IDIAP RESEARCH REPORT

Using Audio and Video Features to Classify the Most Dominant Person in a Group Meeting

Hayley Hung [a] Dinesh Jayagopi [a b]
Chuohao Yeo [c] Gerald Friedland [d]
Silèye Ba [a] Jean-Marc Odobez [a b]
Kannan Ramchandran [c] Nikki Mirghafori [d]
Daniel Gatica-Perez [a b]

IDIAP–RR 07-29

July 2007

to appear in
Association for Computing Machinery - Multimedia (ACM-MM), September 23–28, 2007, Augsburg, Bavaria, Germany.

[a] IDIAP Research Institute
[b] Ecole Polytechnique Federale de Lausanne (EPFL)
[c] University of California
[d] International Computer Science Institute (ICSI)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-29

Using Audio and Video Features to Classify the Most Dominant Person in a Group Meeting

Hayley Hung Dinesh Jayagopi Chuohao Yeo Gerald Friedland
Silèye Ba Jean-Marc Odobez Kannan Ramchandran
Nikki Mirghafori Daniel Gatica-Perez

July 2007

to appear in
Association for Computing Machinery - Multimedia (ACM-MM), September 23–28, 2007, Augsburg, Bavaria, Germany.

Abstract. The automated extraction of semantically meaningful information from multi-modal data is becoming increasingly necessary due to the escalation of captured data for archival. A novel area of multi-modal data labelling, which has received relatively little attention, is the automatic estimation of the most dominant person in a group meeting. In this paper, we provide a framework for detecting dominance in group meetings using different audio and video cues. We show that by using a simple model for dominance estimation we can obtain promising results.

1 Introduction

We address the problem of finding dominance in group meeting scenarios. Applications include corporate recruitment where it is important to observe how candidates contribute to the group dynamic in collaborative team environments. Such a task is also important for analysing employee performance in corporate team building exercises. We focus on extracting a quantifiable measure of dominance from audio and video meeting data, and choose deliberately not to provide a prior definition of dominance for the experiments presented here since it is multi-faceted and can differ across people. Instead, we use multiple annotator consensus to define it and through this, we can investigate if certain audio and/or video features correspond well with human judgement. A more detailed discussion of dominance and its relation to the perception of power can be found in [7]. We present and compare results using fully and partially automated methods for feature extraction to identify the most dominant person in a group meeting.

Early work in the area of automatic modelling of dominance in conversations was suggested by Basu et al. [2] who used networked Markov chains to represent the interrelations between agents in synthetic data. They also showed preliminary results using human-human interaction data where two out of five participants were pre-selected to debate for one minute. Their model was able to detect who the two debating participants were using manually labelled speaker turns, speaking energy, pitch, rate, and also region-based motion energy and blob tracking. This work, however, did not provide a conclusive study of how each of the individual features affected the performance of the influence detection. Furthermore, the conversational setting was rather specific.

Another study of group dominance in scripted meeting scenarios was proposed by Zhang et al. [11] using a two-level dynamic Bayesian network where individual and group-level states were modelled separately. Audio and speech transcription-based features were used, and participant speaking length was shown to perform well as a baseline measure of ranked dominance in 30 five-minute scripted meetings. Semantically higher level features for determining dominance rankings from meetings were proposed in [8], which extracted three categories of features related to individual speech behaviour, aspects of the interaction, and meeting semantics. These were all extracted using manual speech transcriptions of the meetings; the robust automated extraction of some of them would be a non-trivial problem.

Until now, little work has systematically investigated the automatic estimation of dominance in natural, non-scripted group meetings scenarios using audio and video features. Also, how individual audio and video features correlate with perceptions of dominance has not been fully addressed. Here we tackle both problems and hypothesise that dominance in group meetings correlates with high dynamic levels of human activity, which can be represented by relatively primitive features extracted from audio and/or video sensor data. An additional contribution is the generation of a dominance-annotated meeting data set for experimentation.

In the remainder of this paper, Section 2 will describe the features that we have chosen to measure dominance as well as the data generation and annotation process. Section 3 will present our experimental results and discussions and Section 4 concludes and discusses possible future work.

