Modeling Dominance in Group Conversations using Nonverbal Activity Cues

Dinesh Babu Jayagopi[1,3], Hayley Hung[1], Chuohao Yeo[2], Daniel Gatica-Perez[1,3] [1]Idiap Research Institute, Martigny, Switzerland
[2]Department of Computer Science, University of California, Berkeley
[3]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
{djaya, hhung, gatica}@idiap.ch , zuohao@EECS.Berkeley.edu
phone: +41 27 7217 781, fax: +41 27 7217 712

***Abstract*— Dominance - a behavioral expression of power - is a fundamental mechanism of social interaction, expressed and perceived in conversations through spoken words and audio-visual nonverbal cues. The automatic modeling of dominance patterns from sensor data represents a relevant problem in social computing. In this paper, we present a systematic study on dominance modeling in group meetings from fully automatic nonverbal activity cues, in a multi-camera, multi-microphone setting. We investigate efficient audio and visual activity cues for the characterization of dominant behavior, analyzing single and joint modalities. Unsupervised and supervised approaches for dominance modeling are also investigated. Activity cues and models are objectively evaluated on a set of dominance-related classification tasks, derived from an analysis of the variability of human judgment of perceived dominance in group discussions. Our investigation highlights the power of relatively simple yet efficient approaches and the challenges of audio-visual integration. This constitutes the most detailed study on automatic dominance modeling in meetings to date.**

***Index Terms*— Group Meetings, dominance modeling, nonverbal communication, audio-visual activity cues**

I. Introduction

Certain people are consistently successful at dominating conversations and their results. In fact, within a few minutes of interaction among unacquainted individuals, a dominance order or a participation hierarchy often emerges [27]. A concept largely studied in social psychology, dominance is one of the basic mechanisms of social interaction and has fundamental implications for communication both among individuals and within organizations [4]. While dominant behavior could bring benefits to the person displaying it in certain contexts, in others it could negatively affect the social dynamics of a group, impacting its cohesiveness and effectiveness, and eroding social relationships.

The automatic modeling of dominance patterns in groups is a key problem in the emerging domain of social interaction analysis from sensor data [14], [23], which spans research in audio and visual processing, information fusion, human-computer interaction, and ubiquitous computing. The analysis of face-to-face multiparty conversations to extract patterns of turn-taking [6], [7], [20], addressing [18], interest and attraction [15], [24], [30], functional roles [32], or dominance [2], [26] is challenging, given the complex nature of real communication, and the difficulty to model, accurately and efficiently, the behavior of multiple interacting individuals. Automatic dominance estimators from audio-visual media could be part of relevant human-centered applications including self-assessment, training, and educational tools [23], and systems to support group collaboration [10], [19].

A solid body of work in psychology has documented the multi-modal nature of dominance [12], and in particular of the role that nonverbal communicative cues (not involving the spoken words) play in the expression and perception of dominant behavior. Although speech is the main modality in conversations [9], [28] , substantial information is conveyed in the visual modality through body movement, postures, and gestures. It is known that, in terms of *vocalic* and *kinesic* cues, dominant individuals behave more actively (i.e., talk and move more, more often, and with larger ranges) than non-dominant people [4], [12]. Some of these activity cues can be automatically extracted from data, and initial work [2], [25], [26] has mainly investigated perceptual modalities in isolation (where cues were often extracted manually),

This work was supported by the Swiss National Science Foundation through the National Center for Competence in Research on Interactive Multimodal Information Management (IM2), the European Project Augmented Multi-Party Interaction with Distant Access (AMIDA), the US Video Analysis and Content Extraction (VACE) program, and Singapore's Agency for Science, Technology, and Research (A*STAR).

or proposed dominance recognition approaches that were applied to relatively constrained interaction scenarios or that were limited in their validation.

This paper presents a systematic study on fully automated dominance modeling in small group meetings from nonverbal activity cues. Focusing on a common data set of face-to-face interactions recorded with multiple cameras and microphones, our work contains several contributions. First, we investigate a number of easily extracted and efficient audio and visual activity cues for the characterization of dominant behavior. Our cues include a novel set of visual features extracted in compressed-domain video. We consider audio-only, visual-only and audio-visual cases to understand the relative power of each of the modalities and the benefits of using them jointly. Second, we study unsupervised and supervised approaches for dominance modeling, which differ in complexity and needs for training data. Third, through the analysis of the variability of human judgment of perceived dominance in our corpus, we define and study a set of dominance estimation tasks (most-dominant person, least-dominant person) that allow us to objectively quantify the difficulty of each task, as well as the variation in performance as human performance itself varies. Our results highlight a number of relevant issues, including the robustness of basic audio features, the power of some visual activity cues, and the overall advantages of simple approaches. Our best methods are able to achieve 91.2% (resp. 83.9%) accuracy for the classification of the most (resp. least) dominant person in a meeting. To our knowledge, this work constitutes the most detailed study on automatic modeling of dominance in small group meetings from audio and visual activity cues to date.

The paper is organized as follows. Section II reviews the literature on dominance in social psychology and on computational approaches related to our work. Section III presents the components of our work. Section IV describes the data, its annotation process, and the definition of the dominance classification tasks. Section V presents the audio and visual cues. Section VI presents our models for estimating dominance and describes the experimental protocol. Sections VII and VIII present and discuss the results for the studied dominance classification tasks. Section IX summarizes the finding of our work and provides some concluding remarks.

II. Related Work

In the next subsections, we summarize the most relevant work in social psychology and social computing related to our own.

A. Dominance in social psychology

Dominance is a fundamental construct in social interaction [4]. In social psychology, dominance is often seen in two ways, “as a personality characteristic (trait) and to indicate a person’s hierarchical position within a group (state)” [28] (pp. 421). Although dominance and closely related terms like power, status, and influence have multiple definitions and are often used as equivalent, many social psychologists advocate for a clearer distinction, power being ”the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person” [13] (pp. 208), and dominance being a set of “expressive, relationally based communicative acts by which power is exerted and influence achieved”, “one behavioral manifestation of the relational construct of power”, and “necessarily manifest” [13] (pp. 208-209).

The study of dominance has spanned several decades of work in psychology and is obviously too large to review here (for recent accounts, see [4], [13]). However, two main threads of work are key to the development of automated dominance modeling approaches, as both justification and inspiration: the existence of specific social cues used by people to express dominance in conversations, and the ability to correctly infer or perceive dominance by observers of an interaction using such cues.

The first aspect is rich, and has been widely studied. Both verbal and nonverbal cues are indicators of dominance. Being the primary interest of our work, we focus on nonverbal cues, which are known to be effective in predicting behavioral outcomes. Directly related to our work, nonverbal cue categories of interest include vocalic and kinesic [13]. Vocalic cues involve amount of speaking time (or length) [28], speech loudness (or energy), speech tempo, pitch, vocal control, [13], and interruptions [3]. Among these, *speaking activity* as measured by speaking length has shown to be a particularly robust cue to predict dominance [28]. Kinesic cues include body movement, posture, and elevation, and gestures, facial expressions, and eye gaze [13]. In particular, it has been found that, regarding body movement, dominant people are normally more active than non-dominant people (the former move more and with a wider range of motion, the latter tend to be more limited in their amount and range of body activity), and that gestures that accompany speech are positively correlated with dominance [4], [12]. This suggests that *visual activity* (and in particular, activity that correlates with speaking activity) are strong cues for predicting dominance. It should be clear that, although some of the above cues could be measured from audio and visual sensor data

with existing automatic techniques, their corresponding performance and computational complexity vary rather widely. In our work, we focus on features that are easily extracted and computationally efficient.

