlations for four-person groups with video recordings. Due to less number of examples, no significant correlations observed for Individual Performance Statistics and Hierarchy Cues. In Group Looking Cues, *Fraction of Convergent Gaze* has a correlation of 0.562. This implies that the group performance is higher for teams in which active participants are being looked at by all other members most of the time. Personality Trait Statistics cover information for Openness and Conscientiousness. For the former, a group performs well if some of the members of the group are open to experience more than others. A similar situation holds for the latter case. In other words, higher performances are obtained in groups some of whose members are more conscientious than others. Individual Speaking Statistics show that groups in which some participants have much more unsuccessful interruptions than others are more successful in the task. Lastly, it is seen from the Group Perception Statistics and the Influence Cues that the performance of the leaders, dominant people and the influential people are important determinants of the performance of a group.

V. Influence Model

We used the Influence Model [35] to model the interactions between group members. The approach is a modified version of a coupled Hidden Markov Model (CHMM) with the same structural and graphical architecture. That is to say, it is composed of multiple HMM chains coupled through cross-time and cross-chain conditional probabilities. The modification introduced in the CHMM covers an important deficiency by considerably decreasing the number of parameters to be learned. By this way, higher number of chains can be included into the model without incurring substantial increase in computational complexity. The Influence Model makes such improvement possible with a parametrization strategy in terms of the "influence" each chain has on other chains. To be more specific, the conditional probability $P(S_t^i|S_{t-1}^1,\ldots,S_{t-1}^C)$ is simplified by only retaining the first-order transition probability $P(S_t^i|S_{t-1}^j)$, where t is the time stamp and C is the number of chains. As a result, the full conditional distribution is estimated as follows:

$$P(S_t^i|S_{t-1}^1,\ldots,S_{t-1}^C)=\sum_j \tau_{ij}P(S_t^i|S_{t-1}^j).$$

In the equation above, τ's correspond to "influences" that represent the effect of chains on each other depending on their states. The Influence Model only captures pairwise interactions of chains because the model is not fully-connected as a generalized Coupled HMM. Although information on the joint effect of multiple chains is missing, interaction between pairs of chains satisfy our experimental expectations.

Parameters of the Influence Model are learned by a two-step operation. In the first phase, the forward-backward algorithm is applied for the latent state inference. Then, inferred states are used for the maximum likelihood estimation. Let observations and latent states be $\mathbf{X}=(X^{(1)},\ldots,X^{(C)})$ and $\mathbf{S}=(S^{(1)},\ldots,S^{(C)})$ respectively. Then, forward and backward variables are calculated as follows:

$$\widetilde{\alpha}_1^{(c)}(s)=\pi_s^{(c)}P(x_1^{(c)}|s)$$

$$\widetilde{\alpha}_{t+1}^{(c)}(s)=P(x_t^{(c)}|s)\sum_{c_1=1}^{C}\sum_{s_1=1}^{m_{c_1}}\alpha_t^c(s_1)h_{s_1,s}^{(c_1,c)}$$

$$\beta_T^{(c)}(s)=1$$

$$\beta_{t<T}^{(c)}(s)=\frac{\sum_{c_1=1}^{C}\sum_{s_1=1}^{m_{c_1}}h_{s,s_1}^{(c,c_1)}\beta_{t+1}^{(c_1)}P(x_{t+1}^{(c_1)}|s_1)}{\text{scale}_{t+1}^{(c_1)}}$$

where $\pi_s^{(c)}$ being the initial probabilities, and $h_{s_1,s}^{(c_1,c)}=\tau_{c_1,c}\times a_{s_1,s}^{(c_1,c)}$, a being the probability of latent state transitions. Computation of the forward and backward variables are done based on the observation probability $P(x_t^{(c)}|s)$; where $x_t^{(c)}$ is the observation for chain c at time t (see [43]).

Being a sequential approach, the Influence Model takes as input the frame-based features extracted in Section IV-A9. Originally, it can be used to work only on multinomial data for discrete cases. Since our data is multi-dimensional, we proposed an improvement also to cover multivariate binary cases. For this purpose, we made a Naive assumption of independence between observation features and calculated the observation probabilities as follows:

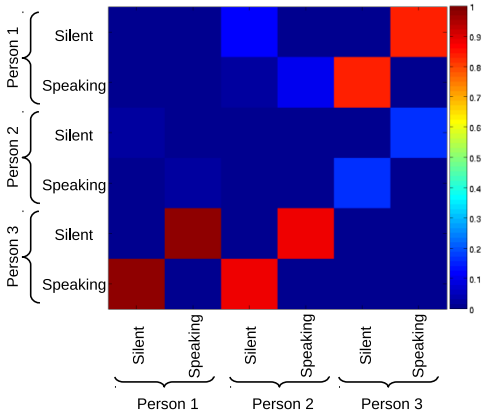

Fig. 4. A sample influence matrix for audio features

$$P(x_t^{(c)}|s) = \prod_{i=1}^{N} P(x_{t,i}^{(c)}|s)$$

As a result, the probability of seeing N observations together is computed as the product of their individual probabilities. Forward and backward variables are then used to estimate model parameters.