2 Framework

The aim of our work is to identify which features are well correlated with the most dominant person in a meeting. First we use speaking length and energy, as suggested by [2]. We also extract features from video since non-verbal cues can also affect perceptions of dominance [9].

Four different types of audio and video features were chosen for classifying dominance. Firstly, audio features were extracted as speaker-turn segmentations and speaking energy for each participant using high-quality audio captured by personal head-set microphones. In addition, we used speaker segmentations estimated from a single audio source by automated speaker diarization. Secondly, both real and binary-valued video features were extracted to measure motion activity from compressed

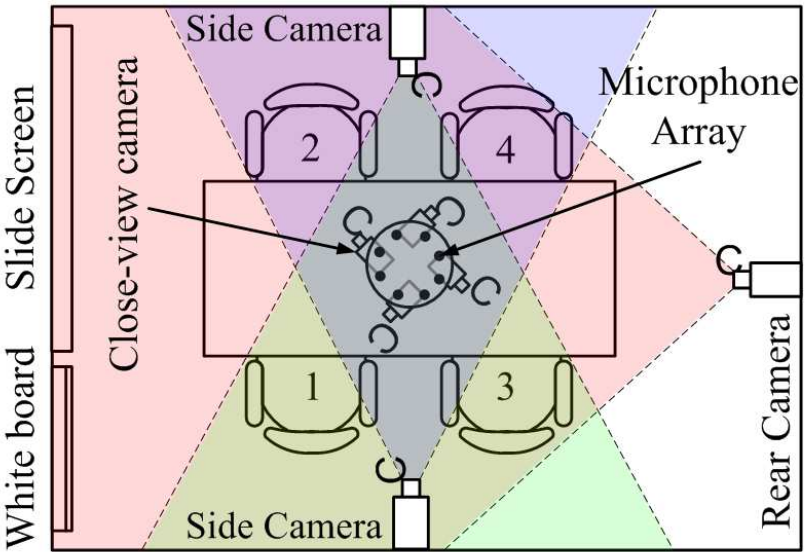


Figure 1: Plan view of the meeting room set up.

MPEG-4 video. For the classification task, we took each of the extracted measures individually and chose the person who had the highest total value to be the most dominant.

2.1 Data and Annotation

The publicly available Augmented Multiparty Interaction (AMI) meeting corpus [3] was used for our work. The data was captured in a room consisting of a table, slide screen, and whiteboard (see Figure 1). A microphone array and close-view cameras were set in the middle of the table to capture localised activity. Headset microphones are also available for each participant.

Natural non-scripted meeting data was produced in a role-play scenario where four members of a team were asked to design a remote control device over a series of sessions. Each participant was given a role such as 'project manager' beforehand. In each case, the participants were not requested to act in a particular way so that natural behaviour and engagement with the task and their respective roles could be captured. In addition, each session was assigned a timetable of activites for the meeting, e.g. one person giving a presentation to the rest of the group or an open discussion.

A total of 59 five-minute meeting segments from 11 sessions were provided for multi-observer annotation. 21 annotators were split into groups of 3 such that each group always annotated the same segments. For each watched segment, annotators were asked to rank the participants, from 1 (most) to 4 (least), according to their level of perceived dominance. They watched each segment using a video player with audio and video streams as shown in Figure 2 where three synchronised videos from the rear and side cameras were amalgamated. Annotators were not given any initial definition of dominance but were asked to provide a free-form verbal description on completion of the annotations.

From the study described above, we targeted the task of automatically classifying the most dominant person in each meeting. Only the meetings where *every* annotator agreed on the most dominant person was used which provided a ground truth data set of 34 five-minute meeting segments representing almost 3 hours of data.

Figure 2: Example screenshot from the meeting videos which were used for annotation.

2.2 Audio Features

Two different approaches were used to find the speaking length of each participant. The first uses audio signals from individual headset microphones to generate a reliable measure of both individual speaking length and energy. The second uses data from a single audio source, which requires the identification of both the speakers and their speaking turns.