The second aspect is also crucial: the fact that people can correctly decode dominance (whether as participants of an interaction or as external observers) provides support for both the expectation of producing reliable human annotations and the hope of designing methods for automatic analysis. The literature here is also rich. Twenty-five years ago, Dovidio et al. showed that people can systematically decode patterns of visual dominance displayed by others [11]. It has been also found that participants and external observers present differences in their perception of dominance [13]. For automatic approaches, this is important for manual data annotation (first-party vs. third-party) in order to generate ground-truth for training purposes. As Dunbar and Burgoon state: "Perhaps coders' perception of dominance correspond more closely with objective measures of verbal and nonverbal dominance than those of participants themselves... However, the coders' observations are limited to the behaviors in a particular interaction, whereas participants are privy to the ongoing interaction that is part of a continuing relationship. Thus, as with many other findings, whose perception you trust depends on what question is being asked." [13] (pp. 228). We believe the third-party option to be an adequate approach for the questions addressed in this paper.

B. Dominance in social computing

Previous research on automatic dominance modeling can be categorized based on the specific group interaction setting, the addressed task, and the technical implementation, including both cues and dominance models. All of the works discussed below studied small groups recorded with multiple cameras and microphones.

For a debating game setting, Basu et al. [2] used 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. Both vocalic cues (manually labeled speaker turns and automatically extracted speaker energy and voicing information) and kinesic cues (region-based motion energy derived from pre-defined regions and skin-color blobs) were used. While promising results were presented, this work neither studied the impact of individual features nor systematically evaluated the performance of the resulting system.

On a small set of meetings from the M4 (MultiModal Meeting Manager) and AMI (Augmented Multi-party Interaction) corpora, Rienks et al. [26] studied 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 combination of nonverbal (e.g. speaker turns, speaking length, floor grabs) and verbal cues (e.g. number of spoken words). However, no study of the annotation quality was conducted, and so a clear understanding or the sources of complexity of the data was missing. Furthermore, labeling the data with a predefined number of dominance levels is, to some extent, arbitrary, and a study of the effect of these choices on the obtained was not done. Rienks et al. [25] extended this approach to a subset of the AMI corpus where the dominance judgements came from the participants themselves.

Finally, Otsuka et al. [22] proposed, following the ideas of [2], 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. While the proposed influence model is simple, and the proposed features are conceptually appealing, neither an objective evaluation nor a comparison to previous approaches were conducted in this work. Our work substantially extends previous research in several ways. First, unlike [2], [22] , we conduct a systematic study of both vocalic and kinesic features and dominance models on a common data set, and present a detailed objective evaluation of the performance of single- and multi-modal cues, and of unsupervised and supervised learning approaches. Second, the specific research tasks we study are distinct, and so complementary, to the ones studied in all previous work. Third, unlike [25], [26] we introduce a set of novel visual activity cues, distinct from those in [2], [22] and computed in the compressed domain with low computational cost. Fourth, unlike [2], [25], [26] , we rely on fully automatically extracted features, and in this sense the presented work is closer to 'what is achievable using computers'. Finally, unlike all previous work, we analyze the annotation of perceived dominance by human judges and are thus able to analyze the implications that the variation of human perception has on the performance of our automatic approaches. A preliminary version, discussing a small part of our work presented here, was reported in [17].

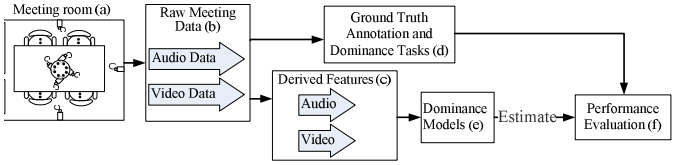


Fig. 1. Flow diagram of our approach.

III. Our Approach

Figure 1 shows a block diagram of the structure of our work:

- **(a,b): Section IV-A.** We use meeting data from the publicly available AMI corpus [5], where multiple microphones and video cameras have been used for audio and video data capture.
- **(d): Sections IV-C, IV-B.** We generated a detailed ground truth annotation of the perceived dominance for each individual in the meetings using multiple human judgments. Through a study of the annotator agreement, we define two sub-tasks to observe the effect on the performance of the dominance models when increased variability in the perception of dominance was present.
- **(c): Section V.** From the raw audio and video data, we derive features which are used to characterize certain nonverbal behaviors. Both the audio and video features have been treated similarly for comparison of the two modalities.
- **(e-f): Section VI.** Two models were considered for estimating dominance; one unsupervised and one supervised. The supervised approach was used for single as well as multi-modal fusion, which allowed us to study the contributions of the audio and video cues to the dominance estimation performance. We evaluated the performance of the models using both hard and soft evaluation criteria, where the latter accounted for the amount of variability in the ground truth annotations.

In summary, our work studies both the underlying variability in perceived dominance by human annotators, and systematically analyzes the objective performance of single and multi-modal dominance estimation models for a number of dominance classification tasks.

IV. Meeting data and Dominance Tasks

A. Meeting Data

We use meetings from the AMI corpus [5] which were carried out in the meeting room shown in Figure 2. 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,

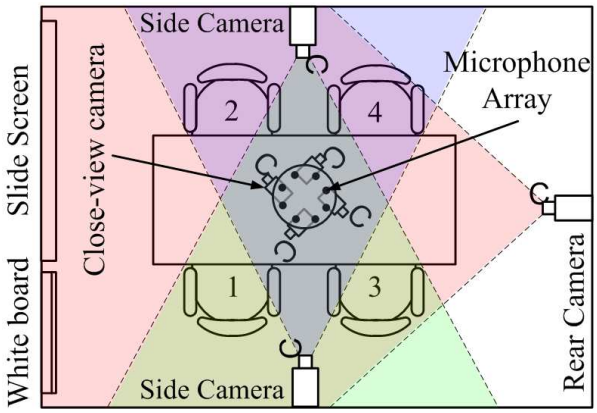


Fig. 2. Plan view of the meeting room set up.

and one with four microphones is set at the ceiling. Participants were also asked to wear both headset and lapel omni-directional 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 cameras mounted on the table captured individual visual activity only, as shown in Figure 3.

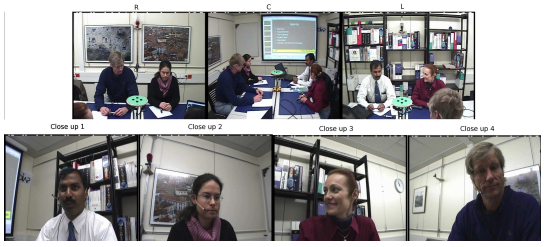


Fig. 3. Examples of the seven camera views available in the meeting room. The top row shows the right, centre and left cameras while the bottom row shows the view from each of the close up cameras.

From the AMI data, a subset of five exclusive team of participants were selected for our meeting data. Each team consisted of 4 participants, who were given the task of designing a remote control over a series of meeting sessions. The level of previous acquaintance among team members varied from being completely unacquainted to knowing each other well. Each participant was assigned distinct roles: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer'. During each session, the team was required to carry out certain tasks, such as a presentation on particular subjects related to the task, or a discussion about a particular aspect. To encourage natural behaviour, the meetings were not scripted and the teams met over several sessions so that they achieved the common goal.

B. Annotating the data

From the AMI data, 11 meeting *sessions* varying from 15 to 35 minutes were divided into 5 minute *segments* for ground truth annotation so that a total of 59 meeting *segments* were used. The segments were chosen to be 5 minutes long, rather than the original full meetings, since this provided more data points for training and testing. There is also evidence that people need a relatively small amount of time to make accurate judgments about the behavior of others [1].

A total of 21 annotators were used and were split into groups of 3 so that each group always annotated the same segments. The annotators were shown a video with views from the side and rear cameras which are shown in the top row of Figure 3. For a given meeting, each annotator viewed only one five-minute segment (in other words, an annotator never judged more than one segment of the same session). Annotators were requested to judge a person's dominance based only on the evidence within each meeting. Importantly, annotators were given neither a prior definition of dominance, nor were told what specific verbal or nonverbal cues to look for in order to make their judgments. Instead, they were requested on completion of the annotations, to provide a free form written description of the personal criteria they used to decode dominance.

