Visual attention, speaking activity, and group conversational analysis in multi-sensor environments

Daniel Gatica-Perez and Jean-Marc Odobez

1 Introduction

Among the many possibilities of automation enabled by multi-sensor environments - several of which are discussed in this Handbook - one particularly relevant is the analysis of social interaction in the workplace, and more specifically, of conversational group interaction. Group conversations are ubiquitous, and represent a fundamental means through which ideas are discussed, progress is reported, and knowledge is created and disseminated.

It is well known that spoken words represent a fundamental communication channel. However, it is also known from research in social and cognitive science - and experienced by all of us everyday - that nonverbal communication is also a key aspect in social interaction, through which a wealth of personal information, ranging from instantaneous internal states to personality traits, gets expressed, interpreted, and weaved, along with speech, into the fabric of daily interaction with peers and colleagues (Knapp and Hall, 2005). In this view, although it is clear that further work on speech and language processing is needed as important modules of ”smart” environments, the computational analysis of nonverbal communication is also a relevant domain and has received increasing attention (Gatica-Perez, 2008).

The coordination of speaking activity (including turn-taking and prosody) and gaze (i.e., visual attention) are key aspects of nonverbal communication. Gaze is an important nonverbal cue with functions such as establishing relationships (through mutual gaze), expressing intimacy, and exercising social control. Furthermore, the role of gaze as a cue to regulate the course of interaction, via turn holding, taking,

Daniel Gatica-Perez
Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne, Switzerland
e-mail: gatica@idiap.ch

Jean-Marc Odobez
Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne, Switzerland
e-mail: odobez@idiap.ch

or yielding, has been established in social psychology (Kendon, 1967; Goodwin and Heritage, 1990). Speakers use their gaze to indicate whom they address and to secure their attention (Jovanovic and Op den Akker, 2004), while listeners use their gaze towards speakers to show their attention and find appropriate times to request the speaking floor (Duncan Jr, 1972; Novick et al, 1996). It has also been documented that the coordination of speaking and looking acts correlate with the social perception of personality traits, like dominance, or of situational power like status (Exline et al, 1975). As we will discuss later, a multi-sensor environment could make use of all this knowledge to infer a number of important facets of a group conversation, with the goal of supporting communication even if some of the participants of the conversation were not physically present.

The chapter reviews the role of these two nonverbal cues for automatic analysis of group conversations. Our goal is to introduce the reader to some of the current technologies and envisioned applications within environments equipped with multiple cameras and microphones. We discuss the state of affairs regarding the automatic extraction of these cues in real conversations, their initial application towards characterizing social interaction patterns, and the some of challenges that lie ahead.

The structure of the chapter is as follows. In Section 2, we motivate the work in this field by discussing two application scenarios in the workplace, briefly reviewing existing work in these directions. In Section 3, we discuss some of the current approaches (with emphasis on our own work) to automatically measure speaking activity and visual attention. In Section 4, we illustrate the application of these technologies to the problem of social inference using a specific case study (dominance). Section 5 offers some concluding thoughts, pointing to open questions in the field.

2 Applications

2.1 On-line regulation of group interaction

The automatic modeling of nonverbal behavior has a clear value for building systems that support face-to-face interaction at work. As described in the introduction, speaking activity and visual attention play key roles in the establishment and evolution of conversations, but also on their outcomes. As one example, the social psychology literature has documented that dominant and high-status people often talk more, more often, interrupt others more, and are looked at by others more too (Burgoon and Dunbar (2006); Dunbar and Burgoon (2005), see more in 4). The effect of this type of behavior at work can be significant. For instance, through dominant behavior people might over-control the floor and negatively affect a group conversation where the ideas of others might be important but overlooked, for instance in a brainstorming meeting or as part of a learning process. It has been documented that people who hold the floor too much are perceived as over-controlling (Cappella, 1985).

Using findings in social psychology and communication about group interaction, recent work has beginning to investigate the automatic recognition of speaking activity and attentional cues as part of on-line systems that actively regulate and influence the dynamics of a group conversation by providing feedback to and increasing the awareness of the group members, in situations where a more balanced participation is key for effective teamwork. Some representative examples are summarized below (Gatica-Perez, 2008).

DiMicco et al (2004) proposed an approach that, in an on-line basis, estimates the speaking time of each participant from headset microphones and visualizes this information publicly, projecting the speaking time proportions on a wall or other large shared surface (Fig. 1(a)). The authors found that this simple feedback mechanism, which represents a minor cognitive load overhead, tends to promote a more even participation of the group members. DiMicco et al. have also explored other ways to display information about the relative participation of team members in an effective way. Bachour et al (Sept, 2008) have also explored a similar idea, but employing a translucent table where the conversation unfolds both as the sensing platform and the display device. Speaking activity for each participant is inferred through a three-microphone array implemented on a the table. A rough estimate of the proportion of speaking time of each participant is then displayed via LEDs located under the surface of the table, and that cover the participant's portion of the table (Fig. 1(b)). In a preliminary study, the authors found that people actually pay attention to the table, and that the displaying strategy can influence the behavior of group members who are willing to accept value the table display, both in the case of participants who might not contribute much, and of people who want to regulate their talking time with respect to the others', that is interested in a more egalitarian use of the speaking floor. Kim et al (2008) opted for a highly portable solution for both sensing and displaying of interaction (Fig. 1(c)). Group members wear a device (dubbed 'sociometric badge') that extracts a number of nonverbal cues, including speaking time and prosody, and body movement through accelerometers. These cues are used to generate a display of the group interaction on each person's cell phone. Evaluating their system on both collocated and remote meetings, the authors reported that this setting improved interactivity by reducing the differences between dominant and non-dominant people, and between co-located and remote participants, without being distractive.

The previous cases rely on speaking time to infer a need for regulation of interaction. Taking one step further, Sturm et al (2007) combined speaking activity and visual attention, building on previous work (Kulyk et al, 2006), and developed a system that automatically estimates speaking time from headset microphones and visual attention from headbands with reflective pieces tracked by infrared cameras. The system builds on the assumption that both cues are strong indicators of the state of a group conversation, and visualizes these cues for each participant aggregated over time on the conversation table (Fig. 1(d)). The system aims at regulating the flow of the conversation by facilitating individual participation.

The above research works are examples of the growing interest in developing systems that can promote better group communication through the automatic under-

standing of nonverbal behavior. Clearly, comprehensive studies about the efficiency and usability of these ideas and prototypes in the real world remain as open research issues. Nevertheless, they suggest that human-centered technology that influences group behavior starts to be feasible, and that such systems could be socially acceptable in multiple group interaction situations.

2.2 Assistance of remote group interaction

Meetings involving remote participants is another area where the better understanding of nonverbal communication and social dynamics can play a role to enhance the quality of interaction. Traditionally, research in this domain has focused on the enabling technological aspects of communication: increasing communication bandwith and/or signal compression, improving audio and video quality, or designing more efficient hardware solutions for tele-conferencing systems. However, in many settings, increasing the perception of presence of remote participants by enhancing the social aspects of communications could be useful (e.g. as is done with emoticons in emails). Let us consider for instance the following scenario: Bob is engaged in a distant meeting with a group of collocated people. He has an audio connection. During the discussion, Bob does not perceive well the nonverbal communicative cues indicating the reactions of the other meeting participants to propositions and comments. As he has no perception of the social behavior related to attention, addressing and floor control:

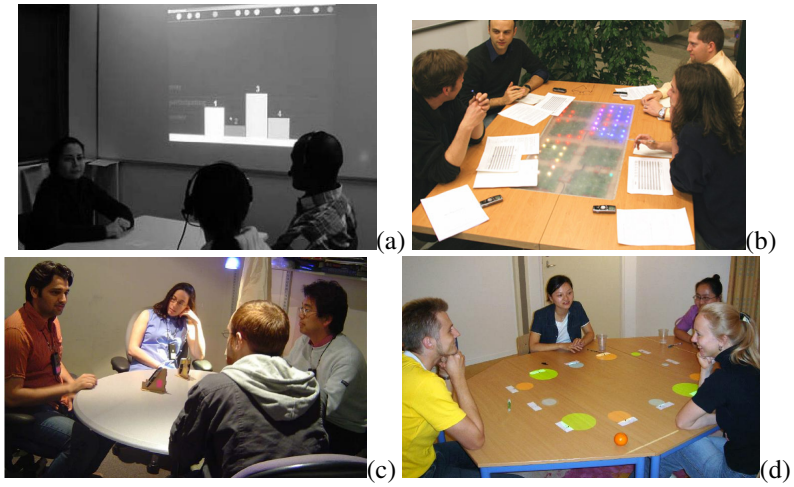

Fig. 1 Systems for on-line regulation of group interaction from speaking activity and/or visual attention: (a) DiMicco et al (2004); (b) Bachour et al (Sept, 2008); (c) Kim et al (2008); (d) Sturm et al (2007). All pictures are reproduced with permission of the corresponding authors.