2.2.1 Speaker Energy and Length Using Individual Sources

Using four sources, 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. Speaker turns were then segmented using a thresholded version of the absolute values where 1 represented speaking and 0 indicated silence.

2.2.2 Speaker Diarization Using a Single Source

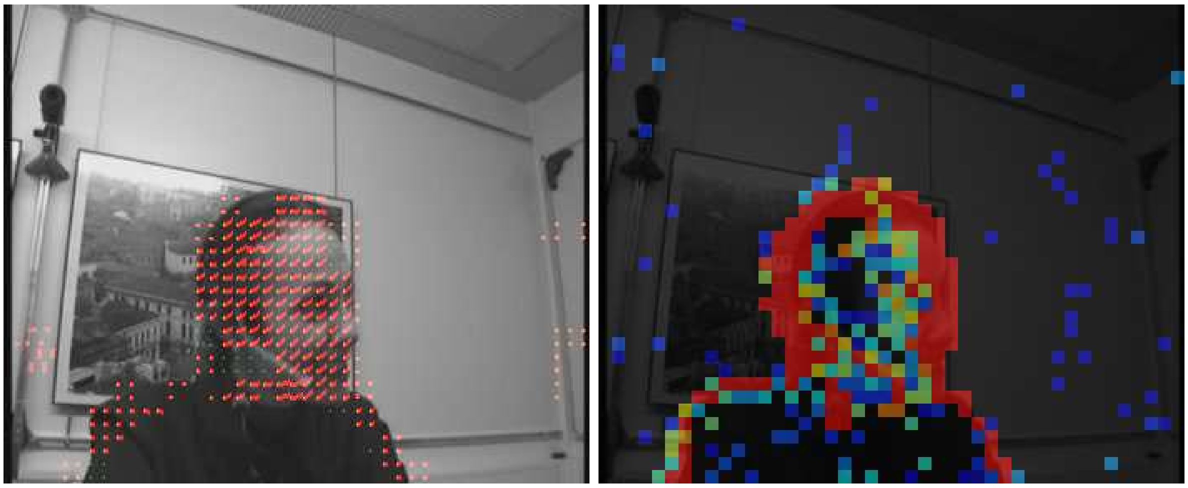
Our approach extends [10] where given a single-channel audio signal, an agglomerative method is applied to perform both the segmentation of the audio track into speaker-homogeneous time segments and grouping into speaker clusters. Although the data was recorded by an 8-microphone array, we wish to solve the problem using a single mono source where all channels are amalgamated by delay summation. The audio track is then converted into 19th order Mel-frequency cepstral coefficients using a frame size of 10 ms. A speech/non-speech detector was then used to filter non-speech regions before subsequent processing.

The algorithm is initialised using a much greater number of clusters than possible speakers (usually 16) where the initial segmentation is generated by randomly assigning equal-length segments of the audio track to each speaker cluster. Gaussian Mixture Models (GMMs) are then trained on these initial segmentations. Each speaker is also represented by a Hidden Markov Model (HMM) where each state is a segment of minimum duration of typically 2.5 s. The states of each speaker HMM are connected together, so that speaker turns are represented by a state transition. The algorithm performs the following loop:

1. (Re-)Segmentation: Compute the likelihood of each segment to belong to one of the GMMs. Find a path in the connected speaker HMMs that maximizes the likelihood for all states using the Viterbi algorithm. Re-segment the audio track.

2. (Re-)Training: Given the new segmentation of the audio track, compute new GMMs for each of them.

3. Cluster Merging: Given the new GMM, find two models that best represent the same speaker by computing the minimum description length/Bayesian information criterion (MDL/BIC) score of each of the individual models and their combination. If the score of the merged GMM is less than or



(a) (b)

Figure 3: Example output from the compressed domain video feature extraction. (a) Motion vectors, (b) Residual coding bit-rate.

equal to the sum of the individual scores, merge the two models and loop back to the segmentation step using the updated GMMs. Continue until no pair is found.

