# Predicting Two Facets of Social Verticality in Meetings from Five-Minute Time Slices and Nonverbal Cues

Dinesh Babu Jayagopi[1,2],Sileye Ba[1],Jean-Marc Odobez[1,2] and Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(djaya,sba,odobez,gatica)@idiap.ch

## ABSTRACT

This paper addresses the automatic estimation of two aspects of social verticality (status and dominance) in small-group meetings using nonverbal cues. The correlation of nonverbal behavior with these social constructs have been extensively documented in social psychology, but their value for computational models is, in many cases, still unknown. We present a systematic study of automatically extracted cues - including vocalic, visual activity, and visual attention cues - and investigate their relative effectiveness to predict both the most-dominant person and the high-status project manager from relative short observations. We use five hours of task-oriented meeting data with natural behavior for our experiments. Our work suggests that, although dominance and role-based status are related concepts, they are not equivalent and are thus not equally explained by the same nonverbal cues. Furthermore, the best cues can correctly predict the person with highest dominance or role-based status with an accuracy of 70% approximately.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

## General Terms

Human Factors

## Keywords

Meetings, Social Verticality, Status, Dominance, Audio-Visual feature extraction

## 1. INTRODUCTION

The scientific and technological value of the automatic analysis of social behavior is undeniable. In particular, the understanding in the workplace of fundamental constructs related to power, hierarchy, dominance, and status (the vertical dimension of social interaction [15]) would open doors to tools to support research in social and organizational psychology and for personal self-assessment [23].

In this paper we focus on two aspects of verticality in group interaction, namely dominance and status. Dominance can be defined as "expressive, relationally based communicative acts by which power is exerted and influence achieved" [12] (p. 208), but also as "a personality trait involving the motive to control others, the self-perception of oneself as controlling others, and/or as a behavioral outcome (success in controlling others or their resources)" [15] (p. 898). On the other hand, status can be defined as "an ascribed or achieved quality implying respect or privilege, [but] does not necessarily include the ability to control others or their resources)" [15] (p. 898). In the workplace, status often corresponds to a person's position in a group or in the organization's hierarchy, and it is often defined by a role (e.g. a project manager or a team leader). Dominance and status are related constructs: dominant-personality people often occupy high positions in an organization; conversely, high-status people are often allowed (even expected) to use dominant behavior with their subordinates. At the same time, these two concepts do not always coincide, and can even contradict: for example, a high-status manager could have an intrinsic non-dominant personality, or fail to control or influence his team [15].

Both dominance and status structure nonverbal behaviour in important ways [20, 12, 15]. From a rich amount of work in social psychology and communication, it is known that several vocalic and kinesic cues [12, 20] are related to dominance and status. For instance, both dominant and high-status people are often more vocally and kinesically expressive than their counterparts, and that both types of people often receive more visual attention. Less clear, however, is whether these cues are correlated in similar amounts with the expression and perception of each construct, and whether automatically extracted cues - likely to be imperfect - would be useful for the prediction of both types of social patterns.

This paper addresses two questions. First, can dominance and role-based status in small-group conversations be automatically explained by the same nonverbal cues? While some social psychology literature has found common ground for the nonverbal display and interpretation of both constructs, and recent computational literature has started to investigate models for automatic estimation of dominance [25, 17] or roles [10, 27] in conversations, no attempt has been made to study these two dimensions of social verticality using common data and nonverbal cues together. Second, is it possible to predict these two aspects of verticality from relatively brief observations and using fully automatic nonverbal cues? Although significant evidence in cognitive science support 'thin-slice' explanations for many aspects of social cognition, and such approaches have started to be used with success in computational methods [23], the question remains open for the two concepts we investigate here.

We present a study of the discriminative power for dominance and status prediction of a number of automatic nonverbal cues (extracted from multiple audio and visual sensors) that characterize speaking activity, visual activity, and visual attention. Many of the investigated cues have empirical support in social psychology


*ICMI'08,* October 20–22, 2008, Chania, Crete, Greece.

for either or both status and dominance. Using five hours of five-minute slices of task-based group meetings, our work shows that (1) although dominance and status might be related in terms of the associated nonverbal behavior, they are in practice better explained by different nonverbal cues; and (2) that the best single nonverbal cues can correctly predict the person with highest dominance or role-based status with an accuracy of around 70%.

The paper is organized as follows. Section 2 summarizes related work in computational models. Section 3 details the data and the research tasks. Section 4 describes the nonverbal cues used in our study. Section 5 presents the prediction model. Section 6 presents and discusses the results and some of the involved challenges for future work. Section 7 offers some concluding remarks.

## 2. RELATED WORK

In this section we review the literature on automatic modeling of dominance and role recognition. Broadly, the literature on modeling dominance or influence can be classified into two categories, i.e. dynamic and static models. The dynamic model approach includes the influence model (IM) - an unsupervised Dynamic Bayesian Network (DBN) that models a group as a set of Markov chains, each of which influences the others' state transitions - to determine the degree of influence a person has on the others on a pair-wise basis [4]. Otsuka et al. [22] proposed, following the ideas of [4], to quantify pair-wise influence from automatically estimated vocalic and kinesic mid-level cues (speaking-turn and gaze patterns, respectively), computed in turn with a complex DBN that integrates low-level features.

