Meeting Modelling in the Context of Multimodal Research

Dennis Reidsma, Rutger Rienks, and Nataša Jovanović

University of Twente, Dept. of Computer Science, HMI Group
P.O. Box 217, 7500 AE Enschede, the Netherlands,
{dennisr,rienks,natasa}@ewi.utwente.nl *

Abstract. This paper presents a framework for corpus based multimodal research. Part of this framework is applied in the context of *meeting modelling*. A generic model for different aspects of meetings is discussed. This model leads to a layered description of meetings where each layer adds a level of interpretation for distinct aspects based on information provided by lower layers. This model should provide a starting point for selecting annotation schemes for layers of the meeting and for defining a hierarchy between individual layers.

1 Introduction

Meetings are an important part of daily life. They are a complex interplay of interaction between participants with each other and their environment, and contain a broad spectrum of multimodal information [1]. The interactions can be (automatically) analyzed to gain knowledge about multimodal human-human interaction. There is also a need for applications that support meetings in several ways. For example, an increasing number of meetings involves remote participation, where effectiveness of the meetings depends on support from the remote conferencing application [2]. Furthermore, people often want access to material from past events. Good summarization and browsing tools are indispensable for this. Simulation tools provide a method for validation of models underlying the technology.

This paper starts by presenting a framework for corpus based research on multimodal human-human interaction which interrelates aspects such as those described above. Part of the framework is subsequently elaborated for meetings.

The work presented in this paper is carried out in the context of the AMI project. Other projects that work on the same subjects are for example the Meeting Room project at Carnegie Mellon University [3], the M4 project [4] and the NIST Meeting Room Project [5].

The structure of this paper is as follows. Section 2 describes a general framework for research on multimodal human-human interaction that will be used to

* This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-13). For more information see http://www.amiproject.org/

position our work. The main part of the paper is presented in Sections 3 and 4, where the annotation and modelling parts of the framework are tailored for the meeting domain. The resulting model will be discussed extensively. The paper ends with a discussion of several issues related to annotations of corpora 5.

2 A General Framework for Corpus Based Research on Human-Human Interaction

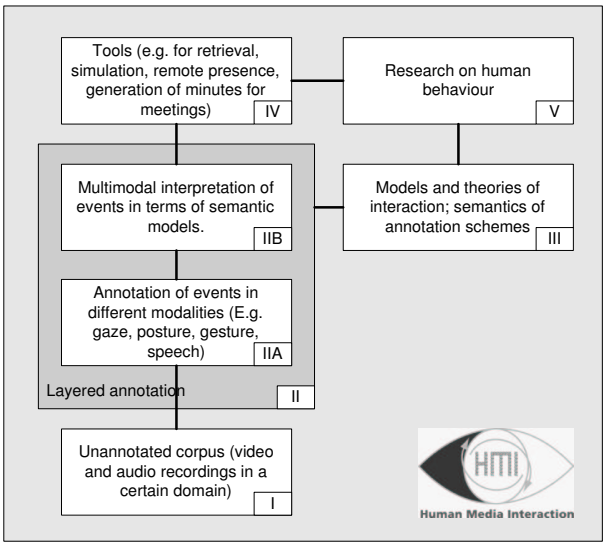

Fig. 1. A framework for corpus based research on human-human interaction

This section discusses our view on research on human-human interaction, of which Figure 1 shows a schematic representation. Based on a corpus of interaction recordings (Box I), manual or automatic recognition processes create layers of annotation (Box II). These annotations are used by the different tools (Box IV). As discussed above, research in social sciences (Box V) contributes to development of techniques and tools, but also makes use of the information provided by them. As will be discussed below, models and theories of human interaction (Box III) are of major importance in this context. The rest of this section discusses the different boxes in more detail. After that the remainder of the article will elaborate on the boxes II and III.

2.1 Box I: Corpus

Research on multimodal interaction often uses a corpus of audio and video recordings. In general, a corpus should be representative of the domain, be large

enough to do relevant research and be accessible. Many projects use smart rooms to record data. These smart rooms are equipped with a range of sensors (visual, aural and other types) that allow, often detailed, capturing of the interactions in the room. Examples of existing corpora containing meeting recordings are those used in the Meeting Room project at Carnegie Mellon University [3], in the The Meeting Recorder Project at ICSI [6] and in the M4 project [4].