The Influence Model that is run for a group outputs a matrix with a dimension of $D \times D$, where D is computed as the product of the number of states and the number of participants in the group. A sample influence matrix is depicted in Fig. 4 for a 3-person group. Here, each participant corresponds to a chain with two states, i.e. *Silent* and *Speaking*. A cell in the matrix represents the influence of a member's state (in rows) on another member's state (in the columns). The influence takes values between 0 (blue) and 1 (red). In our example, Person 1 has influence on Person 3 such that when Person 1 is *Silent*, Person 3 tends to be in *Speaking* state and when Person 1 is in *Speaking* state, Person 3 tends to be *Silent*. Note that, Person 1 affects no one except Person 3. Person 1 and 2 are affected by Person 3 such that they tend to listen Person 3 while she/he is speaking and vice versa. It is clear from the interactions that Person 3 is the most influential member in the group while Person 2 is the least influential one.

VI. Experiments

We addressed the problem of predicting group performances in two ways. First, we performed regression and classification analyses and evaluated the performance of prediction for distinct feature sets representing the entire meeting, i.e. the segment covering the whole duration of the Winter Survival Task. By regression, we aimed at using a predictive model to estimate a single score corresponding to the group performance. In classification, we reduced the problem from predicting an actual value to classifying the group performance as one of the two performance levels: low and high. Second, we approached the classification problem from a different angle. Instead of using the whole meeting as a segment, features extracted from one-second intervals were processed by probabilistic sequential techniques.

A. Meeting-based Evaluation

In the experiments below, four different settings were evaluated based on the dataset size and the selected features to assess the change in group structure and the predictive power of features. We processed the dataset in two ways depending on its size: the full ELEA corpus (40 meetings) and the four-person groups for which video recordings are available (21 meetings). As far as the features are concerned, we used either the full feature set regardless of their correlation values (Section IV-A) or those with the significant correlations (Section IV-B). Their combinations produce (1) full ELEA corpus with full feature set (107 features), (2) full ELEA corpus with correlated features (32 features), (3) 4-person groups with full feature set (112 features), and (4) 4-person groups with correlated features (12 features). Dashes in the tables indicate that either the feature is not available for the current experimental setting (i.e. lack of video recordings) or features in the set are not statistically significant based on the correlation analysis.

In all the experimental procedures, we followed the leave-one-out (LOO) cross validation scheme. The coefficient of determination (R^2) and Mean Squared Error (MSE) are reported for regression experiments; the accuracy is reported for classification experiments. The data was normalized such that each feature has zero mean and one standard deviation. Feature set names in Tables II and V were presented in an abbreviated form but their meanings can be seen from Table I. The combination of all features was shown as *All*.

1) Regression: We used Support Vector Regression to evaluate the power of the group cues to predict the group performance. The model parameters were optimized by applying an internal cross validation procedure within the training folds of the LOO approach. The training data in each fold were again divided into training and test sets by performing the same LOO technique. For each internal validation step, a range of possible parameters was evaluated and those minimizing the MSE were recorded as candidates. The final parameters were chosen as those providing the lowest MSE across the internal cross validation steps. These were then used in the outer cross validation procedure. Radial Basis Function was chosen as the Kernel type. We selected the parameters for Gamma, Epsilon, and C from the range $[2^{-8}, 2^{8}]$, $[0, 8]$, and $[2^{-2}, 2^{11}]$ respectively.

The Coefficient of Determination, R^2, indicates the relative improvement achieved by using the regression model as a predictor instead of the sample mean. The measure is defined as follows:

$$R^2 = 1 - \frac{\sum (y_t - \hat{y})^2}{\sum (y_t - \overline{y})^2}$$

where y_t and $\overline{y}$ are the test variables and their mean; and $\hat{y}$ is the predicted values. Test variables, i.e. group performances, range between 22 and 60 inclusively (see Section III-B). Note that the R^2 can be negative when the baseline model outperforms the regression model. The Mean Squared Error quantifies the difference between the estimator and the test values and is computed as $1/n \sum (\hat{y} - y_t)^2$, n being

TABLE II
REGRESSION RESULTS FOR FULL ELEA CORPUS. R^2 AND MSE VALUES FOR DIFFERENT FEATURE SETS. TEXT IN BOLD INDICATE $R^2 \geq 0.15$.

		Sg	Lg	Prs	S_i	Prf_i	Prc	Inf	H	All
All features	R^2	**0.17**	-	-0.20	-0.28	0.01	0.10	**0.22**	-0.13	0.08
	MSE	**78.40**	-	112.41	119.78	93.01	84.78	**73.10**	105.78	86.26
Correlated features	R^2	-0.41	-	-0.02	-0.11	-0.15	**0.16**	**0.24**	0.09	0.11
	MSE	131.97	-	95.92	104.23	108.03	**78.46**	**71.25**	85.09	83.32

TABLE III
REGRESSION RESULTS FOR GROUPS WITH VIDEO. R^2 AND MSE VALUES FOR DIFFERENT FEATURE SETS. TEXT IN BOLD INDICATE $R^2 \geq 0.15$.

