

Chapter 1

Multimodal analysis of small-group conversational dynamics

1.1 Introduction

The analysis of conversational dynamics in small groups, like the one illustrated in Figure 1.1, is a fundamental area in social psychology and non-verbal communication [Goodwin, 1981, Clark and Carlson, 1982]. Conversational patterns exist at multiple time scales, ranging from knowing how and when to address or interrupt somebody, how to gain or hold the floor of a conversation, and how to make transitions in discussions. Most of these mechanisms are multimodal, involving multiple verbal and nonverbal cues for their display and interpretation [Knapp and Hall, 2005], and have an important effect on how people are socially perceived, e.g., whether they are dominant, competent, or extraverted [Knapp and Hall, 2005, Pentland, 2008].

This chapter introduces some of the basic problems related to the automatic understanding of conversational group dynamics. Using low-level cues produced by audio, visual, and audio-visual perceptual processing components like the ones discussed in previous chapters, here we present techniques that aim at answering questions like: Who are the people being addressed or looked at? Are the involved people attentive? What conversational state is a group conversation currently at? Is a particular person likely perceived as dominant based on how they interact? As shown later in the book, obtaining answers for these questions is very useful to infer, through further analysis, higher-level aspects of a group conversation and its participants.

The chapter is organized as follows. Section 1.2 provides the basic definitions of three conversational phenomena discussed in this chapter: attention, turn taking, and addressing. Section 1.3 then discusses the state of affairs



Figure 1.1: A small-group conversation extracted from the Augmented Multi-Party Interaction (AMI) corpus.

on automatic modeling of each of these concepts. Section 1.4 discusses how these phenomena have an effect on a particular social construct of interest in group interaction, namely dominance. Section 1.5 offers some concluding remarks. Finally, Section 1.6 provides references for further reading.

1.2 Conversational dynamics phenomena: definitions

In this section, we review three fundamental elements of conversational dynamics, namely attention, turn taking, and addressing. Each of these concepts is later analyzed from the computational perspective.

1.2.1 Conversational attention

People in meetings pay attention to their colleagues and the various things that happen, with varying degrees, as a result of their interest. In a group conversation many activities occur: some of them are planned in advance, many are not. If a computer system could estimate the attention level and focus of people, it could inform the team about their collective degree of engagement, and make each individual aware of how attentive they are perceived by others.

In his 1890 monumental work *Principles of Psychology*, William James eloquently described interest and attention. For the first concept: "Millions

of items of the outward order are present to my senses which never properly enter into my experience. Why? Because they have no interest for me. My experience is what I agree to attend to. Only those items which I notice shape my mind - without selective interest, experience is an utter chaos. Interest alone gives accent and emphasis, light and shade, background and foreground - intelligible perspective, in a word." And for the second one: "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are its essence" (both quotes from James [1890], Chapter XI).

In conversations, it is known that listeners show attention by orienting their gaze - their eye's direction - towards speakers. These also use gaze to indicate whom they address and are interested in interacting with Goodwin [1981]. Knapp and Hall, a century later than James, point out that people "gaze more at people and things perceived as rewarding" and "at those whom they are interpersonally involved" Knapp and Hall [2005] (p. 349 and 351, respectively). Conversational attention is therefore inherently multimodal, involving coordinated gaze and speaking activity, and other cues like body pose, gestures, and facial expressions. Conversational attention is also dynamic: the focus of attention constantly shifts in the course of a conversation. Two common examples illustrate this phenomenon: on one hand, materials that are freshly introduced in a conversation have prominence, at least temporarily; on the other hand, mentioning something or placing emphasis on it also turns it into the object of attention and brings it temporarily to the foreground.

1.2.2 Turn taking and conversational floor

As children, we learn the value of letting others talk, and also to speak when appropriate. In a simplified social world, when one speaks in a group conversation, i.e., when one takes a turn, others pay attention to what is being said and actively listen, giving verbal and nonverbal feedback. In this situation, there is a single current speaker holding the floor, who gives others the right to speak by explicitly asking questions or inviting reactions via nonverbal behavior. Discussions can emerge from these exchanges. If an automatic system could infer the state of a group conversation from low-level observations, e.g., a monologue or a discussion, it could then create indexing information based on these states, or use this information as context for other processes. For instance, a heated discussion can be indicative of the overall interest of the group in the topic being discussed. The automatic identification of floor holding states and floor changes could also be useful for meeting summarization. More specifically, detecting who has the floor at any given time could be used to distinguish certain important utterances from background speech or side comments when creating a summary of a

meeting. Finally, floor modeling could be useful to predict both the next speaker and the addressees whom a speaker talks to and expects a response from.

Turn taking is a basic form of organization for conversations. Although organized meetings often have an agenda that organizes the topics on a certain level, as well as a chairman who takes care that the agenda is followed, at the lower level of conversational activities, turn taking is a "locally managed" process Sacks et al. [1974], i.e., it only depends on the current conversational situation – who has what conversational role among speakers, addressees, or overhearers. A well-known model by Sacks et al. Sacks et al. [1974] assumes two types of turn allocation techniques. In the first one, the current speaker selects the next speaker. This can be done by looking at or asking a question to that person. In the second technique, the next turn is allocated by self-selection.

The adequacy of this turn-taking model is debated, given that modeling conversations as clean sequences of contributions of speakers, one after the other, is far too simple to include the complex dynamics of multi-party conversations [Cowley, 1998]. In practice, parallel activities often happen in a meeting, and two or more participants can speak at the same time, contributing to one or multiple simultaneous conversations or talking about different topics, and with different people paying attention to what happens. Conversational floor theories have aimed at describing and explaining these phenomena.

