Predicting Remote Versus Collocated Group Interactions using Nonverbal Cues

Dairazalia Sanchez-Cortes[1,2], Dinesh Babu Jayagopi[1,2] and Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(dscortes,djaya,gatica)@idiap.ch

ABSTRACT

This paper addresses two problems: Firstly, the problem of classifying remote and collocated small-group working meetings, and secondly, the problem of identifying the remote participant, using in both cases nonverbal behavioral cues. Such classifiers can be used to improve the design of remote collaboration technologies to make remote interactions as effective as possible to collocated interactions. We hypothesize that the difference in the dynamics between collocated and remote meetings is significant and measurable using speech activity based nonverbal cues. Our results on a publicly available dataset - the Augmented Multi-Party Interaction with Distance Access (AMIDA) corpus - show that such an approach is promising, although more controlled settings and more data are needed to explore the addressed problems further.

Categories and Subject Descriptors

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

General Terms

Human Factors

Keywords

Remote meetings, Characterizing small groups, Nonverbal behavior

1. INTRODUCTION

Increasingly, teams in the workplace are expected to collaborate across different physical locations. The challenges involved in designing effective remote collaboration systems are many- communication infrastructure, human-computer interfaces, improving the awareness of the participants, etc. Often, the goal of such designs is (implicitly or explicitly) to make remote meetings as close as possible to face-to-face interactions [11].

ICMI-MLMI'09,Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing November 6, 2009, Cambridge, MA, USA.

The difference in the dynamics between collocated and remote meetings is significant and it can lead to poorer performance in distributed teams [9]. Some of the difficulties of the remote participants and the group as a whole in remote meetings include the inability to conduct side conversations, the challenge of occupying the floor because of the lack of eye contact, the inability to utilize posture shifts, and the phenomenon of in-room attendees forgetting about the remote people [7, 13].

Various technological approaches have been proposed to provide feedback to mitigate these differences. Some approaches allow real-time multimodal visualization of conversation analysis to improve the interactivity of group meetings [5, 4, 7].

Quantifying and measuring the difference in the group dynamics of collocated and remote meetings, using behavioral cues and more specifically nonverbal cues, has been done in different ways in the literature although not extensively. In [8], video conferencing systems using Integrated Services Digital Network (with transmission lags and poor quality video), LIVE-NET (with less transmission lags and high quality video) and face-to-face interactions were compared. The three cases were compared by studying backchannels, interruptions, turns, etc., obtained using manual annotation. In [11, 12] the difference between collocated and remote collaboration was captured by using real-time feedback through individual mobile phones. When groups in remote meetings (without feedback) had one or more dominant people, also had more speech overlap. In contrast, much more work on characterizing the groups meeting face-to-face has been done [10, 2]. The findings shows that various individual behaviors like dominance, status or personality and group behaviors like cooperation, competition, and interactivity could be predicted from limited observations (thin slices).

In this work we study two novel research questions, in the context of characterizing group dynamics in collocated and remote meetings. Firstly, can we distinguish between remote and collocated meetings using only nonverbal cues? and secondly, can we predict the remote participant in the remote meetings with the same type of observations?. The nonverbal cues we consider in this work are based on acoustic information (speech activity based).

Sections 2 discusses our approach. Section 3 introduces the experimental setup. Section 4 documents the results obtained, and Section 5 gives the conclusions of our analysis.

2. OUR APPROACH

Nonverbal cues, particularly audio based ones, are known

to contain useful information for social inference of both vertical and horizontal aspects [3]. We extract a number of nonverbal cues to characterize individual participants and the group as a whole, and then use them to learn classifiers for three different tasks (described in section 3).

Figure 1 shows the extraction process and the associated tasks. We used the binary segmentation (speech and

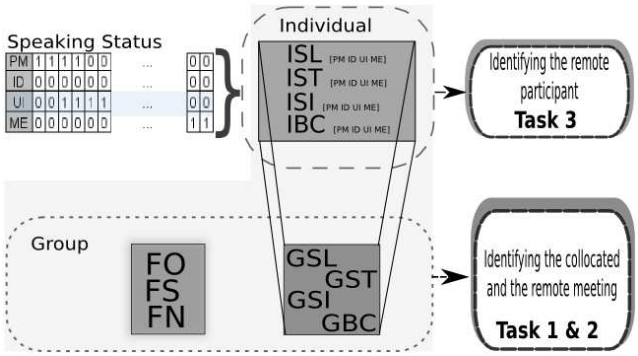

Figure 1: Our approach. Features are extracted, then classifiers are used for three tasks.

non-speech for each participant) available with the data corpus (described in Section 3). This is usually computed by thresholding speaking energy values or using more sophisticated algorithms to combat cross-talk. A turn is a continuous period of time for which the person's speaking status is 1. Then we compute features to characterize individuals and the group as a whole as follows (similar to the work in [2]).