For each meeting segment (simply called meeting from here on for convenience), annotators were asked to rank the participants, from 1 (highest) to 4 (lowest), according to their level of perceived dominance. As well as an absolute ranking, annotators were also asked to rank proportionately by distributing a total of 10 units among the participants, where more units signified higher dominance. To identify segments where the rankings were difficult to allocate, annotators were asked about their confidence in their absolute and proportionate rankings on a seven-point scale. Then, annotators were requested to ascertain specific characteristics of each participant such as their degree of activity, timidness, and talkativeness, also on a seven-point scale.

C. Analysis of the Annotations

From the human annotations, we wished to discover whether there was significant inter-annotator agreement across all meetings. Initial analysis of the meeting data indicated that 12 out of 59 meetings showed full agreement for all 4 absolute rankings of each meeting. This was clearly not enough for a fair representation of dominant behaviour for our experiments. Therefore we decided to relax the agreement condition by considering only the task of estimating the most dominant or the least dominant person. 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. Furthermore, the corresponding self-reported average confidence for the annotation for these meetings was 1.7 (where 1 represents the highest confidence and 7 represents the lowest). This subset represents almost 3 hours of meeting data where the agreement and confidence of the annotators was high. An additional observation of interest is that in 24 out of 34 cases, the most dominant person who was chosen by the annotators played the 'project manager' role.

We conducted further analysis and found that there were 23 additional meetings where 2 out of 3 annotators agreed on the most dominant person, and 54 meetings where atleast two out of the three annotators agreed on the least dominant person. These values and the corresponding average self-reported confidence levels are shown in Table I. This subset contains a larger intrinsic variation in the perceived dominance by human judges.

Finally, a similar analysis showed that there were 31 meetings with full agreement of the least dominant person. Similar to the most dominant case, the confidence decreases as the variability of the data-sets increases (see Table I). It is interesting to note that the confidence in the annotation of the least dominant person was always less than that of the corresponding experiment in the most dominant case. Also, the decrease in confidence as the variability of the data set increased was greater for the least dominant case compared to the most dominant case. We speculate that the behaviour of less dominant people tends to be more difficult to observe since they tend to speak and move less than dominant people [13].

Following the analysis of the annotations, we decided to define a number of dominance classification tasks, one for each of the different subsets discussed above. These are summarized in Table I below. Within each dominance task there are two sub-tasks that correspond to meetings where there is (i) **Full** agreement among annotators who labeled the same meeting, and (ii) **Maj**ority where at least 2 out of the 3 annotators agreed.

V. Audio and visual nonverbal cues for dominance modeling

In order to measure the dominant behaviour of people in meetings, we followed the social psychology literature and hypothesized that activity levels are correlated with dominance. Here we chose to represent activity in terms of audio and visual cues. From the audio sources, we adapted existing analysis techniques to characterize the speaking activity of the meeting participants. From the

Dominance Estimation Task	Sub-Tasks	Average Annotator Confidence	Number of Meetings	Proportion of Total Meetings (%)
Most	**Full**-agreement	1.74	34	57.6
	Majority-agreement	1.85	57	96.6
Least	**Full**-agreement	2.11	31	52.5
	Majority-agreement	2.4	54	91.5

TABLE I
DOMINANCE TASKS AND CORRESPONDING DATA-SETS.

video data, compressed-domain features were extracted from multiple cameras to characterize visual activity. More details are described in the following subsections.

A. Audio cues

Audio cues were extracted from the four close-talk microphones attached to each of the participants (one per person). Firstly we considered time-varying aspects of the speech.

Speaking Energy: The starting point for audio feature extraction is to compute a speaker energy value for each participant, using a sliding window at each time step as described in [33]. 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. For our experiments, the final signal was sub-sampled to a frame rate of 5fps.

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.

Then we considered features accumulated from the entire conversation. These features provided a simple way of quantifying the relative opportunities that participants had to speak. The following list summarizes the features used for our study.

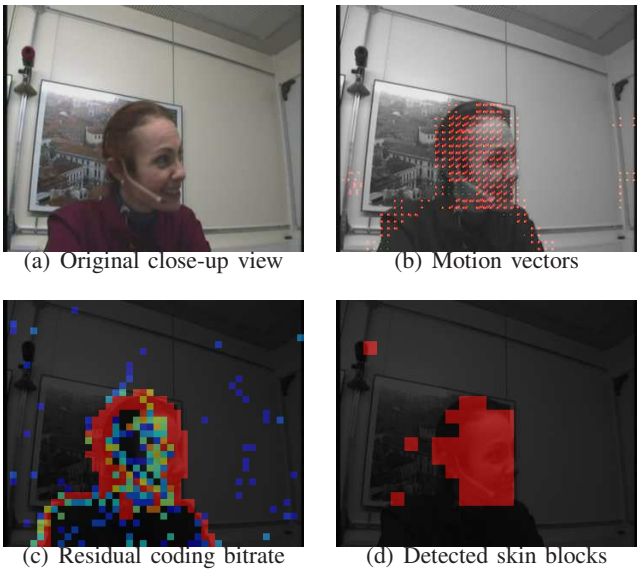
- **Total Speaking Energy (TSE):** Speaker energy accumulated over the entire meeting. This feature follows the findings in psychology that speaker energy is a manifestation of dominant behavior [13]. It is to be noted that the TSE feature captures how much a participant speaks as well as how loud he speaks, and not just how loud he speaks.
- **Total Speaking Length (TSL):** This feature considers the total time that a person speaks [28] according to their binary speaking status.
- **Total Speaking Turns (TST):** A speaking turn is the time interval for which a person's speaking status is active. The total number of speaker turns was accumulated over the entire meeting for each participant.

Several features were then derived to capture more meaningful characteristics of each person's speaking activity.

- **Speaking Turn Duration Histogram (SDHist):** The set of all turn durations is accumulated into a turn distribution or histogram. In all cases, we considered the speaking turn duration histogram with 11 bins, such that 10 bins were equally spaced at one-second intervals, and the last bin included all turns of size greater than 10 seconds for every participant. The bins were chosen in this way to primarily distinguish short turns (some of which are likely to be back-channels) from longer utterances. Empirically, we also found that increasing the number of bins did not lead to significant differences in performance.
- **Total Successful Interruptions (TSI):** This feature encodes the hypothesis that dominant people interrupt others more often [3]. The feature is defined by the cumulative number of times that speaker $i \in \{1, 2, 3, 4\}$ starts talking while another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn before i does, i.e. only interruptions that are successful are counted.
- **Total Speaking 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.

B. Visual cues

In order to capture visual motion activity efficiently, we leverage the fact that meeting videos are already in compressed form to extract visual activity features at a much lower computational cost. These features are generated from compressed-domain information such as motion vectors and block discrete-cosine transform (DCT) coefficients that are accessible at almost zero cost from compressed video [29], [31]. In our data set, there



(a) Original close-up view (b) Motion vectors

(c) Residual coding bitrate (d) Detected skin blocks

Fig. 4. Illustration of compressed domain features *(Best viewed in color).*

is a camera taking a close-up shot of each participant, as shown in the bottom row of Figure 3. Each of these video streams has already been 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, and the rest of the frames are predicted frames [8].