2.2 Box II: Layered Annotation

The box for layered annotation is divided into two sub boxes. This division reflects the difference between direct annotation of objective events and interpretation of these events on a semantic level, or *form* and *function*. Consider e.g. the situation where a user raises his or her hand. In box IIA this is detected and annotated as an event HAND-RAISING. In box IIB this event might be interpreted as e.g. a request for a dialogue turn or as a vote in a voting situation, based on analysis of speech transcripts. The semantic models which underly the interpretations in box IIB belong to box III and are discussed below.

Box IIA This box involves (semi) automatic recognition of events in interaction recordings. Techniques such as computer vision or speech processing are used to obtain representations of the interactions in an efficient way.

Box IIB On this level fusion techniques are applied on the annotations from box IIA to align them and to interpret them in terms of semantic models from box III. This leads to annotations that are semantically oriented, such as description of argumentative structures, intentions of users and other aspects that support more sophisticated processing of the data.

2.3 Box III: Models and Semantics

To enable the interpretations of annotations from box IIB, we need models of human interaction. There is a large variety of examples: models of the dependences between group behaviour and leadership style (see e.g. [7]), a model of the rhetorical relations between utterances (see e.g. [8]), an agent model incorporating emotions (see e.g. [9]) and many more. The actual models depend on the research objectives or the application. Research on the corpus can provide insights into what people may intend with certain behaviour and what types and patterns of behaviour exist.

2.4 Box IV: Tools

Research of human interaction such as described in this paper will also lead to the development tools for supporting this interaction. In the context of meetings,

some of these tools are useful for end users (e.g. a meeting browser, a minute generator, a remote meeting assistant), others will mostly be useful for supporting the annotation process (e.g. custom annotation modules).

Simulation tools can be used for testing hypotheses about certain aspects of human interaction based on these explanatory models in order to validate the models. See for example the work of Carletta et al. [10] where certain mechanisms for turn taking in small group discussions are examined through simulating the resulting floor patterns and comparing them with patterns observed in real life.

Virtual simulation using avatars is another current topic at the University of Twente [11, 12]. One type of virtual simulation may involve virtual replay of a meeting from specific viewpoints such as the viewing perspective of one participant. This could give the impression of watching the meeting through his or her eyes, especially if the participant's head orientation is included in the viewpoint. This might give researchers a unique perspective on the behaviour of meeting participants. This type of simulation can also be used for summarization purposes, creating one virtual meeting that contains the main events of several related meetings in a coherent fashion.

Remote presence tools that allow manipulation of the environment and the representation of participants, adding or exaggerating aspects of behaviour, enable experiments for research on human behaviour in specific (virtual) circumstances. Bodychat by Vilhjálmsson and Cassell [13] provides a good example of simulated nonverbal behaviour in a virtual multi-user environment.

2.5 Box V: Human Behaviour

Research on human behaviour, for example social psychology, provides an insight into human interaction patterns and their components. This in turn is a basis for automatic analysis of this interaction and the retrieval of components.

On the other hand, analysis of interactions opens possibilities for research of human behaviour. In the first place the corpus can be analyzed to discover regularities in human behaviour and construct corresponding models and hypotheses. In the second place the annotations can be used to test and evaluate these models, for example using simulations (see e.g. [10, 14]).

Work that can be placed in the context of this box is for example that of McGrath [15] (group dynamics and collaborative tasks), Carletta et al. [14, 10] (models of group behaviour) and Shi et al. [16] (gaze, attention and meeting dynamics).

3 Meeting Modelling: Overview

The framework in Figure 1 enables us to obtain a structured view on any work that is carried out in the field of multimodal human-human interaction research. In this section we extend the framework for the meeting domain by instantiating the boxes IIA and IIB from the generic framework in a way that covers

the meeting domain. Furthermore we will discuss issues related to the semantic models for box III underlying the higher level annotations of box IIB.

Marchand-Maillet [17] described a first approach of meeting modelling.