		Sg	Lg	Prs	S_i	Prf_i	Prc	Inf	H	All
All features	R^2	-0.71	**0.15**	-0.19	-0.17	-1.89	-0.25	-2.65	-0.61	0.06
	MSE	89.48	**44.79**	62.11	61.40	151.44	65.65	192.10	84.28	49.54
Correlated features	R^2	-	-0.09	0.00	-0.01	-	-0.04	**0.17**	-	**0.28**
	MSE	-	57.05	52.34	53.02	-	54.53	**43.47**	-	**38.01**

the number of test items. The R^2 and the MSE values are reported in Tables II and III. The values in bold font indicate $R^2 \geq 0.15$.

Table II shows the results for the full ELEA corpus by using the full feature sets and the sets consist of those with significant correlations. *Sg* in the former and *Prc* in the latter explain around 17% of the variance in group performance. In both situations *Inf* provide more predictive power. Combining all features together does not outperform any one of the feature sets. Absence of video recordings, hence *Lg*, for the entire dataset is reflected as a dash in the results.

The prediction results given in Table III correspond to the deducted dataset for all the features and the significant ones. The regression model performs better than the baseline mean model in *Lg* for all features and in *Inf* for the significant features. Due to the smaller size of the dataset, we observe less number of significant features. That's why, predictions for *Sg*, Prf_i, and *H* are missing. However, using a small set of important features leads to a high R^2 of 0.28.

In order to investigate the effect of varying dataset size, we compared the full ELEA corpus with the reduced dataset for all features and the significant features, i.e. Table II vs. III. Decreasing the data size causes losing major predictors. This can be seen from the R^2 values in *Sg* and *Inf* for all features, and in *Prc* and *Inf* for correlated features. When all experimental settings are considered, *Inf* can be said to be key predictors of the group performance as they show high predictive power in 3 out of 4 cases. Although combining all features results in the highest overall R^2, it is observed only for one of the settings.

2) Classification: For the classification experiments, we defined two classes based on the performance clusters extracted in [41]. The authors divided the ELEA corpus into three parts depending on the AGS scores of the groups. The resulting clusters correspond to low, average, and high performance levels and include 20, 15, and, 5 groups respectively. In order to create new classes, we used the highest performance in the low performance cluster as the cutoff point. Any group that has a lower performance with respect to the cutoff point was assigned to the low performance class. Similarly, groups having higher performance with respect to the cutoff point were assigned to the high performance class. As a result two balanced classes were obtained, each is composed of 20 groups. Such a selection produces relatively balanced classes also for the reduced dataset, i.e. there are 12 and 9 groups in the low and high performance classes. In the classification phase, we set labels 0 and 1 for scores smaller (or equal) and greater than the cutoff point to represent groups with high and low group performance respectively. Following this classification scheme, random baseline accuracies were determined as 50% and 60% (computed based on the populous class) for the full ELEA corpus and the reduced dataset, respectively.

TABLE IV
CLASSIFICATION RESULTS FOR FULL ELEA CORPUS. ACCURACY VALUES FOR DIFFERENT FEATURE SETS. TEXT IN BOLD INDICATE $Acc \geq 0.655$ FOR $\alpha = 0.05$.

		Sg	Lg	Prs	S_i	Prf_i	Prc	Inf	H	All
All features	Accuracy	0.60	-	0.58	0.33	0.50	0.53	**0.78**	0.58	**0.68**
Correlated features	Accuracy	0.65	-	0.63	0.40	0.40	**0.68**	**0.73**	0.43	**0.75**

TABLE V
CLASSIFICATION RESULTS FOR GROUPS WITH VIDEO. ACCURACY VALUES FOR DIFFERENT FEATURE SETS. TEXT IN BOLD INDICATE $Acc \geq 0.748$ FOR $\alpha = 0.1$.

		Sg	Lg	Prs	S_i	Prf_i	Prc	Inf	H	All
All features	Accuracy	0.71	**0.76**	0.62	0.48	0.43	0.57	0.67	0.33	0.52
Correlated features	Accuracy	-	**0.76**	0.62	0.57	-	0.52	**0.76**	-	0.57

Classification experiments were set similar to the regression case by using Support Vector Classification. Parameters of the model were determined as a result of the internal cross validation process as described before. However, we maximized the accuracy instead of minimizing the MSE. Here, the classification accuracy was computed as the fraction of correct predictions. We performed the experiments with the Radial Basis Function and for Gamma and C parameters that range from $[2^{-8}, 2^{8}]$, and $[2^{-2}, 2^{11}]$ respectively.

Tables IV and V depict classification accuracies. The values highlighted in boldface indicate statistically significant ones. We determined the significance levels by calculating confidence intervals for the baselines. Accuracies above 65.5% and 74.8% were considered significantly different than the 50% (with 95% confidence level) and 60% (with 90% confidence level) baselines respectively. We used a narrower confidence interval for the latter case to compensate the effect of the smaller dataset.

The classification results of the full ELEA corpus for the full feature sets and the important features are given in Table IV. *Inf* for both cases and *Prc* for the latter provide significant classification accuracies. Although merging all features together achieves high classification rates, the highest rate observed for *Inf* outperforms others.