Various floor models have been proposed over time. In Parker's model [Parker, 1998], a floor is a pairwise conversation between two participants of a group conversation. For Edelsky [1981], the floor is a specific type of speaking turn which contains the "acknowledged what's going-on within a psychological time/space", i.e., a psychologically developed, interactional space among people, which allows to distinguish between a main conversation flow and background speech. For Hayashi [1991], the floor is "a means of communicative attention orientation which exists not at the level of turn and move but at a higher level of conversation structure" (Hayashi [1991] p. 2). Hayashi's model involves two main types of floor. One is a 'single conversational floor' in which only one floor is currently occurring in a conversation. The other type is a 'multiple conversational floor' where two or more single conversational floors occur simultaneously. In both cases, the floor is regulated both verbally and nonverbally (though speaking tempo, physical posture, and prosody.) Furthermore, the floor mediates interactions on four levels of conversational structure: utterance, turn, theme, and organization. At the utterance level, the floor constrains "how a speaker says something in a certain setting and what s/he wants to do by saying it". At the turn level, the floor constrains "turn skills such as when and how to take or yield a turn, and what the interactant intends to achieve in doing so". At the level of theme, the floor contributes to determine the "selection,

continuity, and discontinuity of the topic, and to making the flow of topic coherent.” At the organizational level, the floor ”sequences discourse components coherently in a global structure”. Finally, patterns of floor structure are related to social constructs including ”power, solidarity, cooperation, conflict, and competition” (citations from Hayashi [1991], pp. 6-7).

1.2.3 Addressing

When small groups meet, most of what is said by somebody is directed towards everybody else. However, a speaker’s contribution is sometimes meant for a selected audience or even a single participant. This could be due to a variety of reasons: sometimes what is said is only of the addressee’s concern, or because the speaker has a specific interest in the addressee’s attention or feedback; privacy concerns might also be the motivation for choosing a specific addressee. Linguists and conversational analysts define an addressee in two ways: as the listener(s) whom the current speaker selects as the one(s) he expects a response from, more than from other listeners (e.g. see Goffman’s work Goffman [1981]); and as those listeners who are expected by speakers to take up what is being said (e.g. see Clark’s work Clark and Carlson [1982]).

In social psychology, it is known that the addressing phenomenon occurs through different communication channels, including speech, gaze, and gesture, e.g. listeners express attention by orienting their gaze to speakers, who in turn typically gaze at whom they address, and to capture visual attention in order to hold the floor [Goodwin, 1981]. It is also known that participants in group conversations, interacting and exchanging roles as speakers, addressees, and side participants (i.e., those not being addressed), contribute to the emergence of conversational events that characterize the flow of a meeting. A system capable of automatically inferring addressees would be useful, for instance, to extract side conversations, or to detect possible alliances in a given group.

1.3 Automatic modeling of face-to-face small group conversational dynamics

In this section we present a brief discussion of well known approaches towards automatic analysis of conversational dynamics in small groups. Following the sequence of concepts from section 1.2 of this chapter, we discuss works towards the estimation of visual attention, speaking turns and conversational floor patterns, and addressees. Due to space reasons, we do not include an extensive review of the literature, and have often chosen works from our research groups for further discussion. In Section 1.6, we provide a few pointers to additional reading materials. Some of the material presented in

this section has been adapted from Gatica-Perez [2009].

1.3.1 Modeling visual attention

Estimating eye gaze in arbitrary conversational situations is a challenging problem given the difficulty in using eye trackers due to practical issues like camera placement and image resolution. While some solutions using wearable cameras have started to appear [Noris et al., 2008], and other methods have been used to have a proxy for gaze in group conversations in laboratory conditions [Otsuka et al., 2005, Sturm et al., 2007], the problem of estimating gaze in conversations has most often been tackled by using head pose as a gaze surrogate. This has generated an increasing body of work [Stiefelhagen, 2002, Stiefelhagen et al., 2002, Ba and Odobez, 2004, 2006], which has mainly been conducted on laboratory data sets like the AMI meeting corpus [Carletta et al., 2005] or the NTT corpus [Otsuka et al., 2005] where small groups seat down and discuss a variety of topics.

Typically, existing methods for estimation of visual attention assume that each person involved in a group conversation has a finite and usually small number of visual attention targets, corresponding to the other people, certain elements of the environment, and artifacts of common use like tables, screens, etc. The methods often include two stages, where the 3-D head pose of a person, characterized by pan, tilt, and roll angles, is first estimated, and then the discrete visual attention labels are estimated from head pose angles and additional observations.

For head pose estimation, existing techniques initially detect and localize a person's head or face. This is then used to extract a number of visual features related to the appearance and shape of heads and faces. Finally, static or dynamic models based on classifiers or trackers are used to infer the 3-D head pose. In some methods, the head localization problem is solved jointly with the head pose estimation problem.

Once the head pose is estimated, the problem of estimating visual attention is addressed as a sequence recognition problem. Initial works examined the case when the attention focus of each person is assumed to be independent, and no other source of information, besides the head pose, is available. Examples of this approach are the works of Stiefelhagen et al. [2002] and of Ba and Odobez [2006]. It is clear, however, that very important information is left out following this assumption, as the state of a conversation (e.g. a person making a presentation, or an object being used by several group members at the table) effectively constrains the focus of attention of the group. Furthermore, conversational attention is a multimodal phenomenon, as the speaking activity of oneself and the others plays a role in defining who becomes the visual target at any given time. The interplay between speaking activity and visual attention is one of the most interesting aspects of current research for modeling of visual attention in conversations. The

works by Otsuka et al. [Otsuka et al., 2005, 2006, 2007] and Ba and Odobez [2008, 2011] stand out as examples of this research direction and will be discussed in more detail.

In a small group discussion context, Otsuka et al. [2005] proposed a Dynamic Bayesian Network (DBN) approach to jointly infer the gaze pattern for multiple people and the conversational gaze regime responsible for specific speaking activity and gaze patterns (e.g., all participants converging onto one person, or two people looking at each other). This work used, as a proxy for gaze, the head pose derived from a magnetic head trackers physically attached to each person. Furthermore, binary speaking activity was extracted from a manual speaking turn segmentations for each group member. The same model was later used with a more realistic approach, in which head pose angles were estimated from visual observations [Otsuka et al., 2006]. Otsuka et al. later extended this work to explicitly model patterns of the form 'who responds to whom, when, and how' Otsuka et al. [2007], and to incorporate facial expressions as components of the model [Kumano et al., 2009]. Finally, Otsuka et al. developed a real-time automatic group analysis system that integrates head pose tracking and speaker diarization Otsuka et al. [2008].

