Supporting Engagement and Floor Control in Hybrid Meetings

Rieks op den Akker, Dennis Hofs, Hendri Hondorp,
Harm op den Akker, Job Zwiers, and Anton Nijholt

Human Media Interaction Group
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
infrieks@ewi.utwente.nl

Abstract. Remote participants in hybrid meetings often have problems to follow what is going on in the (physical) meeting room they are connected with. This paper describes a videoconferencing system for participation in hybrid meetings. The system has been developed as a research vehicle to see how technology based on automatic real-time recognition of conversational behavior in meetings can be used to improve engagement and floor control by remote participants. The system uses modules for online speech recognition, real-time visual focus of attention as well as a module that signals who is being addressed by the speaker. A built-in keyword spotter allows an automatic meeting assistant to call the remote participant's attention when a topic of interest is raised, pointing at the transcription of the fragment to help him catch-up.

1 Introduction

AMIDA is a research project funded by the EU on automatic analysis and recognition of activities in meetings and the understanding of their outcomes. It is a follow-up of the AMI and M4 projects that targeted the development of meeting support technology, multi-modal meeting browsers, as well as automatic audio-, and video-based summarization systems[1]. AMIDA is aiming at two new targets: the development of *real-time* processing (for online speech recognition, online gesture recognition, etc.), and the application of these real-time methods in cross-modal conversational scene analysis for supporting meetings. Meeting support includes the (semi-) automatic retrieval of information needed for decision processes in meetings as well as technology in the form of meeting assistants that support remote participation in hybrid meetings, in which one or more participants are remote and others are present in a shared room. AMI delivered a multi-modal, multi-layered annotated corpus containing 100 hours of audio and video recordings of face-to-face meetings [1],[2]. Besides natural small group meetings, the corpus contains about 35 series of 4 meetings of design groups that consist of 4 people. These

[1] (AMI(DA) stands for Augmented Multi-party Interaction (Distant Access) - www.amiproject.org

A. Esposito and R. Vích (Eds.): Cross-Modal Analysis, LNAI 5641, pp. 276–290, 2009.

are scenario-based meetings, the people play a role in a group that has the task to design a new remote tv control [3]. The same scenario was used in the follow-up project that collected a new corpus of meetings. In each of the series of 4 design group meetings, now three include a remote participant. In two of them 3 people are located in the same room, and one has a video and audio link with the room. Speech is recorded by individual lapel microphones as well as by a microphone array. Speech has been transcribed by hand. Words have been time-aligned. Words of a speaker have been segmented into intentional units and hand-annotated with dialogue act tags. A part of the corpus was annotated with visual information: hand and head gestures, focus of attention (who or what are they looking at), as well as with addressee information, telling whom the speaker was talking to during a stretch of talk. Other content annotations include abstractive and extractive summaries, topic segmentation, named entities, and subjectivity [4]. The annotations are organized using NXT in layers, all directly or indirectly referring to the time line, a requirement for cross-modal analysis as well as for replaying the meeting. (For NXT and NXT based tools see [5], [6]). The annotated corpora have been used to develop models of multi-modal multi-party meeting behaviors as well as for training and testing automatic recognizers of these behaviors or aspects thereof. They are also used for automatic generation of summaries, and for generation of behaviors of synthetic (conversational) characters in virtual smart meeting rooms [7].

This paper presents the first version of a system that demonstrates how these recognition and generation modules can be used to support remote meeting participation. This *User Engagement and Floor Control* (UEFC) demo is meant to show how AMIDA research can contribute to technology that makes remote meetings more engaging by giving remote participants more control in discussions and decision making processes. The UEFC demo is one system developed in a general Meeting Recorder Framework that is being used as a research vehicle for experimental studies of how outcomes and processes in remote meetings depend on properties of communication channels and how engagement and efficiency are affected by meeting support technology. The Meeting Recorder Framework contains a package for media streaming. Audio and video streams can be produced real-time by devices (in on-line use in a life meeting) as well as from files. This makes it possible to exploit the MRF for building systems that play back recorded audio and video files and that use annotation layers of filed meetings in what we call "off-line" systems, as well as for building real life, "on-line", tele-meeting systems.