Figure 4 summarizes the various compressed domain features which can be extracted cheaply from compressed video. In particular, we consider the use of the *motion vector magnitude* (see Figure 4(b)) and the *residual coding bitrate* (see Figure 4(c)) to estimate visual activity level. Motion vectors, illustrated in Figure 4(b), are generated from motion compensation during video encoding; for each source block that is encoded in a predictive fashion, its motion vectors indicate which predictor block from the reference frame (in this case the previous frame for our compressed video data) is to be used. Typically, a predictor block is highly correlated with the source block and hence similar to the block to be encoded. Therefore, motion vectors are usually a good approximation of optical flow, which in turn is a proxy for the underlying motion of objects in the video [8]. We use the *motion vector magnitude* as one measure of visual activity in this work.

After motion compensation, the DCT coefficients of the residual signal, which is the difference between the block to be encoded and its prediction from the reference frame, are quantized and entropy coded. The *residual coding bitrate*, illustrated in Figure 4(c), is the number of bits used to encode this transformed residual signal. While the motion vector captures gross block translation, it fails to fully account for non-rigid motion such as lips moving. On the other hand, the residual coding bitrate is able to capture finer motion, since a temporal change that is not well modeled by the block translational model will result in a residual with higher energy, and hence require more bits to code it. In combination with the motion vector magnitude, the residual coding bitrate provides complementary evidence for visual activity.

For each meeting participant, we detect when they are in view. To do this, we implement a Gaussian Mixture Model (GMM) based skin-color block detector [21] that can detect face and hand regions. This works in the compressed domain with chrominance DCT DC coefficients and motion vector information, and produces detected *skin-color blocks* such as in Figure 4(d). We then threshold the number of skin-colored blocks in the close-up view to detect when a participant is seated. If a participant is not detected in a frame of the close-up view, he is assumed to be presenting at the projection screen, which is a reasonable assumption in the meeting data. We also assume that a person who is presenting is visually active.

If the participant is visible in the close-up view, we measure his visual activity by using either or both of motion vector magnitude and residual coding bitrate. To meaningfully compare motion vector magnitudes and residual coding bitrate, we normalize the quantities. Consider computing a normalized visual activity from motion vector magnitude for participant i in frame t. We first calculate the average motion vector magnitude, $v_{i,t}$, over all blocks in each frame. For each participant in each meeting, we find the median of the average motion vector magnitude, $\tilde{v}_i$, over all frames where the participant is in the close-up view. We also compute the average of the medians, $\bar{v}$, of all the participants. Normalization is then performed where the visual activity level for participant i in frame t, $v^R_{i,t}$, is computed by normalizing as follows:

$$v^R_{i,t} = \begin{cases} \frac{v_{i,t}}{2\bar{v}} & v_{i,t} < 2\bar{v} \\ 1 & v_{i,t} \geq 2\bar{v} \end{cases} \quad (1)$$

The visual activity level from the residual coding bitrate, $r^R_{i,t}$, is also normalized in a similar fashion.

We use the average of visual activity from motion vector magnitude, $v^R_{i,t}$, and from residual coding bitrate, $r^R_{i,t}$, as another estimate of visual activity. This allows us to quantify both rigid and non-rigid local motion. The combined estimate of visual activity for the ith participant in frame t, $m^R_{i,t}$, is given by:

$$m^R_{i,t} = \frac{1}{2}\left(v^R_{i,t} + r^R_{i,t}\right) \quad (2)$$

After raw visual activity extraction in order to facilitate the comparison between audio and visual cues,

visual cues are derived in an analogous fashion to those for audio cues as described in Section V-A. More specifically, the following cues were derived from the raw motion activity values:

- **Visual Activity.** A binary variable computed from compressed-domain video that indicates whether a participant is visually active or inactive at each time step (extracted at 25 frames per second). Three variations were tested, based on Motion Vectors (Vector), Residual Coding Bitrate (Residue), and the average of both features (Combo).
- **Total Visual Activity Length (TVL).** The accumulated motion activity for a person can be of three types, depending on whether it is estimated from the motion vectors, the residual coding bitrate, or their combination.
- **Total Visual Activity Turns (TVT).** This feature quantifies the number of times someone is continuously moving without breaks. This is analoguous to the total speaking turns feature defined in Subsection V-A.
- **Visual Activity Turn Duration Histogram (VDHist).** This tries to represent the motion turn characteristics of each participant. It is similarly defined as the speaking turn duration histogram.
- **Total Visual Activity Interruptions (TVI).** This captures when one person starts and remains visually active while another stops. While there may not be a meaningful notion of visual activity interruption in daily life, our hypothesis is that visual activity is correlated with speech activity such that speaker interruptions might be reflected in TVI as well. It is similar to the TSI feature defined in Subsection V-A.

Table II provides a summary of all the audio and video cues and their associated acronyms.

Glossary of Feature Acronyms	
Total Speaking Energy	TSE
Total Speaking Length	TSL
Total Speaking Turns	TST
Total Speaking Turns without Short Utterances	TSTwoSU
Total Speaking Interruptions	TSI
Turn Duration Histogram	SDHist
Total Motion Length	TVL
Total Motion Turns	TVT
Total Motion Interruptions	TVI
Motion Turn Duration Histogram	VDHist

TABLE II
GLOSSARY OF FEATURE ABBREVIATIONS

VI. MODELS FOR DOMINANCE ESTIMATION

In this section, we use a simple unsupervised model and a supervised model based on SVMs as prototypical models for dominance estimation. Our goal was to understand the relative predictive power of single cues for the dominance estimation task using the unsupervised model, and to explore whether cue fusion, in the SVM setup, could be useful. Our models, henceforth, are representative, rather than exhaustive.

A. Unsupervised model

In this model, audio or visual cues are accumulated over the duration of the meeting. The unsupervised model computes either the largest or smallest accumulated value of each feature, depending on whether we are estimating the most or least dominant person, respectively. That is, we hypothesize that someone is likely to be more dominant if they speak, move, or grab the floor the most out of all the participants in the meeting. While this model is simple, it showed promising performance in our preliminary work [17]. Similarly, we use the smallest accumulated value of the feature to identify the least dominant person in the meeting. We evaluate the model by comparing the label of the person who is estimated automatically with that of the ground truth annotated data.

B. Supervised Model

We also use a supervised method to investigate both single and multi-modal cue fusion. This allowed us to observe more closely, which cues were complementary or correlated and led to some very interesting findings about the comparative importance of the activity cues for robust dominance estimation. In order to make the cues comparable across meetings, we normalized them before fusion. The supervised approach uses a two-class SVM classifier to discriminate between the ‘most’ and ‘non-most’ dominant participants in each meeting. A second two-class SVM is trained to discriminate between the ‘least’ and ‘non-least’ dominant person. A Gaussian kernel was employed for both experiments. For each task, the SVM score produced for each person’s features are ranked. The rankings are then used to determine which participant is assigned the most (resp. least) dominant person label, by considering the point which is furthest from (resp. closest to) the class boundary. This procedure generates exactly one most (resp. least) dominant person per meeting. Note that as stated in Section III, this is different from the work in [26], [25] where each person independently was labeled as ‘high’, ‘middle’ or ‘low’. The model was evaluated using a leave-one-out approach for each combination of input features.

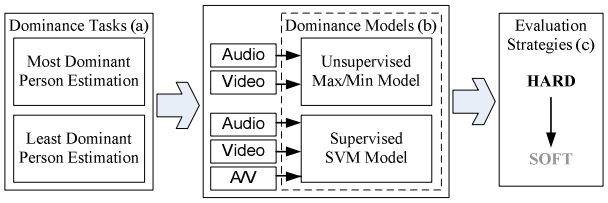


Fig. 5. Flow diagram showing our experimental protocol.

C. Experimental Protocol

Figure 5 shows a summary of the experiments that we carried out. As shown in Fig. 5(a), the experiments were split into two tasks: the estimation of the most dominant and the least dominant person.