Ba and Odobez Ba and Odobez [2008, 2011] also proposed a DBN to infer the joint focus of attention of all group members by integrating individual head pose, speaking activity and the use of meeting artifacts as contextual cues. This work was grounded on the AMI meeting scenario, where four people discuss around a table and use typical objects such as a whiteboard and a projector screen, and defined seven potential visual targets for each participant (the other three participants, the table, the projector screen, the whiteboard, and an unfocused catch-all class). The assumptions of this work, which are backed up by significant empirical evidence, are that the current speaker tends to be looked at depending on the current conversational turn-taking state, and that a change of slide increases the possibility of the screen being looked at temporarily. These assumptions were introduced in the model via statistical dependencies on a graphical model. The system used three types of observations: head pose angles, speaker segmentations, and a binary slide change detector. As output, the model jointly inferred both the visual focus of each person and the conversational group state. Ba and Odobez showed that this model significantly improves the recognition performance on a subset of the AMI corpus, but that the problem is challenging, given the used camera resolution, and the initial, sometimes coarse approximation of true gaze by head pose. A snapshot of the results on AMI data appears in Fig. 1.2.

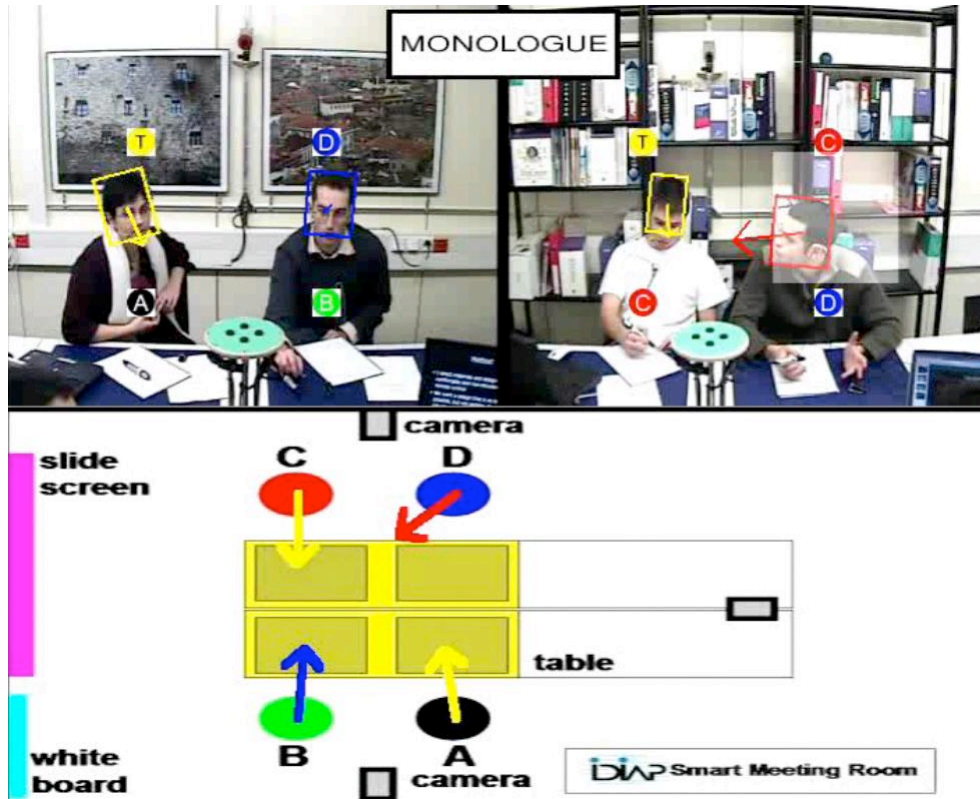


Figure 1.2: Automatic inference of joint focus of attention and conversational state on AMI data, based on the work by Ba and Odobez Ba and Odobez [2008]. A distinct color is used to indicate all information related to each of the four meeting participants, identified by a circle and a letter ID drawn over the person’s body. The top panels (left and right) show the two camera views used to estimate location (color bounding boxes), head pose (color arrows), and visual focus (color circles above people’s heads). The focus target identifiers A, B, C, and D refer to people, while T refers to the table. Note that persons A and C are recognized as looking at the table, while person B looks at person D, and this person in turn looks at C. The current speaker (person B) is highlighted by a gray square. The inferred conversational state (‘monologue’) is displayed on the top part of the image. The bottom panel shows the same information from a top view, where the whiteboard and the slide screen are also shown.

1.3.2 Modeling turn-taking and conversational floor

Most of the existing work on automatic recognition of turn-taking patterns in group conversations has addressed the problem using sequential models that assume a discrete set of turn-taking patterns, and analyze a group

conversation as a sequence of such patterns. As in the previous section, most of the existing work has been conducted on laboratory data sets.

McCowan et al. [2003, 2005] investigated the joint segmentation and recognition of four-person meetings into turn-taking-like patterns, including monologues, discussions, and presentations. The approach used standard Hidden Markov Models (HMMs) and basic audio and visual features extracted from three cameras, lapel microphones, and a microphone array. The features included pitch, energy, speaking rate, and speaking activity for audio, and skin-color blob location and motion for video. A number of HMM variations (multistream, coupled, and asynchronous HMMs) were tested on the MultiModal Meeting Manager (M4) corpus, resulting in promising performance. An example of recognized patterns can be seen in Fig. 1.3). In a second attempt, Zhang et al. [2006] proposed a two-layer HMM framework (see Fig. 1.4), in which activities performed by individuals, like speaking or writing, are recognized in the first layer from raw audio-visual observations, and the group turn-taking patterns are then recognized in the second layer. The layered architecture has several advantages in terms of flexibility and ease of training, and the possibility of using different sequential models for each layer. The results obtained on the M4 corpus confirmed these benefits in practice. Other works have used other hierarchical representations. For example, Dielmann et al. [2007] studied two variations of multilevel DBNs using audio-only cues. Furthermore, a comparison of recognition models on the M4 corpus was conducted by Al-Hames et al. [2005].

