Gathering a corpus of multimodal computer-mediated meetings with focus on text and audio interaction

Saturnino Luz[*], **Matt-Mouley Bouamrane**[*], **Masood Masoodian**[†]

[*]Department of Computer Science
Trinity College Dublin
Dublin, Ireland
{luzs,bouamram}@cs.tcd.ie

[†]Department of Computer Science
The University of Waikato
Hamilton, New Zealand
m.masoodian@cs.waikato.ac.nz

Abstract

In this paper we describe the gathering of a corpus of synchronised speech and text interaction over the network. The data collection scenarios characterise audio meetings with a significant textual component. Unlike existing meeting corpora, the corpus described in this paper emphasises temporal relationships between speech and text media streams. This is achieved through detailed logging and time stamping of text editing operations, actions on shared user interface widgets and gesturing, as well as generation of speech activity profiles. A set of tools has been developed specifically for these purposes which can be used as a data collection platform for the development of meeting browsers. The data gathered to date consists of nearly 30 hours of recorded audio and time stamped editing operations and gestures.

1. Introduction

The study of multimodal meetings currently attracts considerable interest among a wide variety of research communities as it embodies a number of challenges in the fields of groupware, collaborative computing, human-computer interaction, speech and video technologies, and multimedia indexing and retrieval. Despite this high level of interest, corpora of multimodal meetings are not easily available and might not always meet one's requirements. As a result, many researchers have developed their own corpora, tailored to their particular needs. The corpus described in this paper was collected in order to meet one such particular need. However, the corpus and tools used in collecting it were designed so that they can support detailed investigation of a range of phenomena, thus constituting potentially useful resources for the research community.

The primary target for investigation is a set of interaction modalities employed by meeting participants during remote, computer-mediated meetings. These include linguistic modalities (text and speech) but also, crucially in the context of our project, gesturing, pointing and semantic annotation. The assumption behind recording non-linguistic interaction is that this information, which is usually overlooked or lost in common multimedia recording settings, might prove valuable for meeting indexing and information extraction from archived meetings.

This paper is structured as follows. In section 2., we briefly survey recent research on meeting browsing and discuss related work. We then analyse requirements from the perspective of our storage and retrieval model and describe the design of the COWRAT corpus. In section 4. we describe the equipment and software tools used for gathering the corpus. Sections 5. and 6. details the data collection scenarios and the current contents of the corpus. The paper closes with a brief discussion of the role of the corpus in the study of multimedia meetings and in the design of interfaces for browsing and information retrieval, followed by a statement of planned future work.

2. Application background

The development of multimodal meeting browsers is a relatively new area. Increasing volumes of recorded multimedia meeting data are driving the need for efficient tools to quickly access and retrieve important pieces of meeting information. This is a challenging task as the continuous (time-based) components of multimedia recordings (audio and video) lack obvious structure, and salient parts of the data are hard to identify. In scenarios where meetings produce space-based artifacts, such as text documents, drawings and tables, much of the information about the decision making process is contained in the continuous, audio-visual medium. Common approaches to meeting browsing include indexing according to features extracted from continuous media (speakers identity, keyframes etc), and the use of modality translation, particularly automatic speech recognition (ASR), to generate meeting transcripts and summaries (Waibel et al., 2001; Tucker and Whittaker, 2005). Corpora used in these approaches, such as the ICSI Meeting Corpus (Morgan et al., 2001), are tailored to the particular challenges posed by feature extraction and ASR, usually consisting of high quality audio recordings and word-level orthographic transcriptions.

While these systems have attained some level of success, the meeting browsing problem is far from solved. Errors introduced by speech recognition and the lack of intuitive visualisation interfaces make it difficult for the user to contextualise transcripts. An aspect of the problem that is often neglected is the relationships between timing patterns in speech (e.g. speech turns) and non-verbal actions (e.g. pointing, editing, drawing). In (Bouamrane et al., 2006) we presented an approach to meeting browsing that builds on a multimedia retrieval model which targets such aspects of participant interaction (Luz and Masoodian, 2005). Since these aspects focus less on exchanged content and more on interaction, corpus design requirements of our approach to meeting browsing differ from those of standard meeting corpora. Identification of group interactions has also been investigated by (McCowan et al., 2005). Their work is based on extracting low-level audio and visual features from audio recordings and high-level modelling and recognition of a set of specific meeting events using Hidden Markov Models (HMM). Our approach differs significantly from the one adopted in (McCowan et al., 2005) in that it relies on automatic low level metadata generation *during meeting capture* and higher level post-meeting processing using these metadata. In the following section, we specify the nature of the data and metadata gathered in the context of our project.

3. Requirements and corpus design

The meeting scenario we have targeted is one where a group of collaborators (typically two) synchronously write a text document which reflects the results of an oral discussion held simultaneously with the collaborative writing activity. Examples of such scenarios include collaborative writing of minutes, joint preparation of articles, work plans etc.

