Recognition of meeting actions using information obtained from different modalities
-a semantic approach-

Natasa Jovanovic
University of Twente
Department of Computer Science
September 2003

Table of contents

1. Introduction

Meetings play an important role in everyday life. Meeting minutes can serve as a summary of a meeting but they can't provide a trustworthy representation of the meeting. A solution to this problem is to provide audio and video recordings of the meeting. It gives a more realistic representation of meetings but if we need some particular information like *"What was discussed?"* than it is necessary to replay these recordings several times from the beginning.

Therefore, it is very important to develop a system which will enable easy and efficient access to the meetings that have been archived. It is also important to enable the searching in a meeting archive by some criteria like topics, dates, participants, some specific actions during the meeting and also to retrieve a summary according to the user's specification. These are objectives of the M4 (Multimodal Meeting Manager) project. Meetings take place in smart rooms. Smart rooms are environments equipped with multimodal sensors and computers. Smart rooms can automatically identify attendants, transcribe and identify what they say etc. The M4 project is built on the ideas of smart rooms. It is concerned with the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings, using the outputs of a set of multimodal sensors. There are some ongoing projects with similar issues. The ICSI project is also concerned with the development of a system for recording and browsing meetings; based only on audio data [23]. The closest project to M4 is the Meeting Room project at Carnegie Mellon University [24]. It is concerned with the recording and browsing of meetings using audio and video data. The M4 project proposes several innovations: multimodal localization and tracking of meeting focus, automatic multimodal emotion and intent recognition, gesture and action recognition, textual and multimodal summarization and a framework for integration of multimodal data.

A meeting is a dynamic process which consists of group interactions between meeting participants. The group interactions in meetings are called *meeting actions*. In human-human interaction several communication systems are in use. Humans communicate not only by words but also by face expression, gaze, body and hand gestures etc. These verbal and non-verbal signals are highly connected and they together transmit complete information. In this report we will describe our semantic approach in modelling a meeting as a sequence of meeting actions. The semantic approach is based on representing the meaning of multimodal behaviour of a meeting participant using information obtained from different sources, as well as on recognition of meeting actions using semantic features (dialog acts, topics, participants activities, states and roles etc.) extracted from participants multimodal behaviour.

Section 2 describes social and psychological aspects of social human behaviour. Section 3 explains why we need a semantic approach in meeting modelling and compares the audio-visual feature level approach in modelling meetings with a semantic approach. Section 4 proposes a lexicon of meeting actions. Section 5 describes other aspects of meetings as target detection in meetings, user modelling, etc. Section 6 gives the motivation and explanation of the semantic approach as well as a technical specification of the semantic based model. Conclusions and future directions of research are given in Section 7.

2. Analysis of social interaction between participants in one group

In this section we will describe social psychological aspects of social human behavior. First we will describe verbal and non-verbal elements of social behavior. After that we will say something about the organization of verbal and non-verbal elements.

When two or more people are engaged in interaction they communicate using verbal and/or non-verbal elements. Some of them are intentional, some not. Verbal signals are carefully managed while non–verbal signals are more spontaneous and hardly controlled. Language is most useful for describing objects and other people. Non-verbal elements are useful for expressing emotions, interpersonal attitudes, focus of attention, etc. Nevertheless the most powerful and natural communication between humans is the combined use of verbal and non-verbal elements.

2.1 Non-verbal behavior

Argyle ([1]) described the following elements of non-verbal communication:

1) Bodily contact is a basic type of social behavior. Bodily contact can occur in a wide variety of ways and the extent of bodily contact depends of age, relationship between people, culture etc. Greetings or farewells may include shaking hands, kissing; hitting is an act of aggression; keeping hands during a whole interaction process may be a sign of companionship, etc. But all of these bodily contacts are usually done in a way defined by culture.
2) Proximity –people who are involved in an interaction have to choose a certain degree of proximity. Degree of proximity depends on relationship between participants, cultural elements, settings, and physical deficiency of participants. Deaf or shortsighted people will take greater proximity. People in a party are closer to each other than people in a meeting room or classroom.
3) Posture- can be classified in several varieties: sitting, standing, laying, kneeling etc. Each of these varieties can be further classified according to the manner in which it is done**.** Posture is a matter of cultural convention according to the situation. Posture reflects the status or the way a person perceives his status in relation to others. Posture also can express emotional state. For example, if a person stands with stretched legs, crossed arms and looking aside that could mean that he is surprised, suspicious, undecided, dominant etc. Beside that posture can be regarded as an aspect of personality.
4) Physical appearance – plays an important role during a meeting. The components of which attractiveness and other perceptions are based on are clothes**,** physique**,** face, hair, and hands. These aspects of physical appearance which are constant during encounter are partly biologically given and unchangeable and partly under voluntary manipulation of physical appearance.
5) Facial and gestural movements

   Previous mentioned elements of social behavior remain unchanged during the whole period of interaction. Facial and gestural movements are faster moving elements.

   The face is one of the most expressive areas of the body. Facial expression is the main second communication channel used at the same time as speech. In human social interaction facial expression shows the emotional state of person and attitudes to the other participants. It provides feedback on

whether he/she understands or agrees with what is being said, whether he/she is surprised etc.

Hands are expressive and visible part of the body but less expressive than facial expression. Hand movements of one individual vary from movements of the others. Interpretations of these movements are cultural dependent. In human social interaction hands are used for illustrations, accompanying speech. Sometimes gesture can replace speech (communication with a deaf person, communication between people that speak different languages). Hands movements may show emotional states and it is usually unintentional.

Head position and movements are highly visible but the amount of information that they convey is limited. Head position can help in detection of a person's focus of attention. Head movements mean different things in different cultures. Nodding or shaking the head are important elements in verbal communication. Nodding is used for expressing agreement, as a sign of understanding etc.

6) Gaze direction

One of the most important aspects of social interaction is gaze direction. Because gaze direction shows focus of attention, it is important for A whether B looks at him, at some other objects which they are both concerned with or other objects and people. Mutual-gaze or eye contact is the case when A looks at B and B looks at A. Eye-contact is significant when persons want to establish a relationship. They look longer at each other at the same time if they want to establish closer relationship. Looking at the other can be used to establish a special kind of relationship: dominant, dependent, affiliative etc. The process, which affects gaze direction, is the need for feedback. A wants to know whether B is still attending; how his last message was received (whether B understood him, thought it was funny, etc.). One of the important functions of eye-movement is channel control.

Experiments show that a listener looks at a speaker longer than a speaker at a listener. A person looks away when he/she plans a new utterance or when content of an utterance is hard to understand.

7) Non-verbal aspects of speech

In some situation it is more important how something is said than what is said. This is covered by non-verbal aspects of speech (timing of speech, emotional tone of speech, speech errors and accent).

Timing of speech considers length, frequency and total amount of utterances, silences, frequency of interruption, length of pauses before replying etc. Timing of speech is a function of personality but also it is influenced by timing of speech of other persons involved in the conversation.

The physical dimensions of emotional tone of speech are: loudness, pitch, frequency distribution, quality of voice etc. It is very difficult to make a dictionary of emotions and to map non-verbal cues into emotions because different people express the same emotion differently.

Speech errors can be divided in two groups:

ah-errors: 'ah', 'um', 'hmm'

non-ah errors: sentence change, sentence incompletion, repetition, omission (leaving out a word or leaving it unfinished), stutter.

Non-ah errors can be seen as disturbances of speech due to anxiety. On the other hand ah-errors increase when the talking task is more difficult and represent thinking time.

The person's accent can reflect his nationality and social class, his educational or occupational background. A person can pick up a new accent (teenagers use different accents at home and at school). Also, a person can change accent depending on a particular situation. The accent is part of a person's self-presentation.

2.2 Verbal behavior

The unit of verbal behavior is the utterance. Argyle in [1] pointed out several aspects of verbal behavior.

Utterances can be categorized in different ways. The utterances can be distinguished whether they are about events external to the participants in interaction, about participants or about interaction. Furthermore, utterances can be classified as questions, utterance which influence behavior of the others, conveying information (about facts or opinion), to establish social relationship etc. Bales [8] in his interaction analysis uses 12 types of utterances: shows solidarity, shows antagonism, shows tension release, shows tension, agrees, disagrees, gives suggestion, ask for suggestion, gives opinion, ask for opinion, gives orientation, ask for orientation.

The topic of conversation can affect interaction in different ways. A topic can be impersonal, remote, abstract or highly personal. For a participant in an interaction the topic can be cognitively easy or difficult, interesting or uninteresting, embarrassing etc.

The linguistic structure of utterances is an important aspect of verbal behavior. There are many studies how long and short utterances are used in referring objects. Unknown, unfamiliar, infrequently occurring objects needed longer phrases and utterances become shorter with repeated uses. Under some conditions utterances have more repetition of words by type/token ration or by the number of the new words introduced during the period of discussion. It can happen for example when speakers are under pressure to reach an agreement.

2.3 The organization of verbal and non-verbal elements of behavior

Verbal and non-verbal signals are highly connected in social interaction. Argyle in [1] explained that verbal communication depends on the non-verbal background in several ways.

- each participant must signal continuously his attentiveness and responsiveness to the others
- each participant in interaction must signal his attitudes and intention towards the others
- there must be a continuous regulation of speaking and listening
- speech may be illustrated using gestures

Verbal signals are usually used for describing objects, people, and events; to discuss facts, opinions and problems. Non-verbal signals are used for expressing emotions, attitudes etc. Each of them can be used to substitute the other to a limited extent, and under certain conditions. Non-verbal signals may to some limited extent be used instead of words. For instance if two persons don't speak the same language they may use sign language for the communication. The words can be used in place of

non-verbal elements in several ways. Emotional sounds may be replaced with emotional words. Self-presentation may be carried out by means of words.