In other works related to recognition of speaking turn patterns, Banerjee et al. [2004] proposed a simple method to recognize three types of group meeting activities, namely discussions, presentations, and briefings, from close-talk audio. A decision tree was used as classifier of one-second observations windows, where features included the number of speakers, the number of speaker changes, the number of overlapping turns, and the average length of the overlaps. In a different approach, Campbell and Douchamps [2007] used an integrated system composed of a microphone array and a parabolic camera to do an analysis of overlapping speech and back-channeling for three types of conversations (formal meeting, relaxed conversation, and party), finding significant differences in the amounts of overlapping.

Regarding multimodal floor modeling, a systematic approach is due to Chen et al., who used the Video Analysis and Content Extraction (VACE) meeting corpus collected with multiple cameras, microphones, and magnetic sensors [2005, 2006], Chen and Harper [2009]. In Chen et al. [2005], the authors first proposed to combine gaze, gesture, and speech for floor control modeling. In Chen et al. [2006], they reported work on multimodal markers of floor control in VACE meetings, including a scheme for floor control annotation, and the use of a labeled corpus to identify mul-

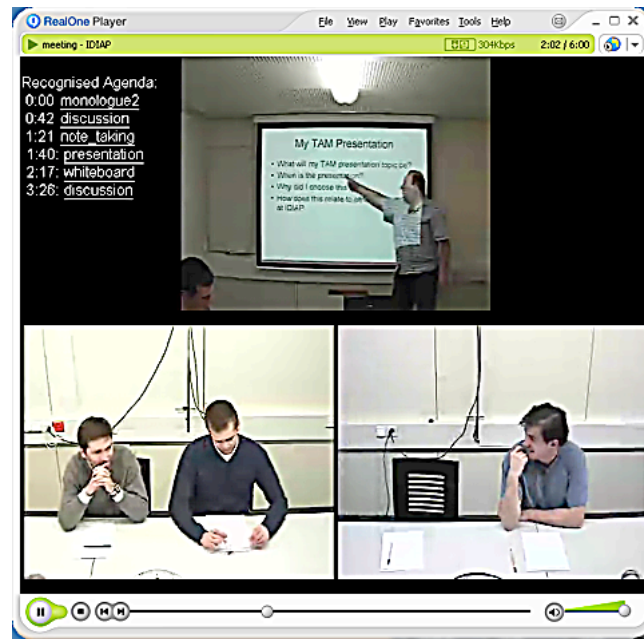


Figure 1.3: The three camera views of the M4 meeting corpus, and a sequence of automatically recognized speaking turn patterns (top left) with the HMM approach proposed by McCowan et al. [2005].

timodal cues correlated with floor changes. A few multimodal cues were identified as helpful for predicting floor control events, including discourse markers, which occur frequently at the beginning of a floor; mutual gaze between the current floor holder and the next one, which occurs during floor transitions; and gestures that relate to floor capturing. Finally, an approach for multimodal floor control shift detection was proposed in Chen and Harper [2009], which involved a combination of verbal and nonverbal features, and Maximum Entropy, Conditional Random Fields, and AdaBoost classifiers.

In a separate research line, Van der Vliet studied the floor ideas of Parker [1998] and Edelsky [1981] described in Section 1.2 of this chapter, and tested their validity on the AMI meeting corpora [der Vliet, 2006]. Van der Vliet explored the floor concept and related it to some meeting properties that could be used as cues to predict the floor, like gestures and gaze, developing a floor annotation scheme. The scheme consisted of two main categories to distinguish between utterances that are or are not part of the floor. Some subcategories were defined to gain more insight into how a floor is established, and which types of floor transitions occur in a conversation. The relation between gestures and floors was manually analyzed in AMI meetings, finding that not all floor transitions are accompanied with gestures, and that floor-giving as well as floor-capturing gestures could be used as

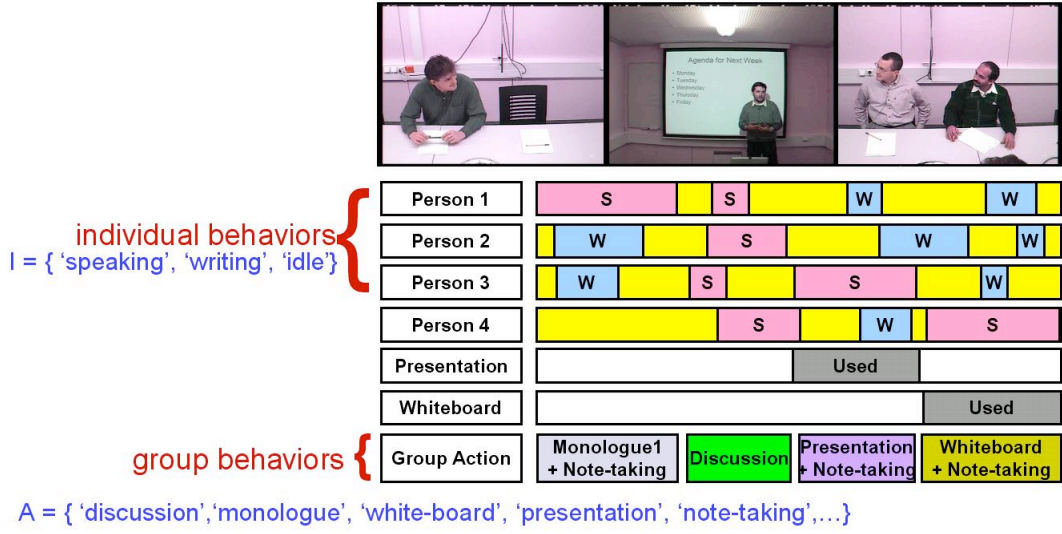


Figure 1.4: Layered HMMs to model turn taking patterns, proposed by Zhang et al. [2006]. In the first layer, a small number of conversational individual states are recognized. In the second layer, these recognized states are used as observations, along with environment contextual features (use of the whiteboard or the projector screen) to recognize turn taking group patterns.

cues for floor transitions.

1.3.3 Modeling addressing

Regarding computational modeling of addressing, the goals of the existing works are, on what hand, the recognition of addressees (i.e., what participants in a conversation the current speaker is talking to), and on the other hand, the exploration of connections between addressing and other conversational activities, like the ones described in Section 1.3.2.