Basic constraints of tele-meeting systems are the capacities of the communication network and the capacities of the audio and video displays. They constrain how information can be presented on the "user's interfaces" display and what the affordances of the interface can offer the user/remote participant to control the meeting. An earlier off-line system (based on play back of a hand-annotated meeting) for remote meeting support developed within the AMIDA project demonstrated how a mobile remote participant could be informed about the meeting by means of a 2D or 3D visualisation of the meeting room [8].

The UEFC system that we present in this paper is made to be used by a remote participant that has a fast internet connection and a desktop computer screen. The use case scenario is that of a participant that cannot or has chosen not to attend the meeting continuously, who may have a special role in a project group and who is interested in some agenda items more than in others and who will either devote "continuous partial attention" to the meeting or is multi-tasking. This scenario becomes more and more common practice in our meeting culture, where also people that meet physically in the same room are often distributing their attention between reading their mails and participating in local meeting activities. Moreover, people sometimes choose to attend a meeting staying in their own office because it makes it easier to selectively attend only parts of the meeting, even if travel time can be neglected.

2 Mediated Communication Problems

Despite its frequency of use, communication in audio and video conferencing systems still suffers from a number of limitations that have been shown to impact the engagement in and the effectiveness of remote meetings. Among these are noisy lines or background noise, difficulties hearing the speaker, not knowing who is speaking, and a lack of the experience of "social presence". In face-to-face meetings the audio and visual "channel" are both used in turn-taking, controlling the conversational flow, or floor control. Floor denotes a cognitively shared attentive space that mediates in the sequential or simultaneous organization of participants' contributions: turn-taking as well as topic management (cf. Edelsky [9] and Hayashi [10]). There can be several floors at the same time as we see in cocktail parties, and in larger group meetings, [11]. Although a floor mostly contains one main speaker at a time, verbal and non-verbal listener feedbacks (laughters, head nods) that occur in the back channel control turn-taking and are important for keeping the pace in the conversation. The physical and psychological barriers that exist in hybrid meetings make that it is often hard for remote partners to attend to a selected floor of which the participants share the same room. It makes it harder for them to take turn, and to initiate a new topic and a new floor.

Ideas about how technology could prevent specific problems in computer mediated communication no doubt rest on a theory of communication. From such a theory we may expect that it explains (or at least makes believable) the effectiveness of introducing technological means. Some theories of communication are information-theoretic or cognitive. The most popular is the sender-message-receiver model (see for example [12]). Other theories of communication emphasize the social dimension. Politeness, dominance, fight for the floor, but also commitment, empathy, social presence and social signals are key elements of these theories. Central is the obligation that co-partners have towards the others to be clear, and to follow and provide feedback (see [13]). On the level of conversation the roles change with the obligation: those between speaker and selected addressees differ from those between speaker and overhearers. Monk et al.

[14] emphasize that on the more global task level "peripheral participants", who are monitoring communicative behavior not explicitly directed at themselves, need to be considered when designing equipment for video mediated cooperative work. A complete theory of human communication will encompass both, the information- and cognitive- processing aspects as well as the social aspects, as well as their relations. These relations concern the various ways that sender, receiver and message are related and involves notions as authenticity, trust, and credibility. Notions that are important when considering group decision making as well as making persuasive technology [15].

In [16] the authors give a list of problems that people experience with communication in hybrid meetings (i.e. meetings where some people are local and some are remote).

- Audio problems:
  - Poor quality speakerphones
  - Too much background noise
  - Multiple speakers speaking at the same time can be difficult to understand
  - People speaking too far from microphones
- Remote attendee problems:
  - Inability to conduct side conversations.
  - In-room attendees forget about remote people
  - Challenging to break into lively conversation
  - Difficult to detect in-room speaker changes
  - Hard to identify people currently in the meeting room
  - Hard to identify the current speaker
  - Difficult to participate in brain-storming sessions
  - Cannot see in-room demonstrations or artifacts
- Meeting room problems:
  - Local people are more emotionally salient than remote participants.
  - Easy to forget about remote participants
  - Often local people do not know who is still connected

Key point is that subjects in mediated communication suffer from increased uncertainty, caused by the spatial and temporal distance between co-participants. Uncertainties are expressed in questions as "Did he receive my message?" or "Is what I made of this message what my partner meant?" Speakers and listeners work parallel in cooperation. While speaking a speaker will be interested in how the message is received and the receiver will send signals if and how he received the message. So they work on shared beliefs in a "grounding" process. What changes with the physical circumstances and the channel properties is the sample time and the granularity of the information units, the packages that are send back-and-forth. Units and sample times are not determined in advance by grammatical rules, they are "interactively determined" by the floor participants. ([17], p.726-727.) When people get used to audio delay they adapt sample time and package size, and explicit verbal control messages become more frequent.