Rienks et al. [25] studied static models a supervised approach based on Support Vector Machines (SVMs). The addressed task was three-way classification of the participants' dominance level (high, normal, low). Audio-only features derived from manually annotated data were used, and included a collection of nonverbal (e.g. speaker turns, speaking length, floor grabs) and verbal cues (e.g. number of spoken words). Recently, Hung et al. reported estimation of the most dominant person on non-scripted meetings using both audio and video features [17]. Dominance annotation was done on 5 hours of meeting data to systematically understand and evaluate dominance behaviour. In [18], this work was extended to the least dominant person task. An SVM-based approach to fuse both audio and visual cues on the most and the least dominant person task was also reported. The task of predicting the dominant clique, employing a similar approach, was attempted in [19].

The literature on automatic role recognition is relatively limited. Vinciarelli studied the problem of role recognition in multiparty audio recordings of radio bulletins [27]. The six roles included an anchorman among others. Unlike a meeting scenario, the conversations in this case are often dyadic in such setups making the task easier when compared to the role recognition in meetings. The reported performance was of approximately 85 % frame-based classification accuracy on programs of 12-minute average duration each, more than twice the duration we analyze in this work. Another role recognition problem was addressed by Zancanaro et al. [31] and Dong et al. [10]. Instead of organizational roles, the authors targeted the recognition of two types of functional roles in meetings: 'task-based' functional roles, which included Orienteer, Giver, Seeker, Procedural Technician, and Follower; and 'socio-emotional' roles, which included Attacker, Supporter, Protagonist, and Neutral. Each meeting was 25-minute long in average, a much longer temporal support than we address here. In their work, the authors explored the use of SVMs [31] and IM [10]. In both [31, 10], the authors reported 60-70 % frame-based classification accuracy for the two role classification tasks.

## 3. EXPERIMENTAL SETUP

### 3.1 Meeting data

Our objective in this work is to study and model social verticality in task-oriented small groups. We are specifically interested in studying dominance and status defined by roles in short meetings (analogous to the thin slices approach [1]). We chose meetings from the Augmented Multi-party Interaction (AMI) corpus [7]. Each meeting had 4 participants, who were given the task of designing a remote control over a series of meeting sessions. 11 meeting sessions varying from 15 to 35 minutes were divided into 5 minute segments (simply called meetings from here on for convenience) making a total of 59 meetings. This corresponds to 5 hours of meeting data.

Meetings in the AMI corpus were carried out in a multi-sensor meeting room as shown in Figure 1. The room contains a table, slide screen, and white board. A circular microphone array containing eight evenly distributed sources is set in the middle of the table; and another one with four microphones is set in the ceiling. Participants were also asked to wear both headset and lapel omnidirectional microphones, which were attached via long cables to enable freedom of movement around the room. Three cameras were mounted on the sides and back of the room to capture mid-range and global views, respectively, while 4 additional ones mounted on the table captured individual visual activity only. The meeting room is shown in Figure 1. Example screen-shots of the seven camera views are shown in Figure 2.

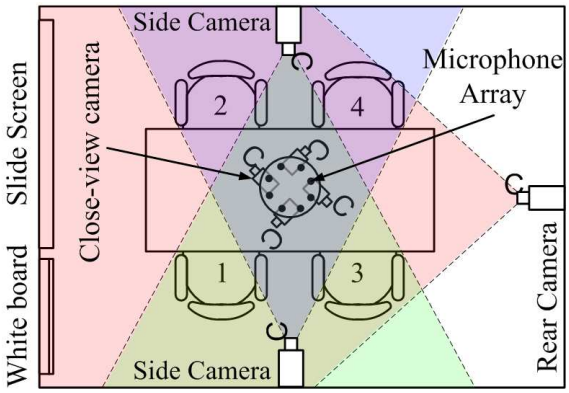



**Figure 1: Plan view of the meeting room set up (from [17]).**

**Figure 2: Examples of the seven camera views in the meeting room. The top row shows the right, centre and left cameras which were used for annotation, while the bottom row shows the view from each of the close up cameras.**

## 3.2 Dominance Task: Predict the most-dominant person

As described in [18] three annotators ranked the participants in every meeting, from highest (1) to lowest (4), according to their level of perceived dominance. 21 participants, 14 men and 7 women, with varying cultural backgrounds were used to annotate the meetings. From the annotations, a significant number of the meeting segments (34) showed full agreement of the most dominant person, i.e. all the annotators agreed on the most dominant participant. We conducted further analysis and found that there were 23 additional meetings where 2 out of 3 annotators agreed on the most dominant person. This subset contains a larger intrinsic variation in the perceived dominance by the annotators. We chose these 57 (34 +23) meetings where there was full or majority agreement for our experiments. It was interesting to observe that only 2 meetings out of 59 had total disagreement on the most dominant person.

## 3.3 Status Task: Predict the Project Manager

In order to study dominance and status together, we use the same 57 meetings for this task. Similar to the most dominant person task, we define the project manager task. As each participant was assigned distinct roles in the AMI corpus: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer', the ground truth is given. In fact, out of the 57 meetings, 37 meetings were such that the Project Manager (PM) was also judged to be the most-dominant person on whom the majority of the annotators agree. This suggests that in many cases (around 65 % of the cases), the project manager also displayed a dominant behaviour.