3. Modeling a meeting as a sequence of interactions between participants

A meeting is a dynamic process that can be represented as a sequence of meeting actions like opening, monologue, discussion, presentation etc. In Section 5 we will say something about other aspects of a meeting. Each meeting action is determined by a participant's activity. The participants communicate multimodal because this is the natural way of human – human communication and behavior of one participant is constrained by the behavior of other participants. These are two aspects of interaction during the meeting that we have to take in consideration. Argyle ([1]) defined behavior as a function of both participant and environment.

Beh=f(P,E)

For example if person A is more nervous than person B and they have to wait half day in a queue (situation E1) than A will be more anxious than B. If the situation E2 is drinking morning coffee than A and B are more relaxed than in the situation E1. For a meeting participant the environment includes the other meeting participants.

In order to include all the aspects of a meeting, it is very important to define a *complete and not-redundant meeting model* that consists of basic states and activities. Combination of these basic entities enables creation of a number of meeting activities. Stephane Marchand-Maillet in [2] has proposed a meeting data model that can be used for a meeting browser tool. In this model a meeting consists of basic meeting activities as opening, discussion, vote, presentation, break, silence and closing. Each meeting activity consists of participant activities which can be divide in two groups. The first group consists of participant's interactivity i.e. the mode of interaction that a participant can have with other meeting participants. For example, monologue, talking, chatting, silent etc. The second group of activities shows participant involvement in the meeting as idle, presenting, voting etc. Beside these two groups of activities each participant is at any moment in a specific physical state as sitting, standing, writing, absent etc.

It is necessary to define a set of questions about meetings on which we want to get answers from our system. This approach is useful from both of the following two perspectives.

- Design of a system (data model)
- Evaluation of the completeness and validation of the data model.

Stephane Marchand-Maillet [2] defines three groups of actors for browsing a system.

1) Participant- a person who attended a meeting
2) Customer- a person who is aware of the topic (for example project member) but absent at a meeting, or a person unaware of the topic
3) Analyst- the entity in charge of meeting post processing

According to these three levels of system analysis the set of information, which we want to get from the system, can be divided in the following categories [2]:

Personal questions:

1. Where did I seat?
2. Whom did I talk (at that time /about what)?

Question about participants:

1. Who was there?
2. Who talked to whom (at that time)?
3. Who presented?
4. What was his/her role?

Question about topics:

1. What was discussed?
2. Was this topic discussed?
3. Was there any conclusion reached?
4. Was the topic changed during the meeting? etc.

Question about actions:

1. Were there any presentations/votes/monologues/breaks etc. (and when)?
2. Were there any decisions taken (and when)?

Post processing tasks:

1. Relate participants/topics across meetings
2. Asses mood during the meeting
3. Asses participant participation in the meeting
4. Classify meetings
5. Create meetings relationship (per date/ per topics/ per participants.) etc.

Our approach in annotating a meeting as a sequence of meeting actions using information obtained from different modalities like speech, gesture, gaze etc. is based on this proposed model and the questions listed above.

The main idea is to use a *semantic approach* on different levels of interpretations and annotations. In human-human interaction in a meeting we have a multidimensional (multilevel) modeling problem**.** First we have to integrate all modalities in order to recognize a participant activity (participant multimodal behavior). Here we take a definition of modality proposed by in [5]. Modality refers to the type of human communication channel used to convey or acquire information. In its broad sense, modality is the way an idea is expressed or perceived, or the manner an action is performed. Multimodal integration can be done on different levels: signal level, feature level and semantic-decision level. In our approach we will use semantic integration of different modalities in order to give a semantic meaning of multimodal behavior of participants. Second, we have to use a multimodal-stream approach for recognition of meeting activities as semantic combination of multimodal persons behavior (MB) recognized on lower level. On this level there is an information stream for every participant. It means that we have to define a *lexicon of meeting actions* and to determine features, which are the parts of MB representation, to be used for the recognition of these meeting actions.

Why do we need a semantic approach in recognition of meeting actions? The same idea is proposed in [3] but on the audio-video feature level. This approach applies a computational framework for automatic meeting analysis that involves three components: a set of multimodal group actions (meeting actions), a set of individual

actions and a model of the interactions. A defined set of eight meeting actions consists of monologue1, monologue2, monologue3, monologue4, presentation, white board and note-taking. An individual action may be either fully recognized or just measured. A fully recognized individual action may be used for browsing or indexing the system. Measurements of individual actions may be used as observable features for recognition of meeting actions. An observation sequence is defined as $O = (o_1, o_2, o_3, ..., o_T)$ where o_t is a set of multimodal (audio-visual) features at time *t* (e.g. location-based speech activity; the pitch and speaking rate of each participant; orientation of each participant's head etc.) This set of features can be divided into multiple streams according to participant *i* and according to modality *m*. They defined a feature vector

$o^{i,m} \in R^{Ni,m}$ where $N_{i,m}$ is the number of features for individual participant *i* and modality *m*.

There are also some participant-independent features. The features that correspond to a single participant can be defined as:

$o_t^{i,1:M} = (o_t^{i,1}, o_t^{i,2}, ..., o_t^{i,M})$ where M is the number of modalities

This approach proposed to model interactions between participants using Hidden Markov Models. The general idea is that for each meeting action $v_j \in V$ estimate parameters θ_j of a distribution over corresponding sequences of observations $P(O \mid \theta_j)$ where sequence of observation *O* would correspond to event v_j. In this framework an HMM is created for each meeting action v_j. Given a training set of observation sequences they created a new HMM for each sequence as a concatenation of sub-model HMMs corresponding to the sequence of meeting actions. The experimental configuration of their model consists of:

1. feature extraction according to an individual participant or modality (Audio-only, visual-only and individual participants features)
2. four HMM systems which represent different approaches in combining these streams.
   - Early integration: a single HMM trained on all features
   - Participant multi-stream: a multi-stream HMM combining the streams for individual participants
   - Audio-visual Multi-stream: multi-stream combining the audio-only and visual-only streams
   - Audio-visual Asynchronous: Asynchronous HMM combining the audio-only and visual-only streams. Asynchronous HMM is a new architecture for modeling joint probability of a pair of asynchronous sequences describing the same sequence of events [4].

The best results were achieved using Audio-visual Multi-stream approach. The action error rate in this case is 5,5 %. The action error rate is equivalent to the word error rate in automatic speech recognition.

The main difference between this approach and our proposed approach is in the information that we use as inputs. In this approach inputs are audio and video signals and in our approach inputs are results from recognition processes that use audio and video signals as inputs. This causes a difference in the level of features, which are used as observations in meeting actions recognition process. This model for automatic

meeting analysis on audio-visual feature level is successful for their defined set of meeting actions. However, there are a number of meeting actions that need more semantic details in recognition tasks (consensus, disagreement, opening, closing, lecturing). For instance, we need information about the topic of a discussion, the roles of participants and their relationship in order to distinguish lecturing from presentation, white board from monologue and global discussion from multi-discussion. If we integrate more semantic details in a meeting representation on different levels than we can contribute more purposeful analysis (post processing) of recorded meetings. This is the main goal of our semantic approach.

4. A lexicon of meeting actions

The first step in modeling a meeting as a sequence of meeting actions is to describe a lexicon of possible meeting actions. It should be the result of analysis of the recorded and scripted meetings (meeting corpus), the real meetings (participating in the meetings, analyzing meetings on a TV etc.). Because of natural and unpredictable behavior of people during the meeting it is very hard to give the precise definition of a meeting action**.** Therefore some level of abstraction in defining meeting events is required. Our main goal is to represent a meeting as sequence of meeting actions, not to determine precise time boundaries of each meeting action. In a definition of meeting action we can include cues as: number of speakers, speaker behavior, participants behavior, key words, dialog acts typical for this meeting action, topics (number or content) etc. It relays on the fact that each individual meeting action has something like a micro grammar**.** The same idea is proposed for the cue-based model for automatic dialog acts interpretation [6]. The main cues are specific lexical, collocation, and prosodic features. For example, *Please* or *would you* are cues for request (lexical and collocation features); loudness and stress can distinguish *yeah* as agreement from *yeah* as backchannel (prosodic features).

We propose a set of meeting actions that can be divided into three groups:

- single speaker dominate actions
- multi speaker actions
- non-verbal dominate actions

The first group is named as single speaker dominate because in these meeting actions a single speaker has a dominate role. We can have more than one participant speaking at the same time but the actions are determined by multimodal behavior of one single speaker. For example, during presentation one speaker is in front of the projector screen, points to the screen and at the same time some of the participants chat. Than we have more than one speaker but the presentation meeting action is determined by the behavior of the speaker in front of the projector screen. Additional information about chatting between two participants is useful for reasoning about their interest in the discussed topic, or their unmannerly behavior etc. Non-verbal dominate actions are actions which are determined by non verbal interactions between participants. They can start with verbal interaction (current speaker proposed something, or says something) and than non verbal interaction dominate. For example, the current speaker proposes that all participants write notes and they all perform it for some period of time. In a real situation during non-verbal behavior they can also talk but non-verbal elements dominate. Fig 1 shows the hierarchical organization of a proposed set of meeting actions.