More detailed descriptions of the method can be found in [1, 10]. Since speaker diarization is an unsupervised learning technique, the clusters need to be labelled with real speaker identities. This was performed by matching the speaker-turn segmentation extracted using individual sources, to the cluster with the longest speaking time by finding the lowest sum of absolute difference.

2.3 Motion Activity from Compressed Video

Some of the video processing used for compression can be reused in video analysis using computationally inexpensive methods [5]. Using compressed videos from the close-view cameras, we extracted the motion vector magnitude and the residual coding bit-rate, as an estimate of personal activity levels.

Motion vectors, as shown in Figure 3(a), are obtained essentially through block matching from motion compensation during video encoding. This can be interpreted as crude approximations of the optical flow [6], where their magnitudes indicate the degree of translational motion at each block location. For our dataset, we used an MPEG-4 encoder with a group-of-picture (GOP) size of 250 frames and a {I-P-P-...} structure where the first frame (I) is Intra-coded, and the rest (P) are predicted frames.

After motion compensation, the transform coefficients of the residual signal (the difference between the block to be encoded and its prediction from the reference frame) are quantized and entropy encoded. The residual coding bit-rate is the number of bits needed to encode the residual signal. While the motion vector captures gross block translation, it may fail to represent non-rigid motion such as the lips moving. However, the residual coding bit-rate captures such motion, since temporal change can be detected at finer spatial granularity resulting in a higher residual energy and resultant encoding rate, as shown in Figure 3(b). In combination with the motion vector magnitude, it provides complementary evidence for motion activity.

For each meeting participant, we would like to detect when they are in view, and also estimate activity levels without dynamic background clutter. To do this, we have implemented 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 [4]. The position of these skin-coloured blocks are then estimated and propagated for the subsequent P frames for the duration of the GOP structure using the motion vector information.

To detect when a participant is not in view, we threshold the number of visible skin-colored blocks with 2% of the total number of blocks in one frame. If the participant is visible, we measure their motion activity by using either the motion vector magnitude or residual coding bit-rate. To compute the normalized motion activity from the motion vector magnitude for participant i in frame t, we first calculate its average, $v_i(t)$, over the skin-colored blocks in each frame. For each participant in each meeting, we then find the median of the average motion vector magnitude over all frames where the participant is in view. Next, we compute the average of the medians, $\bar{v}$, for all the participants. Finally, the motion activity level from motion vector for participant i in frame t, $v_i^n(t)$, is computed by normalizing as follows:

$$v_i^n(t) = \begin{cases} \frac{v_i(t)}{2\bar{v}} & v_i(t) < 2\bar{v} \\ 1 & v_i(t) \geq 2\bar{v} \end{cases} \tag{1}$$

The motion activity level from residual coding bit-rate is also normalized in a similar fashion. Note that if a participant is not detected in a frame, they are assumed to be presenting at the slide screen, and is assigned an activity level of 1 for that frame for both the real and thresholded binary output features. To measure dominance, the motion activity features that we used were (i) motion activity level from motion vectors; (ii) motion activity level from the residual coding bit-rate; and (iii) the average motion activity from the motion vectors and residual coding bitrate.

3 Experiments

The experiments can be divided into three parts: simple audio features, speaking length using speaker diarization, and motion activity from the compressed domain. For the simple audio features, total speaking length and energy were used. From the motion activity features extracted from the compressed domain, we compared the performance of motion activity estimation based on motion vectors, the residual coding bitrate, and also their combination. For these cases, we generated both real-valued and thresholded binary measures of the motion activity for each participant. Table 1 shows the classification performance of the features.