- Bob cannot know well if somebody addresses him unless called by name;
- When trying to request the floor, Bob has no idea whether other people are doing the same.

After a few failed attempts such as interrupting others at a wrong place or not answering promptly, Bob tends to disengage himself from the meeting.

This scenario is one example which examplifies that, like in face-to-face meetings, taking or giving the floor must follow social protocols which to a large extent rely on the appropriate perception and use of nonverbal behavior. Gaze, body postures, and facial expression are codes by which interactors show whether they are attentive or not or ready to speak, and are also used to communicate reactions to propositions and comments (Kouadio and Pooch, 2002; Rhee et al, 1995). In remote meetings involving negotiations, the lack of perception of these codes was shown to lead to a misunderstanding of the other people's priorities and the delay of the achievement of an agreement (Rhee et al, 1995). Thus, in video conferencing situations, investigations have been conducted towards the enhancement of gaze perception, even at a weak level (Ohno, 2005), to favor user engagement.

There are several ways to improve remote communication through the analysis of gaze and turn taking patterns. In Bob's example, such an analysis would be useful to build remote assistants that would improve his participation and satisfaction in the meeting, by providing a better perception of presence and of the social behavior of the remote participants in the main room:

- Addressee assistant, which indicates whether the current speaker (or any other person) is looking at Bob;
- Interruptability assistant, which would indicate when Bob can safely interrupt the meeting in a non-disruptive way.

More generally, the goal would be to allow Bob to have a better perception of the conversation dynamics developing in the meeting room. Matena et al (2008) investigated such issues, by proposing a graphical interface for cell phones which represents the current communication situation in an instrumented meeting room by using the output of multimodal processing tools (Fig.2). In their study, the authors found that users were actually interested by the concept and that would use such display mainly for work meetings. Knowing who speaks to whom was the most liked feature.

When distant people have video access, another way to improve social perception is through automatic video editing. Many video conference and meeting recording systems are nowadays equipped with multiple cameras capturing several aspects of the meeting place, from close-up views of the participants to a panorama view of the whole space. The goal of an automatic video editing method is to combine all these streams into a single one which would convey the dynamics of the conversations and attract the attention of the viewers. This is performed through video cutting, i.e. by defining camera switching rules. Most of the current approaches rely on voice switching schemes and shot transition probabilistic models estimated from TV programs, i.e. they mainly show the main speaker and avoid too small or too long shots

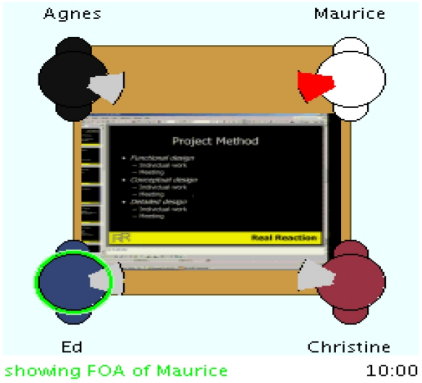

Fig. 2 Example of a cell phone display for a remote participant, showing the current slide being displayed in the main meeting room, and conveying information about the discussion dynamics in the meeting room through visualisation of automatically extracted cues. The head orientation of participants in the main room is represented by a sector in front of their avatar, and this sector is colored in red when a person is speaking (e.g. Maurice in the above example). The person being the current focus of the participant selected by the cellphone user for display (Marice in the example) is marked by a green circle. From (Matena et al, 2008), reproduced with permission of the authors.

to ensure variability. Such models, however, usually fail to provide viewers with contextual information about the other attendees. To address these issues, Takemae et al (2005) proposed alternative gaze-based switching schemes. They showed that in three-person meetings, selecting the view of the person which is mostly looked at was the most effective way of conveying the conversation, and particularly of conveying addressee responses to a speaker's utterance. This work followed the idea that people monitor their gaze towards others to capture their emotions, reactions, and intentions and as such, people receiving more visual attention (even when not speaking) might be more important than others with respect to the conversation. As another exemple, the study in Ranjan et al (2008) showed that using the speaker gaze as switching cue when the speaker holds a long turn, or the majority gaze when multiple people speak, were also useful in conveying the conversational aspects of meetings.

3 Perceptual Components: Speaking Turns and Visual Attention

3.1 Estimating Speaking Turns

The extraction of speaking turns answers the question "who spoke when", i.e. itdetermines for each speaker the periods of time when he is silent and the periods when he is speaking, as illustrated in Fig. 3. This is one of the most fundamental tasks in many speech related applications. This segmentation problem is also called diariza-

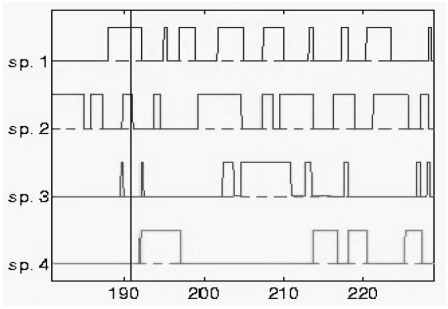

Fig. 3 Speaker turn segmentation of four people in a meeting. The x-axis denotes time. On the y axis, the speaking status (0 or 1) of each speaker (sp.1 to sp. 4) is displayed.

tion, when one has to automatically infer the number of people and their speaking turns in an unsupervised manner. Different techniques have been proposed in the past, which depend on the type of audio signal which is processed, like broadcast news, telephone speech, or dictations systems to name a few examples.

The segmentation of spontaneous speech in meetings is challenging for several reasons. Speech utterances can be very short (i.e. they are sporadic), which makes temporal smoothing techniques difficult. Furthermore, people tend to talk over each other, which generates overlapped speech signals. In (Shriberg et al, 2001), it was shown that such overlap could represent up to 15% of the data. In addition, the presence of concurrent audio signals, such as body noise from the participants (body motion, breathing, chair motion,...) as well as machine noise (fans, projectors, laptop) will be mixed with the speech signal. This will affect the acoustic features used to identify a given speaker.

One approach for speaker segmentation is to use intrusive solutions, i.e. to request each participant to wear a close-talk microphone (lapel near the throat, or headset near the mouth). This allows to know precisely when each speaker is talking, because there is no need to identify *who* is speaking, but only detect *when* the person wearing the microphone speaks. The problem reduces to a Voice-Activity-Detection (VAD) issue (Ramírez et al, 2007), where the problem is simplified since the speech signal of the person is much cleaner than other audio signals due to the distance. The simple thresholding of the short-term energy can provide good results. Nevertheless, in lively conversations, there might still be a non negligeable presence of high energy cross-talk from close by participants (esp. when using lapels), as well as breathing and contact noises. To address this issue, Wrigley et al (2005) used learned Gaussian Mixture model (GMM) classifiers to identify different audio configurations (e.g. speaker alone or speaker plus cross-talk) for each person in the meeting, and performed the multi-channel joint speech segmentation of all people. They obtained segmentation accuracy up to 96% on a large database of meetings. Dines et al (2006) proposed an approach using a discriminant multi-layer perceptron for phoneme plus silence classification, and reported even better results (less than 2%) on similar head-set data.

To reduce the constraints on meeting participants and allow more mobility freedom, microphone arrays (also known as multiple distant microphones (MDM)) can be employed. MDM signals contain spatio-temporal redundancy which can be used for speaker turn extraction. One common way to exploit this redundancy is to produce a single audio stream out of the multiple channels using beamforming techniques, with the benefit of enhancing the audio signal coming from the most active spatial location (i.e in principle the current speaker place) in the room. Diarization techniques which have been developed for other contexts, e.g. broadcast news (Valente, 2006), can be applied to this generated stream. However, adaptation of such techniques to meetings can be difficult due to the short utterances encountered in meetings, the high level of overlapped speech which makes it difficult to build precise acoustic models of the speakers, and the difficulty to detect and identify the presence of several speakers in a single audio stream.
When the geometry of the array is known, the location of the speakers can be extracted. Under the assumption that people remains seated at the same place, the detection of speech activity coming from specific space regions can be used to solve the diarization problem, and provide more accurate results, esp. on overlapped speech, as demonstrated by Lathoud and McCowan (2003). More generally, when moving people are involved, MDM can be used to perform multi speaker tracking. Fig. 4 presents a possible approach to this problem. Alternatively, the use of the time difference of arrival (TDOA) between each channel pairs can directly be exploited as additional features to the acoustic ones. Vijayasenan et al (2008) demonstrated within an information bottleneck diarization framework that this could lead to a reduction of speaker classification error from around 16% to less than 10%. In addition the direct use of the TDOA features presents the advantage to be exploitable even if the array geometry is unknown.