Our analysis of the remote meeting corpus are in line with earlier findings: in remote meetings participants refrain from using verbal backchannel signals, which makes that speakers often experience that they talk in a void. Visual contact may help here. See also [18] for similar conclusions. Addressing in remote meetings is more explicit, since speakers are often not convinced that the remote partner is paying attention. The visual channel is important when people discuss objects, or documents. Moreover, it helps to identify who is speaking and to signal focus of attention of the speaker, which helps understanding verbal referring expressions. The above findings have been leading for the development of the UEFC demonstrator for the scenario of use we discussed earlier.

3 The User Engagement and Floor Control Demo

Figure 1 shows the HMI meeting room and a remote meeting participant using the system described in this paper. This section describes the modules of the UEFC demo, and the graphical user interface that presents the view on the co-participants and the remote meeting room. The remote participant (RP) is present in the meeting room on a large screen. A smaller screen shows the view that the RP has of the meeting room, so that both sides can always see what the other side sees.

(a) (b)

Fig. 1. A remote meeting in which the UEFC system is being used

The modules that receive media input streams, send their respective outputs to a central database application known as *The Hub*, which sends it through to the modules that rely on the data. Figure 2 shows the dependencies of all modules between each other. The Media Streamer records all video and audio from the local and remote participants; the ASR, VFOA and KWS modules (see below) process video or audio data, which is used by the Dialogue Act Recognizer and the Addressing Module. The Media Streamer is a video conference tool developed at the University of Twente. It runs on all participant's computers and that of the meeting room itself. It reads out the data from the webcam

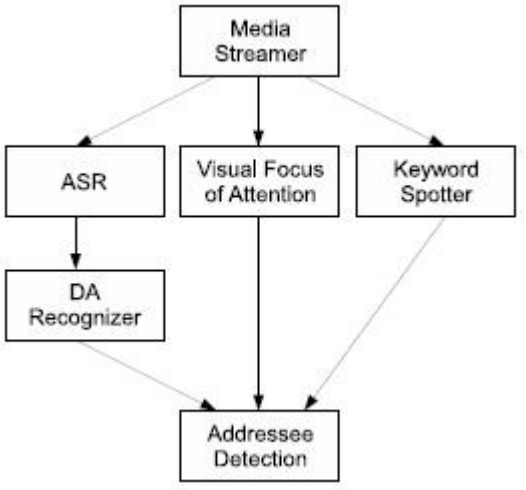

Fig. 2. Dependencies between the modules of the UEFC demonstrator

and microphone, and can stream the data to another PC for further processing. It also takes care of compressing the video stream. The Hub serves as a central point of communication for different software modules developed within AMIDA. Modules can *subscribe* to the Hub as a producer or consumer (or both) of specific types of data. In our UEFC example, the Visual Focus of Attention Module produces "focus" data, which is consumed by the Addressing module, which in its turn produces "addressing" data. The Hub makes sure that every module is aware of new data arriving from other modules. The sections below will explain the details of the other modules in the system.

3.1 Automatic Speech Recognition

The ASR system receives the incoming audio streams from all participants on different sockets, which allows the system to be split between Windows and Linux systems easily. A Java wrapper allows the results of recognition to be streamed via the Java middleware to the Hub. From the Hub, metadata is available to all other consumers. Thus the demonstrator runs on several computers. Audio is captured on one machine in real time from a single microphone or a microphone array and on-line beamformer. For every audio stream it generates the words that are being spoken. It does this in spurts; there needs to be a short silence before the system starts to process the stream. The word data that is send to the Hub, includes start- and end time information, and from which of the participants it came. The system that is used within the UEFC Demonstrator is the webASR system from the University of Sheffield. For the details on this system please refer to [19][2].