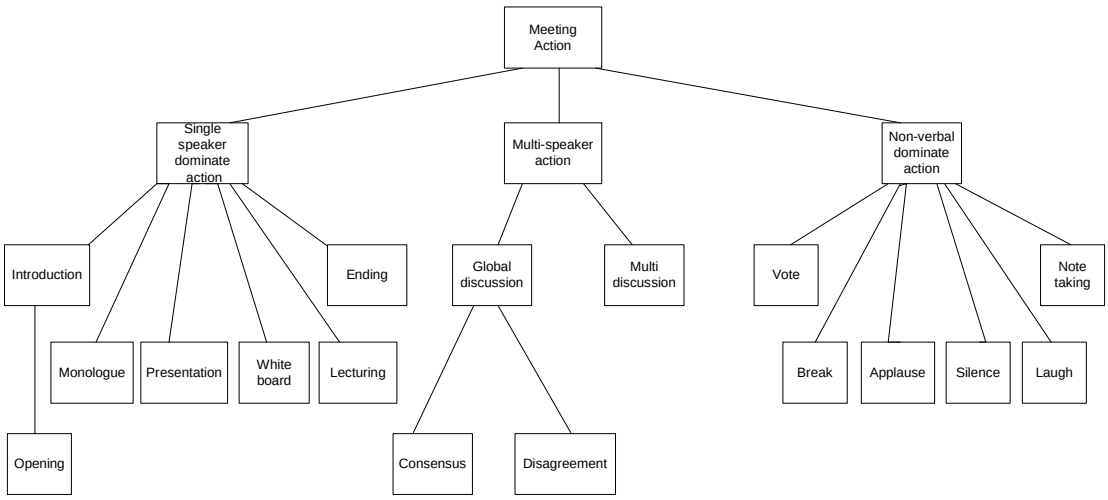

Fig 1. Hierarchical organization of meeting actions

4.1 Definition and main characteristics of meeting actions

It is impossible to completely reflect a real meeting situation in a computational framework. Therefore in definitions of meeting actions we will have some assumptions. When we talk about the time boundaries of a meeting action we only take in consideration the beginning of a meeting action. We will assume that the end point of one action is the beginning of the following action. In what follows we will make references to the DAMSL annotation scheme for dialogue acts [7].

4.1.2 Single speaker dominate meeting actions

Introduction

Definition: Introduction is a meeting action which happens at the beginning of a section of the meeting and summarizes activities which have to be performed or topics which have to be discussed during this section.

Characteristics:

A section of the meeting is a logical whole i.e. an unit of a meeting which is bounded with meeting breaks and meeting end**.** The opening is a special type of introduction which happens at the beginning of the meeting and it relates to a whole meeting. Opening is an introduction in the meeting.

Number of speakers: In our model we will assume that only one speaker introduces the meeting participants in the new section of the meeting. It is usually a chairman but it doesn't have to be the rule.

Time-boundaries:

Start: The introduction starts after a planned break. More about planned and unexpected break see below in the description of the break meeting action.

Topics: - summarization of the previous parts of the meeting (topics that were

discussed, conclusions that were reached)
- agenda of new section
- topics that will be discussed
- information about presence of new meeting participants or groups.

Speaker behavior:
State: sitting, standing or walking
Activity: monologue or presenting
Participants' behavior:
State: sitting, standing, writing, nodding, still
Activity: silent, back channel noise, chatting

Opening

Definition: Opening is a meeting action which happens at the beginning of the meeting and it is an introduction to the meeting.

Characteristics:

Number of speakers: In a real situation it is possible to have more than one speaker but not at the same time. It is usually the chairman but it is not the rule. It depends on the type of the meeting. We will assume that always one speaker opens a meeting.
Topics: introduction to the meeting, agenda, topics that will be discussed during the meeting, about the main participants or groups which are present at the meeting etc.
Time boundaries:
*Start***:** at the beginning of the meeting
Speaker behavior:
Activity: monologue or presenting
State: sitting or standing or walking
Participants' behavior:
State: sitting, standing, nodding, writing, still
Activity: silent, back channel noise, chatting

Ending

Definition*:* Ending is a meeting action which happens at the end of the meeting and it consists of a summarization and the closing of the current meeting.

Characteristic:

Number of speakers: As well as, in the opening meeting action the number of speakers can be more than one but we will assume that one speaker closes the meeting. It is usually the chairman but it is not the general rule.
In a real situation the ending meeting action may be interrupted with some other

meeting action. For instance, if the current speaker proposes the date of the next meeting, it can cause a new discussion which may be finished with agreement (consensus) or disagreement. Considering our assumption that we have only one speaker during the ending meeting action, this event will be recognized as a discussion with consensus or disagreement as possible followers.

Time boundaries:

Start: - when the current speaker starts to talk about topics that are typical for the ending meeting action. Topic is a semantic category that is very important for determining this meeting action. The important cue is that it is the end of the meeting (the last action but can be followed by an applause). The key words typical for closing the meeting may be used for the detection.

Topics: - summarization of the meeting (main topics which were discussed)
- conclusions reached during the meeting (if there was any conclusion)
- plans for the next meetings
- agenda for the next meeting (date/ place/ topics)
- thankfulness to participants for attending the meeting etc.

Speaker behavior: Speaker behavior during the ending meeting action is the same as during the opening meeting action.

State: sitting, standing or walking

Activity: monologue or presenting

Participants' behavior:

State: sitting, standing, nodding, writing, still

Activity: silent, back channel noise, chatting

Monologue

Definition: Monologue is a meeting action where one speaker speaks continuously without interruption.

Characteristics:

Number of speakers: one speaker

Time boundaries:

Start: changing of the speaker is a possible start of the monologue meeting action.

Duration constraint: Speaker speaks without interruption for some period of time. Therefore, a threshold for duration of speech activity should be defined. This threshold can be predefined using recorded meetings in meeting data corpus.

Topics: for recognition of the monologue meeting action a content of topic is not important. It is usually a general (impersonal, abstract, remote) or personal topic which may consist of more subtopics.

Speaker behavior:

State: sitting, standing, walking

Participants' behavior: We will assume that they can do almost everything except talking

Activity: silent, idle

State: sitting, standing, walking, nodding, writing

If during someone's monologue most or all of the other participants take notes than the lexicon can be extended with a new meeting action: monologue with note taking. In order to simplify the problem we will keep the proposed set of meeting actions from the beginning of Section 4.

Presentation

Definition: Presentation is a meeting action where one participant makes a presentation using the projector screen.

Characteristics:

Number of speakers: one speaker. In a real situation it is possible that during presentation other participants talk between themselves but we will assume that we have one speaker with dominate role.

Time boundaries:

Start: There are more possibilities for the recognition of the beginning of presentation

- when speaker stands in front of projector screen and points to projector screen
- when speaker just points to projector (without changing the seat place)
- when speaker stands up (it is not obligate) and says that he/she will present something.

"*I am gonna do a short presentation*"

In this case the speaker commits to do something. For this utterance dialog act using the DAMSL [7] annotation scheme is COMMIT.

The problem is how to recognize time boundaries when after the presentation the same person continues with a monologue without changing the seat. The possible solutions are: topic change, body orientation and gaze detection (he/she doesn't look at the projector screen any more) etc.

Topics*:* Using background knowledge (content of the slides) or/and knowing where the speaker points at some specific moment, may be useful to recognize the topic and subtopics of the presentation. The problem is how to the detect from the video information where a participant exactly points at.

Speaker behavior:

State: sitting, standing, walking, pointing at screen

Activity: talking

Participants' behavior:

During the presentation other participants can do anything except things which are cues for recognition of presentation (standing in front of the projector screen and pointing, etc.). We will allow that during presentation other participants can talk. In this case we have two aspects. The first is when the other participants chat between themselves. In this case the meeting action will be recognized as presentation. The second is when the other participants interrupt the presenter with questions, remarks, comments. In order to develop an exhaustive model this real situation may be represented as discussion or presentation or a new meeting action *presentation with discussion*. More realistic is to represent it as a discussion meeting action, which will be followed with a continuation of the same presentation. The same remarks that we mentioned for taking notes during the monologue are valid for presentation as a speaker dominant meeting action.

White board

Definition: White board is a meeting action where one participant in front of the white-board talks and makes notes on the white-board.

Characteristics:

Number of speakers: We will assume that only one speaker can be in front of the white-board and writes notes on the white-board. In real situations it is possible that more than one participant is in front of the white-board talking at the same time and writing notes. One example of such situation is when two persons argue about something that is written on the white-board and one wants to show that what the other person wrote is invalid, to propose his/her own idea etc.

Time boundaries:

Start: - when the current speaker is in front of the white-board is a sign that this action could be recognized as white-board but it doesn't mean that he/she will write something on the board. Speaker may walk during monologue or discussion and stand for some time in front of the white board. During the discussion or monologue the speaker looks at the other participants (gaze direction). Most of the time, during the white-board meeting action, the focus of speaker attention is the white-board. The most important cue is writing on the white board.

- when the speaker explicitly said that he will write something on the white-board, or something like "let's summarize " or some other key words which are characteristic for the beginning of the white-board meeting action.

Topics: - summarization of current topics or previous discussed topics
- what will be the next topics
- explanation etc.

Speaker behavior:

State: standing or walking in front of the white-board, writing
Activity: talking

Participants' behavior:

Like for the presentation, we will assume that the other participants may be in any state: sitting, standing or walking. They can chat between themselves or perform backchannel noises as laugh or mumbling or may be silent. They can be idle or may take notes for themselves. The same remarks which we mentioned for taking notes during the monologue and presentation are valid for white-board meeting action. Also the remarks we made on discussion during a presentation are valid for the white-board meeting action.

