Who Will Get the Grant ?

A Multimodal Corpus for the Analysis of Conversational Behaviours in Group Interviews

Catharine Oertel
KTH
Royal Institute of Technology
Linstedtsviägen 44
Stockholm, Sweden
catha@kth.se

Kenneth A. Funes Mora
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
kfunes@idiap.ch

Samira Sheikhi
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
samira.sheiki@idiap.ch

Jean-Marc Odobez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
odobez@idiap.ch

Joakim Gustafson
KTH
Royal Institute of Technology
Linstedtsviägen 44
Stockholm, Sweden
jocke@speech.kth.se

ABSTRACT

In the last couple of years more and more multimodal corpora have been created. Recently many of these corpora have also included RGB-D sensors' data. However, there is to our knowledge no publicly available corpus, which combines accurate gaze-tracking, and high- quality audio recording for group discussions of varying dynamics. With a corpus that would fulfill these needs, it would be possible to investigate higher level constructs such as group involvement, individual engagement or rapport, which all require multimodal feature extraction. In the following paper we describe the design and recording of such a corpus and we provide some illustrative examples of how such a corpus might be exploited in the study of group dynamics.

Categories and Subject Descriptors

H5.3 [**Information Interfaces and Presentation**]: Group and Organisation Interfaces—*Theory and models*

Keywords

corpus collection; group dynamics; eye-gaze; involvement

1. INTRODUCTION

In recent years, there has been a growing interest in dyadic and multi-party communication analysis, with the main goals of designing computational models for non-verbal behavior recognition such as the identification of group dynamics [11], [6] and of person relationships [15]. Such models could be exploited for information retrieval and indexing (e.g. for fast browsing of meetings), or for devising the next generation of Human-Computer or Human-Robot interactions (HCI, HRI) systems that have a more social understanding of human behaviors.

To perform this research, more and more multimodal, multiparty corpora were created. This include the AMI meeting corpus [12], the SONVB dataset [13], the D64 corpus [14] or the IDIAP Wolf Database [8]. While most of these corpora are both multimodal and multiparty in nature, they vary a lot in their design, set-up, original goals and research tasks that can be addressed. For instance, the AMI dataset is a corpus of work meeting recordings consisting mainly of scenario driven meetings with 4 participants, as well as real meetings. As a result, participants' visual attention is divided between the whiteboard, the notepads, their interlocutors and others. This in turn has an effect on gaze-patterns with respect to turn-taking behaviour and makes tasks such as automatic detection of turn changes more difficult [9]. It also will alter the proportionate amount of mutual gaze which has been found to be a good predictor of group involvement [16]. On the other end, there are more unconstrained corpora such as the D64 corpus [14] and the IDIAP Wolf Database [8]. Both corpora are multi-modal and multiparty in nature and rich in group dynamics. However, both corpora are limited in that gaze-patterns of all participants cannot be deduced for considerate portions of the recording. The SONVB corpus is composed of dyadic real job interviews and intended to validate whether hireability can be predicted from non-verbal behavior. Automatic gaze annotation was shown to be possible, but limited to two visual targets (other person/looking away) [4]. In addition, due to the very distinct roles of the interviewer and interviewee, their gaze patterns are not symmetric.

With the KTH-Idiap Group-Interviewing corpus we aimed at creating a corpus which encompasses as many different

UM3I'14, November 16, 2014, Istanbul, Turkey.

ACM 978-1-4503-0652-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2666242.2666251.

dynamics on as many different levels as possible while still allowing for verbal and non-verbal feature extraction such as eye-gaze at all times, for all participants, during the entire duration of the recording. To this end we proposed to study and record group interviews in which several participants are jointly interviewed to get a grant, and which can be seen as an extension of job interviews from the dyadic to the multi-party case.

We propose a maximization of variation of group dynamics on three different levels. The first maximization of variation of dynamics is on the level of participants. Three people are interviewees and one person is the interviewer. We expect that verbal and non-verbal behaviour will be more similar between the participants than between the interviewer and any of the participants.

The second one is on the level of engagement. We recruited PhD students as interviewees and asked them to talk about their PhD projects with the assumption that they would have intrinsic motivation to both talk about their projects and exchange ideas with fellow PhD students. We hypothesized that intrinsic motivation would results in higher engagement in the conversation.

The third one is on the level of conversational dynamics. We designed the different tasks within the corpus in such a way that they would encompass both competitive as well as collaborative sections. The competitive sections of the corpus, are the ones in which the PhD students are asked to present their PhD projects and elaborate on the impact of society of their project. The collaborative section is the one in which they have to come up with a joined project.