In summary, it is feasible to extract quite reliably and precisely speaker turns in meeting rooms, where the precision level mainly depends on the intrusiveness of the solution. The required level of precision depends on the applications. For conversational analysis current solutions are usually good enough. Still, analysis of social interactions and behaviors, which can rely on subtle cues such as short backchannels or patterns of interruptions will benefitiate from improvements in speaker turn extractions.

3.2 Estimating Visual Focus of Attention

Recognizing the gaze or visual focus of attention (VFOA) of people in meetings answers the question “who looks at whom or what”, i.e. which visual target the gaze of a person is oriented at. Traditionaly, gaze has been measured using Human-Computer Interaction (HCI) techniques. Usually, such systems require either invasive head mounted sensors or accurate estimates of the eye features obtained from high-resolution iris images (Morimoto and Mimica, 2005). These systems would be very difficult to set up for several people in a room, and would interfere with natural

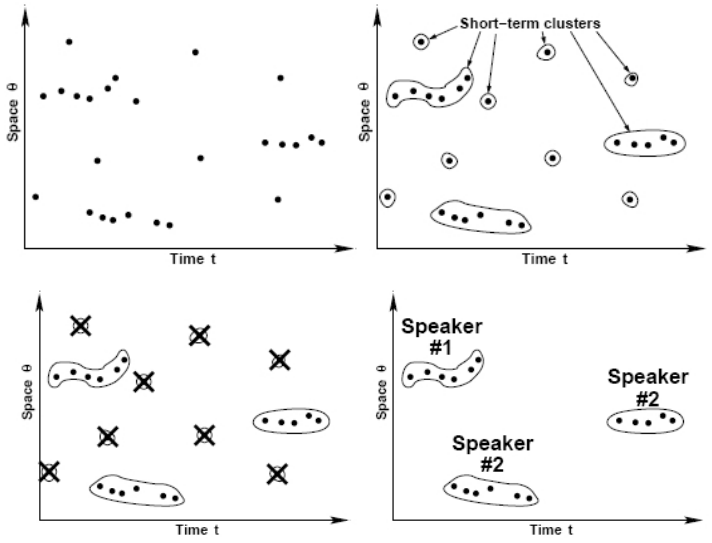

Fig. 4 Example of a processing chain for turn tracking using microphone array. First line: acoustic sources are first detected and localized in space. These are then grouped by spatio-temporal proximity. Second line: non speech sources are discarded. Finally, segments are automatically clustered into individual speakers. Taken from (Lathoud, 2006).

conversation, by restricting head mobility and orientation. As an alternative, head pose can be used as a proxy for gaze, as supported by psychovisual evidence showing that people do exploit head orientation to infer people's gaze (Langton et al, 2000). This was empirically demonstrated by Stiefelhagen et al (2002) in a simple setting with 4 people having short conversations. Social findings on gaze/speaking turn relationships were also exploited for VFOA estimation. Using people speaking status as another information source to predict a person's VFOA, Stiefelhagen et al (2002) showed that VFOA recognition could be slightly increased w.r.t. using the head pose alone. Moving a step further, Otsuka et al (2006a) proposed an interesting framework by introducing the notion of conversation regimes driving jointly the sequence of utterances and VFOA patterns of all people (for exemple a 'convergence' regime towards a speaker). Regimes were estimated along with people's VFOA, and their experiments with 4-participants informal short conversations showed that people VFOA could be reliably estimated and that the conversational regimes matched well with some addressing structures and dialog acts such as question/answers.

At work, most meetings involve the use of documents or laptops put on tables as well as projection screens for presentations. Studying conversations and gazing behaviours in these conditions is more complex than in meetings where people are only conversing (Stiefelhagen et al, 2002; Otsuka et al, 2005, 2006a). The addition of new targets, such as the Table or the slide screen leads to an increase of head pose ambiguities (i.e. the same head pose can be used to focus at different VFOA targets),

making the recognition of VFOA from head pose a difficult task (Ba and Odobez, 2008b). Introducing conversation context like the regimes in Otsuka et al (2006a) is thus necessary to reduce ambiguities. However, the presence of artifacts (laptops, slide screen) needs to be taken into account, as it significantly affects the gaze patterns. This is the *situational attractor* hypothesis of Argyle and J.Graham (1977): objects which play a central role in a task that people are doing attracts the visual attention, thereby overruling the trends for eye gaze behaviour observed in 'direct' human-human conversations. Such impact of artifacts on gaze behaviour was reported in task-based meetings (Chen et al, 2005). For instance, when a new slide is displayed, the 'convergence' regime towards the speaker is no longer observed, since people look at the slide, not the speaker.

To further ground the discussion on VFOA recognition issues, in the following we will describe our work with more details. More precisely, we address the joint recognition of people VFOA in task-based meetings involving table and slide as potential targets, using meeting contextual models in a DBN framework. We first introduce the setting and the different variables involved in our modeling (VFOA, observations, conversational events) (Subsection 3.2.1), and then present our DBN model, detailing qualitatively some of its main components (Subsection 3.2.2). Finally, we report some recognition results, comparing with the existing literature and discussing issues and ideas for future work.

3.2.1 Experimental context and cue extraction

In this section, we set the experimental problem, and then introduce the different variables involved in the modeling of the conversation patterns.

Set-up and task: Our source of data is the AMI corpus[1]. In this corpus, the recorded meetings followed a general task-oriented scenario: four persons having different roles were involved in the creation and design of a new television remote control. Fig. 5 shows the physical set-up of the meeting room. Amongst the different sensor recordings which are available in the corpus, we used the following: the video streams from the two cameras facing the participants (Fig. 5 bottom images) and of a third camera capturing a general view of the meeting room (Fig. 5, top left image). As audio sensors, we used the close-talk microphones.

Our main objective is to estimate people's VFOA. Thus, we have defined the set of interesting visual targets for a given participant seated at seat k, denoted $\mathscr{F}_k$, as comprising 6 VFOA targets: the 3 other participants, $\mathscr{P}_k$, (for example, for seat 1, $\mathscr{P}_1 = \{$seat2,seat3,seat4$\}$), as well other targets $\mathscr{O} = \{$Table, Slide Screen, Unfocused $\}$. The later target (Unfocused) is used when the person is not visually focusing on any of the previously cited targets.

[1] www.idiap.ch/mmm/corpora/ami

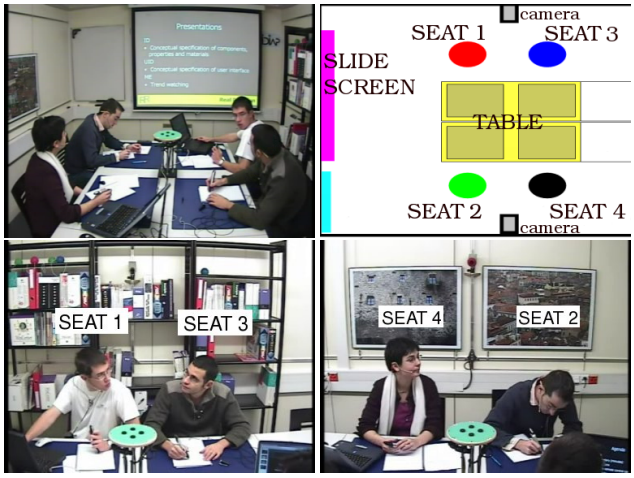

Fig. 5 Meeting recording setup. The first image shows the central view that is used for slide change detection. The second image shows a top view of the meeting room to illustrate seats and focus target positions. Seat numbers will be used to report the VFOA recognition results of people seated at these locations. The third and fourth images show the side cameras that are used for tracking people's head pose.

Fig. 6 Head pose extraction. The head localization (position, scale, in-plane rotation, cf left image) and head pose (discrete index indentifying a specific out-of-plane rotation, cf right image) are jointly tracked (taken from (Ba and Odobez, 2008b)).

Below, we describe the three types of features we used as input to our model: audio features describing people's speaking status, the people head orientations, and a projection screen activity feature.

Speaking status of participants (1 if the person speaks, 0 otherwise) were obtained by thresholding the energy signal of his close-talk microphone. To obtain more stable features, we smoothed the instantaneous speaking status by averaging them over a temporal window of 5 seconds, leading to the audio features s_t^k describing the proportion of time participant k speaks during this interval.