For each of the tasks, we considered the set of experimental conditions illustrated in Figure 5 (b-c). Firstly, we considered each modality separately for both the supervised and unsupervised approaches. The supervised approach also allowed us to compare the performance of audio-visual feature fusion with combining features from the same modality. For each dominance task, we also considered different evaluation criteria, which accounted for increasing variability in the ground truth annotations, where hard (EvH) or soft (EvS) scoring criteria were used (Figure 5 (c)). The criteria themselves are explained in more detail in Subsection VII-B. For each of the two dominance tasks that we investigated, we consider two sub-tasks; full and majority agreement, as illustrated in Table I. It is important to note that for each model and evaluation criterion, the overall performance is calculated based on the estimation for each meeting rather than for each participant. The results are reported as classification accuracies, and discussions regarding the statistical significance of the results are summarized in Section IX.

VII. Classifying the Most-Dominant Person

A. Full-agreement data set

1) Audio cues: Table III shows the results obtained using audio cues. Using the unsupervised model with single features, the total speaking length (TSL) was most effective at 85.3% classification accuracy. This result is important not only because of the simplicity of this automated technique but also because it confirms the findings in social psychology [28], [13] about speaking time being a strong cue for dominance perception by humans. The total speaking energy (TSE) also performed well. While the total number of speaking turns (TST) did not perform as well, removing short utterances, some of which likely correspond to back-channels, (TSTwoSU), performed as well as TSL. Finally, the total number of successful interruptions (TSI) did not perform as well on our meeting data set. All these audio cues performed significantly better than chance (which would result in 25% classification accuracy).

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TSL	**85.3**
	TSE	82.4
	TST	61.8
	TSI	61.8
	TSTwoSU	**85.3**
Supervised	SDHist	82.4
	TSE, TST	88.2
	TSL, TSE, TST	88.2
	TSE, TST, TSI	88.2
	TSL, TSE, TST, TSI	88.2
	SDHist, TSE, TST, TSI	**91.2**
Random Guess	None	25.0

TABLE III
Performance of **Audio** cues for **Most**-dominant person with **Full**-agreement data.

The results with the supervised model trained on multi-dimensional audio cues are shown in Table III. A selection of the best performing feature combinations are displayed. We first observe that the Speaking Turn Duration Histogram (SDHist) did not perform better than the simple speaking length. No variation of performance for SDHist was observed if we discarded short turns.

A closer look at the meetings where TSL or TSE failed indicated that in some cases speaking turns or successful interruptions predicted the most dominant person correctly. This suggested that using the features jointly might improve performance. In practice, fusing these features in the supervised learning setup proved beneficial. We observe that although TST is not very discriminative as a single feature, it helps when combined with TSE alone or with TSE and TSL, yielding a 3% accuracy improvement. The best feature combination (SDHist, TSE, TST, TSI) yield an absolute performance improvement of 6% with respect to the performance obtained with TSL, with 91.2% accuracy.

A direct comparison of these results with the existing literature on automatic dominance detection is not possible as the addressed tasks , the data sets, and the experimental protocols used in each case are different. However, a few observations are still pertinent. First, both our results and [25] suggest that benefits can be obtained with audio fusion. Second, both speaking length and number of turns appear in our work and in [25] as part of the best performing feature combinations, an important difference being that, unlike [25], in our case all features are fully automatic. Third, the best performance figure obtained for our two-class task (around 90%) is considerably higher than the best reported performance obtained for the three-class problem in [25] (around

70%). Hypothetical reasons for this include the larger number of classes but also the fact that the data in [25] was not separated using any knowledge about the variability in perceived dominance. We study the case of higher variability in the human judgments in Section VII-B.

2) Visual cues: Table IV shows the results obtained with visual cues. Regarding single cues in the unsupervised setting, the total visual activity length (TVL), which quantifies how much people move, is consistently the best visual feature (76.5% accuracy), and seems to be the most robust. Motion turns (TVT) quantify how often people move. In practice, we observe that these features are generally 'noisy', presenting spikes of very short duration. However, removing short turns and leaving only those that should correspond to *intentional* motion (and that likely correspond to conversational activity too) results in the same performance as TVL. This is an interesting finding that seems to be supported by evidence in social psychology [4]. It was interesting to observe that, for TVL and TVT, the residual bitrate option performed slightly better than using the motion vectors; for TVT, the combination worked the best. The motion vector and residue cues capture different information. The former, being derived from block motion compensation in video compression, is better at capturing translational motion. The latter is related to the amount of non-rigid motion in the close-view cameras, including finer visual activity that is usually not captured by motion vectors. In contrast, TVI is not an effective cue: the results indicate that the concept of visual activity interruption (i.e., overlap) does not hold for video as clearly as it does for audio. As with audio cues, all the results with single video cues are considerably better than a random guess.

Compared to single audio cues, the best results with single visual cues degrade by 8.8% (76.5% vs. 85.3%). This is interesting since from the free-form verbal descriptions of how annotators perceived dominance, we found that about half of them mentioned the use of how much a person talks. In addition, annotators mentioned audio or language-based cues more than those related to visual activity. Despite this, it is remarkable that without using the audio at all, the most dominant person can still be correctly estimated in more than 75% of the cases with easily computable nonverbal visual cues. Furthermore, it is interesting to note that the use of compressed-domain cues, as compared with similar visual activity cues extracted in the pixel domain, did not lead to any classification performance loss (for more details, please refer to [31]). Also note that TVL performed better than some single audio cues. Figure 6(a) plots the values of TSL and TVL for all meetings in the full-agreement data set. The red crosses correspond to the positive examples (most-dominant) and the black circles to the negative ones. The figure indicates that there is a degree of correlation between the visual activity and speaking activity, but that the discrimination seems to be higher for the audio case.

For the multiple feature case, a small selection of the best performing combinations is also shown in Table IV. The visual activity histogram (VDHist) used in isolation was not a very effective cue, regardless of whether short turns were filtered out or not. The combination of the two best performing single features (TVL and TVT) did not improve performance over the single cues. However, when TVL, TVT, and VDHist were combined, we observe an small improvement of 3% (79.4% accuracy), suggesting that feature fusion in the supervised approach is also beneficial for visual cues. Overall, the best achieved performance with visual cues and supervised learning is 11.8% worse than the corresponding best performance for audio cues (79.4% vs. 91.2%), compare Tables III and IV.

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TVL (Vector)	73.5
	TVL (Residue)	**76.5**
	TVL (Combo)	73.5
	TVT (Vector)	67.6
	TVT (Residue)	70.6
	TVT (Combo)	**76.5**
	TVI (Vector)	52.9
	TVI (Residue)	52.9
	TVI (Combo)	44.1
Supervised	VDHist (Vector)	58.8
	VDHist (Residue)	61.8
	VDHist (Combo)	55.9
	TVL, TVT(Combo)	70.6
	VDHist, TVL (Residue)	73.5
	VDHist, TVT (Residue)	76.5
	VDHist, TVL, TVT (Residue)	**79.4**

TABLE IV
PERFORMANCE OF **VISUAL** CUES FOR **MOST**-DOMINANT PERSON TASK WITH **FULL**-AGREEMENT DATA.

3) Audio-visual fusion: A selection of results obtained with audio-visual cues and the supervised approach are shown in Table V. For the visual cues, we use the Residue option, which was overall the best one for the visual-only case. We also reproduce, for convenience, some of the results using the audio features displayed in Table III. Unfortunately, audio-visual fusion did not yield any further improvement in classification performance compared to using the audio-only cues. The obtained performance is often better than the visual-only case but always worse than or equal to the audio-only case. This

holds in particular for the single-cue case, e.g. the total speaking and visual activity lengths (TSL, TVL), and for the best audio feature combination (SDHist, TSE, TST, TSI). The best obtained performance remains 91.2%. Note that the differences in performance between the best methods are not statistically significant at the 5% level using a standard binomial test, as the number of data points is relatively small. Nevertheless these results show that such features and feature combinations are worth exploring. Figure 7 summarizes the best results obtained for single and multi-modal cases. The correlation between the best audio and visual cues is a likely explanation to the lack of success with audio-visual fusion.