3.1 Box IIA and IIB for the Meeting Domain

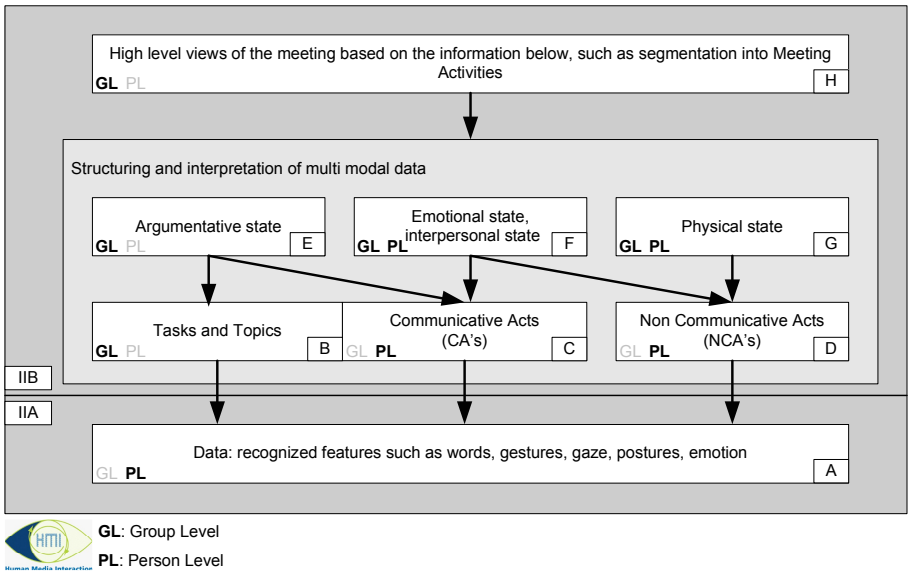

Fig. 2. Box IIA and IIB tailored to the meeting domain.

Figure 2 shows the proposed meeting model. The bottom level of the model corresponds to box IIA in Figure 1. Box IIB is divided into three levels, containing labelling of events with an interpretation on the lowest level, evolving state descriptions on the second level and global meeting parts on the highest level.

On the first level, single or combined events from box IIA are labelled with an interpretation. For instance a posture can be SITTING, a gesture or utterance may be an acknowledgement, etc.

The second level of interpretation consists of different types of incremental states that evolve during a meeting. Each occurrence of an event on the first level adds to one or more of these states. The physical state of a meeting (the setting), for example, is dependent on non-communicative actions of the participants (walking, moving objects, etc). These two levels together constitute an extensive *discourse state* of the meeting.

The topmost level of interpretation is the highest level of semantic interpretation of distinct meeting parts (e.g. a sequence of agenda items or meeting activities).

Clark defines a joint *activity* as an activity with more than one participant that serves a common goal [18]. A joint activity consists of joint subactivities and joint *actions*. A joint action consists of a group of people coordinating with each other in executing a step in the process of achieving the common goal of the joint activity. Each person can have his or her own *role* in that joint action.

Meeting activities are joint activities in this sense. One of the goals of the AMI and M4 projects is to find these kinds of group events in order to detect meeting structures (see e.g. [19]). To detect these highest level components a combination of the first and second level of interpreted information is needed.

Consider for example a situation with one person standing in front of the whiteboard. This person is the only one who is talking. He or she points at the projector screen every now and then. The argumentative structure of the communication is very simple (not many counterarguments and few conflicting statements). This evidence suggests that at the highest level this meeting segment can be interpreted as a 'presentation'. Other possibilities can be 'discussion', 'break' and 'brainstorm'. See [20] for a possible list of meeting activities. Note that in other publications these meeting activities are called *meeting acts* or *meeting actions*. To avoid confusion with the low level meeting actions this paper will employ the term *meeting activity* for the high level activities.

4 Meeting Modelling in More Detail

This section will explain the boxes at each level in Figure 2 in more detail. For each box is indicated whether the components from that box belong to the personal or the group level of the meeting. Non-communicative acts such as MOVING A COFFEE MUG for instance can only occur at a personal level, where as the argumentative state will always be at a group level.

The arrows depict the dependencies between the different boxes: The set of arrows leaving a box represent the input for the recognition and interpretation algorithms that construct the data in that box. The evolution of the argumentative state is dependent on the communicative acts and tasks and topics at a certain segment in the meeting. Vice versa, different communicative acts have a different impact on the argumentative structure that is being constructed by the recognition modules.

4.1 The First Level

Box A: Unimodal Events
The bottom level of the model is the level with the basic, uncombined, uninterpreted events from different modalities. Since this level pertains to *uninterpreted* events there is no reference to complicated interpretation models for this box. It contains unimodal events such as words, non-word sounds, gestures, postures,

gaze direction, etc. on a personal level. Events such as the light falling out, falling chairs or the beamer malfunctioning are also included here.