[2] webASR is located at the following website: http://webasr.dcs.shef.ac.uk/.

3.2 Dialogue Act Recognition

The Dialogue Act Recognition module segments the words from the ASR module into Dialogue Act segments and classifies them with a Dialogue Act Tag from the AMI tag set. At the time of writing the segmentation is done using so-called *spurts*, meaning that a segment boundary is inserted whenever there is a pause of a certain size between two words. In the future, the segmentation algorithm described in [20] will be used. The Dialogue Act Classification, or tagging, is done using the system described in [21].

3.3 On-Line Keyword Spotting

The Keyword Spotting module analyses the audio input stream for the occurrence of certain keywords. The acoustic keyword-spotter is based on an estimation of phone posterior probabilities by neural networks and on the classical tandem of target word model and background model, (cf. [22]). It can be given a list of keywords, that can be modified on the fly, for which it will look. Whenever it detects one of the keywords, it sends a signal to the Hub, indicating the word and the time in the audio stream at which it recognized it. The module can handle a list of up to 100 words, and is, for these words, much more reliable than the standard ASR system. The keywords for which spotter looks are inserted by the Remote Participant, so that he can be warned whenever a topic of his interest is being discussed.

3.4 Visual Focus of Attention Recognition

Visual focus of attention (VFOA) of participants provides important cues to recognize interactions in meetings. But, recognizing the VFOA directly is a difficult task. The main cue to recognize VFOA is the head pose. The Visual Focus of Attention module analyses the video streams of each individual meeting participant. It tracks the pose of the head in terms of tilt (vertical movement) and pan (horizontal movement) and maps these values to predefined targets. The system then sends for every 2 frames of video data (e.g. 15 times per second) the best matching target to the Hub. In the current setup we are only interested in who is looking at the Remote Participant's screen, so there are two targets: *remote participant* and *other*. The system that is used is based on work in [23]. In the UEFC demo the main consumer of the VFOA data is the Addressee Detection Module.

3.5 Addressee Detection

The addressee detection module (ADR) that is used in the UEFC system identifies the addressee of the speaker. In particular ADR will tell if the *remote participant* is being addressed. This information is used to call the RP's attention when he is not actively participating in the conversation. See Figure 2 for the input the ADR module depends on.

The addressee classifier is trained on statistical models of patterns of addressing behavior in the annotated AMI and Amida meeting corpora. Jovanović build

a classifier for addressee prediction trained on the AMI corpus using Dynamic Bayesian Network technology: for predicting the addressee of the current dialogue act it uses the information about the addressee of previous dialogue acts. The method assumes that four participants meet face-to-face and sit at fixed positions at a table. For use in the remote setting with a variable number of local participants, who may move around, a more general addressee detection module was made. The idea is that each participant in the meeting has a dedicated addressee predictor, a binary classifier that tells whether the participant is addressed by the speaker or not. The features used are of the following three types.

1. lexical features, in particular the use of personal pronouns, the number of words of the dialogue act,
2. contextual features, how active was the participant in the last turns and how often was he addressed in the last turns?
3. visual focus of attention of the speaker and other participants, in particular do they gaze at the participant?

A supervised classifier trained and tested on the hand-annotated AMI corpus has an accuracy of 92,53%. The baseline for this classification task is 89.2%, the percentage of all dialogue acts in the test set that are not addressed to one single distinguished participant (group addressed dialogue acts counts as no), so a classifier that labels every dialogue acts as such will receive that score. For more details about the addressee classifier we refer to [24].

3.6 The Graphical User Interface

The GUI of the UEFC demo shows close up view of each of the local partners and an overview of the meeting room. The positions of the video frames reflect positions at the meeting room table. The central overview frame has a border that changes color when the user is called, or addressed by a speaker in the meeting room. An audible signal will catch his attention. The user can indicate his state of attendance. He can add or remove keywords to and from the keyword list. The keywords are send to the Hub which forwards them to the keyword spotter. If the RP is not actively attending the discussion a visual and audible signal will be generated that a keyword of interest has been detected. The keyword is highlighted in the scrollable frame that contains the online produced transcript of the meeting. This allows the RP to easily catch-up with the ongoing meeting. The GUI also indicates whether there are problems with the audio or video channel.