Individual features From the speech segmentation, we compute the following features.

Speaking Length (ISL_i): Considers the total time that a participant i speaks according to his speaking status.

Speaking Turns (IST_i): It is the number of turns accumulated over the entire meeting for every participant i.

Successful Interruptions (ISI_i): The cumulative number of times that participant i starts talking while another participant j speaks, and i finishes his turn before j does, i.e. only interruptions that are successful are counted.

Backchannels (IBC_i): The cumulative number of times that participant i starts talking while another participant j speaks, and i finishes his turn before j does, i.e. only unsuccessful interruptions that are successful are counted and assumed to be backchannels (a simplified assumption).

Group features From the speaking status of all the participants, several features to capture the overlap, silence patterns of a group as whole were computed. Let T be the total number of frames in a meeting, S be the number of frames when no participant speaks, M be the number of frames when there is a monologue (i.e. only one person speaking), and O be the number of frames when more than one participant talks.

Fraction of Overlapped Speech : $FO = \frac{O}{T}$.

Fraction of Silence : $FS = \frac{S}{T}$.

Fraction of Non-overlapped Speech : $FN = \frac{M}{T}$.

Additionally we compute Group Speaking Length (GSL), Group Speaking Turns (GST) and Group Successful Interruptions (GSI) which are accumulated over all the participants from the ISL, IST, ISI, IBC.

3. EXPERIMENTAL SETUP

Dataset: The Augmented Multi-Party Interaction with Distance Access (AMIDA) corpus [6] consists of 10 hours of recorded, transcribed, and annotated four-person meetings recorded at the University of Edinburgh, see Figure 2. Recordings were gathered using 24 microphones (two-circular, eight-microphone arrays, four headset mics and four lapel mics), six cameras (four close-ups, two views of the room-center of the table and corner), and the output from a slide projector. There are three four-person meetings (for a total of 27-meetings) of which two have a remote participant (18-meetings). Figure 3 shows the scenarios of the AMIDA corpus.

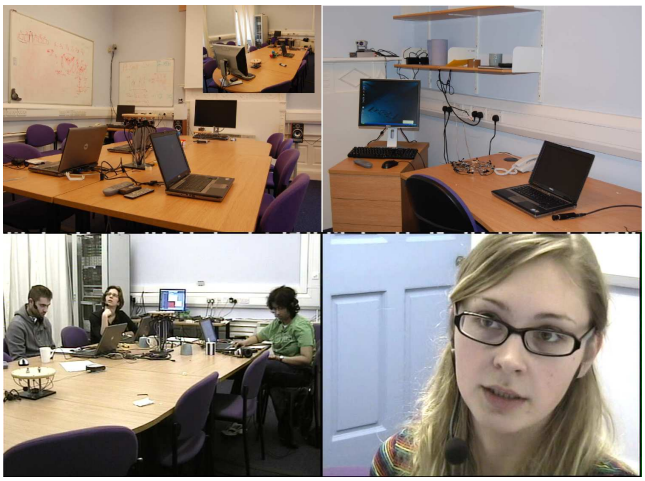

Figure 2: Top: The meeting room of collocated participants (left) and the meeting room of the remote participant (right). The desktop monitor in each of the rooms show the rest of the group members. Bottom: The meeting view that the remote participant (left) and the collocated participants (right) look at during the meetings.

Meetings: The 9 sets of participants in the AMIDA meetings are involved in the design of a remote control and meet three times.

Meeting A - New Project Start: In this meeting participants decide collectively on role allocation (who should do what), and discuss the aim of the project.

Meeting B - Conceptual Design: This meeting consists of individuals presenting their work and the group coming up with a conceptual design via videoconferencing.

Meeting C - Detailed Design: During this meeting participants present their individual work, and present and eval-

uate the clay prototype again via videoconferencing.

Every meeting has a participant with one of the following roles: project manager (PM), industrial designer (ID), marketing expert (ME) and a user interface designer (UD). In all the remote meetings the user interface designer is the remote participant (as shown in Figure 3).

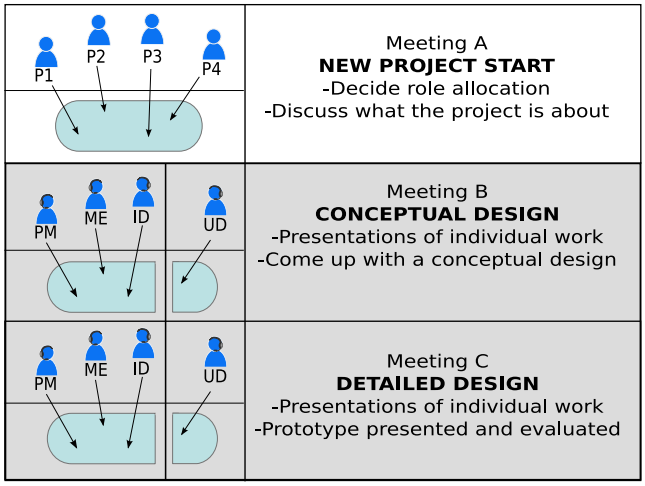

Figure 3: The scenarios of the AMIDA corpus. The user interface designer (UD) is the remote participant in B and C meetings.