In the following paper we mainly focus on describing the design and recording of the KTH-Idiap Group-Interviewing corpus. However, we also provide some illustrative examples of how the KTH-Idiap Group-Interviewing corpus might be exploited for the modeling of joined attention.

2. SCENARIO AND DESCRIPTION

The corpus was recorded with the purpose of providing a database for research into group interaction analysis. In this section, we will provide a full description of the scenario.

Motivation: Going through an interviewing process is something everybody goes through at least once in their life. During the interviewing process it is very important to present oneself in the best light possible and convince the interviewer that one is the most qualified person for the particular job. Depending on the interviewing culture it is important to be extraverted and to show off all skills without appearing to be too arrogant.

The group interview is a widely used technique in the social sciences [3]; also some institutions have been changing their interviewing process to extend from one person to several people being interviewed. In this way the interviewer gets an insight into the interviewees team building skills and also how much ideas the respective interviewees are contributing to the team.

The resulting dynamics can be very diverse. There are several possibilities of how the conversation could potentially unfold. The participants might pair up to show how well they can work in a team, or they might choose to work by themselves. These strategies might be dependent on the person's individual character as well as other influences such as culture and gender and liking of the other participants.

With the KTH-Idiap Group-Interviewing corpus we would like to shed some light onto these dynamics. We intended to control for both the degree of acquaintance and personality.

Participants: As motivated above, our scenario implements a group interview. To this end, each session consisted of four participants: an interviewer and three interviewees.

With two exceptions all participants were PhD students or Postdocs at the Royal Institute of Technology in Stockholm or the University of Stockholm. The interviewer was always played by a Post-Doc, while interviewees were played by PhD students.

Scenario: All participants were made aware of the interview goals: PhD students were told that the Postdoc's purpose in the interview would be to find out who would be the most qualified for a prestigious scholarship. They were told that the interviewer could either choose all of them, two, one or no one at all, which meant that if they were to collaborate well, all of them could be chosen.

There are five phases in the recordings:

- First phase: the three PhD students are left by themselves while all equipment is running. This is done in order to elicit spontaneous speech.
- Second phase: the actual interview starts. Each PhD student is asked to introduce himself in a couple of minutes. This is done in order to achieve a baseline of how participants speech sounds in a neutral condition.
- Third phase: each of the PhD students has to give an elevator-pitch for the respective PhD project.
- Fourth phase: each of the PhD student has to discuss the potential impact their PhD project could have on society.
- Fifth phase: all three students had to come up together with a suggestion for a joined research project.

In addition, there is a calibration phase prior to each recording, which consisted of these steps: i) A chessboard calibration pattern was moved in front of each Kinect/GoPro camera pair, in order to obtain their relative pose (c.f. Sec. 3.3) and; ii) the participant was requested to fixate at the RGB camera of the Kinect while rotating the head.

The last process is used to obtain samples of eye appearance associated to gaze directions. Although this strategy is limited by self occlusions and the diversity and speed of head movements, it could be used to adapt a gaze estimation model and track the participant's gaze during the interview.

3. DATA COLLECTION

In this section, we will describe the technical aspects of the KTH-Idiap Group-Interviewing corpus by explaining the set-up and sensors, and the methods used to synchronize and calibrate them.

3.1 Set-up

The interview set-up can be seen in Fig. 1. All four participants (the interviewer and the 3 interviewees, see next Section) were located symmetrically around a round table, and the following set of sensors per-participant were used to record their behavior:

Figure 1: The KTH-Idiap corpus

- **Close-talking mono-directional microphone.** Audio was recorded using close-talk condenser microphones with cardioid pickup patterns connected to a multi-channel audio interface sampling at 48 kHz, 16bits. They are intended to obtain high-quality speech data with minimal cross-talk.

- **Windows Kinect.** These consumer devices are used to capture visual (RGB) and depth (D) information, with the goal of allowing robust head and facial behavior analysis (see Section 5). They were positioned at around 0.8 meter from each participant, such that the field of view allowed for high-enough mobility. Near-mode was also enabled to sense depth data as close as 0.4m to the sensor.

- **GoPro camera.** GoPro cameras were used to record high-resolution visual data and complement the low-resolution of the Kinect data to perform for instance fine gaze estimation [5]. The GoPro cameras were tied to the Kinect sensor to ensure that their relative position remained fixed throughout the duration of the different recordings.