4.2 The Second Level

The second level of the model contains the boxes B, C and D. Box B concerns *topics* and *tasks*. In box C and D we find the joint and autonomous actions as defined by Clark [18]. Some of these actions have a communicative intent (verbal or non-verbal) and can be annotated as Communicative Acts (box C). The other meeting actions, without communicative intent, are called Non-Communicative Acts (box D).

Box B: Topics, Tasks
Topic and task are closely entwined views on a meeting segment. The question "what is the segment about?" may as well be answered with a task description (e.g. "brainstorming") as with a topic (e.g. "the shape of the appliance that is being designed in this meeting"), but each answer is incomplete, if the other answer is not taken into consideration as well. Tasks and topics can exist both on a person and on group level. They can be part of a hierarchy of subtasks and subtopics. Both may be related to agenda points. A task can be evaluated in terms of its progress. As background knowledge for this box, the meeting agenda gives an indication of several tasks and topics that will appear in the meeting.

Box C: Communicative Acts
A communicative act (CA) is a joint action with communicative intent, i.e. through which a person makes others understand what he or she means. CAs may be verbal or non-verbal or a combination of these. Examples of CAs are asking a question, answering and giving a suggestion. The following aspects are relevant for modelling CAs.

1. Function: the type of CA, see the examples above
2. Agent: the participant performing the CA
3. Addressee: the participant(s) to whom the CA is directed
4. Target: in case of some non-verbal CAs such as pointing there may be a target
5. Relation to other CAs (e.g. question/answer combinations)

A concrete choice for an annotation scheme for CAs is based on the functional aspect and might be based on an adaptation of some dialogue act annotation scheme such as MRDA [21], SWBD-DAMSL [22] or the IPA of Bales [23].

The boxes B and C have been drawn together as we expect a close relation between these boxes. We expect that for various tasks carried out during the meeting there will be specific sets of CAs occurring more frequently for this task than for any other task. We also expect that for every type of CA there will be a set of tasks (within a meeting) where these CAs will appear with a higher frequency than in the other tasks.

We conducted some first experiments to test this hypotheses. The results tentatively confirm our expectations. Five meetings have been annotated with communicative act types, topics, temporal and personal aspects of topics and tasks. Analysis of the results shows a clear relation between certain aspects of the topic and task on one side and the most frequently used CAs on the other side. Some first conclusions from those experiments are the following. Meeting segments about topics dealing with future events contain significantly more action motivators (command, commitment, offer, etc.) than segments about topics dealing with past or present or atemporal events. Segments dealing with topics that directly relate to the speaker or the whole group as opposed to topics about something entirely external also contain significantly more action motivators.

Box D: Non-Communicative Acts
NCAs are much like CAs, but there is no communicative intent and therefore no addressee. Examples are manipulation of objects or walking to a different location.

4.3 The Third Level

On the third level we distinguish three different components: the physical, the emotional and the argumentative states. All of these non overlapping components can be created out of the level two components.

Box E: Argumentative State
For effective summarization and structuring of meeting content the argumentations posed in the meeting should be interpreted and structured. An argumentative state consists of several aspects. In the first place there is the logical *argument structure* describing the relations between statements, evidence, support arguments and counterarguments. Much work has been done in extension of the IBIS system of Kunz and Rittel [8].

In the second place there is the *argument discourse*. This discourse concerns the order and manner in which argumentations are introduced. Pallotta and Hatem describe both argument structure and discourse in one formalism [24]. How people select and introduce their arguments depends on the goals during the current task of both the group and the individuals [25].

Communicative Acts can introduce positions, arguments etc. but can also aim at discussion *control* rather than discussion *content*. For example, the sentence "We will discuss this later in the meeting" does not change the logical argument structure but is certainly relevant to understanding the decision process from a rhetorical perspective.

Another rhetorical technique to reach one's goals in a discussion concerns the use of *emotional* arguments. In its simplest form this could, for example, be the purely emotional rejection of a statement using the expression "I don't feel it that way", which, although it contains no *logical* argument, can influence the outcome of the discussion. Carofiglio and De Rosis [26] state that argumentation

knowledge and emotion should be combined to select the most promising strategy to reach one's goals.