Feature	Class. acc. (%)
TSL, TVL	79.4
TSE, TVL	70.6
TST, TVT	76.5
TSL, TVL, TVT	79.4
SDHist, TSE, TST, TSI, TVL	**91.2**
SDHist, TSE, TST, TSI, VDHist	**91.2**
SDHist, TSE, TST, TSI, VDHist, TVL	82.4
SDHist, TSE, TST, TSI, VDHist, TVL, TVT	82.4

TABLE V
PERFORMANCE OF **AUDIO-VISUAL** CUES WITH **MOST**-DOMINANT PERSON TASK WITH **FULL**-AGREEMENT DATA.

B. Majority-agreement data set

The second task addressed involves the 57-meeting set where at least 2 annotators agree, which corresponds to almost all the data (96%). This data set inherently has more variability with respect to human perceptions of dominance (as further suggested by the lower confidence self-reported by the annotators as discussed in Section IV). The evaluation of this task is therefore aimed at analyzing the performance of models and cues in more challenging conditions.

For evaluation, we used two different ways of computing classification accuracy. 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. A first evaluation criterion, (called EvH for short) computes the classification accuracy as n/N, and a second criterion (called EvS), computes classification accuracy as $(n + m)/N$. The hard criterion assumes that there is only one correctly labeled most-dominant-person for each meeting - the one corresponding to the majority vote by the annotators - and is obviously the correct way to evaluate performance on the full-agreement data set, as done in the previous section. In contrast, the soft criterion assumes that both the 'most-voted' and the 'least-voted' most-dominant-person labeled by the annotators for a given meeting are correct, and thus the prediction of either of them is considered as correct. This evaluation is clearly less stringent, but it is nevertheless important to observe the ability of the algorithms to predict either of the two people perceived by annotators as being most-dominant.

1) Audio cues: Table VI presents a selection of the classification accuracy results obtained for audio cues. For single cues and the unsupervised model, TSL and TSTwoSU are the best performing features for both EvH (77.2% and 75.5%, respectively) and EvS (84.2% for both features). TSE is the third best performing feature, and TST and TSI are not as effective. Interestingly, these findings are consistent with the ones obtained for the full-majority data set (compare to Table III). A consistent decrease in performance (8.1% for TSL) is observed for all cues which suggests that the inclusion of the data that is intrinsically more ambiguous with respect to perceived dominance results in a more challenging task. On the other hand, the results obtained with the soft criterion, which assumes that more than one person can be most-dominant, brings the performance of most features back to the same level they had for the full-agreement data set, which indicates that in several cases the methods guessed the 'least-voted' person as being most dominant. The results for the supervised model and fused audio cues also appears in Table VI. The selection shown is a subset of those in Table III and includes the best performing cases. We observe that, using the EvH criterion, a few feature combinations performed at the same level, but not better, than the best single cue. On the other hand, using the EvS criterion, we observe that the same feature combinations were capable of slightly improving performance (a best performance of 87.7% for the same feature combination that performed the best for full-agreement data). Overall, the supervised approach brought a moderate improvement over the much simpler unsupervised case.

2) Visual cues: Table VII shows selected results obtained with visual cues. Compared to the results obtained for the full-agreement case (Table IV), many observed trends hold: TVL and filtered TVT are the best performing single cues. TVI is a poor predictor, and overall visual-only features perform worse than audio-only. Furthermore, similar to the audio-only results in this section, we observe a general decrease in performance with respect to the full-agreement data set when using the EvH criterion (for the best performing single visual cues, the

Dominance Model	Feature	Class. Acc. %	
		EvH	EvS
Unsupervised	TSL	**77.2**	**84.2**
	TSE	73.7	79
	TST	54.4	64.9
	TSI	52.6	64.9
	TSTwoSU	75.5	**84.2**
Supervised	TSL, TSE, TST	**77.2**	85.9
	TSE, TST, TSI	75.4	84.2
	SDHist, TSE, TST, TSI	**77.2**	**87.7**

TABLE VI

PERFORMANCE OF **AUDIO** CUES FOR **MOST-DOMINANT** PERSON TASK WITH **MAJORITY-AGREEMENT** DATA.

absolute degradation is 6.3%). Furthermore, the results obtained with the EvS criterion for the best visual cues brings the performance back to the same level they had for the full-agreement case. Finally, supervised learning and multiple visual cues did not improve performance over the simple unsupervised, single-cue model.

Dominance Model	Feature	Class. Acc. %	
		EvH	EvS
Unsupervised	TVL (Residue)	66.7	**80.7**
	TVL (Combo)	64.9	**80.7**
	TVT (Combo)	**70.2**	**80.7**
	TVI (Combo)	47.4	61.4
Supervised	TVL, TVT (Combo)	59.7	75.4
	VDHist, TVL (Residue)	**64.9**	**78.9**
	VDHist, TVL, TVT (Combo)	63.1	77.2

TABLE VII

PERFORMANCE OF **VISUAL** CUES FOR **MOST-DOMINANT** PERSON TASK WITH **MAJORITY-AGREEMENT** DATA.

3) Audio-visual cues: The results for the best combinations appear in Table VIII. All visual activity features have been derived with the 'residue' option. We observe that audio-visual fusion did not improve performance over audio-only under either of the evaluation criteria. This is shown in Figure 7. This result holds for both the full-agreement and the majority-agreement data sets.

Feature	EvH	EvS
TSL, TVL, TVT	**75.4**	82.5
SDHist, TSE, TST, TSI, VDHist	**75.4**	**84.2**
SDHist, TSE, TST, TSI, VDHist, TVL, TVT	**75.4**	82.4

TABLE VIII

PERFORMANCE OF **AUDIO-VISUAL** CUES FOR **MOST**-DOMINANT PERSON TASK WITH **MAJORITY-AGREEMENT** DATA.

VIII. CLASSIFYING THE LEAST-DOMINANT PERSON

In this section, we discuss our results for the least-dominant person classification task. The experiments that were carried out were identical to the most-dominant case so the discussion in this section will be more brief. We first conducted experiments on the least dominant person task with full-agreement data (31 meetings) and majority-agreement data (54 meetings). For the unsupervised model, the person that corresponds to the lowest proportion of the feature among all participants is classified as least dominant. The supervised model is trained on the least vs. non-least dominant classes.

A. Full-Agreement data-set

1) Audio cues: The classification accuracy of the cues under the unsupervised and supervised schemes are shown in Table IX. The highest performance of 83.9% was achieved by both the unsupervised and supervised methods so there was no gain from fusing cues.

Like the equivalent case in Section VII-A, the TSI feature performed the worst for the unsupervised case. It was also interesting to see the increase in performance between the TST and TSTwoSU features. This suggests that the short turns were adding noise to the TST features. This was similarly observed for the corresponding set of results in Table III for the most dominant person task.

Unlike the most dominant case, here there is a significant reduction in performance for TSE compared to TSL. We speculate that this is because the total energy is much lower and therefore more sensitive to noise (i.e. the signal-to-noise ratio is lower). TSL showed a slight decrease in performance for estimating the least dominant person, compared to estimating the most dominant person. These results suggest that a similar trend will also be observed with the visual cues; less dominant people are less active, so their measured activity will be more sensitive to noise. In addition, we note that some annotators did comment on how it was more difficult to rank passive participants than active ones.