4 The Meeting Recorder Framework

The UEFC demonstrator is an application that uses a software architecture and framework that we developed for experimenting with remote meetings, one-to-one or hybrid. But it can also be used for off-line applications and for building software agents, and virtual characters. There are three kinds of *client* applications: one for the remote participant, one for the meeting room, the location

where the overview of the meeting room is recorded (and the remote participant is presented), and finally there is an application for each local participant.

The user interface is actually implemented in an integrated application, named *meeting recorder*, which can also be run stand-alone.

4.1 Meeting Recorder

A meeting recorder is a video conferencing application that can record all video and audio to files, and supports the controller marks on the Hub for a synchronized start of all sites. The application can be customized in various ways using an XML configuration file.

First of all, the configuration file describes exactly what streams should be broadcast and what streams should be received. It can broadcast from files or from capture devices. For capture broadcasts, one can specify the media format and the recorded media can be saved to an AVI or WAV file. Video streams can be compressed with DivX or stored as uncompressed RGB video. For each broadcasts stream, you can add custom stream receivers, which are simple Java classes. This is used for example to send an audio stream to the ASR. Such classes can also be added for received streams.

In the choice of the media formats in the UEFC demo, we had to consider limited network bandwidth (high bandwidth but still limited) and the high quality demands of signal processing modules. The ASR and keyword spotter require 16 kHz, 16 bits, uncompressed PCM audio. It appeared no problem to transmit this audio format over the available network, although in the future we may want to compress the audio before transmitting it. The video is another matter. The VFOA requires 15 Hz, 320x240, uncompressed RGB video. This cannot be transmitted over the network and we chose DivX for the compression of transmitted video streams. DivX offers good image quality while the compression is good enough for high bandwidth networking.

The user interface is basically a desktop containing titled frames that can be moved and resized like normal windows. Each broadcast and received video stream has its own frame. With plugin code the frames could be automatically positioned and sized, while any additional frames (not just video frames) can be added as well. An example of the user interface is shown in Figure 3.

The media streaming is done using a separate software package.

4.2 Media Streaming

An open package for low-latency high-bandwidth video and audio streaming has been developed. It uses DirectShow and has basic interfaces in C++ and Java. The Java interface uses JNI to access the C++ interface. The actual network streaming and other functionalities are implemented in Java only on top of the base framework. The base framework stays close to DirectShow concepts with some simplifications. DirectShow leans on Microsoft's Component Object Model (COM). DirectShow has no built-in support to obtain the media data outside COM for custom processing or network transmission. For those purposes, it is necessary to build a DirectShow filter (COM object), which needs to be installed

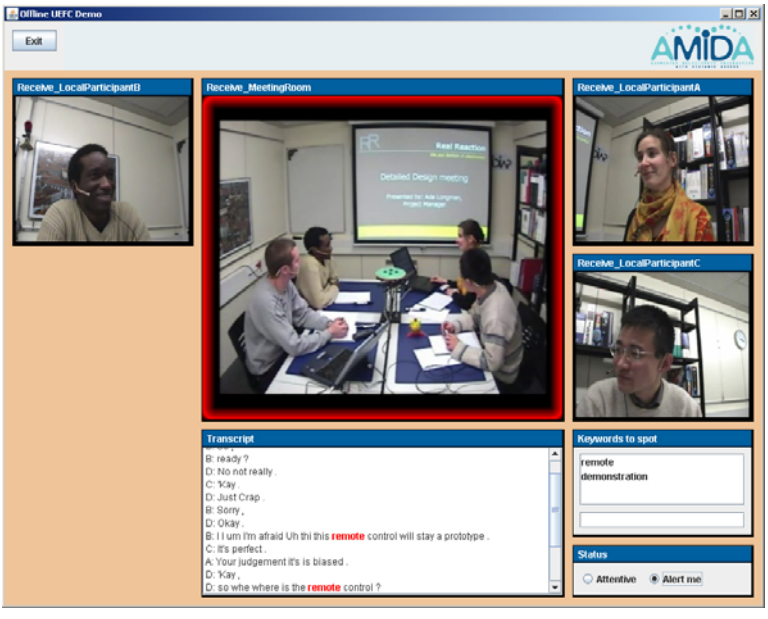

Fig. 3. Meeting recorder

in Windows. This is rather complicated and it can be very difficult to integrate existing software. This media framework hides COM, which allows for more simple interfaces.