Information about previous meetings might, as background knowledge, be relevant for understanding some of the argument structures.

Box F: Emotional state
The second box on this level depicts the emotional state of the meeting on group level and personal level. Aspects such as the mood of the meeting, the emotional state of the individuals or the attitudes of people with respect to each other (interpersonal state) will be modelled here. They can be derived from NCAs (e.g. yawning, posture) as well as CAs (explicitly stating your attitude, or a certain choice of CAs which may signal a specific emotion). People can be indifferent, angry, cooperative, positive, afraid at all times during the meeting.

Background information in this box could be personality, relative status, interpersonal attitudes and meeting roles.

Box G: Physical state
The physical state of a meeting consists of the meeting objects, participants and their locations. Changes to this state follow from observed non-communicative acts.

4.4 The Fourth Level

Box H: High level segmentation
On this level a meeting is segmented into meeting activities, joint activities that are specific for meetings (cf introduction). A subjective approach to define meeting activities would be to use "common sense segmenting". A meeting then consists of a sequence of "typical components" that make up the meeting. A model (cf. introduction) can be developed that explains why certain activities are the highest level components in meetings and how they are related.

Higher level statistical models can be trained to classify combinations of occurrences from lower level components as specific types of higher level components. A recognition module would use such a model to detect the occurrence of e.g. a presentation activity when observing a pattern of certain NCAs, CAs and physical state combined with certain argumentative structures in the data. The training can be done using machine learning techniques like Hidden Markov Models [19] or Dynamic Bayesian Networks, looking for statistical correlations.

This level requires a lexicon of meeting activity types. An example of such a lexicon can be found in [20]. An approach to obtain such a lexicon is the use of unsupervised clustering techniques on the lower level annotation elements, as mentioned in [27]. The resulting clusters can be indicative of possible meeting activity types.

5 Annotations

As meetings can be structured in layers and we wish to label or annotate chunks of data in accordance with these layers, there is a need for an annotation language that supports these structures. An annotation format can be seen as an instantiation of a model. A model describes how the annotation should look like, which annotation structures are possible and what these structures mean. This implies, however, that if the model changes, the annotations are influenced as well and vice versa.

The choice of annotation schemas and structures for the separate boxes should in most applications be inspired by explanatory models of humans interaction and the application goals. Different models or different uses of the models may lead to distinct annotation schemas for the information in the boxes.

5.1 Manual Annotations

The annotations discussed above are not necessarily automatically produced: corpus based work always involves a large amount of manual annotation work as well. There are several reasons for creating manual annotations of corpus material. In the first place ground truth knowledge is needed in order to evaluate new techniques for automatic annotation. In the second place high quality annotations are needed to do social psychology research on the corpus data. As long as the quality of the automatic annotation results is not high enough, only manual annotations provide the quality of information to analyze certain aspects of human behaviour.

It is a well known problem that manual annotation of human interaction is extremely expensive in terms of effort. Annotating a stretch of video with not-too-complicated aspects may easily take ten times the duration of that video. Shriberg et al. report an efficiency of 18xRT (18 times the duration of the video is spent on annotating) on annotation of Dialog Acts boundaries, types and adjacency pairs on meeting recordings [28]. Simple manual transcription of speech usually takes 10xRT. For more complicated speech transcription such as prosody 100-200xRT has been reported in Syrdal et al. [29]. The cost of syntactic annotation of text (PoS tagging and annotating syntactic structure and labels for nodes and edges) may run to an average of 50 seconds per sentence with an average sentence length of 17.5 tokens (cf. Brants et al. [30], which describes syntactic annotation of a German newspaper corpus). As a final example, Lin et al. [31] report an annotation efficiency of 6.8xRT for annotating MPEG-7 metadata on video using the VideoAnnEx tool. The annotation described there consists of correction of shot boundaries, selecting salient regions in shots and assigning semantic labels from a controlled lexicon. It may be obvious that more complex annotation of video will further increase the cost.

The type of research for which the framework described in this paper is developed requires not one or two annotation types on the data but a rich set of different annotations. It is therefore an important task to cut down the time

needed to annotate multimodal data. This section discusses a few approaches to this problem.