Dominance Model	Feature	Class. Acc. (%)
Unsupervised	TSL	**83.9**
	TSE	67.7
	TST	71.0
	TSI	51.6
	TSTwoSU	**83.9**
Supervised	TSE, TST	80.7
	TSL, TSE, TST	80.7
	SDHist,TSE,TST	**83.9**
	SDHist,TSE,TST,TSI	**83.9**

TABLE IX

PERFORMANCE OF **AUDIO** CUES FOR **LEAST-DOMINANT** PERSON TASK WITH **FULL-AGREEMENT** DATA

2) Visual cues: Table X shows some selected results from our experiments using only the visual cues for the majority-agreement data-set. While in the equivalent

results of the most-dominant task in Table IV, both (TVL(Residue)) and (TVT(Combo)) had the best performance, for the least-dominant task, only (TVT(Combo)) performed the best. This is likely to be caused by the removal of the shorter turns, which account for noisy measurements of the visual activity. However, TVT might also eliminate significant amounts of true activity for the most passive person. We also found that the TVI feature performed less well in general. Overall, the visual features are less discriminative than the audio ones, and also less effective compared to the most-dominant task. In terms of statistical significance, the decrease in performance between the best audio and video performance for the full-agreement case was not statistically significant at conventional levels using a standard binomial test. See Figure 7 for a comparison.

Dominance Model	Method	Class. Acc.(%)
Unsupervised	TVL(Vector)	54.8
	TVT(Vector)	58.1
	TVT(Combo)	**64.5**
	TVI(Combo)	54.8
Supervised	VDHist(Vector)	45.2
	TVL, TVT(Combo)	45.2
	VDHist, TVL(Vector)	45.2
	VDHist, TVL(Combo)	48.4
	VDHist, TVL, TVT(Vector)	45.2
	VDHist, TVL, TVT(Combo)	**54.8**

TABLE X

PERFORMANCE OF **VISUAL** CUES FOR **LEAST-DOMINANT** PERSON TASK WITH **FULL-AGREEMENT** DATA.

3) Audio-Visual Fusion: The audio-visual cues performed similarly to the visual-only cues since the best performing feature combinations still performed less well than TSL or TSTwoSU, as shown in Table XI. In general, the results using audio-visual features did not perform as well as those of using audio cues. The drop in performance when using video rather than audio features was also observed with the most-dominant person task, but was not as pronounced as in the least-dominant case. Due to the low levels of visual activity of the least-dominant participant, it is likely that it is more sensitive to noise. In addition, we can see from Figure 6(b) that the audio and visual activity are well correlated and therefore not complementary.

B. Majority-agreement data-set

For this task, there was a total of 54 meetings, which accounted for 91.5% of the total data. We show a selection of performance results for this task in Table XII. The best achieved results are also shown in Figure 7.

Feature	Class. Acc.(%)
TSL, TVL	77.4
TST, TVT	77.4
SDHist, TVL	**80.7**
SDHist,TSE,TST,TSI,VDHist, TVL, TVT	80.0

TABLE XI

PERFORMANCE OF **AUDIO-VISUAL** CUES WITH SUPERVISED MODEL FOR **LEAST-DOMINANT** PERSON TASK WITH **FULL-AGREEMENT** DATA. ALL MOTION FEATURES HAVE BEEN DERIVED WITH THE 'RESIDUAL' OPTION.

Firstly, it was interesting to see that TSL was not the feature that gave the best performance, though it was ranked second behind TSTwoSU. This observation suggests that the adding annotator variability and having proportionately less observations in the captured signal leads to a greater need for noise removal. Furthermore, we found that the shorter turns were not a discriminative feature for estimating dominance and it is likely that for the least-dominant person, they would represent a larger proportion of a person's total speaking turns than that of the most dominant person.

Increasing the variability in the data did not always lead to a drop in performance. We also observed that fusing the TVL feature with other features led to an increased performance when the supervised model was used. However, none of the feature combinations which included visual activity cues could perform as well as those of the audio activity.

Dominance Model	Features	Class. Acc. % EvH	Class. Acc. % EvS
Unsupervised	TSL	59.3	75.9
	TSTwoSU	**68.5**	**83.3**
	TVL(Vector)	53.7	62.9
	TVT(Combo)	48.1	63
Supervised	TSL,TSE,TST,TSI	59.3	77.8
	SDHist,TSE,TST,TSI	**64.8**	**79.6**
	VDHist, TVL, TVT(Combo)	51.8	62.9
	TSL, TVL	61.1	**79.6**
	TSL, TVL,TVT	61.1	**79.6**
	SDHist,TSE,TST,TSI,VDHist, TVL	61.1	75.9

TABLE XII

PERFORMANCE OF **AUDIO, VIDEO AND AUDIO-VISUAL** CUES FOR **LEAST-DOMINANT** CLASSIFICATION TASK WITH **MAJORITY-AGREEMENT** DATA.

IX. FINAL DISCUSSION AND CONCLUSION

Overall, our study has investigated how dominance can be estimated by different audio and video cues, and affected by annotator variability, estimation method and the exact task involved. Our investigation suggests the following:

Audio cues. When taking the cue which performed best in all categories, the audio cues always gave the highest classification accuracy. We observed that TSL gave the best results as a single feature, though was second best for the task of estimating the least dominant person when the data set had majority agreement. In addition, TSTwoSU was found to be more robust to annotator variability by obtaining the highest performance in both most and least dominance tasks. There was a marked improvement in performance between the TST and TSTwoSU features, indicating that much of the noise in the TST feature was caused by the shorter turns, which were not discriminative for our task. We also found that while the SDHist feature was less effective on its own, in all the highest single or multi-modality cases, it was found to be complementary to other features. TSI performed badly in general, suggesting that interruptions are not always a good cue for dominance estimation. One point to note, however, is that this cue was derived using a coarse measure, which did not quantify the quality of the interruption in terms of speaker overlap, for example.

Visual cues. We found that their performance was never able to improve upon those of the best audio cues. However, it was interesting to see that a comparison of the performance of the single audio and video cues (Figure 7) shows that the gap between modalities in some cases is very small even though the visual cues are coarse and fast to compute and the resulting features are noisy. It was particularly interesting to observe that reasonable performance was achievable in the most-dominant case without having to listen to the conversations at all. There were also some single cue cases where the visual cues performed better than the audio cues. It was also relevant to observe that VDHist was effective as a complementary cue, leading to its use in all the best video and audio-visual cue fusion results.

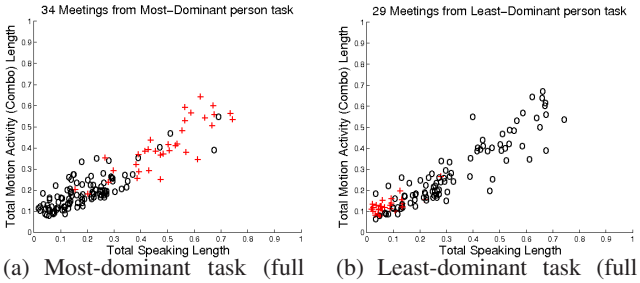
Audio-Visual Cues. In terms of audio-visual cue fusion, we found that in some cases the feature combinations matched the best performing audio-only cues, but was never better. This can be explained by the overall lower performance of the visual cues. One observation we must make here is that the audio signal was extracted from close-talk headset microphones while the video signal was captured from a much further distance from the participants. It would be important to see how the results using audio cues would change if more challenging audio data from far-field microphones was used. Parallel work using a single distant microphone to extract the total speaking length has shown that there is indeed a decrease in performance [16].

Full and Majority Agreement Data. From the two evaluation criteria that were used for the data sets with majority agreement, we found a systematic drop in performance when comparing the performance of the hard evaluation criterion with the full agreement case. However, it was interesting to observe that with the soft criterion, the performance in some cases was equivalent to that of the corresponding full-agreement case.