There is a relation between addressing and turn-taking. In Sacks et al.'s theory of turn-taking, speakers may select the next speaker by inviting them, and if this situation does not occur, other participants in the interaction do self-selection as next speaker [Sacks et al., 1974]. Goffman's definition of addressee, cited in Section 1.2 earlier in this chapter, refers to this next-speaker selection notion of addressing. This implies that the knowledge of the next speaker informs about the addressee of the previous speaker.

One of the most comprehensive studies on automatic addressing modeling in small groups is the one by Jovanovic and op den Akker, conducted on the AMI meeting corpus. [Jovanovic and op den Akker, 2004, Jovanovic

et al., 2005, 2006, Jovanovic, 2007] In [Jovanovic and op den Akker, 2004], the authors proposed a scheme of verbal, nonverbal, and contextual features for addressee recognition, but no experiments were conducted to validate it. In [Jovanovic et al., 2005], the authors later annotated a subset of the AMI corpus with respect to addressee behavior, which included a discrete visual focus for each participant, addressee information, and dialogue acts – speech utterances labeled as questions, statements, backchannels, and floor grabbers. The annotation used dialog acts as units, defining four possible addressing classes (speaker addresses a single person, a subgroup, the whole audience, or if the addressee is unknown) for each act. In [Jovanovic, 2007], an approach for recognition was proposed, based on a mix of manual and automatic features and BNs. This work has been continued by op den Akker and Theune [2008].

In other work, Takemae et al. [2004] also studied the addressing problem in small groups, using manually annotated gaze and close-talk microphones. This work studied the single-person and multi-person addressee cases separately, and reported high classification accuracy for these two addressing classes, using basic features extracted from people’s gaze and speech utterances as units. In other work, based on a small subset of the AMI corpus, Gupta et al. reported an approach for addressee classification that outputs the addressee of those dialogue acts that contain referential uses of “you” [Gupta et al., 2007]. For a given utterance emitted by a speaker, four different class labels are used: one for the potential addressee to speak next; two more for the other two remaining participants based on the order in which they next speak; and a final one to represent addressing to the entire group. Lexical features and features of the conversational history were extracted; no visual information was used. A Conditional Random Field classifier achieved a significant improvement on predicting the previous and next speaker, although the overall performance highlights the complexity of the task. The biggest confusion was found to be between utterances being classified as the next speaker or the entire group.

Addressee detection is a problem that arises when technology makes the move from two-party man-machine natural dialogue systems to systems for multi-party conversations. In this context, the addressing problem has been addressed in the virtual agent literature [e.g. Traum, 2004] and in robotics [e.g. Katzenmeier et al., 2004]). Three examples can illustrate the variety of research problems that are relevant from this perspective, and the potential applications of the technology described in this section. First, Knott et al. described a multi-agent system for second language learning [Vlugter and Knott, 2006]. In this tutoring scenario, a rule-based system for detecting who is addressed by the learner was used. Second, in [Traum and Rickel, 2002], Traum’s rule-based method for addressee prediction, also discussed in Traum [2004], was used in an environment where humans have conversations with virtual characters in real time. More recently, the work on interactive

agents in multi-party situations by Bohus and Horvitz showed the feasibility of real-time inference of addressing patterns, based on audio-visual input and DBN reasoning [Bohus and Horvitz, 2009].

1.4 Towards social inference: dominance in small groups from conversational dynamics

The factors that determine the patterns in attention management, addressing, and floor management go beyond simple rules of managing the conversation and making it go smoothly. Attention, addressing, and floor management can also be used in strategic games, to exert control: who is getting the attention, who is allowed to speak, etc.

Within the context of meetings, dominance is typically viewed as exerting control over the social interaction together with an ability to influence. As Argyle puts it, dominant people want to talk a lot and to be influential in decisions [Argyle, 1994].

While variations in the dominance of individuals is natural, research on groups and group dynamics has shown that the best results for task completion and decision making come when all members of the group are able to give voice to their opinions and ideas [Nunamaker et al., 1991]. When one or two participants seek to dominate a discussion to the exclusion of others, the overall performance of the group is diminished. Thus, the idea is that dominance detection can be used to provide feedback, either during a meeting [Rienks and Heylen, 2006, Sturm et al., 2007] or as coaching afterward [Pianesi et al., 2008], to improve both the involvement of group members and the quality of decision making. Several studies have started to examine the automatic detection of dominance in small group meetings. Before the studies are surveyed below, the operational definition of dominance as it is laid down in annotation schemes is discussed first.

1.4.1 Annotating dominance in meetings

Judgments about the dominance and influence of meeting participant can either be first-hand, provided by the participants themselves in questionnaires after a meeting [Rienks et al., 2006], or annotated by observers of the meeting or meeting recordings [Rienks and Heylen, 2006, Jahagopi et al., 2009]. If the meeting participants know each other and have an ongoing relationship, first-hand dominance annotations have the advantage of being able to take this knowledge into account; on the other hand, annotations by external observers may be more likely to correspond to known verbal and nonverbal dominance cues [Dunbar and Burgoon, 2005]. It is an open question which type of dominance annotation provides more reliable data to support learning.

Dominance annotation, whether first-hand or by external observers, typically involves ranking meeting participants according to their perceived dominance or influence. The work by Rienks et al. ranked participants across entire meetings [Rienks and Heylen, 2006, Rienks et al., 2006]; the work by Jahagopi et al. annotated dominance judged over segments of meetings rather than meetings in their entirety [Jahagopi et al., 2009]. In the work by Rienks and Heylen [2006], ten annotators were used to rank the participants of eight meetings from the AMI and M4 corpora, with each annotator judging at most four meetings. Overall dominance rankings for each meeting were determined by summing up the individual rankings for each annotator. In further work by Rienks et al. [2006], first-hand judgments were used to obtain influence rankings for 40 meetings. In questionnaires, meeting participants were asked to rank all of the meeting participants. The final influence ranking for each meeting participant was determined by first summing up all the rankings that he or she received, normalizing, and then binning the normalized value into one of three influence categories. In the work by Jahagopi et al. [2009], 59 five-minute segments from 11 AMI meetings were annotated using 21 annotators. Each meeting was judged by a group of three annotators, with each annotator providing two types of judgments: the ranking of the dominance of the participants from 1 to 4, and a proportional ranking that split 10 units among all participants. A similar approach was taken more recently by Aran et al. [2010], which resulted in a corpus of 125 five-minute AMI meeting segments with dominance annotations.