The underlying information retrieval model we aim to test using the corpus is based on the idea of establishing *temporal and contextual neighbourhoods* for each speech and text segment (Luz and Masoodian, 2005). Briefly, a temporal neighbourhood describes text and speech segments recursively linked to a target segment through partial concurrence. A contextual neighbourhood describes links in terms of co-occurrence of keywords. Temporal and contextual patterns are common in collaborative meetings and characteristic of collaborative writing. In order to investigate those patterns, we require consistent annotation of the recorded data with the following types of metadata elements:

- basic *segmentation* elements to establish text and speech units,
- *time stamps* to keep track of actions (write, delete, copy etc) performed by each participant on each text segment of the resulting text,
- detailed *action descriptions*, including actions performed on segments deleted from the final text, and
- user-defined *keywords* with which participants can highlight text they consider relevant.

These metadata elements are encoded in XML in the COWRAT corpus. Figure 1 shows a simplified version of the document type definition (DTD) used for text encoding. The time stamped XML document is synchronised with a *audio profile*. Audio profiling at the moment serves the single purpose of distinguishing silence and speech-filled time intervals. At first, audio profiles are not directly connected to text metadata. The two media are connected as a result of a post-processing stage that generates a Multimodal Activity Matrix (MAM). An MAM is a Boolean matrix which summarises multimodal activity over time: rows represent discrete time units (1 second, for simplicity) and columns represent participants' interaction channels (speech and text) so that the total number of columns is the number of communication channels (two, for the meetings collected so far) times the number of meeting participants. Such matrices provide a convenient way of visualising meeting activity and have been explored in several applications (Bouamrane et al., 2006).

```
<!ELEMENT comapdoc (meeting|section|segment|actions)*>

<!-- Meeting metadata: venue, description etc -->
<!ELEMENT meeting (venue,description,partlist)>
<!ATTLIST meeting date CDATA #REQUIRED>
<!ELEMENT venue (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT partlist (participant+)>
<!ELEMENT participant (#PCDATA)>
<!ATTLIST participant id CDATA #REQUIRED>

<!-- Meeting data: some basic structure -->
<!ELEMENT header (#PCDATA | timestamp)*>
<!ATTLIST header
          level CDATA #IMPLIED>
<!ELEMENT segment (#PCDATA|keyword|header|timestamp)*>
<!ATTLIST segment
          id CDATA #REQUIRED>
<!ELEMENT section (segment+)>
<!ATTLIST section
          level CDATA #IMPLIED>

<!-- timestamp tag; It can be further constrained -->
<!-- so that actionids match action id values -->
<!ELEMENT timestamp EMPTY>
<!ATTLIST timestamp agent CDATA #IMPLIED
                    action CDATA #IMPLIED
                    actionid CDATA #REQUIRED
                    start CDATA #REQUIRED
                    end CDATA #REQUIRED>

<!ELEMENT actions (action*)>
<!ELEMENT action (#PCDATA)>
<!ATTLIST action id CDATA #REQUIRED
                 type CDATA #REQUIRED
                 startT CDATA #REQUIRED
                 endT CDATA #REQUIRED
                 points CDATA #IMPLIED>

<!ELEMENT keywordTypes (keywordType*)>
<!ELEMENT keywordType #EMPTY>
<!ATTLIST keywordType id CDATA #REQUIRED
                      name CDATA #REQUIRED>
<!ELEMENT keyword #EMPTY>
<!ATTLIST keywordType id NMTOKEN #REQUIRED
```

Figure 1: Document type definition for collaborative text

We assume text segmentation units to be paragraphs or, more precisely, strings separated by two consecutive newline characters. Metadata are recorded at the level of segments, as illustrated in Figures 2 and 3. Our aim is to keep track of these segmentation units from their creation, following their evolution during the collaborative writing process. The time stamping model employed for corpus collection handles common editing operations such as insertions, cutting, pasting and deletions in order to coherently manage paragraph timestamps sets in face of structural change to the document (Bouamrane et al., 2005). This mechanism also records gestures (pointing and transient freehand drawing) generated through a telepointer.

Time stamps are associated with each segment, as shown in Figure 2. This limits their reach to text segments that appear in the final version of the document. Association is by content rather than position in the text. Once a set of timestamps is associated with a segment it will follow that segment if the segment is moved (cut and pasted) to a different location in the document, or disappear if the segment is permanently deleted.