5 Does It Work?

The UFCD system shows the working of the various technical modules developed in the AMIDA project, such as the real-time head pose tracker that informs the system about the visual focus of attention and the addressee detection module that decides whether the remote participant is being addressed. Moreover, the media streaming software allows for synchronized audio and video streaming over the internet with an acceptable delay of less than 1 sec. People that have worked with the system appreciated that you can see the faces of the individuals and also have an overview of the meeting room. The overview makes it easier to follow discussions performed in the remote meeting location, whereas the individual videos improve one-to-one contact in a dialogue between the remote participant and someone located in the meeting room. The participants in the meeting room found it hard to establish mutual gaze in a one-to-one interaction with a remote participant. This is caused by parallax, the distance between the positions of the cameras and the screen position where the faces are shown.

The main question, however, concerning evaluation is if the system can be used so that it indeed improves the engagement of remote participants in hybrid meetings and their control over the floor. To answer these questions requires more

field tests in realistic settings where participants do have personal interests in the outcomes of a meeting. Another issue is whether people will use a teleconference system more when it supports them in distributing their attention over different tasks by automatic attention call and the possibility to catch-up. Do these techniques improve efficiency of meetings? More engagement doesn't always imply improved efficiency in terms of task performance and vice versa. In an interesting study about the effect of various forms of interfaces in a video mediated cooperative task with a shared tabletop display Hauber et al. [25] used a variety of social and performance measures to see how the interface design affects social presence and efficiency. In this study, the face-to-face condition is included as a gold-standard control. They conclude that "there were important differences between the 2D and 3D interfaces. In particular, the 3D interface positively influenced social- and co-presence measures in comparison to 2D, but the task measures favored the two-dimensional interfaces." The important lesson is that we need to be careful in defining what we are aiming at: improve engagement and social interaction or efficiency of task completion. Moreover, the impact that meeting support technology has on engagement and effectiveness of task completion will depend on the type of task that the group has to perform. McGraths group task classification system separates tasks into the following four quadrants: (a) generating, (b) choosing, (c) negotiating, and (d) executing. The quadrants and the tasks they contain are related to one another within a two-dimensional space. One dimension reflects the degree to which the task entails cognitive (choosing) versus behavioral performance requirements (executing). The other dimension reflects the degree and form of interdependence among group members (generating versus negotiating). It has been found that a "communication medium is more likely to affect group outcomes when there is a need for the expression and perception of emotions, when tasks require coordination and timing among members activities, when one is attempting to persuade others, or when tasks require consensus on issues that are affected by attitudes or values of the group members." (citation from [26]). The intellective and decision making tasks and the cognitive conflict and mixed-motive tasks thus seem to require a greater amount of coordination and group member interaction than the other tasks. If this holds, it is a real challenge to apply meeting support technology for use in decision making tasks and cognitive conflict tasks and to design experiments to measure the impact of these technologies on engagement and effectiveness of remote meetings.

6 Conclusion and Future Work

The latest results in automatic conversational scene analysis in face-to-face meetings - see [27] for an up-to-date overview - make it possible to build systems that select automatically the best camera view for meeting observers. These could be remote meeting participants in a teleconference situation or users of meeting archive systems ([28], [29], [30]). Because of the real-time constraint the most challenging is the use of these technologies by remote participants in an ongoing meeting. What is the best view presented to the remote participant and

how does user control over meeting support technology affect the cognitive load and distract the attention from the real issues so that it hampers engagement and participation? Meeting assistants enter smart meeting rooms in the form of software agents that aim to assist the meeting process and to facilitate more effective and efficient meetings. An investigation by Rienks et al. discusses how "pro-active meeting assistants" could be exploited [31]. As soon as these new technologies get into use they change meeting practices. Feedback to the group during meetings about the talkativity of participants has shown to affect social behavior [31], people change their language and use more explicit signals if they believe that helps the technology to identify it. Ehlen et al. [32] give an example of the latter, where participants started to mention "*action items*" explicitly in meetings, after they found out that the meeting browser they used and that identified these items automatically in previous meetings listed them under this name. In remote meetings people use more explicit procedures to make clear whom they want to address, so that it becomes easier for outside observers, including meeting assistants, to identify these signals. As soon as meeting technology, based on studying regularities in human conversational behavior in meeting practice, becomes a part of that practice ("technology in the loop"), people adapt their behavior in sometimes unforseen ways.