Head pose is the most important cue to recognize people VFOA. The head pose of

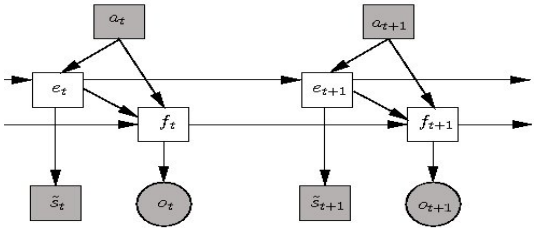

Fig. 7 Dynamic Bayesian Network graphical model. Squares represent discrete variables, circles represent continuous variables. Observations are shadedand hidden variables are unshaded.

a person was estimated by applying an improved version of the tracking method described in Ba and Odobez (2005) to our mid-resolution head video sequences (center images of Fig. 5). The approach is quite robust especially because head tracking and pose estimation are considered as two coupled problems in a Bayesian probabilistic framework. More precisely, the *joint* tracking of the head location (position, scale, in-plane rotation) and pose (represented by a discrete index denoting an element of the out-of-plane head pose set, as shown in Fig. 6) was conducted.
Texture (output of one Gaussian and two Gabor filters) and skin color head appearance models were built from a training database, and used to evaluate the likelihood of observed features. An additional pose-independent likelihood model based on background subtraction feature was used to provide better localization information and reduce tracking failures. More details on models and estimation procedure can be found in Ba and Odobez (2005).

A slide screen activity feature a_t characterizing the degree of involvements of participants into a slide-based presentation on the gaze patterns. Intuitively, the effect will be more important when a slide has just been displayed, and will decrease as time goes by. Thus, we used as slide activity feature a_t the time that elapsed since the last slide change occurred [2]. To detect the slide changes, we applied a very fast compressed domain method proposed by Yeo and Ramchandran (2008) to the video stream from the central view camera (see Fig 5).

Conversational events: to model the interactions between people's VFOA and their speaking statuses, we follow a similar approach to Otsuka et al (2005). We introduce a set $\mathscr{E} = \{E_i, i \in 1...16\}$ of 'conversational events' uniquely characterized by the set of people currently holding the floor. Since we are dealing with meetings of 4 people, the set comprises 16 events which can be divided into 4 types: silence, monologue, dialogue, or discussion.

[2] A slide change is defined as anything that modifies the slide-screen display. This can correspond to full slide transitions, but also to partial slide changes, or to switch between a presentation and some other content (e.g. the computer desktop).

3.2.2 Joint VFOA and conversational event DBN model

To estimate the joint VFOA of people, we rely on the DBN model displayed in Fig. 7, and according to which the joint distribution of all variables is given by:

$$p(f_{1:T}, e_{1:T}, \lambda, o_{1:T}, \tilde{s}_{1:T}, a_{1:T}) \propto \\ p(\lambda) \prod_{t=1}^{T} p(o_t|f_t) p(\tilde{s}_t|e_t) p(f_t|f_{t-1}, e_t, a_t) p(e_t|e_{t-1}, a_t), \quad (1)$$

where $f_t = (f_t^1, f_t^2, f_t^3, f_t^4)$, $o_t = (o_t^1, o_t^2, o_t^3, o_t^4)$ and $\tilde{s}_t = (\tilde{s}_t^1, \tilde{s}_t^2, \tilde{s}_t^3, \tilde{s}_t^4)$ denotes respectively the joint VFOA of all participants, their head poses and their speaking proportion, and λ represents a set of model parameters. This DBN expresses the probabilistic relationships between our random variables and reflects the stochastic dependencies we have assumed. In the current case, one important assumption is that the conversational event sequence is the primary hidden process that governs the dynamics of gaze and speaking patterns, i.e. gaze patterns and people utterance are independent given the conversational events. Indeed, given a conversational event, who speaks is clearly defined. At the same time, since people tend to look at the current speakers, the conversational event will also directly influence the estimation of people gaze through the term $p(f_t|f_{t-1}, e_t, a_t)$, which increases the prior probability of looking at the VFOA targets corresponding to the people who are actively involved in the current conversational event. This prior, however, is modulated by the duration a_t since the last slide change to account for the impact of slide presentation on gaze patterns. In the following, we introduce all the terms appearing in Eq. 1, and describe qualitatively their impact on the model.

The Conversational events dynamics is assumed to factorize as $p(e_t|e_{t-1}, a_t) = p(e_t|e_{t-1})p(e_t|a_t)$. The first term was modeled to introduce temporal smoothness on the event sequence. The second term, learned from training data, introduced prior knowledge about the type of conversational events given the slide activity. Essentially, it was observed that some seconds after a slide change, monologues occur slightly more frequently than in a normal conversation.

The VFOA dynamics is assumed to have the following factorized form:

$$p(f_t|f_{t-1}, e_t, a_t) \propto \Phi(f_t) p(f_t|f_{t-1}) p(f_t|a_t, e_t) \quad (2)$$

where the three terms are described below.

The shared focus prior $\Phi(f_t)$ is a group prior which favors joint VFOA states for which the number of people looking at the same VFOA is large, thereby reflecting statistics collected from our data which showed that people look more often at the same focus than if they would behave independently.

The joint VFOA temporal transition term assumes conditional independance of people VFOA given their previous focus, i.e. $p(f_t|f_{t-1}) = \prod_{k=1}^{4} p(f_t^k|f_{t-1}^k)$. Individual

VFOA transitions were defined to enforce temporal smoothness, with high probability to keep the same focus and lower probability to transit to other focus.

The VFOA meeting contextual priors $p(f_t|a_t,e_t)$ is the most interesting part. It models the prior of focusing at VFOA targets given the meeting context (conversational event, slide activity). We first assumed the conditional independance of people VFOA given the context, i.e. $p(f_t|a_t,e_t) \propto \prod_{k=1}^{4} p(f_t^k|a_t,e_t)$, and then defined the prior model $p(f_t^k = l|a_t = a, e_t = e_i)$ of any participant by following the intuition that this term should establish a compromise between the known properties that: (i) people tend to look at speaker(s) (ii) during presentations, people look at the projection screen (iii) speaker's gazing behaviour might be different than listener's one. This was done by defining this term as:

$$p(f_t^k = l|a_t = a, e_t = e_i) \approx g_{l,ivs,e_i}(a) = \vartheta_1 e^{-\vartheta_2 a} + \vartheta_3.$$

where *ivs* denotes the involvement status of person k (speaker or listeners) into the conversational event e_i, and the set of parameters $\vartheta_j,\ j = 1,...,3$ (one set for each configuration (l, ivs, e_i)) were obtained by fitting the training data.
Fig 8 illustrates the estimated model when the conversational event is a dialog. For the 'table' target, we can see that its probability is always high and not so dependent on the slide context a. However, whether the person is involved or not in the dialog, the probability of looking at the projection screen right after a slide change is very high, and steadily decreases as the time since last slide change ($a_t = a$) increases. A reverse effect is observed when looking at people: right after a slide change, the probability is low, but this probability keeps increasing as the time a increases as well. As could be expected, we can notice that the probability of looking at the people involved in the dialog is much higher than looking at the side participants. For the later target, we can notice a different gazing behaviour depending on whether the person is involved in the dialog or not: people involved in the dialog focus at the side participants, whereas a side participant almost never looks at the other side participant.

Observation models relate the hidden variables to the speaking statuses and head orientation observed data.

The speaking observation models assumes that, given the conversationnal event, the proportion of time that people speak are independent and can be set *a priori*. For instance, if E_j represents a monologue by person 2, we can define that person 2 will speak around 95% of the time during the predefined window size, while other participant utterances are assumed to correspond to backchannel, and will thus speak only 5% of the time.

The head pose observation model is the most important term for gaze estimation. Assuming again conditionnal independence ($p(o_t|f_t) = \prod_{k=1}^{4} p(o_t^k|f_t^k)$), the head pose distribution when person k looks at target i was modelled as a Gaussian:

$$p(o_t^k|f_t^k = i) = \mathcal{N}(o_t^k, \mu_{k,i}, \Sigma_{k,i}). \quad (3)$$

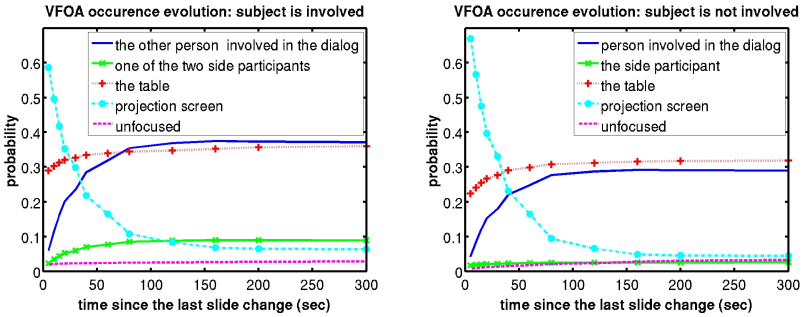

Fig. 8 Fitted contextual prior probabilities of focusing to a given target (occurence probability), in function of time a since last slide change, when the conversational event is a dialog, and for: Left: a person involved in the dialog. Right: a person not involved in the dialog. Taken from (Ba and Odobez, 2008a).

with mean $\mu_{k,i}$ and $\Sigma_{k,i}$. Setting these parameters is indeed not a trivial task, especially for the means. While researchers have shown that there exists some *correlation* between the gazing direction and the head pose, this does not mean that the head is usually oriented in the gaze direction. Indeed, relying on findings in cognitive sciences about gaze shift dynamics (Freedman and Sparks, 1997; Hayhoe and Ballard, 2005), we defined a gazing model (Odobez and Ba, 2007) which exploits the fact that, to achieve a gaze rotation towards a visual target from a reference position, only a constant proportion κ of the rotation is made by the head, while the remaining part is made by the eyes. This is illustrated in Fig 9. In other words, if μ_k^{ref} denotes the reference direction of person k, and $\mu_{k,i}^{gaze}$ the gaze direction associated with the VFOA target i, the head pose mean $\mu_{k,i}$ can be defined by:

$$\mu_{k,i} - \mu_k^{ref} = \kappa(\mu_{k,i}^{gaze} - \mu_k^{ref}). \tag{4}$$

where the proportion κ is usually around 0.5. Grossly speaking, the reference direction can be defined as looking straight in front of the person. This also corresponds to the direction which roughly lies at the middle between the VFOA target extremes, as people orient themselves to be able to visually attend all VFOA of interest.