# 4. NONVERBAL CUES

Various nonverbal behaviours that indicate dominance and status or role have been reported in the literature [6, 11, 12, 15, 20, 24, 26]. We employ speech activity, visual activity and visual focus of attention for predicting the most dominant person and the project manager. More details of the cue extraction techniques are described in the following subsections. We have explicitly chosen not to use language-based cues since we wished to make a detailed study of nonverbal cues, some of which are relatively fast and easy to compute. All the cues are generated at a frame rate of 5 fps for further analysis.

## 4.1 Vocalic cues

Vocalic cues correlated with dominance and high status involve amount of speaking time (or length) [26], speech loudness (or energy), speech tempo, pitch, vocal control, [12], speaker turns and interruptions [21]. It is also reported in the literature that high-status people talk more, speak first or respond quickly in conversations, attempt more interruptions, have a greater fluency, higher speaking rate, and a 'confident' voice tone [24, 20]. Among these, speaking activity as measured by speaking length has shown to be a particularly robust cue to perceived dominance [26].

In this work, we extract a number of vocalic cues as defined below, from the four close-talk microphones attached to each of the participants. Firstly, we extract speaking energy and speaking status.

*Speaking Energy*: The starting point is to compute the real-valued speaker energy for each participant using a sliding window at each time step as described in [32]. Speaking energy was extracted using the root mean square amplitude of the audio signal over a sliding time window for each audio track. A window of 40 ms was used with a 10 ms time shift.

*Speaking status*: From the speaking energy, a binary variable was computed by thresholding the speaker energy values. This indicates the speaking / non-speaking (1/0) status of each participant at each time step.

We then derive various other cues as summarized in the following list. These cues were accumulated over the entire 5-min slices and hence provide a simple way of quantifying their *relative* contribution. All the cues defined can be broadly classified into non-relational and relational cues(dependent on other participants like the interruption based cues).

### *4.1.1 Non-Relational cues*

The non-relational cues defined for a participant $i \in \{1, 2, 3, 4\}$ are the following:

- **Total Speaking Energy (TSE):** Speaking energy accumulated over the entire meeting
- **Total Speaking Length (TSL):** This feature considers the total time that a person speaks [26] according to their binary speaking status.
- **Total Speaker Turns (TST):** We define a turn as a length of a continuous period of time for which the person's speaking status is 'true'. The total number of speaker turns was accumulated over the entire meeting for each participant.
- **Total Speaker Turns without Short Utterances (TSTwoSU) :** This is a variation of the TST feature, computed as the cumulative number of turns that a speaker takes such that the speaker turn duration is longer than one second. The goal is to retain only those turns that are most likely to correspond to 'real' turns, eliminating all short utterances that are likely to be back-channels (like 'yes' or 'no' answers).

### *4.1.2 Relational cues*

The following relational cues are defined for a participant $i \in \{1, 2, 3, 4\}$ with respect to another participant $j \in \{l : l \neq i\}$.

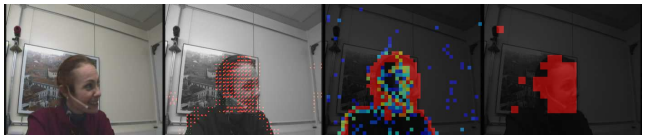
- **Total Successful Interruptions (TSI):** This feature encodes the hypothesis that dominant or high status people interrupt others more often [21]. The feature is defined by the cumulative number of frames that participant $i$ starts talking while another participant $j$ speaks, and $j$ finishes his turn before i does, i.e. only interruptions that are successful are counted.
- **Total number of times being successfully interrupted (TBI) :** The feature is defined by the cumulative number of times that while participant $i$ is talking another participant $j$ starts talking, and $i$ finishes his turn before $j$ does i.e. only successful "being interrupted" events are counted.
- **Total Unsuccessful Interruptions by the speaker (TBC):** This feature encodes the hypothesis that dominant or high status people give more / less backchannels. The feature is defined by the cumulative number of times that participant $i$ starts talking while another participant $j$ speaks, and $i$ finishes his turn before $j$ does, i.e. only interruptions that are unsuccessful are counted.
- **Total number of times being unsuccessfully interrupted (TBBC):** The feature is defined by the cumulative number of times participant $i$ starts talking while another speaker $j$ is speaking and $i$ finishes his turn before $j$ does i.e. only unsuccessful "being interrupted" events are counted.

- **Total number of times speaking first after another speaker (TSF):** This feature encodes the hypothesis that dominant or high status people respond to others first [24, 20]. The feature is defined by the cumulative number of times that participant $i$ speaks first (before other participants by backchannelling or successfully interrupting), after another participant $j$ started talking.

## 4.2 Visual activity cues

Kinesic cues related to dominance and status include body movement, posture, and elevation, and gestures, facial expressions, and eye gaze [15, 12, 24]. Regarding body movement, it has been found that dominant people are often more active. Literature on status describes high-status people as one who claims more space with their bodies and intrude upon their partners noticably.

In this work, we extract and employ visual activity based cues to approximate some of the cues described above. Visual activity cues were extracted efficiently in the compressed domain, leveraging the fact that meeting videos are already in compressed form [30]. Close-up view camera video data, one for each participant was used (as shown in Figure 2).