We use these meetings to study the group interactions in remote versus collocated settings. The average duration for collocated (also called A-meetings) is 18.6 minutes; for remote meetings (called B and C meetings) the average duration is 37.7 minutes. While the collocated meeting was a pure discussion type meeting, the remote meeting had presentations followed by discussions. In order to have a fair comparison, we consider only the last five minutes of B and C meetings (which mostly were discussions) for our subsequent analysis.

Tasks: In order to model the difference between collocated and remote meetings, we define three tasks.

Task 1: The first task is to distinguish between collocated and remote meetings given the nonverbal behavior of the entire group. For this classification task, A meetings belong to the first class and, B and C meetings belong to the second class.

Task 2: The goal of the second task is to infer the collocated meeting given three meetings (1 collocated and 2 remote), where the participants are exactly the same. This task is simpler when compared to the first task.

Task 3: The third task is to predict the remote participant in the remote meetings.

4. RESULTS

Task 1: For the first task, we learnt a Gaussian mixture model using the Expectation Maximization (EM) algorithm [1] for each of the group features. Figure 4 shows cross-validation performance of this task with one and two Gaussians. For this task, the Group Speaking Turns (GST) with two Gaussians had the best performance (70%) which is slightly above random performance (66%), but the difference is not statistically significant (given that the size of the data set is small). Figure 5 suggests that while learning a global threshold to classify collocated and remote meetings might be difficult (due to inter-group variations), making a local decision (when looking at only the meetings with the same participants) would still be possible. The results of task 2 verify that this is true.

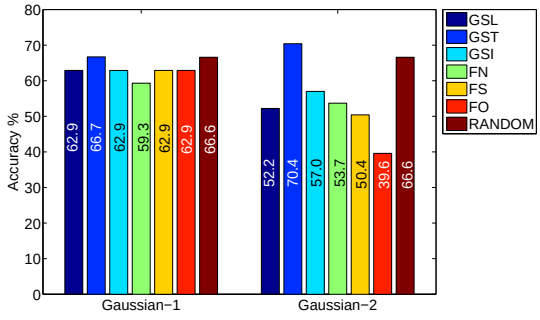

Figure 4: Performance of group features on predicting the collocated and remote meeting (Task 1).

Task 2: For the second task, we have only 3 meetings with the same participants (9 such sets). Therefore, we use a simple unsupervised approach to predict the collocated meetings - hypothesizing that collocated meetings have either the minimum or the maximum value of the group nonverbal cue. Figure 6 shows average performance for the various features. For this task the Group Speaking Turns (GST) again performed the best (81% accuracy), showing that the collocated meetings have lesser number of turns as compared to the remote meetings. One possible reason might be that because of the presence of a remote participant the group needs more turns per unit time to achieve their desired objective. Also, this result is statistically significant ($p = 0.01$) compared with random performance (33%). From Figure 5 we can observe that collocated meetings are the ones with less speaking turns.

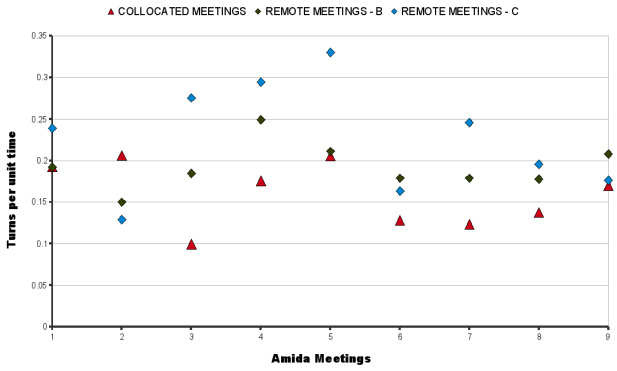

Figure 5: Group Speaking Turns for each of the 9 sets of AMIDA meetings.

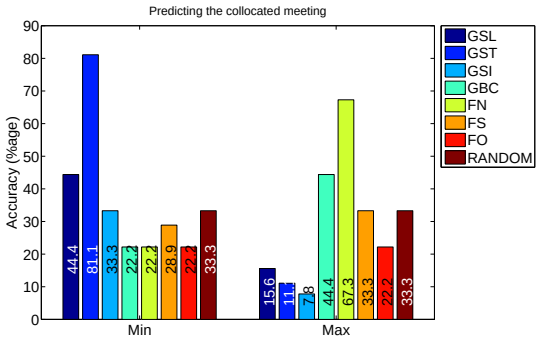

Figure 6: Performance of group features on predicting the collocated meeting (Task 2).