Lecturing

Definition: Lecturing is a meeting action where one participant, who has a dominate role in a relation to the other participants, lectures the other participants about a particular subject.

Characteristic:

Lecturing can be presented as the following meeting action (in the way we defined them above) or a combination of them

- monologue
- presentation
- white-board

All characteristic of these meeting actions are also valid in the case of lecturing.

Participants' roles and relations:

Lecture is a meeting action typical for a meeting which main goal is acquiring new knowledge (lessons in a school, at the language courses, tutorials). Additional information about relations between participants and their roles in the meeting are needed for recognition of this meeting action. For instance, teacher and students, lecturer and hearers, instructor and candidates etc.

Speech style:

The form of speaker's talk as well as dialog acts of speaker utterances can be of great importance for distinguish the lecture meeting action from presentation, monologue and white-board. During, for example, a monologue a speaker talks in the first singular or plural form (I- form, We-form) speaking about a personal topic (my research area, my favorite move etc) or a group topic (programming practice at the University of Twente, proposal for new projects etc.). The speaker gives his/her opinion about the discussed topic, proposals, offers to do something, plans group activity etc. In the case of lecturing, a speaker most of the time talks in third singular or plural for (He/She/It –form or They-form).

Dialog acts:

The ASSERT dialog act is dominant in this meeting action. During lecturing a lecturer may give some exercises to the other participants .This is marked as a ACTION-DIRECTIVE or OPEN-OPTION dialog act.

4.1.2 Multi-speaker actions

Global Discussion

Definition: Global discussion is a meeting action that involves more than one speaker in discussion about the same topics.

Characteristics :

Number of speakers: more than one speaker at different or at the same time (overlapping).

Time boundaries:

Start: - speaker changing is a potential start of a discussion. Turn-taking is a very important element in determining action boundaries. Discussion is characterized with frequent turn taking, speech overlapping, interruptions etc.

- in a real situation discussion usually follows some other meeting actions. It may start with an INFORMATION_REQUEST dialog act. For instance, after speaker dominant meeting action (monologue, presentation, white–board etc.) discussion can start with

So, do you have any questions about …? [YES/NO question]

- often, the current speaker is interrupted and the discussion is continued on the same topic
- sometimes key words can mark the beginning of the discussion. *Let's talk about*; *We will discuss about* etc.

Topics: - discussion is a meeting action that usually follows some other more specific meeting action. Therefore, topic of discussion may be topic from the previous meeting actions.

- planning some future action
- making conclusions etc.

It is characteristic for global discussion that all speakers discuss the same topics.

Participants behavior:

- during the discussion all participants are potential speakers.
- during a speaker turn the others can be in any state: sitting, standing , walking, nodding, writing
- perform some other activity
  - silent, back channels (laughing, mumbling)
  - talking at the same time

Multi-discussion

Definition: Multi-discussion is a meeting action which involves more than one speaker (usually several groups of speakers) discussing different topics. It consists of more global discussions.

Characteristics:

Number of speakers: more than one speaker in the specific time range. Speakers are divided in different groups according to the discussed topic. If there are only two speakers in the specific time range topics are different and we have two different monologues.

Time boundaries:

Start: The main difference between global discussion and multi discussion is in a grouping participants according to the topics. In a global discussion it is possible to have more than one topic but participants discuss these topics. The same remarks for the detection of the beginning of global discussion are valid in multi discussion case but additional request is detection of more than one topic in the same time frame.

Topics:

Multi-discussion consists of group discussions and because of that all observations about global discussion topics are valid in this case. Discussed topic may be connected in some way. Participants who attend the same meeting are together at the same place and the same time to solve some problems, to plan group actions, to learn something new etc. Therefore, it is natural that in multi-discussion the discussed topics are related. For example, if the goal of the meeting is to make project proposals, than the participants may be divided in different groups to discuss the main objectives of the different parts of the projects.

Participants' behavior:

Behavior of the participants in a multi-discussion is the same as in the global discussion case. Participants during the meeting can be grouped according to their interests, their relationship, their roles etc. Therefore, knowledge about participants is a very important source of information.

Consensus

Definition: Consensus is a meeting action where all meeting participants express consensus.

Characteristics:

Number of speakers: All participants express consensus in a short time frame. Agreement can be expressed in a verbal or non-verbal way. Mostly the combination of verbal and non-verbal elements is used. Theoretically speaking, it is possible that all participants express consensus with non-verbal element (nodding), |speakers|>= 0, but it is very rare. Interpretations of head movements are culturally dependent. In some cultures nodding means disagreement. Because of that we cannot relay only on non-verbal elements.

In the definition of consensus we said that all participants have to express consensus. The question is does the speaker, who proposed something, has also to express agreement? We can take into account the addressees of the proposal. If addressees are marked as '*you*' than the current speaker doesn't have to express agreement. But, if addressees are marked as '*we*' then the current speaker also has to express agreement in order to recognize the current meeting action as consensus.

Time boundaries:

Start: - when the current speaker proposes something, or asks a question to check whether the other participants agree with his idea, proposal, and opinion. It is usually tagged with OPEN-OPTION, INFORMATION-REQUEST or CHECK dialog act.

- sometimes consensus doesn't follow an explicit proposal. It can be a natural continuation of a previous meeting action and it is related to the topic that was discussed

Topics: There is a wide range of verbal elements to express consensus. These utterances are usually marked as ACCEPT dialog act.

Key words: yeah, okay, agree, good idea, perfect, of course, good, me too, yes, sure, it sounds good, it's good., that makes sense, I think so, definitely, I think that's alright, alright, excellent etc.

At the beginning of this section we explained semantic aspects of the word "yeah". Stress or loudness may help distinguish "yeah" as agreement of "yeah" as backchannel . Also, if " yeah" follows a proposal than it is probably an agreement but if it follows information it is a backchannel [6].

Disagreement

Definition: Disagreement is a meeting action where all participants express disagreement.

Characteristics:

Disagreement is a meeting action that is in a structural way the same as consensus but with opposite meaning.
Verbal elements are utterance usually marked as REJECT.
Non-verbal element for expressing disagreement is shaking head that is opposite to nodding.
Key words: no, disagree, not, I don't think so, I don't agree, not necessary, not important, not at all, etc.

4.1.3 Non-verbal dominate meeting actions

Note taking

Definition: Note taking is a meeting action where all participants write notes.

Characteristics:

In a real meeting, a situation where all participant write notes at the same time is very rare. Writing notes is an action on individual (participant) level and can be part of the other meeting actions. Like we have already mentioned, if other participants during someone's presentation, monologue, white-board etc. write notes, this meeting action may be recognized as a new meeting action: presentation with note taking, monologue with note taking, white-board with note taking.

The question is how will we represent a meeting action where some of the participants write notes (not all) and no other meeting action is recognized? In order to simplify our model we will assume that during the note-taking meeting action all participants take notes.

Number of speakers: Note-taking is a non-verbal dominate meeting action. It can start with the current speaker proposal or directive for taking notes. Besides that, we will assume that participants may speak, laugh or mumble during the note-taking meeting action.

Time boundaries:

Start: - note taking may start with the current speaker's command for taking notes. For example:

All of you note down. Write down a few notes.

All these utterances are marked as ACTION-DIRECTIVE dialog acts

- when all participants write notes and no other meeting action is recognized.

If the current speaker, who ordered the other participants to write notes, doesn't write notes we will still recognize this meeting action as note taking (teacher-students).

Participants' behavior:

State: sitting and writing

Activity: silent, backchannel (laughing, mumbling), talking

Vote

Definition: Vote is a meeting action where some of the participants or all participants express their agreement with a proposal by raising their hand or by nodding.

Characteristics:

Number of speakers*:* The current speaker proposes to vote. In a real situation the

other participants may express their agreement or disagreement with proposal by verbal elements. In our model we will suppose that voting is performed with non-verbal elements (rising hand or/and nodding).

Usually the current speaker regulates the whole process of voting. It may be a chairman but it doesn't have to be the rule. Voting is usually divided in a few steps (proposal, raising hand or nodding, counting and making conclusion). If the number of participants who voted for a proposal is greater than a defined threshold than the proposal is accepted, otherwise it is rejected.

Time boundaries:

Start: -when a current speaker expresses a proposal for voting. This is marked as INFORMATION-REQUEST,OPEN-OPTION or CHECK or dialog act. The proposal can be expressed using voting key words :
"Who is for…?", "Who is against…?", "Who is reserved… ?"

Topics: Topic is a content of the proposal. A proposal is usually connected with performing some actions (individual or group actions) and it may be connected with the previous discussed topics.

Participants' behavior:

State: sitting, standing, writing, nodding, still

Activity: rising hand (voting), backchannel (laughing, mumbling), silent

Silence

Definition: Silence is a non-verbal meeting action where all participants are silent for some period of time and they don't perform any other non-verbal activity which may influence recognition of some other non-verbal meeting action (note-taking, break etc).

Characteristics:

Duration constraint:

The question is how to determine the time period for a silence meeting action. If the time threshold is too small, there is a great chance that a silence meeting action is recognized almost after each meeting action. In other words, the sequence of recognized meeting actions could consist of numbers of silence meeting activity which is not our goal. It means that we have to determine such time threshold that only significant silence meeting actions are recognized.

Number of speakers: zero

Time boundaries:

Start: - if no one performs talking or laughing activity and no other non-verbal meeting activity is recognized.