3.2 Synchronization

We synchronized the different modalities as follows. A device delivering a master audio signal (a sine wave) and turning a LED on was used as the main synchronization mechanism.

A multichannel audio interface allowed for the automatic synchronization between all mono-directional microphones. The sine wave was picked up by the mono-directional microphones as well as the microphones in each GoPro camera, which allowed for automatic audio and video synchronization in post-production (e.g. with PluralEyes 3 from Red Giant)

As a result, cross-audio synchrony and audio-GoPros synchronization was achieved. The Kinects were in turn synchronized with their respective GoPro videos using visual information, in particular, the observation of the LED turning ON, which served as a frame-level event for fine video alignment.

3.3 Calibration

Different levels of calibration are necessary to automatically process the data. First, each Kinect was calibrated as a stereo camera pair between the RGB and depth (D) sensors using Herrera's publicly available toolbox [7]. Then, a calibration pattern was used to obtain the GoPro pin-hole model intrinsic parameters and, moreover, the relative pose (extrinsics) between the GoPro and Kinect cameras.

The final needed calibration is the 3D pose of the Kinect sensors w.r.t. the world coordinate system (**WCS**). An extension of [4] was used. Notice the Kinect tilt can be retrieved by fitting a plane to the background wall's depth observations. The **WCS** is thus defined from a reference Kinect, correcting its tilt.

Each other Kinect's pose is obtained as follows: i) the tilt angle is obtained from the background wall's plane; ii) the roll angle is 0 whereas the yaw angle is defined by setup design (c.f. Fig. 1); iii) the Kinects are assumed to be at the same height (in a table); iv) finally, from manual measurements of distance between the Kinects, their translation along the table's plane was computed.

4. CORPUS

The corpus consists of five interactions of groups of four. Each interaction lasted for about an hour, which results in approximately 5 hours of recordings of multi-modal and multi-party data. In the following subsections we mention the annotations which are available for the corpus together with the extracted features and some description on the structure of the recorded interactions.

4.1 Gaze annotations

In order to aid the evaluation of the automatic gaze estimation, a small part of the corpus was manually annotated for eye-gaze. Gaze annotation was carried out on the frame level. We thus defined 4 different gaze targets for a given subject: each of the other 3 participants and the "other" class. This latter group encompasses all phenomena such as looking at the table, looking up, looking down or unfocusing.

4.2 Discourse level annotations

The corpus has been annotated for voice activity as well as different kinds of very short utterances. These utterances include backchannels, hesitations and fillers.

4.3 The Questionnaire

Two distinct sets of questionnaires were given out to participants. One for the interviewer and one to each participant. The questionnaire for the participant consisted of the Big Five Questionnaire [10].

The interviewer was asked to rate each interviewee separately at the end of the whole interviewing process on the following items which were presented on a 7 point likert scale and to make a final recommendation whether this person should be chosen to receive funding.

- was the interviewee a team-player
- was he/she interested
- was he/she fascinated by his/her own research
- was he/she interested in topics beyond his/her research

- was he/she able to sell the potential of his/her work for society
- was he/she capable of contributing ideas to the group

In addition to the questionnaire the interviewer was given precise instructions about the interviewing process.

The interviewer was told that his job during the interview was to find the person, who would not only excel in his or own specific field of research but who could also see beyond that. Someone with a passion and who might be able to benefit society as a whole at a later stage in his/her career. He was ask to look out for someone who can both work independently but who can also be a team player. In a nutshell someone who does not only have excellent technical skills but also extraordinary social and leading skills.

4.4 Corpus description

The corpus was recorded with the purpose of providing a database for research into group interaction analysis. In the following subsection, we describe and discuss in more details the structure of each of the 5 recordings.

As a measure of hierability or task success we asked the interviewer to assess the participants performance.

Each interaction followed the same protocol to allow for comparison across groups. Nevertheless, interactions were very diverse, and to give an idea about their content, we illustrate below some of the highlights for each of them.

Interaction 1.
Group composition. Besides the interviewer, it consists of 2 men (persons A and B) and 1 women (person C), all working in different institutions. Persons A and B have met before but do not know each other well, while person C has never met any of the other participants.

Participants have different backgrounds (computer science, linguistics, physio therapy). Person B and C score the highest in terms of extraversion amongst all participants.

Highlight. When person C describes her research project, person B shows sudden interest, asking more specific questions and describing in turn his own research. He initiates the idea of starting a collaboration together. Both of them then engage in an animated discussion about possibilities of collaboration including funding options.