5.2 Efficient Interfaces for Manual Annotation

In the first place, whenever manual annotation is inevitable, the efficiency of creating this annotation is heavily dependent on the user interface of the annotation tool. Different aspects (dialogue act labelling, hand gestures, head orientation) are best annotated with different strategies. Figure 3 shows an example of such a specifically targeted strategy, an annotation module for head orientation. The figure shows the video of a person whose head orientations are to be annotated. A coffee mug is used as a proxy for his head. A flock-of-birds 6 DOF position and orientation sensor is attached to the side of the mug, keeping track of its orientation. Rotating the mug in response to head movements in the video results in a real-time annotation of the head orientation of the person in the video. A small 3D display of a virtual head follows the mug for verification of the annotation quality. For fragments with fast head movements the video speed can be slowed down. On those fragments an efficiency ratio of 1:2 per annotated person can be achieved.

Fig. 3. Realtime annotation of head orientation

5.3 Semi-Automatic Annotation

Apart from interface improvements in order to increase the annotation efficiency, tools are developed to perform annotations automatically. Although many of

these efforts result in low-quality annotations, some automatic annotation procedures can provide support for for manual annotation. These kinds of semi-automatic annotation techniques are already applied for audio transcriptions and video segmentation (e.g. manual correction of automatically detected boundaries). Manual correction of automatic annotations is much faster than complete manual annotation [29].

For example, the labelling of hand gestures can be facilitated with a semi-automatic procedure. Gesture labelling schema's such as that of McNeill [32] describe gestures using attributes such as hand shape, symmetry, category of gesture and the relative location of the gesture with respect to the subject's body. McNeill uses the gesture space of Pedelty (see figure 4) to distinguish gesture locations. There are reliable automatic recognition procedures for hand and face tracking under the right conditions (see e.g. [33]). Therefore the annotation environment can reliably fill in the relative location attributes. Furthermore the information contained in this tracking data can give an indication of which gesture type occurred, allowing the annotation environment to give a default suggestion for that attribute. Syrdal et al. show that, at least in some cases, default suggestions can speed up manual labelling without introducing much bias in the annotation [29].

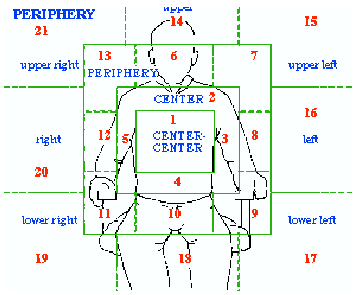

Fig. 4. The gesture space of Pedelty, 1987

5.4 Integrating Annotations

The annotations should be stored in a suitable data format, supported by a powerful API toolkit. Such a toolkit should meet several requirements. It should support development of custom annotation modules such as described above. Different layers of annotation should be easy to integrate. Transformation of annotations from other formats into the desired format should be supported. The data formats should allow explicit expression of the layered structure of the annotations and their derivational relations. There are many publications dealing with these issues, the reader is referred to [34] for more information.

We chose to work with the layered stand-off annotation format from The University of Edinburgh, the Nite XML Toolkit [35]. The layered structure of the annotations described before is very easy to map on a data model in this toolkit. It enables each of the component classes from the various layers to be supplied with their own models and annotation schemas. Higher level annotations can relate to or be dependent on lower level annotations.

6 Conclusions

This paper presents a view on corpus based research in the domain of human-human interaction in general. Various kinds of research can be positioned in this framework, relating them to each other. Part of the framework is elaborated for meetings. Different aspects of meeting modelling are discussed. Annotations are an important aspect of corpus based research. Continuing focus on development of efficient annotation tools and schemas is needed, where the underlying models should guide the development.