Table 1 shows that total speaking length and energy generally perform well. They did not return the same output for each meeting, even though their overall performance was similar, indicating a possible complementary aspect of the two features. The total speaking length estimated from the speaker diarization shows that, while there are improvements to be made, the results are already competitive. Errors that occurred are due to estimation problems within the temporal overlap between

<table>
<tr><th colspan="2">Method</th><th>Classification Accuracy</th></tr>
<tr><td colspan="2">Total Speaking Length</td><td>85%</td></tr>
<tr><td colspan="2">Total Speaking Energy</td><td>82%</td></tr>
<tr><td colspan="2">Total Speaking Length Using Speaker Diarization</td><td>74%</td></tr>
<tr><td rowspan="3">Total Motion Activity: Real Values</td><td>Motion Vectors</td><td>55%</td></tr>
<tr><td>Bitrate</td><td>47%</td></tr>
<tr><td>Combination</td><td>53%</td></tr>
<tr><td rowspan="3">Total Motion Activity: Binary Values</td><td>Motion Vectors</td><td>59%</td></tr>
<tr><td>Bitrate</td><td>56%</td></tr>
<tr><td>Combination</td><td>62%</td></tr>
<tr><td colspan="2">Chance</td><td>25%</td></tr>
</table>

Table 1: Most dominant person detection results.

speaker turns.

Regarding the motion activity features, they performed less well compared to the other features, but the results were considerably better than chance (which would result in 25% classification accuracy). This indicates that these features also have discriminative power. While motion, bitrate, and their combination performed similarly, a closer look at the results revealed that there were some differences in which meetings the misclassifications occur. Also, in the meetings where all the features misclassified the most dominant person, the two biggest source of errors were: (i) false detection of when a participant is not in the close-up view due to skin-color detection errors; and (ii) that the most dominant person was not the one who moved the most.

Overall, it was interesting to observe that in some cases where the audio features failed the motion features were successful. This suggests that there may be some complementary aspect of the features which should be further investigated in the context of feature fusion.

The automated feature extraction process could also have been improved. For example, with the automated speaker diarization, we found that the number of clusters that was extracted did not always correspond to the number of meeting participants. This is unlikely to have affected the classification accuracy greatly since we only needed to find the cluster with the longest total speaking time. In addition, the gap in performance between the speaker segmentations using the headset microphones indicates that there are still inaccuracies in the diarization which could be improved. Furthermore, automatic assignment of the speaker clusters to named participants would be desirable. The motion activity estimation can also become inaccurate if hands enter the frame sporadically or if people are at the slidescreen, away from their seats.

4 Conclusion and Future work

Our work indicates that, for the 34 meeting segments in which the most dominant person was reliably decoded by all annotators, our investigated audio cues were able to classify the most dominant person with relatively good accuracy. Also, our compressed-domain video features performed less well but still provided some discrimination, especially when the motion vectors and residual bit-rate measures were combined. While the results using speaking length or energy perform quite well, we emphasise that only a subset of all possible meetings was selected and more work needs to be done in analysing more subtle group behaviours related to dominance.

This initial study serves as the starting point for further research on perceptual cues and models for dominance. We plan to reduce the complexity of the speaker diarization algorithm since it is computationally intensive. For the compressed domain video features, it may be possible to better estimate when someone is not in their seat by processing other camera views. Furthermore, other higher-level video features could be promising (e.g. gaze) and will be investigated. The issue of multimodal fusion is another open area.

Acknowledgments

This research was partly funded by the U.S. Government VACE program, the EU project AMIDA (pub. AMIDA-30), the Swiss NCCR IM2, and the German Academic Exchange Service (DAAD). We thank Bastien Crettol (IDIAP) for his support with data annotation.

References

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proc. IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[2] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. In *NIPS*, 2001.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Proc. MLMI*, 2005.

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

[5] S.-F. Chang. Compressed-domain techniques for image/video indexing and manipulation. In *Proc. IEEE ICIP*, pages 314–317, 1995.

[6] 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.

[7] 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.

[8] 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. ACM Press, 2006.

[9] J. Rosip and J. Hall. Knowledge of nonverbal cues, gender, and nonverbal decoding accuracy. *Journal of Nonverbal Behavior*, 28(4):267–286, December 2004.

[10] B. P. X. Anguera, C. Wooters and M. Aguilo. Robust speaker segmentation for meetings: The icsi-sri spring 2005 diarization system. In *Proc. of NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[11] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting Markov chains. In *NIPS*, 2005.