Model parameters adaptation and Bayesian inference. In the Bayesian framework, we can specify some prior knowledge about the model by defining a distribution $p(\lambda)$ on the model parameters. Then, when the observations are given, during inference, optimal parameters of the model are estimated jointly with the optimal VFOA and conversational event sequences. In our work, we showed that this was important for parameters involving the VFOA state variables and in particular for $\mu_{k,i}$ since we observed large variations between the head pose means computed empirically using the VFOA ground truth data for different people. These variations can be due to either people specific ways of looking at a given target (people turn more or less their head to look at targets), or the introduction of a bias in the esti-

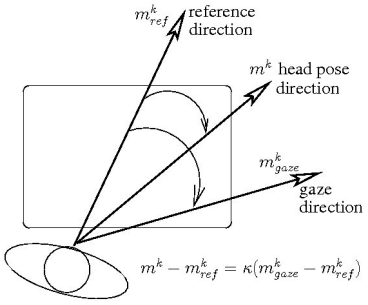

Fig. 9 Head pose-VFOA relationship. Assuming that the participant has a prefered reference gazing direction (usually in front of him), the gaze rotation towards the VFOA target is made partly by the head and partly by the eye.

mated head pose by the head pose tracking system. The inference problem consists of maximizing $p(f_{1:T}, e_{1:T}|a_{1:T}, s_{1:T}, o_{1:T})$, the posterior distribution of the conversational events $e_{1:T}$, the joint VFOA $f_{1:T}$, and the model parameters λ, given the observation variables (see Eq. 1). Given the model complexity, direct and optimal estimation is untractable. Thus, we exploited the hierarchical structure of our model to define an approximate inference method which consists in the iteration of two main steps: First, estimation of the conversational events given the VFOA states, which can be conducted using Viterbi algorithm. Second, estimation of the VFOA states and models parameters given the conversational events using a Maximum A Posteriori (MAP) procedure (Gauvain and Lee, 1992). These iterations guarantee at each time step to increase the joint posterior probability distribution. More details can be found in (Ba and Odobez, 2008a).

3.2.3 Results

Dataset and protocol. To evaluate the above model, we used 4 meetings of the AMI corpus, with duration ranging from 17 to 30 minutes for a total of 1h30. The performances of our algorithms are measured in term of frame based VFOA recognition rate (FRR), i.e. the percentage of frames which are correctly classified, and a leave-one-out protocol was adopted to allow the learning of some of the model parameters (mainly $p(f^k_t|a_t, e_t)$).

Main results. Table 1 provides the results obtained with our contextual model. Results with an 'independent' approach, i.e. with VFOA recognition done independently for each person solely from his head pose (in this case, the VFOA dynamics only contains the term $p(f^k_t|f^k_{t-1})$) are also given. From the performance of this latter model, we see that our task and data are quite challenging, due mainly to our complex scenario, the meeting length, the behaviour of people, and the errors in the head pose estimation (some heads are rather difficult to track given our image reso-

Model	seat 1	seat 2	seat 3	seat 4	average
Contextual	61.7	56.1	48.5	55.9	55.6
Independent	41.3	45.5	29.7	36.4	38.2
Contextual: no adaptation	53	52.5	44.3	53.9	50.9
Contextual: speech/conversational events only	62.5	48.8	43.6	48.1	50.7

Table 1 VFOA recognition results for different modeling factors. Taken from (Ba and Odobez, 2008a).

lutions). We can also notice that the recognition task is more challenging for people at seat 3 and 4, with on average a FRR of 8 to 10% less than for seat 1 and 2. This is mainly explained by the fact that the VFOA targets spans a pan angle of around only 90% degrees for seats 3-4 vs 180% for seats 1-2, thus increasing the level of VFOA ambiguities associated with each head pose.
When using the contextual model, the recognition improves by a large amount, passing from 38.2% to 55.6%. The influence of the speaking activities through the conversational events and of the slide context on the VFOA estimation (and adaptation process), is beneficial, esp. for seat 3 and 4, by helping in reducing some of the inherent ambiguities in focus.

Indeed, improvements can be attributed to two main interconnected aspects: first, estimated conversational events and slide activity variables introduce direct informative prior on the VFOA targets when infering the VFOA states; second, head pose model parameters for each VFOA are adapted during inference, allowing to account for specifities of each person's gazing behaviour. To evaluate the impact of the adaptation process,we removed it during inference. This lead to a recognition rate drop of almost 5%, as shown in Table 1, demonstrating the benefit of adaptation. Notice that the removal of adaption in the independent case actually lead to a slight increase of 1.8%. The reason is that in this case, adaptation is 'blind', i.e. the pose model is adapted to better represent the observed head pose distribution. In practice, as head poses for some VFOA targets can be close, we noticed that the head pose mean of one VFOA target could drift to the head poses of another VFOA, resulting in significant errors. In the contextual approach, observed head poses are somehow 'weighted' by the context (e.g. when somebody speaks), which allows to better associate head pose with the right VFOA target, leading to better adapted head pose means $\mu_{k,i}$ and avoiding drift.

Finally, Table 1 also shows the results when only the conversation events are used to model the utterance and gaze pattern sequence, i.e. we remove the slide variable in the DBN model. The performance drop of around 5% clearly indicates the necessity to model the group activity in order to more accurately recognize the VFOA.

3.2.4 Discussion

The above results (55%) show that recognizing the VFOA of people is not an easy task in general. According to our experience and the literature, there are several factors which influence the accuracy we can obtain. In the following, we highlight the main ones and propose some future directions for improvements.

Physical setup and scenario. Our recognition results are not as high as those reported in the literature. Stiefelhagen et al (2002) reported a recognition rate of 75% using audio-visual features, while Otsuka et al (2006a) reports results of 66%. In those two cases, meetings with 4 participants were considered. They were sitting in a circular fashion, at almost equal space positions, and only other participants were considered as VFOA targets[3]. In addition, participants were only involved in discussions, and meetings were much shorter (less than 8 minutes), which usually implies a more active use of the head to gaze at people. Overall these meetings were quite different from our longer task-based meetings involving slide presentation, laptop and artefact manipulations, and thus 'more complex' gaze behaviour. Even in our setting, we can notice results ranging from 48% to 62% for the different seats, showing that good VFOA recognition can only be achieved if the visual targets of a person are well separated in the head pose angular space. Overall, our results indicate that obtaining in more complex meeting the same performance as published in earlier work on VFOA recognition (Stiefelhagen et al, 2002) is definitively problematic, and that there are limitations in using the head pose as approximation to the gaze.

Head pose estimation. Accurate head pose estimation is important for good results. Ba and Odobez (2008b) addressed the VFOA recognition from head pose alone on 8 meetings recorded in the same setting described in this chapter. Head pose was either measured using flock-of-birds (FOB) magnetic sensor, or estimated with a visual tracker. This allowed to evaluate the recognition errors due to using head pose as a proxy for gaze, and errors due to pose estimation. The results showed that VFOA recognition errors and head pose errors were correlated, and that there was an overall drop of around 13% when using the estimates instead of the FOB measurements. Note that Otsuka et al (2006a) only reports a drop of 2% when using estimates rather than magnetic sensor measurements. This smaller drop can be explained by three facts: their setting and scenario are simpler. Each participant face was recorded using one camera directly looking at him, with high resolution. This resulted in lower tracking errors. Finally, they were using an utterance/gaze interaction model similar to what we presented in this chapter (this was not the case in Ba and Odobez (2008b)) which may already correct some of the pose errors made by the tracker.