At some stage the interviewer interrupts the exchange and explicitly encourages person A to talk about himself and his projects as he has remained silent throughout the whole exchange. The two other participants follow up by encouraging the third person and asking questions.

In the final phase, where participants have to come up with a joint project, both person B and person C try to incorporate person A into the project, but the later remains skeptical about the feasibility and does not offer ideas of incorporation by himself.

Outcome. Person B and C were chosen. This is the only example of an interaction where two participants got chosen.

Interaction 2.
Group composition. Participants are three men working in the same department. Person A and C being office mates working on the same project, while person B works on a related research topic but work on a different project. Person A scored higher than B and C in terms of extraversion.

Highlight. Person C was struggling with advertising and explaining his project. Person A tried to encourage the participant to talk more about his PhD. He gave cues and asked helpful questions. He also was helpful towards Person B, although not to the same degree. He advertised his research very animately. He showed that he had interests and ideas outside his own research and also tried to encourage the others to talk about respective other ideas.

Outcome. Person A got chosen.

Interaction 3.
Group composition. Participants were two men (A and C) and one woman (B), who did not know each other before. Person A was a fresh PhD student in computational biology. Person B was at the end of her PhD in linguistics and person C was half-way in her PhD in dialogue systems. Person A had the highest extraversion score.

Highlight. All participants performed similarly in the introduction part and research project statement. However, when discussing the impact on society of their research, person A got quite animated and volunteered to go first. He, however, only received polite feedback without much enthusiasm. Different from the other participants, he addressed a lot of ideas and problems which are outside the direct scope of the research but show how all their research might benefit society, and also tried to engage the other participants into a more philosophical discussion about their research.

Outcome. Person A got chosen.

Interaction 4.
Group composition. This interaction featured three men, who all knew each other. They are all doing their PhDs in the same field but with a different specialization. They all scored similarly low extraversion scores.

Highlight. All participants followed the instructions of the interviewer. However, their interaction remained quite limited, engaging in long monologues and only in very rare occasions asked each others questions or showed signs of interest, even in the interview phase related to the building of a collaborative project.

Outcome. None of the participants got chosen.

Interaction 5.
Group composition. There were 2 women (A and B) and one man (C), all working in the same department, knowing each other well, with the same research background. Person A and C had high extroversion scores while Person B did not

Highlight. All of them showed interest at each others research projects and also tried to come up with ideas on how to collaboratively work together. However, none of them stuck out particularly.

Outcome. None of the participants got chosen.

5. PRELIMINARY RESULTS

In this study, we want to investigate whether similar gaze patterns around turn-taking events in dyadic conversations can also be observed during multi-party conversation. These preliminary results build on and in part extend the study in [14] that reported gaze patterns in dyadic turn-taking. Studying gaze patterns in multi-party turn-taking is interesting as more and more research goes towards building multi-party dialogue systems. In order, however, to be able to build such systems it is important to understand and model gaze patterns in multi-party turn-taking.We are particularly interested in the gaze patterns of the observer, as a

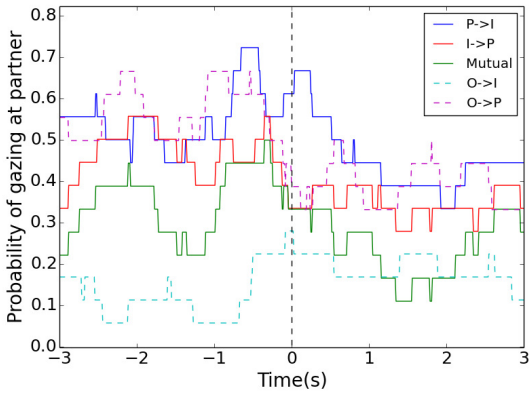

Figure 2: Gaze Probabilities for "Overlap with Backchannel"). "P" denotes previous speaker, "I" the incoming speaker and "O" the observer speaker

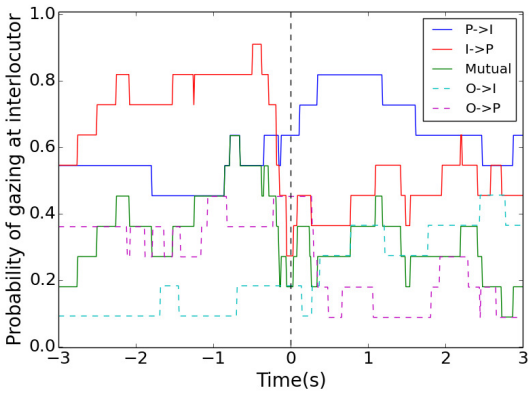

Figure 3: Gaze Probabilities for "Overlap with-Speaker Change". Here "P" denotes previous speaker, "I" the incoming speaker and "O" the observer speaker

description of the observer's gaze patterns is novel and could not be studied in [14].