Table V presents the results of the classification task for

the smaller dataset with the full feature sets and the important features. Significant classification rates are observed for *Lg* in both settings and for *Inf* in the second one. Binding feature groups (*All*) does not improve the results over the best performing feature set(s), i.e. *Lg* in the former and, *Lg* and *Inf* in the latter.

Joint analysis of the regression and the classification tasks tells us that *Inf* and *Lg* are major predictors of the group performance. *Sg* are observed to be significant only in one of the settings of the regression task. Only in the experimental setup in which the full ELEA corpus is used with the important features, *Prc* show significant predictive power.

B. Frame-based Evaluation

In this part, we focus on classifying the group performances by using probabilistic sequential approaches for the frame-based features introduced in Section IV-A9. The classification experiments were performed for three different feature sets: (i) audio, (ii) video, and (iii) visual focus of attention. For the audio and video feature sets, we analyzed the effect of the turn taking behavior by evaluating successful interruptions. Audio-Successful Interruption feature set for a currently speaking participant is composed of the *Speaking Status* of the person and 3 features that correspond to the *Successful Interruptions-Audio* computed over other group members, resulting in a 4-dimensional feature vector. Video-Successful Interruption feature set for a currently moving participant includes two motion features of the person, i.e. *Head Motion Status* and *Body Motion Status*, 3 features for the *Successful Interruptions-Head*, and 3 more for the *Successful Interruptions-Body*, resulting in an 8-dimensional feature vector. Visual focus of attention feature set for a participant, *Focus on Speaker*, is a 3-dimensional vector, each dimension of which represents another group member's gaze to the person.

Classification in sequential models consists of two basic phases. First, a model is learned for each class from the training examples of that class; then, a new sequence to be classified is evaluated by using each model. The class that corresponds to the model producing the highest likelihood is selected as the predicted one. We adopted the same class definitions presented in the meeting-based analysis and used the leave-one-out approach for the experiments.

The frame-based analysis was carried out only for four-person groups for which video recordings are available, i.e. 21 meetings. We introduced such a constraint for computational simplicity in the course of training and testing stages. This is especially relevant for the Influence Model in which each participant corresponds to a distinct Markov Chain. Usage of the whole ELEA corpus leads to a change in the group size and the corpus does not contain enough examples of three-person groups to perform a statistical analysis of different group sizes on the prediction performance. Fig. 5 illustrates the main steps of our frame-based analysis.

1) Influence Model: The regression and classification results given in the meeting-based analysis showed that the features extracted from the Influence Model provided significant information in predicting the group performance. With this motivation, we focused on developing a classifier with the Influence Model. As explained before, the IM is characterized by multiple Markov Chains and the effect of chains on each other. Knowing the fact that each chain corresponds to a group member, the frame-based feature sets created for each participant constitute the observation vector of the relative chain. We considered an ordering of the chains based on the influence values of the team members such that the first chain of the model corresponds to the most influential participant, and the last one links to the least influential member. The aim of this ranking is to account for the interaction patterns and to eliminate the randomization effect in the model building. In order to reflect the temporal flow of the discussion features, the left-to-right architecture was adapted as the chain topology. Finally, we solved the classification problem for varying number of hidden states, i.e. 2 to 4.

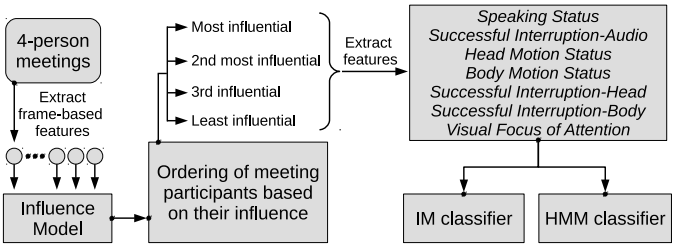

Fig. 5. The flowchart of the frame-based analysis

Building a classifier mainly equates to estimating the model parameters, which include the components of the Markov Chain, i.e. the prior, transition, and emission probabilities, and the influence values. The estimation process starts with an initialization of the parameter values. As experimentally observed, the quality of the classifier is less dependent on the initial values of the prior and transition probabilities than the observation probabilities. Considering our assumption on the chain topology, we assigned relatively high values to the first state in the initial state distribution and to the self-transitions. The emission probabilities, on the other hand, were computed from a sampled data. For this purpose, we first selected a group from the training set randomly. Then, the observation vector of the group was uniformly segmented based on the number of hidden states. The emission probabilities were calculated from the segmented observation and hidden state pairs. We assumed that, in the beginning, the influence of each member on another is the same and that the self-influence is zero. Initial estimation of parameters is followed by the update procedure, i.e. Baum-Welch algorithm. We set the number of iterations in the expectation-maximization (EM) step to 50.

Table VI shows the classification accuracies of the models generated for different approaches (IM vs. HMM), modalities, features sets, and the number of hidden states. We repeated each experimental condition ten times and reported the averaged results. With multiple runs, we aimed at avoiding any potential bias that may stem from the randomization step in the emission probability estimation. The results of the IM indicate that the features extracted from the video recordings, i.e. *Video* and *VFOA*, outperform those extracted from the audio recordings. Even if the extracted features

TABLE VI
FRAME-BASED CLASSIFICATION ACCURACIES FOR IM AND HMM. TEXT IN BOLD INDICATE $Acc \geq 0.748$ FOR $\alpha = 0.1$.