1.4.2 Automatic dominance detection

Researchers have used a wide variety of features and machine learning methods to model dominance automatically. Many of the features that are typically used were inspired by work in social psychology on dominance and group dynamics. Bales [1950], Argyle [1994], and Dovidio and Ellyson [1982] all discuss the types of behavior associated with dominance. According to the literature, dominant individuals are more active, speak more, use eye contact when speaking, tend to interrupt other speakers, and so on. These characteristics can be encoded using speech and audio features such as speaking time, number of speaking turns, number of successful interruptions, number of times interrupted [Rienks and Heylen, 2006, Rienks et al., 2006, Jahagopi et al., 2009], and visual features such as measurements of visual activity and the number of visual activity interruptions [Jahagopi et al., 2009], and also looking time and number of looking events [Hung et al., 2008].

To estimate the dominance of meeting participants, Rienks and Heylen [2006] used support vector machines (SVM). Although the number of samples in their study is small, their results do suggest that it is the speakers

with the lowest and highest dominance that are easiest to predict, with the number of speaker turns and the number of successful interruptions being important features. Rienks et al. [2006] later compared this approach with other supervised learning methods, including multi-layered perceptrons, decision trees, and naive Bayes, and also with an unsupervised DBN approach. Among the findings in the study was that features representing turn information performed well for both supervised and unsupervised methods.

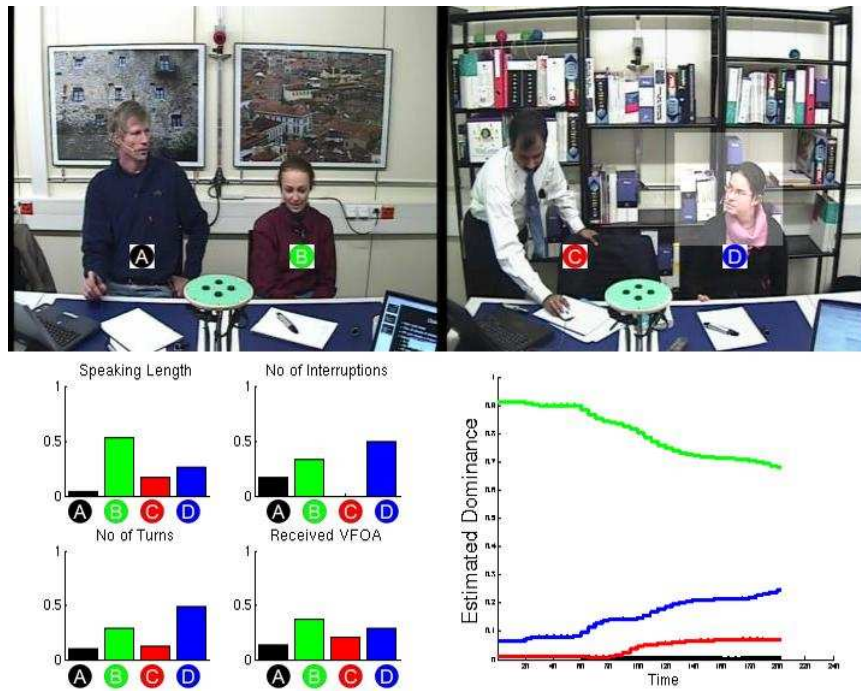


Figure 1.5: Automatic inference of dominance on AMI data, based on the work of Jahagopi et al. [2009]. The top panels show the four meeting participants, each represented with a letter and a circle of distinct color. The current speaker (person D) is highlighted by a gray square. The bottom left panel shows four normalized features (speaking time, speaking turns, interruptions, and received visual focus) extracted for each participant and accumulated over time. Person B talks the most and is the focus of visual attention the longest time. The bottom right panel shows the SVM-based, estimated dominance values for each participant over time. The x-axis spans the five-minute duration of the meeting segment. Person B is inferred to be the most dominant one, and after some time the inferred dominance patterns tend to stabilize.

Jahagopi et al. [2009] modeled the problem of dominance estimation a bit differently. Rather than trying to predict the ranking of individual par-

ticipants, they focused on predicting the most and least dominant participants. They considered two methods. The first was a simple, unsupervised approach that estimated dominance by summing a given feature (e.g., speaking time) over the full meeting segment. The participant with the highest feature sum was inferred to be the most dominant, and the one with the lowest was considered the least dominant. This method was first proposed by Hung et al. [2007]. For their second method, they trained SVM models using various sets of audio and visual features. Their results showed that for the clear cases (meetings in which there was 100% agreement among annotators, which corresponded to roughly 50% of the data for the most dominant and least dominant cases), the most and least dominant participants were best predicted using only audio features. However, when the less clear cases were included (meetings in which there was only majority agreement, corresponding to over 90% of the data), it was the combination of audio and visual features that achieved the best performance. An example of estimated dominance over time for a specific meeting can be seen in Figure 1.5). Work by Hung et al. [2008] further investigated the estimation of most and least dominant people from joint features of visual attention and speaking activity. Finally, Aran and Gatica-Perez [2010] studied score-level and rank-level fusion strategies to improve dominance estimation from audio and visual features.

The work discussed above shows the importance of the automatic identification of attention and floor management features. In most studies, the features that relate to turn-taking also have a high positive impact on dominance estimation. Importantly, the interest for developing recognition methods for other social perception concepts (i.e., leadership, personality, and roles) using features related to conversational dynamics has grown in the last years; this hints at the relevance of future computing research in small-group analysis and understanding.

1.5 Open issues

The previous sections have shown that initial progress in various directions of conversational modeling in small groups has been achieved, but many problems are still open. We believe that further progress in this domain will require of advances in at least two areas.