In Figure 2 and 3 we present preliminary results on gaze patterns during two distinct turn taking events, namely "Overlap with Backchannel" and "Overlap with Speaker Change". The point 0 is understood as either the onset of a backchannel, or the onset of speech which leads to a speaker change.

We distinguish between "previous speaker" (P), "incoming speaker"(I) and "observer"(O). In the case of "Overlap with Backchannel" "previous speaker" is the speaker who holds the turn, whereas "incoming speaker" is the speaker who produces the backchannel and "observer" is the third interviewee who is not verbally involved in this exchange.

In the case of "Overlap with Speaker Change" "previous speaker" is the speaker who held the turn prior to the intervention of a second speaker. "Incoming Speaker" is the speaker who successfully grabs the turn and "observer" is the third interviewee who is not verbally involved in this exchange. For this preliminary analysis we exclude the interviewer from out analysis and only concentrate on the three interviewees.

For "Overlaps with Backchannels" in dyadic conversations a substantial increase in "previous speaker's" partner oriented gaze was observed. Our preliminary results indicate that the same is also true for multi-party conversations.

It is interesting to note that the probability that the observer looks towards the incoming speaker also increases prior to the production of the backchannel. This increase in the probability of the observer looking towards the incoming speaker might be a reaction to the previous speaker looking at the incoming speaker. In order to investigate this hypothesis however further more data will be needed.

For "Overlap with Speaker Change" in dyadic conversations it has been found that the previous speaker looks towards the "incoming speaker" while the "incoming speaker" averts his/her gaze before the speech onset of the "incoming speaker".

The same holds also for multi-party conversations. It is however interesting to note that the probability of the observer looking towards both the incoming speaker and the previous speaker increases. This might indicate that the observer is unsure who will take the turn. However, this hypothesis as well will need further investigation based on more data.

6. DISCUSSION

In this study we described the motivation for recording the KTH-Idiap Group-Interviewing corpus; a multi-modal corpus of group interviews.

The corpus we collected is different from other corpora in many ways, but perhaps its most salient characteristic is that participants were intrinsically motivated to talk, that it is rich in dynamics and that it has been recorded in such a way that both gaze and audio information can be extracted automatically. If one was to try and situate it in the landscape of already existing multimodal-corpora it is probably most similar to the SONVB corpus.

The corpus portrays the conversational behaviour of people interacting with up to 3 interlocutors (or 4 at some moments, if including the postdoc evaluator). The participants differ in gender, cultural background and extraversion but are similar in age and social status. They belong to the category of PhD students.

The KTH-Idiap Group-Interviewing corpus captures the conversational behaviour of people in both the listener as well as in the speaker role. It also portrays those people in less demanding conditions (e.g. during spontaneous condition and introductions) as well as more demanding situations (e.g. during the discussion on impact of research on society as well as collaborative condition).

A significant advantage of this corpus is that it is based on conversations of people being -in principle- intrinsically motivated to contribute to the conversation. This suggests a use for the corpus in the field of dialogue modeling that can not be obtained from structured task corpora. Moreover, the use of RGB-D sensors (Kinects) allows for the large scale study of three or even four–party gaze distributions and multimodal back channeling behaviour.

The spatial distribution of participants, their number, the multi-modality of the data and the dynamics of the interaction makes this corpus also a very valuable corpus for investigating the problem of automatic gaze coding or visual focus of attention (VFOA) estimation.

When considering only head pose for estimating the VFOA from head pose, it is possible to directly map the head poses to the visual targets, either by manual setting or using training data for learning the parameters [1]. However, these approaches are only applicable in static set-ups (as in this case) thus still requires user intervention.

As an alternative, other studies [17], which take inspirations of the head pose contribution in human gaze shift dynamics, directly provide an explicit mapping between the gaze directions needed to look at a given VFOA target and the head poses expected to be used for looking in that direction. These expected poses can then be used in a Hidden Markov Model decoder for estimating the VFOA states from sequence of head poses.

Furthermore, the decoding can also take advantage of the speaker-gaze pattern relationships (as priors) exhibited during conversation to improve VFOA recognition [2]. In this view, finer patterns, as those documented in the previous section could help improving the accuracy of such systems.