Modality	Feature Set	IM			HMM		
		Number of States			Number of States		
		2	3	4	2	3	4
Audio	*Speaking Status + Successful Interruptions*	0.52	0.55	0.54	0.53	0.42	0.40
Video	*Motion Status + Successful Interruptions*	0.66	**0.75**	0.63	0.67	0.61	0.53
VFOA	Focus on Speaker	0.67	0.64	0.61	0.61	0.56	0.59

are different, in terms of modality, this finding is consistent with the classification results of the meeting-based analysis in which *Lg* is better in classifying the group performance than *Sg*. Note also that, *VFOA* and *Fraction of Convergent Gaze* are just two different representations that characterize the same looking behavior. More importantly, successful interruption for the video modality with the 3-hidden-state model has a significant classification accuracy of 75%.

2) Hidden Markov Model: We performed the same classification experiments presented in the previous section by using the Hidden Markov Model for comparison purposes. The problem to be solved by the HMM is a pruned version of the IM in which the interaction between participants is omitted. Hence, parameters to be learned are composed of the HMM components only. As a basic HMM is characterized by a single chain, we merged the chains of the IM into one by concatenating the participants' frame-based features together. The same order of participants based on the influence values has been applied. Apart from this change, we remained faithful to the parameter estimation procedure and the experimental setting, i.e. the number of hidden states and repetitive runs, the maximum iteration count in EM were kept the same as the IM-based approach.

The classification accuracies obtained by the HMM are presented in Table VI-HMM. As previously observed, the features extracted from the audio recordings perform worse than those extracted from the video data. The setting that produces the best accuracy, i.e. 67%, is the one for the Video modality with the 2-state-model. Despite being the highest score, it is not significant. Note also that, the Influence Model outperforms the HMM in almost all the experimental conditions when compared one-by-one.

It is seen from the results in Table VI that the classification performance varies from one modality to another. Since each modality addresses to a distinct set of group attributes, a classifier trained for one modality misses important information from other sources that also play role in prediction. In order to overcome this deficiency, we built ensembles of classifiers from the classification results obtained for Audio, Video, and Visual Focus of Attention. We created two types of combinations by using different fusion strategies, i.e. *Sum* and *Voting*. Let ll_i^{low} and ll_i^{high} denote, respectively, the likelihood values of the low and high performance classes computed for modality $i; i \in Audio, Video, VFOA$. In *Sum* fusion, the likelihood values were summed over the modalities for each class, i.e. $Sum^{low} = \sum_i ll_i^{low}$ and $Sum^{high} = \sum_i ll_i^{high}$. Class assignments were done based on the fused likelihoods such that the class producing the highest cumulative probability was picked. In *Voting*, however, the final classes were selected by applying majority voting over the predicted classes, i.e. a class that had been selected by at least two out of three modalities was picked as the resulting class.

TABLE VII
FRAME-BASED CLASSIFICATION ACCURACIES FOR THE FUSION OF CLASSIFIERS. TEXT IN BOLD INDICATE $Acc \geq 0.748$ FOR $\alpha = 0.1$.

Fusion Type	IM			HMM		
	Number of States			Number of States		
	2	3	4	2	3	4
Sum	0.62	**0.76**	0.62	0.62	0.62	0.48
Voting	0.71	**0.76**	**0.81**	0.71	0.52	0.57

The advantage of using the fusion techniques can be seen from the IM column of Table VII. We obtained significant classification accuracies for both the *Sum* and *Voting* schemes such that the produced values are higher than the best one found by the single modality, i.e. Video. Although the 4-state models have accuracies of 54%, 63%, and 61% for Audio, Video, and VFOA respectively, their fusion (*Voting*) has the classification rate of 81%, which is the highest result achieved up to this point. This suggests that the classification criteria vary from one group to another depending on the modality, e.g., a group that is successfully classified by using the audio-based features may not be classified by using the video-based features. In such situations, the fusion procedure benefits from the individual strengths of each classifier and provides improvement in the performance.

HMM column in Table VII reports the results of the fusion procedure using HMM as the classifier. The *Sum* ensemble is able to provide improvement over the best of its components only with the 3-state model. The *Voting* scheme results in higher classification accuracies, i.e. 71%, in comparison to the *Sum* but neither fusion type satisfies the significance condition over the random baseline.

We finalize our experiments by evaluating the joint effect of features introduced in the meeting-based and the frame-based analyses on predicting the group performance. As the extracted features are in different format, i.e. single value representing a meeting vs. time series data, we applied decision level fusion instead of feature level fusion. In this process, we fused Audio, Video, and Visual Focus of Attention features in the frame-based analysis (see Table VI) with the best performing features in the meeting-based analysis, i.e. *Lg* and *Inf* (see Table V). We created different fusion schemes based on the combinations of the latter. The *Voting Fusion* was then applied to the selected predictions to determine the final classes. Note that, the *Sum Fusion* is not applicable here as the probability information provided by the Support Vector Classifier and sequential classifiers (IM and HMM) are not compatible with each other, i.e. the former gives probabilities of a testing instance to belong to each class and the latter provides the likelihood of a given observation sequence. Also

TABLE VIII
CLASSIFICATION ACCURACIES FOR THE FUSION OF MEETING-BASED AND FRAME-BASED CLASSIFIERS. TEXT IN BOLD INDICATE $Acc \geq 0.748$ FOR $\alpha = 0.1$.