Participants' behavior:

State: sitting, standing, walking, still, writing

Activity: silent, mumbling

Laugh

Definition: Laugh is a meeting action where all participants or most of the participants laugh at the same time

Characteristics:

.

Laugh at individual or group level may occur during some other meeting action. In that case it may be recognized as a new meeting action (discussion with laughter, or presentation with laughter etc) . In order to simplify the problem we will keep the proposed set of meeting actions from the beginning of Section 4.

How to define "most of participants"? We can assume that only if all participants laugh it will be recognized as a laugh meeting action or we can determine from experimental data a threshold depends on the number of participants.

We determine laugh at a participant level and take it into account during the recognition of other meeting activities. If no other action has been recognized and participants laugh it should be recognized as a laugh meeting action. Also, if all (or more than some threshold) participants laugh than it has importance for the meeting (on meeting action level) more than on participant level. Laugh can mean that someone told joke. Laugh may be useful for assessing the mood, atmosphere during the meeting etc.

Number of speakers*:* zero
Time boundaries*:*
Start: - when all or the most of participants laugh and no other non-verbal meeting action is recognized.
Participants' behavior:
State: sitting, standing, walking, still, writing
Action: backchannel (laughing), silent (some of participants)

Applause

Definition: Applause is a non-verbal meeting action where all participants or some of the participants applaud at the same time.

Characteristics:

It is usual that if someone starts to applaud than the other participants also start to applaud. If there are two or more groups that are rivals or with opposite opinions about a discussed topic than applaud may be relate to a particular group. Threshold that determine "some of participants" can be defined as the minimum of numbers of members of one group. In order to determine membership to one group we need knowledge about relations between meeting participants.

Applause is an action that follows speaker dominant meeting actions as opening, ending, presentation, monologue etc. Applause can interrupt any of these meeting actions if a speaker says something important or if a speaker introduces important participants (invited speakers, sponsors etc.).

Number of speakers: Applause is a non-verbal dominate meeting action. In our model we will allow that a participant applauses and talks at the same time.

Time boundaries:

Start: -when some or all of the participants start to applaud.

Participants' behavior:

During applause meeting action we will allow that a participant can be in any state: sitting, standing, walking, applaud, writing (if he/she doesn't applaud), still, and perform any activity: laughing, mumbling, talking, silent.

Break

Definition: Break is a short interruption of the meeting**.**

Characteristics:

An interruption can be a result of different occasions. Interruption can be *planned* (according to an agenda, as a result of agreement about a break during the meeting, etc.) or *unexpected* (someone suddenly enters the room and with the questions interrupts current meeting action, a sudden phone call may interrupt a current meeting action).

Planned break is a meeting action where all participants are out of the meeting room. It is also possible that all participants or some of them are in the meeting room but in some informal conversation (chatting) or idle. Usually they are not at their seats. They can walk, change seats etc. In order to simplify our model, we will assume that a planned break is a meeting action where all participants are out of the room.

An unexpected break is characterized by some external events. For example, a person enters a room and takes the turn from the current speaker. Taking turn and topic change are very important cues for detecting break. An external event is not enough for causing a break. For instance, receiving a SMS is an external event, which can momentarily change a participant's focus of attention, but very soon the previous action will be continued.

Number of speakers: In our proposal for a planned break number of speakers is zero. In the case of an unexpected meeting break the number of speakers can be more than one.

Time boundaries:

Start: - planned break can start with the current speaker's proposal for a break. This proposal is marked with OPEN-OPTION dialog tag. Key words as "break", "pause" etc. also can determine the beginning of the break

- when participants start to leave the meeting room
- an unexpected break starts with an external event (a person enters a room and takes the turn, phone call etc.)

Participants' behavior:

- During a planned break participants are not present in the meeting room. The state for all participants is absent.
- During an unexpected break a participant can be in any state and perform any action. The behavior of the participants depends on the meeting action that has been interrupted.

4.2 Priority of meeting actions

Some of the meeting actions from our proposed set of possible meeting actions are highly specific and easy to recognize (planned break, applause, vote etc.). Some of the meeting actions are hard to distinguish from similar actions (lecture from presentation, opening from monologue, etc.). We saw that turn taking is a potential beginning of a number of meeting actions. Than we have to check if there are more specific and obvious details that can help us to recognize an appropriate meeting action. Because of that it is useful to define priorities of meeting actions. The meaning of priority in this case is a priority in the recognition process and the level of specificity of a meeting action. Priority is marked as a natural number in scale form 1-12. A meeting action with a less number has higher priority.

Break	1
Applause	1
Vote	1
Note-taking	2
Lecturing	3
White-board	4
Open	5
Ending	5
Presentation	6
Introduction	7
Monologue	8
Consensus	9
Disagreement	9
Multi-discussion	10
Discussion	11
Laugh	12
Silence	12

5. Other aspects of meetings

A meeting is more than a sequence of meeting actions. There are other aspects of a meeting which are interesting for modeling and browsing. A meeting is determined by the behavior of the meeting participants. For example, personal characteristics of meeting participants (age, gender, profession, speech style) and their relationships can influence dynamism and formality of the meeting. Knowledge about users may be useful on the individual and group level of meeting modeling.

The user profile can be explicitly specified during the registration process or be learned during the processing of the recorded meetings. Before the beginning of the meeting all participants have to provide some personal data as name, age, gender, native or non-native English speaker, profession, membership to a specific group (University of Twente, MIT laboratory). Information about a user that may be learned is speech style (dynamic, slow etc.), presentation style, which modalities are usually

used in the specific situation, his/her role during the meeting etc. Speech style can be learned from audio and video data processing. If a speaker during the his/her speech uses his hands to illustrate what was said, to explain some situation, changing posture and focus of attention in a very short time interval than it can be characterized as a dynamic speech style. Using processing of a recorded meeting we can learn more about the participants behavior, their characteristics, etc. In other words, besides tracking the speaker in order to determine and learn his/her speech style we can track the behavior of participants during the whole meeting. For example, what they are doing during the speech of an other participant, whether the discussed topic is interesting or not for the participants etc.

At each level of abstraction background knowledge plays an important role. Background knowledge may include previous knowledge about the meeting: agenda, written notes, presentation slides, content of white-board etc. This can be useful for detecting topics and sub-topics that were discussed at the meeting. Tracking meeting participants starts from their entering in the meeting room. The number of participants, their identification and seat position are also part of background knowledge.

Our goal is to develop a system which will provide a deep analysis of the recorded meeting. The question: *"What John said to Peter about the programming standard?* " contains three very important aspects of the meeting.

- source of the messages (John)
- discussed topic (programming standard)
- target of the messages (Peter)

Target detection needs a multimodal approach. Modalities that are used in target detection are: speech, gaze and gesture. Sometimes the speech transcript contains enough information to detect who is the addressee of the message. For instance, during the monologue a meeting participant may speak about the group actions and the addressee is detected from the utterances: *we, all of you, everybody, all of us, you guys* etc. The most important point where target detection plays a crucial role is at the end of the turn of the current speaker when he/she asking the explicit question selects the next speaker or when the current speaker expresses weak suggestion or command which has to be executed etc.

The turn of a speaker can be finished with the question: "What do *you* think? ". The addressee of the question is determined with the pronoun *you*. In this case the addressee can be one participant, or group of participants or all meeting participants. In order to resolve this ambiguity we need additional information obtained from different sources. A speaker's focus of attention (gaze direction) is very important in target resolution. Pointing to the person can resolve target ambiguity. The name detection is also a very powerful method. In the example: "What do *you* say, *John*?" detection of the name John and previous knowledge about participants are together used to determine the addressee of the question.

Like we said the target of the message doesn't have to be a particular person. It can be the group of the participants or all participants. It is conversation context dependent and also depends on the current meeting action. Usually, after monologue or presentation a question is addressed to all participants. Also if a question is addressed to everyone than usually the persons in front of the speaker are focus of speaker attention.

In human-human communication a speaker is usually the focus of attention of listeners. If a listener for some longer period doesn't look at the speaker it can be a

sign that he/she is absent, not interested in the topic, sleeping, or thinks about something else etc. The speaker's focus of attention is more uncertain. The speaker's focus of attention depends of the meeting action. If it is a presentation meeting action the speaker's focus of attention will be mainly projector screen. Sometimes a speaker may pay attention to one person for a longer period. It could mean that this person is in front of speaker (in his visible area) or the person is interesting for him/her for some reasons or the person asked the question etc.

6. Semantic Model

6.1 Motivation and explanation of the model

Modeling human-human interaction is more difficult than modeling multimodal human-computer interaction. Usually in human-computer interaction we have an action-reaction model. For instance, a user asks the computer for some information (action) and if the computer recognized the request it provides the user with the requested information (reaction). The communication is domain dependent. In human-human communication we have an interaction model. It means that A can't predict what will B say or do next. It is impossible to define the strict rules of human-human interaction. There are some possible expectations of the subsequent interaction. For instance, a question lead to an answer, an order may lead to an action, open-ended questions lead to longer answers etc.

Like we mentioned before in human-human communication in the meeting we have a multidimensional problem. First we have for each meeting participant to integrate all information obtained from different modalities in order to recognize a participant's multimodal behavior. In the following step we have to combine the recognized multimodal behavior of each participant in order to recognize an appropriate meeting action. On this level the participants are streams.