Supervised and Unsupervised Models. It was interesting to observe that while the best performance of 91.2% for the estimation of the most dominant person was obtained using the SVM method, the best performance with the unsupervised model and a single cue was already 85.3%. For the task of estimating the least dominant person, the best performance was 83.9%, which was obtained from using both the unsupervised and supervised approaches. This is an interesting result since the unsupervised model does not require training data and has a much lower computational overhead compared to the supervised model.

Most and Least Dominant Tasks. It was interesting to observe that there was a consistent drop in performance between the two tasks as shown in Figure 7. Closer inspection also shows that there is a more significant decrease in performance between the audio and video cues for the least dominant task compared to that of the most dominant. This is an interesting finding that highlights the inherent increase in uncertainty when trying to identify people who have a lower level of activity. While the most dominant person in a meeting might be considered the most active and therefore more observable, finding the least-dominant person is closer to identifying the most passive or someone with the least observable cues. This seems to be reflected in the self-reported annotator confidence values (see Table I). Such a problem may be better solved with more sophisticated visual cues where for instance attention can be measured.

Evaluation advantages and limitations. Our work has produced novel evaluation resources (data annotation, research tasks, and corresponding data sets) that build upon and enrich the publicly available AMI meeting corpus. We also plan to make these resources public. Finally, as the size of the data set is relatively small, many of the observed performance differences are not statistically significant at conventional levels. In this view, the results presented here need to be interpreted with care, specially from the view of generalization. While the social psychology literature has validated, over multiple studies, the robustness of certain nonverbal cues for dominance perception [28], similar work to ours would have to be done in other scenarios to thoroughly validate such cues in automatic systems, using larger and varied data sets.



(a) Most-dominant task (full agreement)

(b) Least-dominant task (full agreement)

Fig. 6. Scatter plots of the total speaking and visual activity length where the red crosses show the ground truth annotated person with the corresponding audio and visual cues, and the black circles show the negative class in each case.

Future work. One of the limitations of our work is its reliance on high-quality audio (derived from close-talk microphones) to extract cues. We have taken initial steps to address some of these limitations by investigating the extraction of nonverbal cues (such as speaking turns) from single distant microphones [16]. The results suggest that the most-dominant person classification performance degrades, as compared to the head-set microphones, but the degradation is not drastic. We believe that the extraction of audio nonverbal cues from far-field microphones is a relevant area of future work. In the second place, the nonverbal communication literature also refers to various cues related to body-language as cues for dominance (e.g. postures and gestures) and this would be interesting to explore. In the third place, we plan to address the dominance problem in terms of cliques rather than dominant individuals since there are occasions when multiple people can be perceived as similarly dominant. Finally, the performance measures considered in this paper are simply a few of the various possible options. In the future, it would be interesting to examine the effect of various cues on the speed of detecting dominance, or other measures of importance to different applications.

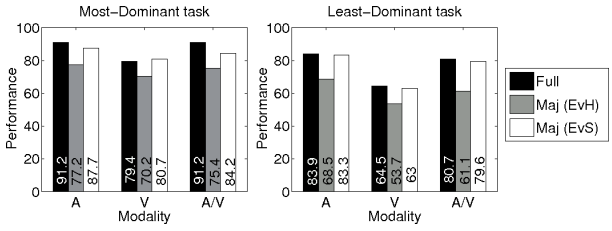


Fig. 7. Comparison of the best performance values in each modality and each dominance sub-task. A:Audio, V:Video, A.V:Audio-Visual.

Acknowledgments: We thank Kannan Ramachandran (UC Berkeley), Jean-Marc Odobez and Sileye Ba (IDIAP) for discussions.

REFERENCES

[1] N. Ambady, F.J. Bernieri, and J.A. Richeson. Toward a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of the Behavioral Stream. *Advances in Experimental social psychology*, 32:201–257, 2000.

[2] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *Proc. IEEE CVPR Int. Workshop on Cues in Communication (CVPR-CUES)*, Kauai, Dec. 2001.

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

[4] 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.

[5] 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. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh,UK, Jul. 2005.

[6] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek. A multimodal analysis of floor control in meetings. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington, DC, May 2006.

[7] T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed memory Markov process. In *Proc. NIPS*, Dec. 2004.

[8] M.T Coimbra and M. Davies. Approximating optical flow within the MPEG-2 compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):103–107, 2005.

[9] J.P. Dillard and K.J. Tusing. The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, 26(1):148–171, 2000.

[10] J. DiMicco, A. Pandolfo, and W. Bender. Influencing group participation with a shared display. In *Proc. ACM Conf on Computer Supported Cooperative Work (CSCW)*, New York, NY, USA, Nov. 2004.

[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. Measuring nonverbal dominance. In V. Manusov, editor, *The sourcebook of nonverbal measures: Going beyond words*. Erlbaum, 2005.

[13] 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.

[14] D. Gatica-Perez. Analyzing human interaction in conversations: a review. In *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Sep. 2006.

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

[16] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Mar. 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 Int. Conf. on Multimedia (ACM MM)*, Augsburg, Sep. 2007.
[18] N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *Proc. Conf. European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Apr. 2006.
[19] O. Kulyk, J. Wang, and J. Terken. Real-Time Feedback on Nonverbal Behaviour to Enhance Social Dynamics in Small Group Meetings. *Proceedings of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, 3869:150–161, 2006.
[20] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):305–317, Mar. 2005.
[21] S J McKenna, S Gong, and Y Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.
[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] A. Pentland and A. Madan. Perception of social interest. In *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, Beijing, Oct. 2005.
[25] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small group meetings. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264, New York, NY, USA, 2006. ACM Press.
[26] R.J. Rienks and D. Heylen. Automatic dominance detection in meetings using easily detectable features. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.
[27] E. Rosa and A. Mazur. Incipient status in small groups. *Social Forces*, 58(1):18–37, Sep. 1979.
[28] M. Schmid Mast. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research*, 28(3):420–450, Jul. 2002.
[29] 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.
[30] B. Wrede and E. Shriberg. Spotting hotspots in meetings: Human judgments and prosodic cues. In *Proc. Eurospeech*, Geneva, Sep. 2003.
[31] 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.
[32] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *in Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Banff, Nov. 2006.
[33] 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.

Dinesh Babu Jayagopi graduated with a B.Tech degree in Electronics from Madras Institute of Technology, Chennai in 2001. He then obtained a masters in System Science and Signal Processing from Indian Institute of Science in 2003. Between 2003 and 2006, he worked as a senior research engineer at Mercedes-Benz Research and Development India. Since 2007, he is a research assistant at Idiap Research Institute, Switzerland and a doctoral student at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His research interests include human behaviour analysis and modeling, machine learning and signal processing.

Hayley Hung graduated with a MEng degree in Electronic and Electrical Engineering from Imperial College in 2002. She received a PhD in Computer Vision from Queen Mary University of London in 2007, which was supervised by Professor Shaogang Gong. She is currently a Post-doctoral researcher at Idiap Research Institute in Switzerland. Her research interests include human-human interactive behaviour analysis, video saliency, and unsual behaviour detection. She is an associate member of the IET and also a member of the IEEE.

Chuohao Yeo (S'05) received the S.B. degree in electrical science and engineering and the M.Eng. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 2002. He is currently working towards the Ph.D. degree in electrical engineering and computer science at University of Califonia (UC), Berkeley, CA. His research interests include image and video processing, distributed source coding, computer vision and machine learning. Mr. Yeo is a student member of SPIE. He was a recipient of the Singapore Government Public Service Commission Overseas Merit Scholarship from 1998-2002, and a recipient of Singapore's Agency for Science, Technology and Research Overseas Graduate Scholarship since 2004. He received a Best Student Paper Award in SPIE VCIP 2007.

Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001. He joined the Idiap Research Institute in 2002, where he is now a senior researcher. He is an Associate Editor of the IEEE Transactions on Multimedia. His interests include multimedia signal processing, social computing, and machine learning. He is a member of the IEEE.