We presented the Meeting Recorder Framework, a general framework for media streaming and for building off-line and on-line remote meeting support tools. An application of the MRF is the User Engagement and Floor Control demo for demonstrating technology for conversational scene identification that has been developed in the AMIDA project, in a configuration that is meant to support engagement of a "multi-tasking" remote participant in a hybrid meeting. The framework is currently used to study the effect of audio delay on the one-to-one audio only interaction where subjects have to negotiate about the best meeting date. Another aim of this study is to see if visual feedback about the transfer rate of the audio signal affects the time that it costs subjects to adjust their interactive behavior to the audio delay. Future work will consist of the evaluation of the remote meeting support technology in decision and negotiation tasks as we discussed in the last section.

One of the outcomes of a market assessment investigation that was carried out in the AMIDA project says that potential users would welcome a remote meeting assistant that helps the user to catch-up with a meeting when they arrive late, or miss a part of it for other reasons. Users also like to see that the system provides links to relevant documents or other information resources related to the issues being discussed in the meeting. These functions have been implemented in the Automatic Content Linking Device (ACLD), a real-time query-free document retrieval system [33]. The ACLD is a system that constantly retrieves items from a document repository, which includes meeting related documents together with excerpts from previous meetings of the group, and displays them to a participant or to all of them; the device can be used online, during a meeting, or off-line, integrated in a meeting browser. The idea is to see how functionalities of the ACLD could be usefully integrated with the UEFC system. Further research

should reveal where engagement and participation in meetings is still improved by information and communication technology and where the technology stands in the way instead of supporting interaction.

Acknowledgments

We are gratefull to the anonymous reviewers for their comments on an earlier version of this paper. We acknowledge our AMIDA partners from IDIAP in Martigny, DFKI in Saarbrücken and the Universities of Brno, Sheffield, and Edinburgh for their contributions to the UEFC demonstrator. The work reported in this paper is sponsored by the European IST Programme Project FP6-0033812 (AMIDA). This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

References

1. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., Wellner, P.: The AMI Meeting Corpus. In: Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, Netherlands (2005)
2. Carletta, J.C.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Resources and Evaluation 41(2), 181–190 (2007)
3. Post, W., Cremers, A., Henkemans, O.: A research environment for meeting behavior. In: Nijholt, A., Nishida, T., Fruchter, R., Rosenberg, D. (eds.) Social Intelligence Design, Enschede, The Netherlands (2004)
4. Wilson, T.: Annotating subjective content in meetings. In: Proceedings of LREC (2008)
5. Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann, H.: The nite xml toolkit: Flexible annotation for multi-modal language data. Behavior Research Methods, Instruments, and Computers 35(3), 353–363 (2003)
6. Reidsma, D., Hofs, D., Jovanovic, N.: A presentation of a set of new annotation tools based on the nxt api. In: Proceedings of Measuring Behavior, Wageningen, The Netherlands (2005)
7. Nijholt, A., Rienks, R., Zwiers, J., Reidsma, D.: Online and off-line visualization of meeting information and meeting support. Visual Comput. 22, 965–976 (2006)
8. Matena, L., Jaimes, A., Popescu-Belis, A.: Graphical representation of meetings on mobile devices. In: MobileHCI 2008 Demonstrations (10th International Conference on Human-Computer Interaction with Mobile Devices and Services), Amsterdam (2008)
9. Edelsky, C.: Who's got the floor? Language in Society 10, 383–421 (1981)
10. Hayashi, R.: Floor structure of english and japanese conversation. Journal of Pragmatics 16, 1–30 (1991)
11. Aoki, P., Romaine, M., Szymanski, M., Thornton, J., Wilson, D., Woodruff, A.: The mad hatter's cocktail party: A social mobile audio space supporting multiple conversations. In: Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI 2003), pp. 425–432 (2003)