The goal of a multimodal system is building a coherent meaning representation of exchanged information. Roamry and Kumar in [9] defined the requirements of a common framework for meaning representation. Some of the requirements are:

1) structure information in a granular, typological and abstract form which can be used in information extraction application
2) provide a modular multi-modal system design
3) enable multi-level annotation in multimodal corpora

Our idea is to develop a modular multimodal system which will use a semantic approach on the participant level and meeting action level. Before we give a technical specification of the model we will explain the main idea of the model.

Inputs in the model are results of recognition processes of different modalities (speech, actions, gaze etc.). Using additional knowledge (user model, background knowledge etc.) we will first try to interpret modality information i.e. to extract meaning of information and to represent this meaning in an appropriate way. For instance, if for participant *A* we get information that at the specific time *t* he looks at the right and we know that at his right side is person *B* we will represent this information as A looks at *B* at time *t*. On this level is very important to filter information that is sufficient and useful for the next level i.e. for the other modules in the system which have this information as input. Combining this information for each participant we get as a result a representation of the multimodal behavior of a

participant. The modalities fusion technique depends on modality meaning representation. More about fusion techniques will be said in the following section.

The fusion of modalities is just the first step in the recognition of the multimodal behavior of each participant. Some of the required information may still be missing or may still be ambiguous. To resolve these problems we need an additional processing step and additional information from background knowledge, from user profiles etc. These are modules on the participant level. The next level is the combination of multimodal behavior of participants in order to recognize an appropriate meeting action. We will use well-known technique for modeling temporal processes – Hidden Markov Models. The main difference between our usage of Hidden Markov Models for recognition of a sequence of meeting actions and the approach proposed in [3] and explained in Section 2 is at the feature level. In our approach features are semantic features extracted from multimodal behavior of each participant. Video and audio extracted features are used in the framework proposed in [3]. Using semantic features it is possible to recognize high level meeting actions and to encode more semantic details in a meeting annotation which will enable us to find answers on questions easier and faster.

The similar idea for a semantic based multimodal interpretation framework is proposed in [10]. MIND (Multimodal Interpretation for Natural Dialog) is a semantic based framework for identifying meanings of user multimodal input. It is comparable with our identifying meaning of multimodal behavior of meeting participants. MIND is embodied in an intelligent infrastructure RIA (Responsive Information Architecture) for aiding user in their information seeking process. The communication is domain dependent. MIND uses three processes: unimodal understanding, multimodal understanding and a discourse understanding process. In MIND representation of users input and representation of context (domain context, conversational context, visual context) are crucial. In the information seeking process understanding of each other's information needs is important. Information needs are characterized by motivation for seeking information and information itself. MIND uses an *intention model* to capture the first aspect (the purpose of a message) and an *attention model* to capture the second aspect (objects and relations that are salient in a message). Also, MIND uses a *constraint model* to capture the different types of constraints that are important for an information seeking process (reference constraint and data constraint). MIND represents these models using a combination of typed feature structures. MIND represents intention, attention and constraints from user inputs as result of both unimodal understanding and multimodal understanding. During the unimodal understanding MIND applies speech and gesture recognizers. The result is represented as a *modality unit* which has intention, attention and constraint features. During the multimodal understanding MIND combines the semantic meanings of modality units (unification based modality fusion) and uses context information to derive an overall understanding of multimodal user inputs (context-based inference). Such understanding is captured in a representation called *conversation unit*. A conversation unit also has the same representation i.e. the same type of intention, attention and constraint feature structure. MIND uses conversation history to represent conversation context which is based of goals and sub-goals of user inputs and RIA outputs. MIND uses a representation called *conversation segment* to group together inputs that contribute to a same goal or sub-goal. A conversation segment also has the attention and intention feature structure. MIND semantic based interpretation has three main characteristics:

- fine-grained semantic model

- flexible composition of feature structures which enables complex user inputs
- consistent representation on different level that facilitates context-based interpretation.

We will use the similar idea in our process of recognition of multimodal behavior of participants. We will also have the processes of unimodal and multimodal understanding. For capturing salient information on each level we will use a combination of typed feature structures. The idea is that for each modality unit we use the similar interpretation which will enable a flexible composition of features structures.

The difference between a meeting domain where we have human–human interaction and an information seeking domain where we have a human-computer action-reaction model causes the difference in a definition of features and their possible values and the difference in fusion methods and in using the results of each step in the processing.

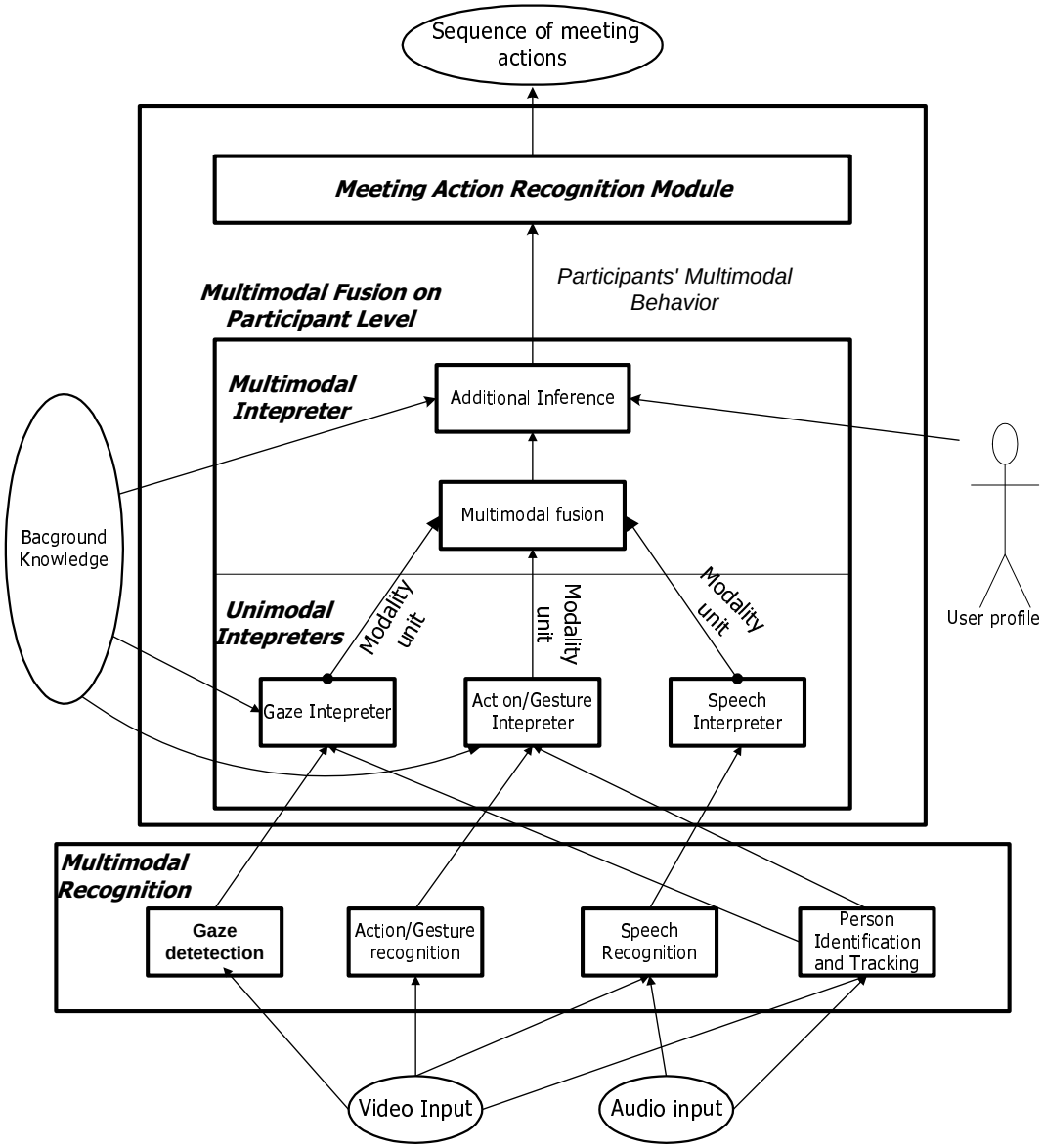

Figure 2. Architecture of the system for recognition of meeting actions

6.2 Technical specification of the semantic based model

The inputs in our semantic based model are the results of recognition modules that use audio and video records of the meetings. Several partners in M4 projects currently work on recognition tasks. The results of the recognition processes should be annotations of recognized information. Because of lack of knowledge about of the exact format of inputs and value set of specific type of information (for instance, set of recognized actions) we will assume that we have some needed information provided by the recognition process. Figure 2 describes the design of our semantic based model.

6.2.1 Specification of inputs of the model

Speech Recognition

Speech transcription is for now obtained using Transcriber [11] i.e. Channeltrans-multi-channel version of Transcriber. Transcriber is a tool for assisting the manual annotation of speech signals. Work on Automatic Speech Recognition (ASR) using SWITCHBOARD recognizer is still in progress. A result of speech transcription using Channeltrans is an XML file, which contains transcription of multiple speakers. The file doesn't contain dialog acts tagging and topic detection. However, we will assume that in our model we have this information.