Fused with (+)	IM Audio+Video+VFOA			HMM Audio+Video+VFOA		
	Number of States			Number of States		
	2	3	4	2	3	4
SVC: all Lg	0.71	0.71	**0.81**	0.62	0.57	0.57
SVC: significant Lg	0.71	**0.76**	0.71	0.62	0.52	0.57
SVC: significant Inf	**0.76**	**0.76**	**0.81**	0.62	0.52	0.48
SVC: significant Lg + Inf	**0.91**	**0.91**	**0.91**	**0.81**	0.67	0.67

note that, we run the experiments only on the groups for which video recordings were available to be consistent with the experimental design.

Table VIII shows the classification results computed from the *Voting* procedure. In the first three rows, three classifiers from the frame-based analysis (Audio, Video and Visual Focus of Attention) were fused with a classifier from the group-based analysis, i.e. the one generated via Support Vector Classifier by using the Influence Cues or Group Looking Cues (either with all features or only with significant ones). In majority voting, a class that had been selected by at least three out of four classifiers was picked as a resulting class. The results indicate that the fused model with *Lg* features shows similar classification accuracies to those obtained from the voting results of the frame-based analysis (see TableVII). This is reasonable since *Lg* features convey similar information to Video and VFOA features. On the other hand, an improvement is observed with the introduction of a new information source, i.e. *Inf*. In this case, classification performance increases for the 2-state IM. Although there is a small decrease in the classification performance of the 2-state HMM, results lay below the significance level. The major improvement was achieved by fusing all five classifiers, i.e. Audio, Video, VFOA, *Lg*, and *Inf*, shown in the last row. In this case, the fused model reaches the highest classification rate in our experiments, i.e. 91%, regardless the number of states in the Influence Model. The model also provides significant performance enhancement for the 2-state HMM. These results prove that the fusion process is indeed useful in incorporating information from different sources and that a combination of two sets of features (meeting-based and frame-based) provides more accurate predictions.

VII. CONCLUSION

In this study, a computational framework for the prediction of decision-making performance of small groups was presented. We used as predictors a wide range of self-reported and automatically extracted features that reflect not only non-verbal communication patterns of groups but also discussion dynamics and group structure. To our knowledge, this work is the most extensive qualitative assessment of the performance prediction of groups in terms of the covered features, and the first to adapt the Influence Model for classification purposes.

In order to gather information on decision-making processes, 40 groups were asked to perform the winter survival task. Discussion of each team was recorded by using audio and video devices. In addition, individual and inter-personal attributes were collected from questionnaires filled before and after the discussion session. We first analyzed the data based on the features extracted from the whole meeting. We determined the factors affecting the group performance by performing a correlation analysis. Among the highest correlated features were the performance of the group member with a specific role (e.g., leader, the most influential one) as well as the amount of silence and one-directional gaze in the discussion. Following the correlation analysis, we investigated the power of the features in predicting the group performance via classification and regression models. The results showed that the Influence Cues and the Group Looking Cues are major predictors of the group performance.

We continued our evaluation with the features extracted from the thin-slices of the meeting. We focused on participants' speaking and looking behavior along with their interruption patterns and conducted classification experiments by using the Influence Model and HMM. Experimental results indicated that the former outperformed the latter in almost all the settings and that only the Influence Model produced classification accuracies satisfying the significance level. In our experiments, we built separate models for audio, video, and visual focus of attention features and used a fusion mechanism to reflect the joint effect of these features on the classification task. We observed that multimodal approach based on the fusion procedure was successful in combining the predictive power of multiple classifiers and provided improvement in the classification performance.

We would like to draw our readers' attention to several interesting findings. In the correlation analysis, it was observed that the groups with lower speaking time-to-meeting duration ratio performed better than those with higher ratios. This may be due to the information overload as referred in [17]. When the time spent on discussing a topic increases, there is a growing chance that the people involved focus too much on the details and deviate from the main subject; which leads to the confusion in making the final decision. The fusion procedure showed us the feasibility of performance enhancement by combining classifiers that cover the different aspects of group interaction. In our opinion, there is still room for performance improvement as long as classifiers linked to non-overlapping characteristics of data are included into the ensemble.

Small group meetings can be classified under four headings: information-oriented, skill-building, problem-solving (decision-making), and brainstorming (creative). In the first two, groups are generally composed of a informant and the audience. The audience learns by listening and by doing in information-oriented and skill-building meetings respectively. In such settings; information flows in one direction, interruptions occur when someone poses a question, and focus is mostly on the informant. Hence, the level of interaction is minimum. That's why, the performance affecting factors we determined become invalid. Besides, using the Influence Model does not bring any additional advantages as the influ-

ential person is always the same, i.e. the informant. However, in the last two types of meetings, each participant is expected to make contributions to meeting targets. Different speaking and looking behaviors can then be observed which results with highly variable forms of interaction. Our approach indicated promising results in such a setting and is applicable to any kind of meetings with similar properties. As the IM allows incorporating many chains, it can be used to model larger meetings with more people without any changes.