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

**Figure 3: Compressed domain video feature extraction. (a) Original image. (b) Motion vectors. (c) Residual coding bitrate. (d) skin-colored regions. (from [17])**

Each video is compressed by a MPEG-4 encoder with a group-of-picture (GOP) size of 250 frames and a GOP structure of I-P-P-..., where the first frame in the GOP is Intra-coded (I frame), and the rest of the frames are predicted frames (P frames). Compressed-domain information such as motion vectors and block DCT coefficients that are accessible at almost zero computational cost from compressed video [28] was thresholded to obtain visual activity (as illustrated in Figure 3). Only skin colored regions were considered for the computation, by implementing a block-level skin-color detector working entirely in the compressed domain which can detect head and hand regions. The chrominance discrete cosine transform coefficients in the I frames were applied to a skin-color detector [8]. The position of these skin-colored blocks are then estimated and propagated for the subsequent P frames for the duration of the GOP structure using the motion vector information.

For each frame where the participant is visible in the close-up view, the average motion vector magnitude or residual coding bitrate over all the estimated skin blocks is calculated and used as a measure of individual visual activity. The two quantities differ in the information that they capture. While the motion vectors capture the rigid body motion like translation, bitrate attempts to capture the non-rigid motion. To meaningfully compare motion vector magnitudes and residual coding bitrate, we need to normalize the quantities [30]. For those frames where the participant is not detected in a frame of the close-up view, he or she is assumed to be presenting at the projection screen, and so is assumed to be visually active.

In an analogous fashion to the audio cues, we define a number of visual activity cues from the raw motion values:

- **Visual Activity**: A binary variable computed as explained in the previous paragraph, that indicates whether a participant is visually active at each time step. Three variations were tested, based on Motion Vectors (M), Residual Coding Bitrate (B), and the combination (C), i.e. the average of both features.

- **Total Visual Activity Length (TVL), Total Visual Activity Turns (TVT), Total Visual Activity Interruptions (TVI), Visual Activity**. All these cues are the visual counterpart of the audio cues defined previously and are used to test the hypothesis of whether visual activity can be treated as vocalic activity in terms of turns.

## 4.3 Visual attention cues

Visual attention based cues have been studied extensively in the social psychology literature. Early research by Efran showed that high-status persons receive more visual attention than low-status people [13]. Cook et al. showed that people who very rarely look at others in conversations are perceived as weak [9]. Further studies have shown that the joint occurrence of visual attention and speaking activity patterns are correlated with social verticality concepts. For instance, Exline et al. showed that high-power people exhibit a relatively high ratio of looking-while-speaking to looking-while-listening periods [14, 11].

In this work, head pose is used to infer visual attention. An extention of the recent work by Ba and Odobez [3, 2] to estimate the joint focus state of all participants is used for this work. Visual attention is estimated using a Dynamic Bayesian Network (DBN), by modeling the relationship between people's visual attention, their head pose, their speaking status, and other contextual cues related to the group activity. These contexual cues include slide-screen activity and conversational events like silence or monologue or dialogue or discussion. Head pose was estimated by jointly tracking the head and head-pose using side-view cameras (as illustrated in Figure 4). There were seven visual attention targets i.e. the four participants, the slide-screen, the white-board and the table and one unfocussed label. The accuracy of the Visual Focus Of Attention (VFOA) estimation was around $52\%$. Unlike [22], as the AMI meetings had objects that distract the visual attention of participants like the slide-screen, the white-board, and the table, the task of VFOA estimation was more difficult. The seating arrangement was also not circular, rather it was rectangular with 2 people facing each other, making the VFOA estimation of certain seat positions (seats numbered 3 and 4 in Figure 1) more difficult than the others.

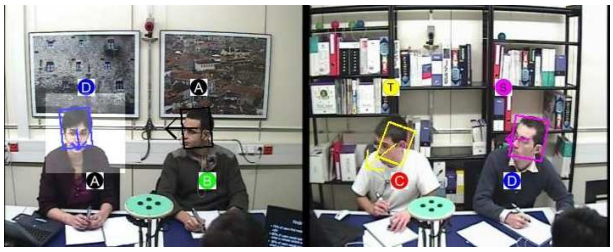


**Figure 4: Estimated visual focus of participants using side-view camera views. Each of the participants is labeled and their focus of attention is displayed above their head. Head location and head pose are also displayed. The white transparent box placed on participant A shows that her speaking status is 'true'.**

From the visual attention of individual participants, along with the speech activity cues, we experimented a number of features that capture the gazing behaviour of participants as follows:

#### 4.3.1 Overall attention cues

- **Total Received Visual Attention (TRVA):** This feature encodes the hypothesis that dominant or high status people are looked at longer [13]. The feature is defined by the cumulative number of frames that a participant $i$ is looked at by the other participants.
- **Total Looking-At-Others Length (TLOL):** This follows the hypothesis that dominant or high status people look at others longer. The feature is defined by the cumulative number of frames that a participant $i$ looks at other participants.
- **Total Looking-At-Others Turns (TLOT):** This follows the hypothesis that dominant or high status people look at others more often, by inverting the hypothesis of Cook et al, that weak people rarely look at others [9]. The feature is defined by the cumulative number of times a participant $i$ looks at other participants.