Action/Gesture Recognition

Action recognition algorithms deliver a stream of actions for each person. Each action can be labeled with a confidence measure and a time stamp that indicate the beginning and the end of it. Until now a set of ten person actions is tested and for their test the body gesture recognizer is ported in the M4 domain. The defined test set consists of :

-entering observed area
-leaving observed area
-taking seat
-rising
-shaking head
-nodding
-voting with left hand
-voting with right hand
-pointing at the white board with the left hand
-pointing at the white board with the right hand

We will use as input in the model a little bit modified set of actions/gestures: entering, leaving, taking seat, rising, shaking head, nodding, rising left/right hand, pointing, applause (clapping hands) and writing

Emotion recognition

The multimodal (audio-video) approach should be used in emotion detection tasks. Our assumption is that we have recognized a set of basic emotions (happy, angry, fear, disgust, surprise and neutral). More about detection of emotion in human-

human dialogue and integration of emotional cues into a dialogue framework can be found in [12] and [13].

Gaze detection

No one of our partners works on detection of gaze direction of participants in the meeting. Gaze direction is a very important cue in the detection of focus of participant's attention. In a meeting application information who is looking at whom can be useful for identifying the addressee of a speech and for monitoring activities in meetings.

Stiefelhagen and Zhu in [14] showed in their experiments that head orientation is very helpful in determination of focus of attention. They found out that head orientation contributes to the overall gaze direction in the meeting 68,9% and that we can predict a user visual focus of attention based on head orientation 88,7%. In our model we will assume that at any time we have information about gaze direction of each participant.

Multimodal person and speaker identification and tracking

For person identification and tracking our partners use video processing and for speaker detection and tracking they use a multimodal (audio-video) approach. We will suppose that at any time we have obtained information about the location of each participant.

Other available inputs

-projector slides
-agenda
-white board
-participants written text

6.2.2 Specification of outputs of the model

The output of the model is an annotated sequence of meeting actions. Annotation of each action has to include more semantic details that are characteristic for the specific meeting action as well as for all meeting actions. For instance, topics and subtopics, taking a note during presentation etc. These semantic details are very useful for information retrieval, information extraction and other processes.

In order to find the best output representation we will start with the questions on which we want to find the answers. According to these questions and using the available data we will define a meeting annotation scheme.

6.2.3 Modules

Our idea is to develop a modular multimodal system. Figure 2 shows that our system consists of the two main modules *Multimodal Fusion on Participants Level* and *Meeting Action Recognition Module*. We will give a detail explanation of these modules in the following sections

6.2.3.1 Multimodal fusion on participant level

This module consists of two sub-modules:

1. Unimodal Interpreters for each participant
2. Multimodal Interpreter for each participant

Unimodal Interpreters

-Speech (Language) interpreter
-Action/Gesture interpreter
-Gaze interpreter

Inputs: inputs in the semantic model described above
Output: semantic meaning representation of specific modality – *modality unit*

The inputs in the semantic model obtained from the recognizers contain data for each participant. For instance, a speech transcript contains data for each meeting participant. The unimodal interpreter has to extract data for each participant and to represent the meaning of each input. For modality unit representation we will use typed feature structure. The question is: *What is the meaning of a non-verbal communication signal*? Isabbela Poggi in [15] shows that any communication signal of any modality is by definition meaningful and because of that its meaning can be paraphrased in a verbal language. Raising the eyebrows can be paraphrased as "*I am surprised*" or looking aside can be paraphrased: "*I am not interesting in the discussed topic*". We saw that the non-verbal signals are culturally dependent and therefore their meanings have to be considered according to culture. The possibility of translation of any communication signal into words and sentences give us a possibility to represent the meaning of different modalities in a similar way using similar representations. It enables us to combine meanings of different modalities on a flexible way in order to represent their grouped meaning. J.C.Martin in [18] proposed six types of cooperation between modalities:

- Complementarity - different chunks of information belonging to the same command are transmitted over different modalities.
- Redundancy – the same chunk of information is transmitted over more than one modalities (e.g. someone says "Give me the red glass in front of you, second on the left " and point at the same time to this glass)
- Equivalence- a chunk of information may be transmitted using more than one modalities (e.g. options for selecting data using mouse or voice command)
- Specialization- a specific chunk of information is always transmitted using the same modality
- Concurrency- independent chunks of information are transmitted using different modalities at the same time (e.g. at the same time a person types a document and speaks on a mobile phone).
- Transfer- a chunk of information produced by a modality is used by another modality (e.g. in an information retrieval application a request may be expressed in one modality (speech) and retrieved information in another modality (video).

Multimodal Interpreter

Inputs: Modality units for a participant
Output: Semantic representation of the multimodal behavior of a participant (Multimodal unit)

Multimodal interpreter consists of two parts

1. Fusion of modality units on the semantic level
2. Use of background knowledge, user model and other available data in order to improve integration, to obtain missing data, to detect the target of a message, etc. In MIND this is called context-based inference [10].

Unimodal Interpreters

Action/Gesture Interpreter

Interpretation of actions is based on the model proposed in [2]. During the meeting a meeting participant may be in any state and may perform any activity. A state is a physical state (sitting, standing, walking, absent) or an individual action - acting state (a participant involvement in the meeting). An activity is a participant interactivity during the meeting.

The question is how to define a criterion for interpreting an action as individual action or participant interactivity? We propose two interpretations based on the following criterions:

1) A participant interactivity is the participant verbal interactivity during the meeting. It is the participant involvement in a conversation during the meeting. It is based on audio sources. A participant individual action (acting state) is based on video sources.
2) A participant interactivity is the verbal or non-verbal mode of interaction that the participant may have with other participants. A participant individual action is action that the participant performs for himself (writing, nodding etc.). These individual actions are not significant for recognition of meeting actions. They are useful for assessing behavior of the participants during the meeting, for assessing their mood during the meeting etc.

Table1 and Table 2 contain a set of possible states and activities based on the first criterion. Our lexicon of meeting actions is based on this interpretation. Table 3 and Table 4 contain a set of states and activities based on the second criterion.

Participant State	Location	Start Time	End Time
Absent		ss:ms	ss:ms
Sitting	Seat X	ss:ms	ss:ms
Standing	Seat X	ss:ms	ss:ms
	White Board Projector Screen Unspecified		
Walking		ss:ms	ss:ms
Acting			
Still		ss:ms	ss:ms
Writing		ss:ms	ss:ms
Nodding		ss:ms	ss:ms
Shaking head		ss:ms	ss:ms
Applaude		ss:ms	ss:ms
Pointing		ss:ms	ss:ms
	Person Y White Board Projector Screen Unspecified		
(*)Voting		ss:ms	ss:ms

Table 1-Defined set of possible participants states (1. criterion)

Participant Activity	Addressee	Start Time	End Time
Silent		ss:ms	ss:ms
Back channel		ss:ms	ss:ms
Laugh			
Mumble			
Talking		ss:ms	ss:ms
Monologue			
Privately-Chatting	Participant X		
Openly-Discuss			

Table 2 –Defined set of the participant interactivities (1. criterion)

Participant State	Location	Start Time	End Time
Absent		ss:ms	ss:ms
Sitting	Seat X	ss:ms	ss:ms
Standing	Seat X	ss:ms	ss:ms
	White Board Projector Screen Unspecified		
Walking		ss:ms	ss:ms
Acting			
Still		ss:ms	ss:ms
Writing		ss:ms	ss:ms
Nodding		ss:ms	ss:ms
Shaking head		ss:ms	ss:ms

Table 3-Defined set of possible participants states (2. criterion)

Participant Activity	Addressee	Start Time	End Time
Silent		ss:ms	ss:ms
Back channel		ss:ms	ss:ms
Laugh			
Mumble			
Talking		ss:ms	ss:ms
Monologue			
Privately-Chatting	Participant X		
Openly-Discuss			
Applaude		ss:ms	ss:ms
Pointing		ss:ms	ss:ms
	Person Y		
	White Board		
	Projector Screen		
	Unspecified		
(*)Voting		ss:ms	ss:ms

Table 4 –Defined set of the participant interactivities (2. criterion)

*- Raising hand can be a cue for asking a question or voting. Asking a question can be a part of discussion or a sign for the beginning of discussion (after presentation, white board or monologue). It is one way how participants express their intentions for taking a turn. More participants may raise hand at the same time, but only one or no one may take a turn. Only for the participant who takes a turn raising hand is recognized as asking a question but for the rest it is recognized as an intention for asking a question. Intention for asking a question is not important for browsing the meeting. The current speaker will follow his act of raising hand with a question. Therefore, the information obtained from non-verbal element (raising hand) is not important for recognition of asking a question.

Gaze Interpreter

Very important information in a multimodal dialogue is gaze direction. People look at each other to signal attention and interest (focus of attention), to monitor a listener acceptance and understanding, to coordinate turn-taking. People look away to plan an utterance, to concentrate on a complex cognitive task [16].

The gaze interpreter has as an input data from the recognizer that will give information about the gaze direction for each participant at a specific time frame. In order to give a semantic meaning to this information we will use context information e.g. information about the position of participant which focus of attention we want to determine, gaze direction information (left, right…), position of other objects of interest and participants in the meeting room. Table 5 contains a description of the possible interpretation of gaze direction

Action	Focus of Attention	Start Time	End Time
Look at	Person X	ss:ms	ss:ms
	White Board		
	Projector Screen		
Look away		ss:ms	ss:ms

Table 5. Interpretation of gaze direction

Speech (Language) Interpreter

The input of the speech interpreter is a meeting speech transcript. This file contains the speech recognition results for each participant together with information about time frame, backchannel noise (laugh, door slam, mumbling), overlapping, fragments (restarts, interruptions) and speech errors. For the recognition of meeting actions we need additional information at the higher level (dialog acts, detected topics, etc.). In our model we will assume that we have this information.

The speech interpreter has the task to determine semantic data relevant for recognition of meeting actions for each turn of each participant from an input speech transcript. For each participant we will represent the semantic meaning not on utterance level but on turn level or turn array level.