There are a number of points that needs to be taken into account about our approach. In the ELEA corpus, each meeting took approximately 15 minutes in which we assumed that the interactions would remain the same. However, for those meetings with longer durations, the person who has influence on another may change. In such a situation, using the Dynamic Influence Model would be more suitable. In order to remove emotional effects, The ELEA groups were formed by different subjects who are total strangers to each other. Influences may be biased in groups involving acquaintances. For example, people may approach or avoid each other depending on friendly attitudes or hostile behavior, which affects the interaction structure. So careful consideration should be taken in forming the discussion groups. Groups with children is another exception in which our current inferences may fail. Since they show different speaking and looking behavior than adults, performance affecting factors and the prediction model need to be reevaluated.

For the future work, we plan to extract fine-grained speaking cues and gaze cues as coarser features showed significant performance in our experiments. We also consider making changes in another important part of our study, i.e. the Influence Model which not only allowed us to extract informative features but also performed well as a classifier. Current version of the IM accounts only for the influence of a person on another. However, joint effect of multiple people on another is a common situation we face in many meetings. As incorporating such effect into the model may increase the overall performance, we may look at the ways to modify the model. On the other hand, the information overload hypothesis that we made for the relationship between the amount of silence and group performance could be tested by performing a semantic analysis of the meeting.

ACKNOWLEDGMENT

This research has been supported by the Swiss National Science Foundation (SNSF) Ambizione fellowship under the SOBE project (PZ00P2_136811).

REFERENCES

[1] A. Haslam. *Psychology in Organizations: The Social Identity Approach*. SAGE, 2004.

[2] A. Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2014.

[3] A. Sapru, and H. Bourlard. Automatic Recognition of Emergent Social Roles in Small Group Interactions. *IEEE Transactions on Multimedia*, 17(5):746–760, 2015.

[4] 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.

[5] B.G. Bell, and N.J. Cooke. Cognitive Ability Correlates of Performance on a Team Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47 (9):1087-1091, 2003.

[6] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi. Automatic prediction of individual performance from "thin slices" of social behavior. In *MM*, pages 733–736, 2009.

[7] B. Raducanu, and D. Gatica-Perez. Inferring Competitive Role Patterns in Reality TV Show Through Nonverbal Analysis. In *Multimedia Tools and Applications*, pages 207–226, 2012.

[8] C.L. Ridgeway. Nonverbal Behavior, Dominance, and the Basis of Status in Task Groups. In *American Sociological Review*, 52(5):683–694, 1987.

[9] D. Frauendorfer, M.S. Mast, D. Sanchez-Cortes, and D. Gatica-Perez. Emergent Power Hierarchies and Group Performance. In *International Journal of Psychology*, 2014.

[10] D.B. Jayagopi, and D. Gatica-Perez. Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models. *IEEE Transactions on Multimedia*, 12(8):790–802, 2010.

[11] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *ICMI*, pages 433–440, 2012.

[12] D. Matsumoto. Culture and nonverbal behavior. In *The Sage Handbook of Nonverbal Communication*, pages 219–235, 2006.

[13] D. Sanchez-Cortes, O. Aran, D.B. Jayagopi, M.S. Mast, and D. Gatica-Perez. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1):39–53, 2013.

[14] D. Sanchez-Cortes, O. Aran, M.S. Mast, and D. Gatica-Perez. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia*, 14(3-2):816–832, 2012.

[15] E. Sundstrom, K.P. deMeuse, and D. Futrell. Work teams: Applications and effectiveness. *American Psychologist*, 45(2):120–133, 1990.

[16] G.A. Neuman, and J. Wright. Team effectiveness: Beyond skills and cognitive ability. *Journal of Applied Psychology*, 84(3):376–389, 1999.

[17] G.M. Spreitzer, D.S. Noble, A.K. Mishra, and W.N. Cooke. Predicting process improvement team performance in an automotive firm: Explicating the roles of trust and empowerment. *Research on managing groups and teams*, 2:71–92, 1999.

[18] G. Varni, G. Volpe, and A. Camurri. A System for Real-Time Multimodal Analysis of Nonverbal Affective Social Interaction in User-Centric Media In *IEEE Transactions on Multimedia*, 12(6):576–590, 2010.

[19] H. Hung, and D. Gatica-Perez. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[20] J.A. Hall, E.J. Coats, and L.S. LeBeau. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychological Bulletin*, 131(6):898-924, 2005.

[21] J.A. Lepine, J.R. Hollenbeck, D.R. Ilgen, and J. Hedlund. Effects of Individual Differences on the Performance of Hierarchical Decision-Making Teams: Much More Than g. *Journal of Applied Psychology*, 82(5):803–811, 1997.

[22] J.B. Shaw and E. Barrett-Power. The effects of diversity on small work group processes and performance. *Human Relations*, 51(10):1307–1325, 1998.