Exploitation of contextual cues. Our results showed that information on the slide activity was improving VFOA recognition. A more closer look at the recognition rates by target class, showed that the joint use of conversational events and slide context further increases the slide screen recognition accuracy w.r.t. the sole use of the slide context (77% vs 70%). This indicates that the model is not simply adding priors on

[3] Note that a 52% recognition rate is reported by Stiefelhagen (2002) when using the same framework as in (Stiefelhagen et al, 2002) but applied to a 5-person meeting.

the slide or people targets according to the slide and conversational variables, but that this is the temporal interplay between these variables, the VFOA and adaptation process which makes the strength of the model.
Of course, during the meeting, slide presentation is not the single activity that attracts gaze. People use their laptops, or manipulate objects on the table. Any information on these activities could be used as further contextual cues in the model, and would probably lead to a performance increase. Issues are how and with what accuracy can we extract such contextual cues. When considering laptops for instance, we could imagine directly obtain activity information from the device itself. Alternatively, we can rely on video analysis, but this will probably result in an accuracy loss (e.g. it might be difficult to distinguish some hand fidgeting or simple motion from hand laptop usage).

Improving behavioral models. There are several ways to improve the behavioral modeling. One way consists of introducing timing or duration models in the DBN to account for phenomenon reported by psychologists, like the timing information existing between speaking turns and VFOA patterns: for instance, people avert gaze at the beginning of a turn, or look for mutual gaze at the end of a turn. In another direction, since we show in the next section that people's VFOA is related to people traits like dominance, one may investigate whether the knowledge of these traits (e.g. introversion vs extroversion) could be used to define a more precise gaze model of the given person. The overall question when considering such modeling opportunities is whether the exploited properties would be recurrent enough (i.e. sufficiently predictable) to obtain a statistical improvement of the VFOA recognition.

4 From speaking activity and visual attention to social inference: dominance modeling

As an example of the potential uses of the nonverbal cues presented in this chapter to infer social behavior, we now focus on dominance, one fundamental construct of social interaction (Burgoon and Dunbar, 2006). Dominance is just one example of the myriad of complex phenomena that emerge in group interaction in which nonverbal communication play a major role (Knapp and Hall, 2005), and whose study represents a relatively recent domain in computing (Gatica-Perez, 2006, 2008). In this section, we first summarize basic findings in social psychology about nonverbal displays of dominance, briefly review existing work in computing in this field, and focus on a particular line of research that investigates the joint use of speaking activity and visual attention for dominance estimation, motivated by work on social psychology. Most of the material presented here has been discussed elsewhere (Hung et al, 2007, 2008a,b; Jayagopi et al, 2009).

4.1 Dominance in social psychology

In social psychology and nonverbal communication, dominance is often seen in two ways, both ”as a personality characteristic (trait)” and ”to indicate a person’s hierarchical position within a group (state)” (Schmid Mast, 2002) (pp. 421). Although dominance and related terms like power are often used as equivalent, a distinguishing approach taken in (Dunbar and Burgoon, 2005) defines power as ”the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person” (pp. 208), and dominance as 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” (pp. 208-209).

Nonverbal displays of dominance involve voice and body activity. Cues related to the voice include amount of speaking time, energy, pitch, rate, and interruptions (Dunbar and Burgoon, 2005). Among these, speaking time has shown to be a particularly robust (and intuitively logical) cue to perceive dominance: dominant people tend to talk more (Schmid Mast, 2002). Kinesic cues include body movement, posture, and elevation, and gestures, facial expressions, and visual attention (Dunbar and Burgoon, 2005; Burgoon and Dunbar, 2006).

From kinesic behavior, visual attention represents particularly a strong cue for nonverbal display and interpretation of dominance. Efran documented that high-status persons receive more visual attention than low-status people (Efran, 1968). Cook et al. found that people who very rarely look at others in conversations are perceived as weak (Cook and Smith, 1975). Furthermore, there is solid evidence that joint patterns of visual attention and speaking activity are correlated with dominance (Exline et al, 1975; Dovidio and Ellyson, 1982). More specifically, Exline et al. found that in dyadic conversations, high-status interaction partners displayed a higher looking-while-speaking to looking-while-listening ratio (i.e., the proportion between the time they would gaze at the other while talking and the time they would gaze at the other while listening) than low-status partners (Exline et al, 1975). This ratio is known as the *visual dominance ratio*. Furthermore, (Dovidio and Ellyson, 1982) found that patterns of high and low dominance expressed by the visual dominance ratio could be reliably decoded by external observers. Overall, this cue has been found to be correlated to dominance for a variety of social factors including gender and personality (Hall et al, 2005).

4.2 Dominance in computing

The automatically estimation of dominance in small groups from sensor data (cameras and microphones) is a relatively recent research area. In the following, we present a summary of existing work. A more detailed review can be found in (Gatica-Perez, 2008).

Some methods have investigated the relation between cues derived from speaking activity and dominance. Rienks and Heylen (2005) proposed an approach based on manually extracted cues (speaking time, number of speaking turns, and number of successful floor grabbing events) and Support Vector Machines (SVMs) to classify the level of a person's dominance as being high, normal, or low. Rienks et al (2006) later compared supervised and unsupervised approaches for the same task.

Other methods have investigated the use of both speaking and kinesic activity cues. Basu et al (Dec. 2001) used speaking turns, speaking energy, and body activity estimated from skin-color blobs to estimate the most influential participant in a debate, using a generative model called the influence model (IM). As part of our own research, we addressed the estimation of the most-dominant person using automatically extracted speaking activity (speaking time, energy), kinesic activity (coarse visual activity computed from compressed-domain video), and simple estimation models (Hung et al, 2007). We later conducted a more comprehensive analysis where fusion of activity cues was studied (Jayagopi et al, 2009). The results suggested that, while speaking activity is a more discriminant modality, visual activity also carries information, and also that cue fusion in the supervised setting can be useful.

Visual attention represents an alternative to the normally coarser measures that can be obtained to characterize kinesic activity with tractable methods. In the following section we described in more detail work that explicitly fuses visual attention and speaking activity to predict dominant people in group conversations.

4.3 Dominance estimation from joint visual attention and speaking activity

The application of findings in nonverbal communication about the discriminative power of joint visual attention and speaking activity patterns for automatic modeling of group interaction is attractive. In addition to the technical challenges related to the extraction of these nonverbal cues, as discussed in previous sections, there are other relevant issues related to the application of the nonverbal communication findings to multi-sensor environments. In real life, group interaction takes place in the presence of multiple visual targets in addition to people, including common areas like the table, whiteboards, and projector screens, and personal artifacts like computers and notebooks. In contrast, most nonverbal communication work has investigated cases where people are the main focus of the interaction.

Otsuka et al. proposed to automate the estimation of influence from visual attention and speaking activity (Otsuka et al, 2006b), computing a number of intuitive measures of influence from conversational patterns and gaze patterns. While the proposed features are conceptually appealing, the authors presented neither an objective performance evaluation nor a comparison to previous methods.

In contrast, (Hung et al, 2008b) recently proposed the automation of two cues discussed in Section 4.1, namely received visual attention (Efran, 1968) and visual

dominance ratio (Exline et al, 1975), and study their applicability in small group meetings on a systematic basis (Hung et al, 2008b). For a meeting with N_P participants, lasting N_T time frames, and where people can look at N_F different focus targets (including each person, the slide screen, the whiteboard, and the table), the total received visual attention for each person i, denoted by rva^i is given by

$$rva^i = \sum_{t=1}^{N_T} \sum_{j=1, j\neq i}^{N_P} \delta(f_t^j - i), \tag{5}$$

where f_t^j denotes the target focus of person j at time t, and $\delta(.)$ is the standard delta function.

The visual dominance ratio was originally defined for dyadic conversations. (Hung et al, 2008b) proposed a simple extension of the concept to the multi-person case, revisiting the 'looking-while-speaking' definition to include all people whom a person looks at when she/he talks, and the 'looking-while-listening' case to include all cases when a person does not talk and looks at any speaker (approximated in this way given the difficulty of automatically estimating listening events). The resulting multi-party visual dominance ratio for person i, denoted by $mvdr$, is then given by

$$mvdr^i = \frac{mvdr_N^i}{mvdr_D^i}, \tag{6}$$

where the numerator (looking-while-speaking) and the denominator (looking while listening) are respectively defined as

$$mvdr_N^i = \sum_{t=1}^{N_T} s_t^i \sum_{j=1, j\neq i}^{N_P} \delta(f_t^i - j), \tag{7}$$

$$mvdr_D^i = \sum_{t=1}^{N_T} (1 - s_t^i) \sum_{j=1, j\neq i}^{N_P} \delta(f_t^i - j) s_t^j, \tag{8}$$

where s_t^i denoted the binary speaking status (0/1: silence/speaking) of person i.