Yet, the data resolution of this corpus is also high enough to allow for actual gaze estimation, here understood as the continuous gaze direction within the 3D environment. Provided the gaze measurements it is possible, in theory, to infer the visual focus of attention from a geometric analysis of the gaze measurements (similar to [4]). Nevertheless, this strategy requires person-independent gaze models or the training of person-specific gaze models, where the latter is difficult to obtain, but expected to generate more accurate results.

All in all there are, however, also inherent limitations to the KTH-Idiap Group-Interviewing corpus. One inherent limitation lies in the relatively small number of group interactions. In order to generally investigate the effect of extraversion in group interviews, for example, further interactions would be needed.

One further limitation lies in the fact that the KTH-Idiap Group-Interviewing corpus has not been further controlled for the degree of acquaintance, culture, or gender. In order to draw conclusions about the impact of these variables, further recordings would be needed.

7. CONCLUSION

The advantage of the KTH-Idiap Group-Interviewing corpus over other corpus collections is that it was recorded using a range of sensors (close-talking microphones, high resolution cameras, RGB-Depth sensors - Kinect). This allows for the automatic retrieval of both eye-gaze as well as voice activity annotations which makes very expensive and time consuming manual annotations mainly superflucious. Due to the configuration of participants and sensors, the KTH-Idiap Group-Interviewing corpus allows for the fine grained analysis of multi-party, multi-modal turn taking behaviors manifested in for example eye-gaze patterns, head–nods and feedback tokens. The multi-modal analysis of turn-taking behaviour in turn is essential for the modeling of group involvement, individual engagement and joined attention.

Acknowledgments. The authors are thankful to everyone who participated in the recordings and in particular to Kalin Stefanov and Mattias Heldner for their invaluable advise and help in the set-up of the recordings. The authors would also like to acknowledge the support from the Swiss National Science Foundation(Project G3E, 200020_153085) `www.snf.ch`.

8. REFERENCES

[1] S. Ba and J. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2006.

[2] S. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 2008.

[3] J. H. Frey and A. Fontana. The group interview in social research. *The Social Science Journal*, 28(2):175–187, Jan. 1991.

[4] K. A. Funes Mora, L. S. Nguyen, D. Gatica-Perez, and J.-M. Odobez. A Semi-Automated System for Accurate Gaze Coding in Natural Dyadic Interactions. In *ICMI*, Sydney, Dec. 2013.

[5] K. A. Funes Mora and J.-M. Odobez. Geometric generative gaze estimation (G3E) for remote RGB-D cameras. In *Computer Vision and Pattern Recognition*, Ohio, June 2014.

[6] D. Gatica-Perez, I. McCowan, and S. Bengio. Detecting Group Interest-Level in Meetings. In *in IEEE ICASSP*, 2005.

[7] D. Herrera C., J. Kannala, and J. Heikkilä. Joint Depth and Color Camera Calibration with Distortion Correction. *in IEEE Trans. on PAMI*, 34(10):2058–2064, 2012.

[8] H. Hung and G. Chittaranjan. The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In *Proc. of the Int. Conference on Multimedia*, Firenze, Italy, 2010. ACM.

[9] M. Johansson, G. Skantze, and J. Gustafson. Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In *Social Robotics*, pages 351–360, 2013.

[10] O. P. John, L. P. Naumann, and C. J. Soto. Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O. P. John, R. W. Robins, and L. A. Pervin, editors, *Handbook of personality: Theory and research*, pages 114–158. NY: Guilford Press, New York, 2008.

[11] C. Lai, J. Carletta, S. Renals, K. Evanini, and K. Zechner. Detecting summarization hot spots in meetings using group level involvement and turn-taking features. In *INTERSPEECH*, 2013.

[12] I. McCowan, J. Carletta, and W. Kraaij. The AMI meeting corpus. In *Proc. Methods and Techniques in Behavioral Research*, pages 137–140, 2005.

[13] L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire Me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans on Multimedia*, 2014.

[14] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell. D64: a corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28, Sept. 2012.

[15] C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *International Conference on Multimodal Interaction*, 2013.

[16] C. Oertel, S. Scherer, and N. Campbell. On the use of multimodal cues for the prediction of involvement in spontaneous conversation. pages 1541–1544, 2011.

[17] S. Sheikhi and J.-M. Odobez. Investigating the Midline Effect for Visual Focus of Attention Recognition. In *Int Conf. on Multimodal Interaction (ICMI), Santa Monica*, Oct. 2012.