Turns are the key of modeling how a conversation “works” and how it is organized. A turn is a fundamental construction unit of conversation. It is bounded by turn

transition points. Turn transition points are places where speaker changes may occur. Turns are organized into a larger unit called a turn array. A turn array is a series of consecutive turns produced by one speaker and bounded by the turns of the other speakers [17]. During the conversation the participants may quickly figure out who will talk next and when they have to talk. Turn-taking behavior is governed by a set of turn-taking rules [6]:

a) If during the turn the current speaker selects the A as the next speaker than A has to take a turn.
b) If the current speaker doesn't select the next speaker than any other speaker may take a turn.
c) If no one else takes the next turn the current speaker may take the next turn.

Selection of the next speaker can be done directly or indirectly by menas of adjency pairs: question-answer, greeting-greeting, request-consent/deny and suggestion-acknowledgement etc. Next speaker may be selected by non-verbal elements such as gesture or gaze. However, these rules are not completely applicable in human-human dialogue because of a number of variations such as:speakers who never take a turn despite an extended pause, or speakers who speak to much such that it is impossible to take a turn, etc.

Semantic features of an array (turn)

Table 6 contains some of the features that are important for representing the meaning of a speaker's turn array and also for the recognition of a meeting action.

FEATURE	VALUE		EXAMPLE
Start Time	mm:ms		
End Time	mm:ms		
Topic	Base	what we are talking about	my holiday
	Subtopics	aspect of base	destination
Act	Dialog Acts	DAMSL dialog acts	ASSERT, CHECK, OFFER
	Addressee	targets of the messages	
	Keywords		
	Number	for each key words	
Form	Number	sg. or pl.	
	Person	1, 2, 3	
Backchannel	Type	defined set of values	Laugh, Mummbling
	RelativeTime	in relation to Start Time	
	Duration		
Overlapping	Actor	person who speak at the same time	

Table 6. Semantic features of speaker's array (turn)

Multimodal Interpreter

Modality fusion on the semantic level

There is a number of approaches in solving the multimodal integration problem on the semantic level

- Unification-Based Multimodal Integration [19]
  This approach uses language processing techniques for multimodal fusion. Type feature structures are used to provide clearly defined and well understood meaning representation of modalities and unification is used for multimodal integration.
- Finite-state methods for multimodal parsing, integration and understanding [20]
  This approach employs a finite-state device to parse multiple streams and to combine their content into a single semantic representation using multimodal context-free grammar (MCFG).
- Melting-pot fusion [21]
  This approach is based on mapping each input invent into a time-stamped set of structural parts or semantic slots and combining them in a domain-independent way.
- Frame Merging Approach [22]
  A frame is a unit of knowledge source describing an object. Each frame has a number of associate slots .The slots are properties of the object, action of the object and relationship of the object with other frames. This relationship with other frames enable designing semantic networks for particular context.

In our approach we will use typed feature structures for meaning representation and because of that we will use an unification based approach for the integration of different modalities. It is also possible to use a rule-based approach for the integration of modalities represented by typed feature structures or to combine these approaches. A result of the fusion is the semantic meaning representation of the multimodal behavior of a participant that contains information about activity, state, speech, gaze direction of meeting participants, etc.

Additional inference

Modality fusion is not the final process in multimodality interpretation. Some data may still be missing or ambiguous. Therefore, we need additional information from background knowledge, a user profile to resolve ambiguities and to derive unspecified data. For instance, we need information about the role of a speaker in order to distinguish lecture from monologue, presentation or white-board.

6.2.3.2 Meeting Action Recognition Module

A meeting action is determined by (inter)activity of meeting participants. In the previous section we saw what kinds of information are important for representing participants behavior during the meeting. Part of the information encoded in this representation is not crucial for recognition of meeting actions but they may be significant for post processing tasks. Therefore we need to extract observation

features crucial for recognition of meeting actions according to the well-defined lexicon of meeting actions. Besides participant specific features there are some features that are common for all participants. Some of the participant specific features are state, location, activity, talking duration, topic, speech form, participant's role, dialog acts, key words, etc. Participant common features are: previous dialog act (the last dialog act from previous turn), previous key words (but only key words marked as action recognition key words), previous addressee (target of the message) etc.

We will use a Hidden Markov Model for modeling a sequence of meeting actions. States in HMM are meeting actions and observations are semantic features described above. We want to find the most probable sequence of meeting actions $A = a_1, a_2, ...a_N$ for a given observation sequence of semantic features O which represents a combination of multimodal behavior of participants

$$\hat{A} = \arg\max_{A \in \alpha} P(A \mid O)$$

The Viterbi algorithm will be used for finding the most probable sequence of meeting actions.

The main problem is the insufficiency of data for training HMM. Another problem is that IDIAP meeting data corpus, which we use in our model, contains meetings that consist of a small set of meeting actions (monologue, dialogue, presentation, white-board, note-taking, consensus and disagreement).

7. Conclusion and Future work

In this paper we proposed a semantic approach for modeling a meeting as a sequence of meeting actions based on a lexicon of meeting actions. The main goal of our approach is to encode more semantic details in each level of meaning interpretation (modality level, participant (multimodal) level, and meeting action level) in order to enable easier and purposeful browsing and querying of an archive of recorded meetings. We defined a lexicon of meeting actions that are divided in three groups: single speaker dominate actions, multi-speaker actions and non-verbal dominate actions.

In order to prove our approach for all low-level and high-level meeting actions defined in our lexicon we need a larger and more natural meeting corpus. Ongoing work will be based on extracting a richer set of semantic features for recognition of meeting actions. We will try to apply and compare different techniques for representing meetings as sequences of meeting actions (HMM, Dynamic Bayesian Networks etc.) or to use a combination of these techniques. Inputs in our semantic model consist of information obtained from different sources (speech, gaze, gesture/action, participant's location etc.). Future work will involve encoding detected emotion in the representation of participant's multimodal behavior. This will enable assessing the mood during the meeting on the group and on the individual level.

8. References

[1] Michael Argyle - Social Interaction, Tavistock Publications, 1973.
[2] Stephane Marchand-Maillet - Meeting Modelling for Enhanced Browsing, 2003.
[3] McCowan, I, Gatica-Perez, D, Bengio, S and Lathoud, G - Automatic Analysis of Multimodal Group Actions in Meetings, IDIAP-RR 03-27, 2003
[4] Samy Bengio - Multimodal speech processing using Asynchronous Hidden Markov Models, MIT Press, 2003.
[5] J. Coutaz - Multimedia and Multimodal User Interfaces : A Taxonomy for Software Engineering Research Issues, East/West Human Computer Interaction'92, St Petersburg, 1992.
[6] Daniel Jurafsky and James H. Martin - Speech and Language Processing – An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition, Prentice-Hall, 2000.
[7] James Allen and Mark Core – Draft of DAMSL: Dialog act markup in several layers, Unpublished manuscript, 1997.
[8] R. F. Bales – Interaction Process Analysis, Addison-Wesley, 1950.
[9] Ashwani Kumar and Laurent Romary – A Comprehensive Framework for MultiModal Meaning Representation, International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands, 2003.
[10] J. Chai, S. Pan and M. Zhou - MIND: A Semantic–based Multimodal Interpretation Framework for Conversational Systems, International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialog Systems, 2002.
[11]http://www.etca.fr/CTA/gip/Projets/Transcriber/
[12] Devillers, Laurance – Annotation and detection of emotion in a task-oriented Human-Human dialog corpus, ISLE Workshop on Dialogue Tagging for Multi-modal Human Computer Interaction, Edinburgh, 2002.
[13] Hartwig Holzapfel, Christian Fügen, Matthias Denecke, Alex Waibel - Integrating Emotional Cues into a Framework for Dialogue Management, ICMI, Pittsburgh, 2002.
[14] Rainer Stiefelhagen and Jie Zhu - Head Orientation and Gaze Direction in Meetings. Conference on Human Factors in Computing Systems (CHI2002), Minneapolis, 2002.
[15] I. Poggi - Towards the Alphabet and the Lexicon of Gesture, Gaze and Touch. Multimodality of Human Communication. Theories, problems and applications, Virtual Symposium. Ed. P. Bouissac, 2002.
[16] Novick, David G., Hansen, David G., Ward, Karen - Coordinating turn-taking with gaze. Proceedings of the International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, PA, 1996.
[17] Herbert M. Isenberg - The Organization of Conversation
[18] J.C.Martin - Towards "intelligent" cooperation between modalities. IJCAI Workshop on "Intelligent Multimodal Systems", Nagoya, Japan. 1997
[19] M. Johnston, P. R Cohen, D McGee, S. L Oviatt, J. A. Pittman, and I. Smith, Unification-based multimodal integration, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997.
[20] M. Johnston and S. Bangalore- Finite-state Methods for Multimodal Parsing and Integration, Proceedings of COLING-2000, Saarbruecken, Germany, 2000
[21] L. Nigayand and J. Coutaz. - A Generic Platform for Addressing the Multimodal

Challenge," In Conference on Human Factors in Computing Systems (CHI '95), ACM Press, 1995.
[22] R. Sharma, V. I. Pavlovic, and T. S. Huang - Toward Multimodal Human-Computer Interface, Proc. IEEE special issue on Multimedia Computing and Communication, 1997.
[23] http://www.icsi.berkeley.edu/Speech/mr/
[24] http://www.is.cs.cmu.edu/meeting_room/