#### 4.3.2 While-Speaking attention cues

- **TRVA** while the participants speak.
- **TLOL** while the participants speak [14].
- **TLOT** while the participants speak.

#### 4.3.3 While-not-Speaking attention cues

- **TRVA**, **TLOL**, **TLOT** while the participants are silent.

#### 4.3.4 Visual Dominance Ratio

The Visual Dominance Ratio (VDR) was defined in [11] as the ratio between the total looking-while-speaking to the total looking-while-listening periods for dyadic pairs. We generalize it to multi-party conversations, by approximating 'looking while listening' as 'looking while someone else is speaking' and 'looking while not speaking' and hence define the following two ratios. The new ratios are called Multi-Party Visual Dominance Ratios (MVDR) [16].

- $MVDR_1$**:** Defined as the following ratio

$$MVDR_1 = \frac{TLO - while - speaking}{TLO - while - someone - else - speaks} \quad (1)$$

- $MVDR_2$**:** Defined as the following ratio

$$MVDR_2 = \frac{TLO - while - speaking}{TLO - while - not - speaking} \quad (2)$$

# 5. PREDICTION

Predicting the most-dominant or the project manager and its evaluation are done as follows. Firstly, the vocalic cues, visual activity cues, and visual attention cues are accumulated over the duration of the meeting (as explained in Section 4). Then, depending on whether the relation of the feature to the task is assumed to be direct or inverse, either the largest or smallest accumulated value of each feature is taken. It is to be noted that unless specified otherwise, the largest value is chosen and whenever the smallest value is chosen, '(min)' appears next to the feature name like TBI(min). That is, we hypothesize that someone is likely to be more dominant if they speak, move, look, or grab the floor the most out of all the participants in the meeting. We evaluate the method by comparing the predicted person with that of the ground truth for both tasks, and computing the classification accuracy as percentages. It is important to note that we predict outcomes for full meetings, rather than for individuals or frames [25, 27]. For the dominance task, when there is full agreement on the most dominant person, the accuracy is computed as normal. When there is majority agreement, a weighting scheme is used to compute the accuracy in order to accomodate the judgements of all the three annotators. Let $N$ denote the total number of meetings, and $A_i$ and $B_i$ be the most-dominant-person ground-truth labels corresponding to the 'most-voted' (two votes) and 'least-voted' (one vote) cases, respectively, for meeting $i$, $1 \leq i \leq N$. Furthermore, let $n$ be the number of times the automatically predicted most dominant person is $A_i$, and $m$ be the number of times the predicted most dominant person is $B_i$. We compute the classification accuracy as $(2/3 * n + 1/3 * m)/N$. We have also experimented with other evaluation methods in our previous work on the same dataset [18]. The maximum achievable performance is less than 100%. In our case it is of 86.5%. It is important to note that the dominance models considered are unsupervised and therefore do not involve any training.

# 6. RESULTS

We conducted experiments using the vocalic cues (see Section 6.1), visual activity based cues (see Section 6.2) and visual attention based cues(see Section 6.3) on the two tasks - most-dominant person and the project manager. In the tables in this section, the column titled MD gives the classification performance in percentages, for the most dominant person on the 57 meetings set. The classification performance for the project manager task is shown in the column titled PM. It is important to note that, though the tasks are independent, the ground truth for both tasks have overlaps i.e. 65% of the project managers are also the most dominant. We also report the results on the overlapping and non-overlapping subsets of meetings, corresponding to the columns titled PM $=$ MD (37 meetings) and PM $\neq$ MD (20 meetings). The results on the subsets helps us understand how specialized these features are for each of the tasks.

## 6.1 Vocalic cues

Table 1 shows the results obtained using vocalic cues. The results are separated into non-relational and relational features. For the most-dominant person task, the total speaking length (TSL) and total number of speaker turns removing short turns (TSTwoBC) were most effective in classifying the most dominant person with a classification accuracy of around 70%. Social psychology literature [26] supports the results that speaking time is a very strong cue for dominance perception by humans. It is to be noted that the same cues predict the most dominant person on a cleaner dataset, with full-agreement on the most-dominant person with an accuracy of 85% [17]. The total speaking energy (TSE) also performed well. For the project manager task, the total number of speaker turns (TST) and the total number of times speaking first after a speaker (TSF) were the best indicators, with a classification accuracy of 66.7%. Also, it is interesting to observe that including the short utterances (of duration around 1 sec) is useful to predict the project manager and not the most-dominant person. For PM $\neq$ MD case, TSL and TSE totally failed as a predictor of the status. This highlights some of the differences between dominance and status.

Regarding the relational cues, the successful interruption cue performed significantly better than random. Also, the hypothesis that dominant or high status people get less interrupted by others did not hold good. Rather it was observed that dominant or high status people get more interrupted by others. This might be due to the fact that less dominant people talk less and hence get inter-

rupted less too. Also in absolute terms, the 'not dominant' project managers were being interrupted less often (as they speak less), as shown by TBI(min) cue. It is important to notice that in the AMI data, groups were gathered with volunteers, and each person was randomly assigned a role. So it might be the case that the people assigned the PM manager does not have a naturally dominant personality.