For both rva and $mvdr$, the most dominant person is assumed to be the one with the highest value for the corresponding cue. We also studied the performance of the numerator and the denominator terms of the multi-party visual dominance ratio, estimating the most dominant person as the one having the highest (resp. the lowest) value. Speaking activity is estimated from close-talk microphones attached to each meeting participant, and visual attention is estimated through the DBN approach, following the techniques described earlier in this chapter.

This approach was tested on two subsets of the AMI corpus (Carletta et al, 2005). The first one consists of 34 five-minute meeting segments where three human observers agreed on the perceived most dominant person (referred to as having full agreement in the following discussion). The second set is a superset of the first one, and consists of 57 five-minute meeting segments where at least two out of the three observers agreed on the perceived most dominant person (referred to as hav-

ing majority agreement). The two subsets model different conditions of variability of human judgment of dominance. The results are shown in Table 2.

Method	Full agreement (%)	Majority agreement (%)
rva	67.6	61.4
mvdr	79.4	71.9
$mvdr_N$	70.6	63.2
$mvdr_D$	50.0	45.6
Random	25.0	25.0

Table 2 Meeting classification accuracy for the most-dominant person task for both full and majority agreement data sets (taken from Hung et al (2008b)).

Considering the performance on these four cues on the full-agreement data set, it is clear that all the tested cues perform better than random, with the visual dominance ratio performing the best, suggesting that there is a benefit on combining visual attention and speaking activity (79.4%) compared to the case of only received visual attention (67.6%) as a predictor for the most dominant person. Nevertheless, it is remarkable that by using only visual cues (i.e., not using what is being said at all), relatively good performance is obtained, more even so if we consider that the automatically extracted visual attention is far from being perfect, where VFOA performance is around 50% as described earlier in the chapter. It is important to note that these two cues are applied to the group setting, which differs from the original use of these cues in psychology not only due the presence of multiple people but also due to the presence of meeting artifacts (table, screen) that significantly affect the group dynamics. Analyzing the results for the components of the MVDR, it appears that the numerator is more discriminant that the denominator taken in isolation, but also that their combination is beneficial. Finally, similar trends are observed for the performance on the majority-agreement data set, with a noticeable degradation of performance in all cases, which seems to reflect the fact that this data set is inherently more challenging given the larger variation in human perception of dominance. Current work is examining other measures that combine visual attention and speaking activity (see Fig. 10).

Overall, the work discussed in this section can be seen as complementing the work oriented towards on-line applications presented in Section 2.1, providing further insights into the value of nonverbal cue fusion for automatic estimation of social constructs in conversations that take place in multi-sensor environments.

5 Concluding remarks

In this chapter, we have shown that there is a growing interest in exploiting conversation analysis for building applications in the work place. We have presented an overview and discussed some key issues on extracting speaking activity and visual

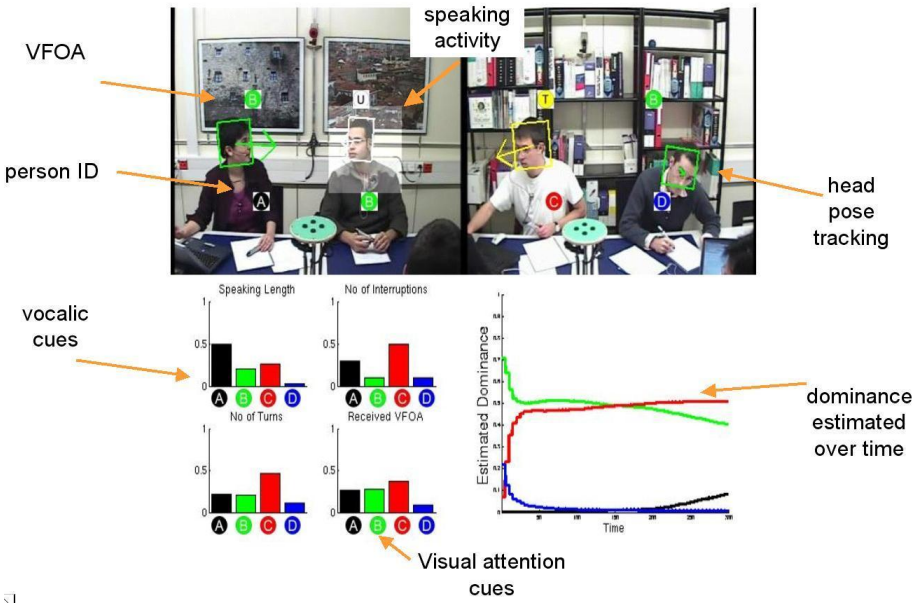

Fig. 10 Estimating dominance over time from multiple cues. A distinct color is used to indicate all information related to each of the four meeting participants (identified by a circle and a letter ID drawn over the person's body). The top panels (left and right) show the two camera views from which location (colored bounding boxes), head pose (colored arrows) and visual focus (colored circles above people's heads) are estimated for each person. Notice that person B is recognized as being unfocused, while person C is recognized as looking at the table. The current speaker (person B) is highlighted by a light gray square. The bottom left panel shows speaking activity cues (speaking time, speaking turns, and speaking interruptions) and visual attention cues (received visual focus) estimated over time. The bottom right panel shows the dominance level estimated over time for each participant. In this case, two people (B and C) dominate the conversation.

attention from multi-sensor data, including through the modeling of their interaction in conversation. Finally, we have described their combined use for estimating social dominance.

Despite the recent advancements in scene conversation analaysis, there are several inter-related issues that still needs to be investigated to move towards easy processing in real-life situations or address the inference of different behavioral and social cues.

A first key issue is the robustness of the cue extraction process with respect to the sensor setting, and the constraints thet this setting may impose on people mobility or behaviour. Regarding audio, most of the existing works on social interaction analysis in meetings rely on good-quality audio signals extracted from close-talk microphones. A more simple and desirable setting would involve non-obtrusive, possibly single, distant microphones, but this is at the cost of larger speaker segmentation errors. VFOA extraction faces several challenges as well, as discussed in

Section 3.2. While cameras are not intrusive, they need to continuously monitor all people's head and face which clearly requires a multiple camera setting in scenarios involving moving people. In addition, relating head orientation to VFOA targets depends on both people and VFOA target locations. Estimating this relationship can be done using learning sequences, but this quickly becomes burdersome when multiple VFOA target models for multiple locations have to be estimated. The use of gaze models, as we presented in this chapter, is more versatile, but still requires gross camera calibration and estimated of people 3D head position and VFOA target locations estimates.

Still, not all applications of interest may need accurate cue extraction for all people. For instance, in the Remote Participant application, one might only be interested at detecting when the majority of collocated participants -or even only the chairman- look at the screen display of the remote participant to detect if the later is addressed. Also, some applications rely on measurement statistics rather than instantaneous cues, and therefore may tolerate important individual errors. As an example, the use of the recognized VFOA for dominance estimation, as reported in Subsection 4.3, leads to a performance decrease w.r.t. use of the ground truth (Hung et al, 2008b), but not as much as what could be expected given the obtained 55% VFOA recognition rate. Similarly, Hung et al. Hung et al (2008a) who studied the problem of predicting the most-dominant person for the single distant microphone case, using a fast speaker diarization algorithm and speaking time as only cue, reported a similar behaviour: performance was lower than using cleaner close-talk microphone signals, but larger decrease in diarization performance did not have the same degree of negative impact in the estimation of the most-dominant person.

Another important issue which has to be taken into account is the group size, which is known to have impact the social dynamics of the conversation. Floor control, gaze, and other non-verbal cue pattern may significantly differ in large group than in small groups. Fay et al (2000) reports that people speaking behaviours is quite different in 10-people group sizes, where communication evolves as a serie of monologues, than in 5-people meetings where communication is dialogue, and that dominance is perceived differently in the two situations. Most of the work so far on automated conversation analysis has been done on small groups from 2 to 5 people. Indeed, for larger groups sizes, the automatic cue extraction becomes very difficult, especially when using non obtrusive techniques. This is the case of the audio when using distant microphones, and even more for the VFOA extraction where the approximation of gaze with the head pose will no longer be sufficient, since gazing might be achieved only by eye motion. From this perspective, investigation of automated conversation and social analysis in large group meetings is very challenging and remains relatively unexplored.

Acknowledgements Our work has been supported by the EC project AMIDA (Augmented Multi-Party Interaction with Distant Access), the Swiss National Center for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2), and the US research program VACE (Video Analysis and Content Extraction). We thank Sileye Ba, Hayley Hung, and Dinesh Jayagopi (members of our research groups at Idiap) for their contribution to the some of the research described here. We also thank Pierre Dillenbourg, Taemie Kim, Joan Morris DiMicco, An-

drei Popescu-Belis, and Janienke Sturm for giving permission to reproduce the pictures presented in Figures 1 and 2.