The first one is sensing. Current work has shown that the analysis of conversational dynamics is feasible given high-quality sensors and controlled - in practice almost always fixed - sensor setups. However, in order to achieve robust and accurate performance in the real world, more powerful and flexible sensing platforms are needed. Both industry and academia are currently pursuing this direction, which includes portable microphone arrays for audio capture, camera arrays that are easy to deploy and reconfigure, and

wearable solutions. Microsoft Kinect is the first of a new generation of commercial sensors that will largely influence future research on conversational modeling [Shotton et al., 2011]. Smartphone-based sensing is another area that will steer research on conversational modeling in daily life [Lane et al., 2010].

The second area is mathematical modeling of small-group interaction. While social science theories for some of the phenomena discussed in this chapter are firm, there is a clear need for computational models that better integrate such concepts and frameworks. As shown in the chapter, several machine learning models can currently be applied to conversational recognition or discovery tasks, but they often have built-in oversimplifying assumptions, and so learning methods that can more accurately describe human communication as interacting streams of multimodal information will likely have impact on future research.

Applications are the ultimate drivers for the technology discussed in this chapter. The automatic identification of conversational attention, turn taking, and addressing can be useful, both as stand-alone modules and as part of larger systems related to multi-party interaction, including indexing and summarization for off-line meeting support, real-time meeting assistance, and self and group awareness. We expect to see them integrated, in personally and socially acceptable forms, in many group interaction spaces of the future.

1.6 Summary and further reading

This chapter introduced three basic phenomena that contribute to shape conversational dynamics in small groups: attention, turn taking, and addressing. These mechanisms are multimodal, i.e., they are expressed and perceived via multiple channels – verbally and nonverbally, through sound and vision. Each of these concepts was then discussed from the computational viewpoint, by reviewing some of the current automatic methods that identify these conversational patterns from audio and video data. As one example of the relevance of these patterns in social perception, we reviewed existing work on automatic recognition of dominance in group conversations. Finally, some of the multiple open problems in this area were briefly discussed.

The literature on audio-visual methods for analyzing group conversations has steadily grown in the past years. A first-hand recount of nearly a decade of research on conversational sensing and social inference is provided by Pentland [2008]. Additional reviews related to small group conversational analysis include [Gatica-Perez, 2009, Vinciarelli, 2009, Shivappa et al., 2010]. A more detailed treatment of computational inference of dominance and related concepts in small groups can be found in [Aran and Gatica-Perez,

2011]. Finally, we have not discussed multimodal computational approaches for dyadic conversations; the interested reader can refer to Morency [2010] for a concise introduction.

1.7 Acknowledgments

We thank the support of the Augmented Multi-Party Interaction with Distant Access (AMIDA) European Project. D. Gatica-Perez thanks the support of the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2). R. op den Akker thanks the support of the European Network of Excellence SSPNet. Finally, we thank Jean-Marc Odobez (Idiap) for the picture used in Figure 1.2.

Bibliography

- M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, pages 52–63, May 2005.
- O. Aran and D. Gatica-Perez. Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*, Istanbul, Turkey, Aug. 2010.
- O. Aran and D. Gatica-Perez. Analysis of social interaction in group conversations: Modeling social verticality. In A. Salah and T. Gevers, editors, *Computer Analysis of Human Behavior*. Springer, 2011.
- O. Aran, H. Hung, and D. Gatica-Perez. A multimodal corpus for studying dominance in small group conversations. In *Proc. LREC Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, Malta, May 2010.
- Michael Argyle. *The Psychology of Interpersonal behaviour*. Penguin Books, Fifth Edition, 1994.
- S. O. Ba and J-M Odobez. A probabilistic framework for joint head tracking and pose estimation. In *Proc. Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4*, pages 264–267, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2128-2.
- S. O. Ba and J-M Odobez. A study on visual focus of attention modeling using head pose. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, pages 75–87, Washington, DC, USA, 2006. Springer Berlin / Heidelberg. ISBN 978-3-540-69267-6.
- S. O. Ba and J-M Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP' 08)*, pages 2221–2224, Las Vegas, NV, USA, March 2008. IEEE Computer Society. ISBN 978-1-4244-1483-3.

- Sil  ye O. Ba and Jean-Marc Odobez. Multi-person visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(1):101–116, January 2011.
- R. Bales. *Interaction Process Analysis*. Addison-Wesley, 1950.
- S. Banerjee and A. Rudnický. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2189–2192, Jeju Island, Korea, October 2004. IEEE Computer Society.
- Dan Bohus and Eric Horvitz. Dialog in the open world: platform and applications. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI-MLMI)*, pages 31–38, 2009.
- N. Campbell and D. Douchamps. Processing image and audio information for recognizing discourse participation status through features of face and voice. In *Proc. INTERSPEECH*, pages 730–733, Antwerp, Belgium, August 2007.
- J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, , and P. Wellner. The ami meeting corpus: A pre-announcement. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh,UK, July 2005. Springer Berlin / Heidelberg.
- L. Chen, T. R. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, , and T. Huang. Vace multimodal meeting corpus. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh,UK, July 2005. Springer Berlin / Heidelberg.
- L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, , and F. Quek. A multimodal analysis of floor control in meetings. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington, DC, USA, Mayo 2006. Springer Berlin / Heidelberg.
- Lei Chen and Mary P. Harper. Multimodal floor control shift detection. In *Proc. ICMI-MLMI*, Edinburgh, UK, 2009.
- H.H. Clark and T.B. Carlson. Hearers and speech acts. *Language*, 58(2): 332–373, June 1982.
- S.J. Cowley. Of timing, turn-taking, and conversations. *Journal of Psycholinguistic Research*, 27(5):541–571, 1998.
- N. Van der Vliet. Who’s got the floor? Master’s thesis, University of Twente, Netherlands, 2006.