Figure 5 shows the histogram of speaking length for both the most-dominant task and the project manager task. We observe that TSL is more discriminant for the dominance task. Similarly, Figure 6 shows the histogram of TSF. It is interesting to observe the difference between the histograms of the project manager and the others, showing that the manager responds first more often than the others, as he has the role of anchoring the meeting. This can be seen from the mean of the TSF feature for the project manager being higher than that of the others.

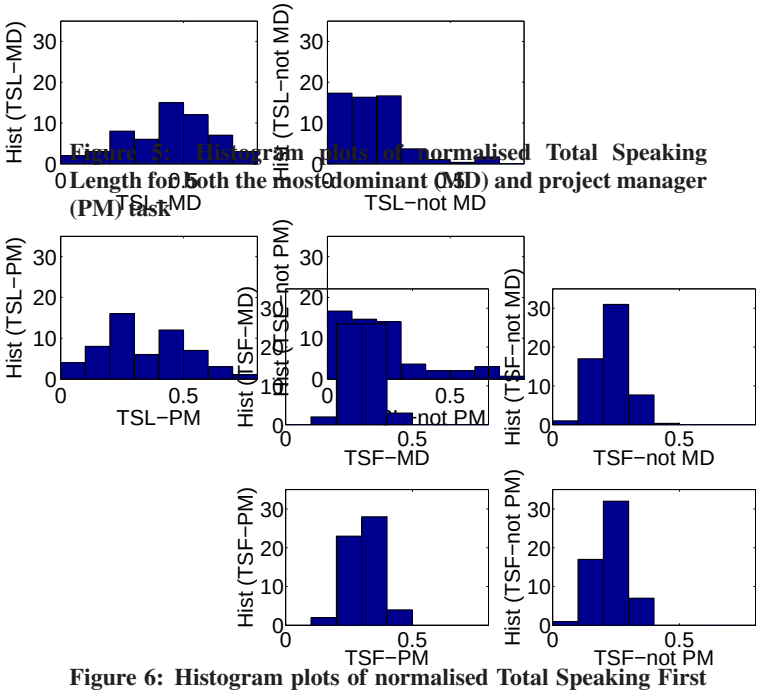



**Figure 5: Histogram plots of normalised Total Speaking Length for both the most-dominant (MD) and project manager (PM) task**

**Figure 6: Histogram plots of normalised Total Speaking First after another participant (TSF) for both the most-dominant (MD) and project manager (PM) task**

## 6.2 Motion cues

Table 2 shows the results obtained with visual activity cues. We experimented with the three options, the motion vector, the residual bitrate, and their combination, to compute visual activity and three types of features: the total visual activity length (TVL), the total motion turns (TVT) and the total motion interruptions (TVI).

For the MD task, the total visual activity length (TVL) that quantifies how much people move, and total motion turns (TVT) that quantifies how often people move (removing the very short turns that we assume to be noise), performed relatively well, with a classification accuracy of $62.6\%$ percent. The social psychology literature supports the value of similar features [6]. All the three options - motion vector, residual bitrate and their combination performed similarly. Compared to the speaking length, the motion length was $12\%$ worser for the MD task. But for the PM task, the difference was not much. For the meetings where PM $\neq$ MD, the TVL cues were much better than TSL. The total visual activity turns (TVT), both bitrate and combo, has some ability at predicting the project manager, similar to their vocalic counterparts, the total speaker turns (TST) cues (a classification accuracy of $52.6\%$). .

| Features | MD (57) | PM = MD (37) | PM ≠ MD (20) | PM (57) |
|---|---|---|---|---|
| **Non-relational features** | | | | |
| TSL | **70.8** | 75.7 | 0 | 49.1 |
| TSE | 67.3 | 70.3 | 0 | 45.6 |
| TST | 52.0 | 73.0 | **45.0** | **63.2** |
| TSTwoSU | **70.2** | **78.4** | 10 | 54.4 |
| **Relational features** | | | | |
| TSI | 51.5 | 56.8 | 30.0 | 47.4 |
| TBI(min) | 12.9 | 13.5 | **50.0** | 26.3 |
| TBI | 48.0 | 51.4 | 10.0 | 36.8 |
| TUI | 26.9 | 35.1 | 25.0 | 31.6 |
| TUI(min) | 27.5 | 24.3 | 5.0 | 17.5 |
| TBUI | **59.1** | 56.8 | 0 | 36.8 |
| TSF | 50.3 | **75.7** | 50.0 | **66.7** |

**Table 1: Prediction accuracy (%) of vocalic cues for predicting the most-dominant person and the project manager.**

| Features | MD (57) | PM = MD (37) | PM ≠ MD (20) | PM (57) |
|---|---|---|---|---|
| **Non-relational features** | | | | |
| TVL(M) | 59.6 | 59.5 | 30.0 | 49.1 |
| TVL(B) | **62.6** | 62.2 | 15.0 | 45.6 |
| TVL(C) | 61.4 | 62.2 | 25.0 | 49.1 |
| TVT(M) | 59.1 | 59.5 | 25.0 | 47.4 |
| TVT(B) | **62.6** | 70.3 | 20.0 | **52.6** |
| TVT(C) | 61.4 | 70.3 | 20.0 | **52.6** |
| **Relational features** | | | | |
| TVI(M) | 46.2 | 54.1 | **40.0** | 49.1 |
| TVI(B) | 49.7 | 59.5 | 25.0 | 47.4 |
| TVI(C) | 49.1 | 64.9 | 30.0 | **52.6** |

**Table 2: Prediction accuracy (%) of visual activity cues for predicting the most-dominant person and the project manager.**

## 6.3 Visual Attention cues

Table 3 shows the results obtained with visual attention cues. We systematically explored being-looked-at (passive) and looking-at (active) cues, as single events as well as jointly with speech activity and silence.