References

Argyle M, JGraham (1977) The central europe experiment - looking at persons and looking at things. Journal of Environmental Psychology and Nonverbal Behaviour 1:6–16

Ba S, Odobez JM (2008a) Multi-person visual focus of attention from head pose and meeting contextual cues. Tech. Rep. 47, Idiap Research Institute

Ba S, Odobez JM (2008b) Recognizing human visual focus of attention from head pose in meetings. accepted for publication in IEEE Transaction on Systems, Man, and Cybernetics: part B, Man

Ba SO, Odobez JM (2005) A Rao-Blackwellized mixed state particle filter for head pose tracking. In: Proc. ACM-ICMI-MMMP, pp 9–16

Bachour K, Kaplan F, Dillenbourg P (Sept, 2008) An interactive table for regulating face-to-face collaborative learning. In: Proc. European Conf. on Technology-Enhanced Learning (ECTEL), Maastricht

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

Burgoon JK, Dunbar NE (2006) The Sage Handbook of Nonverbal Communication, Sage, chap Nonverbal expressions of dominance and power in human relationships

Cappella J (1985) Multichannel integrations of nonverbal behavior, Erlbaum, chap Controlling the floor in conversation

Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, T Hain JK, Karaiskos V, Kraaij W, Kronenthal M, Lathoud G, Lincoln M, A Lisowska IM, Post W, Reidsma D, Wellner P (2005) The AMI meeting corpus: A pre-announcement. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh

Chen L, Harper M, Franklin A, Rose T, Kimbara I (2005) A Multimodal Analysis of Floor Control in Meetings. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)

Cook M, Smith JMC (1975) The role of gaze in impression formation. British Journal of Social and Clinical Psychology

DiMicco JM, Pandolfo A, Bender W (2004) Influencing group participation with a shared display. In: Proc. ACM Conf. on Computer Supported Cooperative Work (CSCW), Chicago

Dines J, Vepa J, Hain T (2006) The segmentation of multi-channel meeting recordings for automatic speech recognition. In: Int. Conf. on Spoken Language Processing (Interspeech ICSLP)

Dovidio JF, Ellyson SL (1982) Decoding visual dominance: atributions of power based on relative percentages of looking while speaking and looking while listening. Social Psychology Quarterly 45(2):106–113

Dunbar NE, Burgoon JK (2005) Perceptions of power and interactional dominance in interpersonal relationships. Journal of Social and Personal Relationships 22(2):207–233

Duncan Jr S (1972) Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23(2):283–292

Efran JS (1968) Looking for approval: effects of visual behavior of approbation from persons differing in importance. Journal of Personality and Social Psychology 10(1):21–25

Exline RV, Ellyson SL, Long B (1975) Advances in the study of communication and affect, Plenum Press, chap Visual behavior as an aspect of power role relationships

Fay N, Garod S, Carletta J (2000) Group discussion as interactive dialogue or serial monologue: the influence of group size. Psychological Science 11(6):487–492

Freedman EG, Sparks DL (1997) Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. Journal of Neurophysiology 77:2328–2348

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

Gatica-Perez D (2008) On social interaction analysis in small group conversations from nonverbal cues. under review

Gauvain J, Lee CH (1992) Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. Speech Communication 11:205–213

Goodwin C, Heritage J (1990) Conversation analysis. Annual Review of Anthropology pp 981–987

Hall JA, Coats EJ, LeBeau LS (2005) Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. Psychological Bulletin 131(6):898–924

Hayhoe M, Ballard D (2005) Eye movements in natural behavior. TRENDS in Cognitive Sciences 9(4):188–194

Hung H, Jayagopi D, Yeo C, Friedland G, Ba SO, Odobez JM, Ramchandran K, Mirghafori N, Gatica-Perez D (2007) Using audio and video features to classify the most dominant person in a group meeting. In: Proc. of ACM Multimedia

Hung H, Huang Y, Friedland G, Gatica-Perez D (2008a) Estimating the dominant person in multi-party conversations using speaker diarization strategies. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas

Hung H, Jayagopi D, Ba S, Odobez JM, Gatica-Perez D (2008b) Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: to appear in the International Conference on Multimodal Interfaces

Jayagopi D, Hung H, Yeo C, Gatica-Perez D (2009) Modeling dominance in group conversations using nonverbal activity cues. IEEE Trans on Audio, SPeech and Language, Special Issue on Multimodal Processing for Speech-based Interactions

Jovanovic N, Op den Akker H (2004) Towards automatic addressee identification in multi-party dialogues. In: 5th SIGdial Workshop on Discourse and Dialogue

Kendon A (1967) Some functions of gaze-direction in social interaction. Acta Psychologica 26:22–63

Kim T, Chang A, Holland L, Pentland A (2008) Meeting mediator: Enhancing group collaboration with sociometric feedback. In: Proc. ACM Conf. on Computer Supported Cooperative Work (CSCW), San Diego

Knapp ML, Hall JA (2005) Nonverbal Communication in Human Interaction. Wadsworth Publishing

Kouadio M, Pooch U (2002) Technology on social issues of videoconferencing on the internet: a survey. Journal of Network and Computer Applications 25:37–56

Kulyk O, Wang J, Terken J (2006) Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)

Langton S, Watt R, Bruce V (2000) Do the eyes have it ? cues to the direction of social attention. Trends in Cognitive Sciences 4(2):50–58

Lathoud G (2006) Spatio-temporal analysis of spontaneous speech with microphone arrays. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Lathoud G, McCowan I (2003) Location Based Speaker Segmentation. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong

Matena L, Jaimes A, Popescu-Belis A (2008) Graphical representation of meetings on mobile devices. In: MobileHCI conference, Amsterdam, The Netherlands

Morimoto C, Mimica M (2005) Eye gaze tracking techniques for interactive applications. Computer Vision and Image Understanding 98:4–24

Novick D, Hansen B, Ward K (1996) Coordinating turn taking with gaze. In: International Conference on Spoken Language Processing

Odobez JM, Ba S (2007) A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of Focus of Attention from Head pose. In: Proc. of ICME

Ohno T (2005) Weak gaze awareness in video-mediated communication. In: Proceedings of Conference on Human Factors in Computing Systems, pp 1709–1712

Otsuka K, Takemae Y, Yamato J, Murase H (2005) A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: Proc. of ICMI, pp 191–198

Otsuka K, Yamato J, Takemae Y, Murase H (2006a) Conversation scene analysis with dynamic bayesian network based on visual head tracking. In: Proc. of ICME

Otsuka K, Yamato J, Takemae Y, Murase H (2006b) Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In: Proc. ACM CHI Extended Abstract, Montreal

Ramírez J, Górriz J, Segura J (2007) Robust speech recognition and understanding, I-Tech, I-Tech Education and Publishing, Vienna, chap Voice activity detection: Fundamentals and speech recognition system robustness

Ranjan A, Birnholtz J, Balakrishnan R (2008) Improving meeting capture by applying television production principles with audio and motion detection. In: CHI

'08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA, pp 227–236, DOI http://doi.acm.org/10.1145/1357054.1357095

Rhee HS, Pirkul H, Jacob V, Barhki R (1995) Effects of computer-mediated communication on group negotiation: Au empirical study. In: Proceedings of the 28th Annual Hawaii International Conference on System Sciences, pp 981–987

Rienks R, Heylen D (2005) Automatic dominance detection in meetings using easily detectable features. In: Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh

Rienks R, Zhang D, Gatica-Perez D, Post W (2006) Detection and application of influence rankings in small-group meetings. In: Proc. Int. Conf. on Multimodal Interfaces (ICMI), Banff

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

Shriberg E, Stolcke A, Baron D (2001) Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In: ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding (Prosody 2001)

Stiefelhagen R (2002) Tracking and modeling focus of attention. PhD thesis, University of Karlsruhe

Stiefelhagen R, Yang J, Waibel A (2002) Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans on Neural Networks 13(4):928–938

Sturm J, Herwijnen OHV, Eyck A, Terken J (2007) Influencing social dynamics in meetings through a peripheral display. In: Proc. Int. Conf. on Multimodal Interfaces (ICMI), Nagoya

Takemae Y, Otsuka K, Yamato J (2005) Automatic video editing system using stereo-based head tracking for multiparty conversation. In: ACM Conference on Human Factors in Computing Systems, pp 1817–1820

Valente F (2006) Infinite models for speaker clustering. In: Int. Conf. on Spoken Language Processing (Interspeech ICSLP)

Vijayasenan D, Valente F, Bourlard H (2008) Integration of tdoa features in information bottleneck framework for fast speaker diarization. In: Interspeech 2008

Wrigley SJ, Brown GJ, Wan V, Renals S (2005) Speech and crosstalk detection in multi-channel audio. IEEE Trans on Speech and Audio Processing 13:84–91

Yeo C, Ramchandran K (2008) Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Tech. Rep. UCB/EECS-2008-79, EECS Department, University of California, Berkeley