Task 3: For the task of predicting the remote participant in a meeting, we hypothesized that the remote participant has either the minimum or the maximum value of the individual features. Figure 8 shows the results. The minimum value of speaking length (ISL) is the individual feature that better predicts the remote participant (50% accuracy) higher but not statistically significantly, rather than random performance (25% accuracy). Figure 7 shows the average speaking length per role on remote meetings. We can see that the role with less speaking length is the user interface designer(UD) followed closely by the industrial designer(ID).

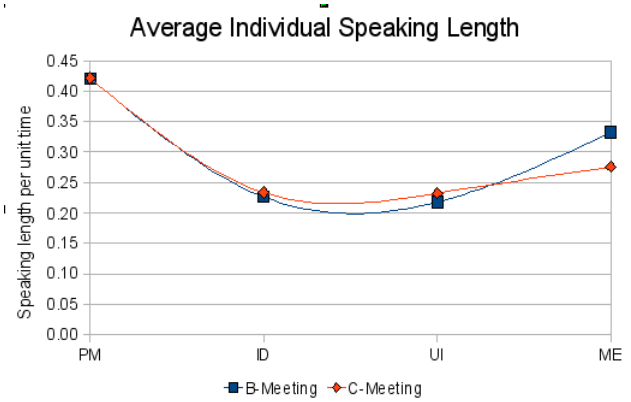

Figure 7: Average of Individual Speaking Length for each of the roles in remote meetings. The user interface designer is always the remote participant

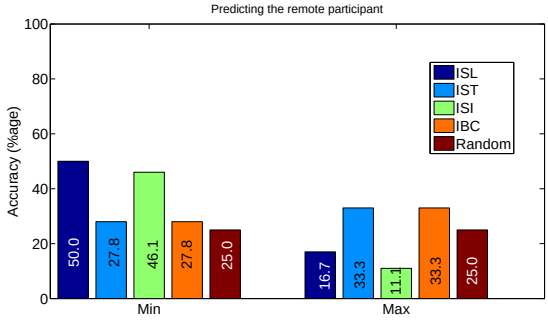

Figure 8: Performance of individual features on predicting the remote participant (Task 3)

5. CONCLUSIONS

In this work, we attempted to characterize the differences in group dynamics between collocated and remote meetings on a very recent data that is publicly available (the AMIDA corpus), by computing speech activity based nonverbal cues. We evaluated the effectiveness of these cues in predicting the remote versus collocated meetings, as well as the remote participant. Based on the results we noticed that collocated meetings have less turns and the remote participant talks less. Clearly, such observations are limited given the size of our dataset. Furthermore, it is noteworthy that the AMIDA corpus was neither designed nor collected with the task of classifying remote and collocated meetings in mind. Therefore the scenarios in these meetings make our tasks challenging. In the future, we would like to expand our dataset to study the generality of our findings, and to extract more relational and visual features.

Acknowledgments: This research was partly supported by the EU project AMIDA, the Swiss NCCR IM2, and through a doctoral scholarship from CONACYT-Mexico.

6. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] D. Jayagopi et al. Characterizing conversational group dynamics using nonverbal behavior. In *Proc. ICME*, New York, June 2009.

[3] J. Hall et al. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological bulletin*, 131(6):898–924, 2005.

[4] J. M. DiMicco et al. Influencing group participation with a shared display. In *Proc. CSCW '04*, pages 614–623, New York, NY, USA, 2004. ACM.

[5] K. Otsuka et al. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proc. ICMI'08*, Chania, October 2008. ACM.

[6] M. Lincoln et al. The amida data recording environment. In *Proc. Workshop MLMI*, 2007.

[7] M. Poel et al. Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. In *Proc. FG 2008*, pages 1–6. IEEE Computer Society Press, September 2008.

[8] O'Conaill et al. Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, 8(4):389–428, 1993.

[9] P. J. Hinds et al. Out of sight, out of sync: Understanding conflict in distributed teams. *Organization Science*, 14(6):615–632, Nov – Dec 1982.

[10] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, Special Issue on Human Behavior*, 1(12), December 2009.

[11] T. Kim, A. Chang, L. Holland, and A. Pentland. Meeting mediator: enhancing group collaboration with sociometric feedback. In *Proc. ACM '08: Conference on CSCW*, pages 457–466, 2008.

[12] T. Kim and A. Pentland. Understanding effects of feedback on group collaboration. *Association for the Advancement of Artificial Intelligence*, pages 25–30, 2009.

[13] H. J. A. op den Akker et al. Supporting engagement and floor control in hybrid meetings. In A. Esposito and R. Vich, editors, *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, Prague*, July 2009.