The hypothesis that dominant or high status people are looked at longer [13] was verified as the TRVA (Overall) performed significantly better than chance. TRVA while not speaking (glancing while someone else speaks), seems to carry more information about both dominance and status than TRVA while speaking. The hypothesis that dominant or high status people look at others more often was also verified with the TLOT cue [9]. Also, 'looking at others while speaking' correlates with both tasks, as seen by the TLO (while speaking) cue. The 'looking at others while not speaking', correlates negatively (using the min option) with both tasks,

as seen by the TLO feature while not speaking. The best performing cues were the MVDR ratios for the dominance task (67.3%) and the 'looking at others while speaking' turns (TLOT) for the Project Manager task (59.6%). The second fact suggested that in out data the project manager frequently observes at his team members, while he is speaking. The visual attention based features were slightly better than the motion features for the dominance task.

| Features | MD (57) | PM = MD (37) | PM ≠ MD (20) | PM (57) |
|---|---|---|---|---|
| **Overall attention cues** | | | | |
| TRVA | **58.5** | 62.2 | 15.0 | 45.6 |
| TLO | 24.0 | 24.3 | 20.0 | 22.8 |
| TLOT | 45.0 | 62.2 | 30.0 | **50.9** |
| **While-Speaking attention cues** | | | | |
| TRVA | 24.0 | 27.0 | 20.0 | 24.6 |
| TLO | **59.6** | 67.6 | 15.0 | 49.1 |
| TLOT | 55.6 | **73.0** | **35.0** | **59.6** |
| **While-not-Speaking attention cues** | | | | |
| TRVA | **60.2** | 64.9 | 15.0 | 47.4 |
| TLO(min) | 47.4 | 48.6 | 25.0 | 40.4 |
| TLOT | 38 | 59.5 | **35.0** | **50.9** |
| **MVDR** | | | | |
| $MVDR_1$ | **66.7** | **73** | 10.0 | 50.9 |
| $MVDR_2$ | **67.3** | **75.7** | 10.0 | **52.6** |

**Table 3: Prediction accuracy (%) of visual attention cues for predicting the most-dominant person and the project manager. Note: all the visual attention based cues are *relational*.**

## 6.4 Centrality measures

The Social Network Analysis literature has studied interaction among people in social environments [29]. Various network centrality measures exist for different relationships. Wasserman et al. [29] discuss measures in which the centrality or status of positions are recursively related to the centrality or status of the positions to which they are connected.

Such measures of centrality can be readily applied where relational data exists. We applied two such measures on some of the relational features. We use an eigenvector-like measure based centrality [5], which we refer to as $Centrality^1$, and another measure of centrality as defined below, called $Centrality_i^2$ **:**

$$Centrality_i^2 = \frac{K-1}{\sum_{j=1}^{K} d_{ij}}, \forall i = 1, 2, 3, ..K \qquad (3)$$

where K is the number of participants (the number of nodes in the social network), and $d_{ij}$ is the distance between nodes $i$ and $j$. Maximizing $Centrality^2$ is equal to minimizing $\sum_{j=1}^{K} d_{ij}$.

We investigated whether centrality measures could be used to predict status or dominance, using it on two representative relational data (arranged as a matrix):

The two relational data matrix considered are defined as follows:

- **Total 'number of times speaking first after another speaker' matrix (TSF matrix) :** Each matrix element $a_{ij}$ is defined by the cumulative number of times that a participant $i$ speaks first (before other participants), after another participant $j$ started talking.
- **Total 'number of frames looking at others' matrix (VFOA matrix) :** The matrix element $a_{ij}$ is defined by the cumulative number of times that a participant $i$ looks at $j$.

We approximate $d_{ij}$ as $a_{ij}^{-1}$, which means that the larger the interaction between people the smaller the distance between them.

| Features | MD (57) | PM = MD (37) | PM ≠ MD (20) | PM (57) |
|---|---|---|---|---|
| $Centrality^1$ | | | | |
| using TSF matrix | 49.7 | 70.3 | 40.0 | **59.6** |
| using VFOA matrix | **56.1** | 64.7 | 20.0 | 49.1 |
| $Centrality^2$ | | | | |
| using TSF matrix | 50.3 | 75.7 | **55.0** | **68.4** |
| using VFOA matrix | 48.5 | 56.8 | 30.0 | 47.4 |

**Table 4: Prediction accuracy (%) of centrality measures for predicting the most-dominant person and the Project Manager.**

In Table 4, we observe that the most central person, as predicted using both the measures, has significant correlation with the most-dominant person and the project manager. The $Centrality^2$ measure using the TSF matrix, predicts the manager with an accuracy of 68.4%, which makes it the best performing feature for the project manager task.