- A. Dielmann and S. Renals. Dynamic bayesian networks for meeting structuring. *IEEE Transactions in Multimedia*, 9(1):25–36, 2007.
- J.F. Dovidio and S.L. Ellyson. Decoding visual dominance: attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, 1982.
- N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.
- C. Edelsky. Who’s got the floor? *Language and Society*, 10(3):383–421, 1981.
- D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, Special Issue on Human Naturalistic Behavior*, 2009.
- E. Goffman. Footing, 1981.
- C. Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York, NY, 1981.
- S. Gupta, J. Niekrasz, M. Purver, , and D. Jurafsky. Resolving “you” in multi-party dialog. In *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September 2007. Association for Computational Linguistics.
- R. Hayashi. Floor structure of english and japanese conversation. *Journal of Pragmatics*, 16(11):1–30, 1991.
- H. Hung, D. Jayagopi, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Chania, Greece, 2008.
- Hayley Hung, Dinesh Jayagopi, Chuohao Yeo, Gerald Friedland, Sileye Ba, Jean-Marc Odobez, Kannan Ramchandran, Nikki Mirghafori, and Daniel Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *MULTIMEDIA ’07: Proceedings of the 15th international conference on Multimedia*, pages 835–838, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-702-5. doi: <http://doi.acm.org/10.1145/1291233.1291423>.
- Dinesh Jahagopi, Haylen Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations from non-verbal activity cues. *Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

- W. James. *The Principles of Psychology*, volume 1–2. Dover Publications, 1890.
- N. Jovanovic. *To Whom it May Concern: Addressing in Face-to-Face Meetings*. PhD thesis, Department of Computer Science, University of Twente, Netherlands, March 2007.
- N. Jovanovic and R. op den Akker. Towards automatic addressee identification in multi-party dialogues. In *Proc. SIGdial Workshop on Discourse and Dialogue*, Boston, MA, USA, April 2004. Association for Computational Linguistics.
- N. Jovanovic, R. op den Akker, and A. Nijholt. A corpus for studying addressing behavior in multi-party dialogues. In *Proc. SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal, September 2005.
- N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *Proc. Conf. European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, April 2006.
- M. Katzenmeier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human human robot interactions based on head pose and speech. In *Int. Conf. on Multimodal Interfaces (ICMI)*, State College, PA, 2004.
- M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. 6th ed., Wadsworth Publishing, 2005.
- Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, and Junji Yamato. Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI-MLMI)*, pages 99–106, 2009.
- Nicholas Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150, Sep 2010.
- I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interactions in meeting. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Meeting Recognition Workshop (ICASSP’ 03)*, Hong-Kong, China, April 2003. IEEE Computer Society.
- I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

- Louis-Philippe Morency. Modeling Human Communication Dynamics. *IEEE Signal Processing Magazine*, 27(6), September 2010.
- B. Noris, K. Benmachiche, and A. Billard. Calibration-free eye gaze direction detection with gaussian processes. In *Proc. Int. Conf. on Computer Vision Theory and Applications*, January 2008.
- J. F. Nunamaker, Alan R. Dennis, Joseph S. Valacich, Douglas Vogel, and Joey F. George. Electronic meeting systems. *Commun. ACM*, 34:40–61, July 1991.
- H. op den Akker and M. Theune. How do i address you? modelling addressing behavior based on an analysis of a multi-modal corpus of conversational discourse. in *Proc. AISB Symposium on Multimodal Output Generation (MOG 2008)*, April 2008.
- K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI'05)*, Trento, Italy, October 2005. ACM.
- K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Conversation scene analysis with dynamic bayesian network based on visual head tracking. In *Proc. IEEE Int. Conf. on Multimedia (ICME, Toronto, Ontario, Canada, July 2006)*. IEEE.
- K. Otsuka, J. Yamato, and H. Sawada. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI'07)*, Nagoya, Aichi, Japan, November 2007.
- K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI'08)*, Chania, Greece, October 2008. ACM.
- K.C.H. Parker. Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965–971, 1998.
- A. Pentland. *Honest Signals: How They Shape Our World*. MIT Press, Cambridge, MA, 2008.
- Fabio Pianesi, Massimo Zancanaro, Elena Not, Chiara Leonardi, Vera Falcon, and Bruno Lepri. Multimodal support to group dynamics. *Personal and Ubiquitous Computing*, 12:181–195, 2008.
- R.J. Rienks and D.K.J. Heylen. Automatic dominance detection in meetings using easily obtainable features. In H. Bourlard and S. Renals, editors,

- Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI 2005*, volume 3869 of *Lecture Notes in Computer Science*, pages 76–86, Berlin, 2006. Springer Verlag. ISBN=978-3-540-32549-9.
- Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wilfried Post. Detection and application of influence rankings in small group meetings. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 257–264, New York, NY, USA, 2006. ACM.
- H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.
- Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao. Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A Survey. *Proceedings of the IEEE*, 98(10):1692 – 1715, October 2010.
- Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from a single depth image. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.
- R. Stiefelhagen. Tracking focus of attention in meetings. In *Proc. IEEE Int. Conf. on Multimodal Interfaces(ICMI '02)*, pages 273–280, Los Alamitos, CA, USA, 2002. IEEE Computer Society. ISBN 0-7695-1834-6. doi: <http://doi.ieeecomputersociety.org/10.1109/ICMI.2002.1167006>.
- R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.
- J. Sturm, O.H. Herwijnen, A. Eyck, and J. Terken. Influencing social dynamics in meetings through a peripheral display. In *Proc. of the 9th Int. Conf. on Multimodal Interfaces (ICMI '07)*, pages 263–270. ACM New York, NY, USA, November 2007. ISBN 978-1-59593-817-6. doi: <http://doi.acm.org/10.1145/1322192.1322238>.
- Y. Takemae, K. Otsuka, and N. Mukawa. An analysis of speakers’ gaze behavior for automatic addressee identification in multiparty conversation and its application to video editing. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication*, pages 581– 586, September 2004.
- D. Traum. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*. Springer, 2004.

- David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proc. Int. Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, Jul. 2002.
- Alessandro Vinciarelli. Capturing order in social interactions. *IEEE Signal Processing Magazine*, 26(5), September 2009.
- P. Vlugter and A. Knott. A multi-speaker dialogue system for computer-aided language learning. In *Proc. Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, Potsdam, Germany, Sep. 2006.
- D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered hmms. *IEEE Trans. on Multimedia*, 8(3):509–520, 2006.