References

1. Armstrong, S., et al, A.C.: Natural language queries on natural language data: a database of meeting dialogues. In: Proc. of NLP and Information Systems-NLDB'03. (2003) 14–27
2. Vertegaal, R.: Who is looking at whom. PhD thesis, University of Twente (1998)
3. http://www.is.cs.cmu.edu/meeting_room/.
4. http://www.m4project.org/.
5. http://www.nist.gov/speech/test_beds/mr_proj/.
6. http://www.icsi.berkeley.edu/Speech/mr/.
7. Hersey, P., Blanchard, K.: Management of Organizational Behavior: Utilizing Human Resources. Prentice Hall (1988)
8. Kunz, W., Rittel, H.W.J.: Issues as elements of information systems. Working Paper WP-131, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung (1970)
9. Pelachaud, C., Bilvi, M.: Computational model of believable conversational agents. Communication in MAS: background, current trends and future (2003)
10. Padilha, E., Carletta, J.: Nonverbal behaviours improving a simulation of small group discussion. In: Proc. 1st Nordic Symp. on Multimodal Comm. (2003) 93–105
11. Nijholt, A.: Meetings, gatherings and events in smart environments. In: Proc. ACM SIGGRAPH VRCAI2004. (2004) to appear.
12. Nijholt, A.: Where computers disappear, virtual humans appear. Computers and Graphics **28** (2004) to appear.
13. Vilhjalmsson, H., Cassell, J.: Bodychat: Autonomous communicative behaviors in avatars. In: Proc. of the 2nd Annual ACM Int. Conf. on Autonomous Agents. (1998)
14. Carletta, J., Anderson, A., Garrod, S.: Seeing eye to eye: an account of grounding and understanding in work groups. Cognitive Studies: Bulletin of the Japanese Cognitive Science Society **9(1)** (2002) 1–20
15. McGrath, J.: Groups: Interaction and Performance. Prentice Hall (1984)

16. Shi, Y., Rose, R.T., Quek, F., McNeill, D.: Gaze, attention, and meeting dynamics. In: ICASSP2004 Meeting Recognition Workshop. (2004) to appear.
17. Marchand-Maillet, S.: Meeting record modelling for enhanced browsing. Technical Report 03.01, Computer Vision and Multimedia Laboratory, University of Geneva (2003)
18. Clark, H.: Using language. Cambridge University Press (1996)
19. McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G.: Automatic analysis of multimodal group actions in meetings. Technical Report IDIAP-RR 03-27, IDIAP (2003)
20. Jovanovic, N.: Recognition of meeting actions using information obtained from different modalities. Report TR-CTIT-03-48, CTIT (2003)
21. Dhillon, R., Bhagat, S., Carvey, H., Shriberg, E.: Meeting recorder project: Dialogue act labeling guide. Technical report, ICSI Speech Group, Berkeley, USA (2003)
22. Jurafsky, D., Shriberg, E., Biaska, D.: Switchboard SWBD-DAMSL shallow-discourse-function annotation (coders manual, draft 13). Technical report, Univ. of Colorado, Inst. of Cognitive Science (1997)
23. Bales, R.F.: Interaction Process Analysis. Addison-Wesley (1951)
24. Pallotta, V., Hatem, G.: Argumentative segmentation and annotation guidelines. Technical report (2003)
25. Gilbert, M.A.: Goals in argumentation. In: Proc. of the conf. on Formal and Applied Practical Reasoning. (1996)
26. Carofiglio, V., de Rosis, F.: Combining logical with emotional reasoning in natural argumentation. In: 3rd Workshop on Affective and Attitude User Modeling. (2003)
27. Dong Zhang, Daniel Gatica-Perez, S.B.I.M., Lathoud, G.: Multimodal Group Action Clustering in Meetings. In: ACM 2nd International Workshop on Video Surveillance & Sensor Networks in conjunction with 12th ACM International Conference on Multimedia. (2004) IDIAP-RR 04-24.
28. Shriberg, E., Dhillon, R., et al., S.B.: The icsi meeting recorder dialog act (mrda) corpus. In: Proc. HLT-NAACL SIGDIAL Workshop. (2004) to appear.
29. Syrdal, A.K., Hirschberg, J., McGory, J., Beckman, M.: Automatic tobi prediction and alignment to speed manual labelling of prosody. Speech communication **33** (2001) 135–151
30. Brants, T., Skut, W., Uszkoreit, H.: 5. In: Syntactic annotation of a German newspaper corpus. Kluwer, Dordrecht (NL) (2003)
31. Lin, C.Y., Tseng, B.L., Smith, J.R.: Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In: Proc of the TRECVID2003. (2003)
32. McNeill, D.: Hand and Mind: What gestures reveal about thought. University of Chicago Press (1995)
33. Brethes, L., Menezes, P., Lerasle, F., Hayet, J.: Face tracking and hand gesture recognition for human-robot interaction. In: Proc. of the International Conference on Robotics and Automation. (2004)
34. http://www.ldc.upenn.edu/annotation/.
35. Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann, H.: The NITE XML toolkit: flexible annotation for multi-modal language data. Behavior Research Methods, Instruments, and Computers **35(3)** (2003) 353–363