12. Bettinghaus, E.P., Cody, M.J.: Persuasive Communication. In: Wadsworth Thomson Learning, 5th edn. (1994)
13. Clark, H.H., Schaefer, F.E.: Dealing with overhearers. In: Arenas of language use. University of Chicago Press, Chicago (1992)
14. Monk, A., Watts, L.: Peripheral participation in video-mediated communication. Int. J. Human-Computer Studies 52, 933–958 (2000)
15. Fogg, B., Tseng, H.: The elements of computer credibility. In: Proceeding of CHI 1999, pp. 80–87 (1999)
16. Yankelovich, N., Kaplan, J., Simpson, N., Provino, J.: Porta-person: telepresence for the connected meeting room. In: Proceedings of CHI 2007, pp. 2789–2794 (2007)
17. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735 (1974)
18. Donath, J.S.: Mediated faces. In: Beynon, M., Nehaniv, C.L., Dautenhahn, K. (eds.) CT 2001. LNCS, vol. 2117, pp. 373–390. Springer, Heidelberg (2001)
19. Hain, T., El Hannani, A., Wrigley, S.N., Wan, V.: Automatic speech recognition for scientific purposes - webasr. In: Proceedings of the international conference on spoken language processing (Interspeech 2008) (2008)
20. Op den Akker, H., Schulz, C.: Exploring features and classifiers for dialogue act segmentation. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 196–207. Springer, Heidelberg (2008)
21. Germesin, S., Becker, T., Poller, P.: Determining latency for on-line dialog act classification. In: Poster Session for the 5th International Workshop on Machine Learning for Multimodal Interaction, vol. 5237 (2008)
22. Szöke, I., Schwarz, P., Burget, L., Fapšo, M., Karafiát, M., Černocký, J., Matějka, P.: Comparison of keyword spotting approaches for informal continuous speech. In: Interspeech 2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, pp. 633–636 (2005)
23. Ba, S., Odobez, J.M.: Recognizing human visual focus of attention from head pose in meetings. IEEE Transaction on Systems, Man, and Cybernetics, Part B (Trans. SMC-B) 39, 16–33 (2009)
24. Op den Akker, H.: On addressee prediction for remote hybrid meeting settings. Master's thesis, University of Twente (2009)
25. Hauber, J., Regenbrecht, H., Billinghurst, M., Cockburn, A.: Spatiality in videoconferencing: Trade-offs between efficiency and social presence. In: Proceedings ACM Conference on computer-supported cooperative work, CSCW 2006 (November 2004)
26. Baltes, B.B., Dickson, M.W., Sherman, M.P., Bauer, C.C., LaGanke, J.S.: Computer-mediated communication and group decision making: A meta-analysis. Organizational Behavior and Human Decision Processes 87(1), 156–179 (2002)
27. Popescu-Belis, A., Stiefelhagen, R. (eds.): MLMI 2008. LNCS, vol. 5237. Springer, Heidelberg (2008)
28. Takemae, Y., Otsuka, K., Yamato, J., Ozawa, S.: The subjective evaluation experiments on an automatic video editing system using vision-based head tracking for multiparty conversations. IEEJ Transactions on Electronics, Information and Systems 126(4), 435–442 (2006)
29. Takemae, Y., Otsuka, K., Yamato, J.: Effects of automatic video editing system using stereo-based head tracking for archiving meetings. In: IEEE International Conference on Multimedia and Expo., ICME 2005, July 2005, pp. 185–188 (2005)

30. Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M., Yamato, J.: A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In: Proceedings of ICMI 2008 - International Conference on Multimodal Interfaces (2008)
31. Rienks, R., Nijholt, A., Barthelmess, P.: Pro-active meeting assistants: Attention please? In: Proceedings of the 5th workshop on Social Intelligence Design (2006)
32. Ehlen, P., Fernandez, R., Frampton, M.: Designing and evaluating meeting assistants, keeping humans in mind. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 309–314. Springer, Heidelberg (2008)
33. Popescu-Belis, A., Boertjes, E., Kilgour, J., Poller, P., Castronovo, S., Wilson, T., Jaimes, A., Carletta, J.: The amida content linking device: Just-in-time document retrieval in meetings. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 272–283. Springer, Heidelberg (2008)