# 7. DISCUSSION

Our study appears to verify several of the hypotheses related to the nonverbal cues, for both the dominance and the status tasks. Overall, the vocalic cues performed slightly better than the visual cues. The best cues for both the tasks were vocalic. Intutively this makes sense, as speech is the principal modality of communication. Also we make use of head-set microphones for the experiments, which is much less "noisier" as compared to the visual modalities. Total Speaking Length is the best nonverbal cue to predict the most dominant person. The hypothesis that high-status people respond first (by back-channeling or attempting to grab the floor) seems to be supported. Dominant or high-status people are active, as verified by the motion length and motion turns. Finally, received visual attention, looking at others while speaking, and the visual dominance ratios also appear to indicate status and dominance.

# 8. CONCLUSION

In this paper we investigate the problem of predicting the most-dominant person and the project manager. Such problems are challenging and are beginning to be investigated. We employed automatic, nonverbal activity cues for doing the prediction in a static framework. At the level of human perception, we found that 65% of the time a project manager was also perceived as the most dominant. This was also revealed in the results as some of the nonverbal cues had comparable classification accuracies for both the tasks. It was interesting to observe that certain cues reveal the dominance behaviour aspect better, whereas certain others capture the status better. Though the audio modality was the best, the visual attention based and the visual activity based cues are promising. The study shows that some of the most difficult cases are when high-status people did not showed dominant behavior through the measured nonverbal cues. Predicting in these cases is a very interesting open issue. Centrality measures, used in social network analysis, also correlate well with both tasks. In the future, we would like to explore the possibility of fusing cues, to exploit any complementary information that these single features could carry. We would also like to expand our set of cues to other easily extractable and correlated to the social verticality, for example speaking rate, pitch, etc.

**Acknowledgments:** This research was partly supported by the US VACE program, the EU project AMIDA, the Swiss NCCR IM2. We thank Hayley Hung (IDIAP) for discussions.

# 9. REFERENCES


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

[2] S. Ba and J.-M. Odobez. Multi-person visual focus of attention from head pose and meeting contextual cues. In *IDIAP-RR 08-47*, 2008.

[3] S. O. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proc. of ICASSP*, 2008.

[4] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *Proc. IEEE Workshop CVPR-CUES*, Kauai, Dec. 2001.

[5] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.

[6] J. K. Burgoon and N. E. Dunbar. Nonverbal expressions of dominance and power in human relationships. In V. Manusov and M. Patterson, editors, *The Sage Handbook of Nonverbal Communication*. Sage, 2006.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *Proc. MLMI Workshop*, Edinburgh,UK, Jul. 2005.

[8] D. Chai and K.N Ngan. Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):551–564, 1999.

[9] M. Cook and J.M Smith. The role of gaze in impression formation. *Br J Soc Clin Psychol*, 14(1):19–25, 1975.

[10] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proc. ICMI*, pages 271–278, New York, NY, USA, 2007. ACM.

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

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

[13] J.S Efran. Looking for approval: effects on visual behavior of approbation from persons differing in importance. *J Pers Soc Psychol*, 10(1):21–5, 1968.

[14] R.V. Exline, S.L. Ellyson, and B. Long. Visual Behavior as an Aspect of Power Role Relationships1. *Nonverbal Communication of Aggression: Proceedings of the Fourth Annual Symposium on Communication and Affect Held at Erindale College, University of Toronto, March 28-30, 1974*, 1975.

[15] 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.

[16] H. Hung, D. Jayagopi, S. Ba, J-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proc. ICMI*, Chania, Greece, Oct. 2008.

[17] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proc. ACM MM*, Augsburg, Sep. 2007.

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

[19] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Predicting the dominant clique in meetings through fusion of nonverbal cues. In *Proc. ACM MM*, Vancouver, Oct. 2008.

[20] A. Leffler, D.L. Gillespie, and J.C. Conaty. The effects of status differentiation on nonverbal behavior. *Social Psychology Quarterly*, 45(3):151–161, 1982.

[21] L. Smith-Lovin and C. Brody. Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review*, 54(3):424–435, Jun. 1989.

[22] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proc. ACM CHI Extended Abstract*, Montreal, Apr. 2006.

[23] A. Pentland. Socially aware computation and communication. *IEEE Computer*, pages 63–70, Mar. 2005.

[24] C.L. Ridgeway. Nonverbal behavior, dominance, and the basis of status in task groups. *American Sociological Review*, 52(5):683–694, 1987.

[25] R.J. Rienks and D. Heylen. Automatic dominance detection in meetings using easily detectable features. In *Proc. MLMI Workshop*, Edinburgh, Jul. 2005.

[26] M. Schmid Mast. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research*, 28(3):420–450, Jul. 2002.

[27] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(6), 2007.

[28] H. Wang, A. Divakaran, A. Vetro, S.F. Chang, and H. Sun. Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, 14(2):150–183, 2003.

[29] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[30] C. Yeo and K. Ramchandran. Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley, Jun 2008.

[31] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *in Proc. ICMI*, Banff, Nov. 2006.

[32] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered hmms. In *IEEE Transactions on Multimedia*, volume 8, pages 509–520, June 2006.