[23] J.E. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[24] J.H. Bradley and F.J. Hebert. The effect of personality type on team performance. *Management Development*, 16(5):337–353, 1997.

[25] L.L. Carl, S.J. LaFleur, and C.C. Loeber. Nonverbal Behavior, Gender, and Influence. *Journal of personality and social psychology*, 68(6):1030–1041, 1995.

[26] M.A.G. Peeters, H.F.J.M. van Tuijl, C.G. Rutte, and I.M.M.J. Reymen. Personality and team performance: a meta-analysis. *European Journal of Personality*, 20(5):377–396, 2006.

[27] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina and V. Murino. Generative Modeling and Classification of Dialogs by a Low-level Turn-taking Feature. *Pattern Recognition*, 44(8):1785–1800, 2011.

[28] M. Higgs, U. Pewina, and J. Ploch. Influence of team composition and task complexity on team performance. *Team Performance Management*, 11(7/8):227–250, 2005.

[29] M.H. Roy. Small group communication and performance: do cognitive flexibility and context matter? *Management Decision*, 39(4):323–330, 2001.

[30] M.J. Zaki and W. Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 12 2014.

[31] M.N. Stolar, M. Lech, and I.S. Burnett. Using the influence model coefficients and the random walk to predict emotional interactions in parent-child conversations. In *ICSPCS*, 2014.
[32] R. Bednarik, S. Eivazi, and H. Vrzakova. A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data. *Eye Gaze in Intelligent User Interfaces*, 24(7):111-134, 2013.
[33] R. van Dick, J. Stellmacher, U. Wagner, G. Lemmer, and P.A. Tissington. Group membership salience and task performance. *Managerial Psychology*, 24(7):609–626, 2009.
[34] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. In *Technical Report 539, MIT Media Laboratory*, 2001.
[35] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *CUES*, 2001.
[36] S. Eivazi, and R. Bednarik. Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In Workshop on Eye Gaze in Intelligent Human Machine Interaction, pages 9–16, 2011.
[37] S. Escalera, X. Baro, J. Vitria, P. Radeva, and B. Raducanu Social network extraction and analysis based on multimodal dyadic interaction. In Sensors, 12(2):pages 1702–1719, 2012.
[38] S.L. Kichuk, and W.H. Wiesner. The big five personality factors and team performance: implications for selecting successful product design teams. *Journal of Engineering and Technology Management*, 14(3-4):195–221, 1997.
[39] S. Luz. Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus Through Analysis of Speech Interaction. In ICMI, pages 575–582, 2013.
[40] T.A. O'Neill, and N.J. Allen. Personality and the prediction of team performance. *Journal of Engineering and Technology Management*, 25(1):31–42, 2011.
[41] U. Avci, and O. Aran. Effect of Nonverbal Behavioral Patterns on the Performance of Small Groups. In *ICMI Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 9–14, 2014.
[42] W. Dong and A. Pentland. Quantifying group problem solving using social signal analysis. In *ICMI-MLMI*, pages 40–43, 2010.
[43] 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 *ICMI*, pages 271–278, 2007.
[44] W. Dong, A. Mani, A.S. Pentland, B. Lepri, and F. Pianesi. Modeling Group Discussion Dynamics. In *IEEE Transactions on Autonomous Mental Development*, 2009.
[45] W. Dong, B. Lepri, F. Pianesi, and A.S. Pentland. Modeling Functional Roles Dynamics in Small Group Interactions. In *IEEE Transactions on Multimedia*, 15(1):83–95, 2013.
[46] W. Dong, B. Lepri, and A. Pentland. Automatic prediction of small group performance in information sharing tasks. In *CoRR*, 2012.
[47] W. Dong, T. Kim, and A. Pentland. A quantitative analysis of the collective creativity in playing 20-questions games. In *C & C*, pages 365–366, 2009.
[48] W. Pan, W. Dong, M. Cebrian, T. Kim, and A.S. Pentland. Modeling Dynamical Influence in Human Interaction. In *IEEE Signal Processing Magazine*, 29(2):77–86, 2012.
[49] Y. Zhaojun, A. Metallinou, and S. Narayanan. Analysis and Predictive Modeling of Body Language Behavior in Dyadic Interactions From Multimodal Interlocutor Cues. In *IEEE Transactions on Multimedia*, 16(6):1766–1778, 2014.

Umut Avci received his PhD degree from University of Trento, Italy in December 2013 and is currently a faculty member at Izmir University of Economics, Turkey. He has also been a visiting researcher at the Idiap Research Institute in Switzerland, working on the analysis of small-group conversational dynamics for performance modeling. His research mainly focuses on machine learning, pattern recognition, and information retrieval.

Oya Aran received her PhD degree in Computer Engineering from Bogazici University, Istanbul, Turkey in 2008. She was awarded a EU FP7 Marie Curie IEF fellowship in 2009 and a Swiss National Science Foundation Ambizione fellowship in 2011. Currently, she is a SNSF Ambizione research fellow at the Idiap Research Institute, working on the multimodal analysis of social behavior in small groups. Her research interests include pattern recognition, computer vision, and social computing. She is a member of the IEEE.