However, information pertaining to deleted text segments is not entirely lost. The action description mechanism (`actions` list and `action` tags) records each action performed in the course of a collaborative writing session. Actions are linked to time stamps of surviving segments and can be used to retrace the document construction process if necessary. Unlike ordinary `timestamps` action elements can span several segments, identified through the `paragraphs` attribute. Action description tags also encode relative position of the segment(s) on which the action was effected though the `startOffset` attribute. In addition to the above, actions of type *gesture* contain the precise coordinates of telepointer movements over the text. Unlike time stamp information, offset and point coordinates only make sense in the context (document state) in which they were performed. The usefulness of an action element is therefore conditioned to a sequential replay of all action elements up to the element in question. An example of action description elements is shown in Figure 3.

4. The data collection environment

The COWRAT corpus was collected through a set of tools specifically implemented for this purpose. These tools comprise a shared audio and text environment for remote collaboration using desktop computers as well as a recording server and post-processing tools. Since in this environment all user communication and actions are computer-mediated, interaction data and metadata can be easily collected. Our main goals in designing these collaborative tools were to meet the corpus design requirements described above while providing a usable shared workspace users might conceivably use in real-world situations. A usability study was conducted which showed that users indeed found working on the shared workspace to be natural and straightforward (Masoodian et al., 2005).

```
<segment id=’4.1’>
  <timestamp actionid=’17’ agent=’2’
             action=’nlinsert’ start=’215’ end=’215’/>
  <timestamp actionid=’19’ agent=’2’
             action=’Insert’ start=’215’ end=’217’/>
  <timestamp actionid=’20’ agent=’2’
             action=’Delete’ start=’220’ end=’222’/>
  <timestamp actionid=’21’ agent=’2’
             action=’Insert’ start=’221’ end=’221’/>
  <timestamp actionid=’22’ agent=’2’
             action=’Insert’ start=’222’ end=’226’/>
  <timestamp actionid=’24’ agent=’1’
             action=’Insert’ start=’231’ end=’231’/>
  <timestamp actionid=’57’ agent=’2’
             action=’Insert’ start=’486’ end=’495’/>
    budget of 3000 from the student union
</segment>
<segment id=’4.1.1’>
  <timestamp actionid=’58’ agent=’2’
             action=’Insert’ start=’496’ end=’498’/>
  <timestamp actionid=’59’ agent=’2’
             action=’Delete’ start=’498’ end=’499’/>
  <timestamp actionid=’60’ agent=’2’
             action=’Insert’ start=’499’ end=’502’/>
  <timestamp actionid=’61’ agent=’2’
             action=’Delete’ start=’503’ end=’503’/>
  <timestamp actionid=’62’ agent=’2’
             action=’Insert’ start=’504’ end=’505’/>
  <timestamp actionid=”63” agent=”2”
             action=”Gesture” start=”510” end=”518” />
    maybe charge people more?
</segment>
```

Figure 2: Paragraph level timestamp management

```
<action id=”57” type=”Insert” startT=”486” endT=”495”
        paragraphs=”4.1,4.1.1” startOffset=”15”>
   from the student union

</action>
<action id=”58” type=”Insert” startT=”496” endT=”498”
        paragraphs=”4.1.1” startOffset=”0”>
   maybe chga
</action>
<action id=”59” type=”Delete” startT=”498” endT=”499”
        paragraphs=”4.1.1” startOffset=”8” endOffset=”9”>
   ga
</action>
<action id=”60” type=”Insert” startT=”499” endT=”502”
        paragraphs=”4.1.1” startOffset=”8”>
   arge people
</action>
<action id=”175” type=”Gesture” startT=”1279” endT=”1283”
        paragraphs=”16.4” startPar=”16.4”
        points=”(233,17),(222,14),(274,17),(233,17)”>

   Booking
</action>
<action id=”176” type=”Insert” startT=”1286” endT=”1298”
        paragraphs=”16.3,16.3.1” startOffset=”40”>
      by the student union)

</action>
```

Figure 3: Action description timestamp

Collaborative document writing is done with RECOLED, a REcording COLlaborative EDitor (Bouamrane et al., 2004; Masoodian et al., 2005). RECOLED implements real-time editing functionality comparable to that found in simple text editors. At the moment, RECOLED’s text structuring and formatting capabilities are limited to basic sectioning. The shared document is replicated on each participant’s site, with updates performed locally, in order to preserve low response time, and then broadcast to the various collaborators. All editing and gesturing operations are captured locally and transparently incorporated to the structure of the document stored in the recording server. Gestures form an important part of RECOLED’s awareness feedback mechanisms and are naturally integrated into the shared environment, as recommended in the computer supported cooperative work literature (Dourish and Bellotti, 1992; Gutwin and Greenberg, 1999), serving therefore two complementary purposes: improvement of the system’s usability and tracking of the user’s focus of interest in a non-intrusive way (Masoodian et al., 2005).

Meeting participants also communicate through speech with the aid of a multicast audio tool. Audio communication is mediated through the Real Time Protocol (RTP). A server (RECPLAJ) records RTP data and control packets exchanged during the meeting. Recorded RTP control packets contain reports of reception quality as well

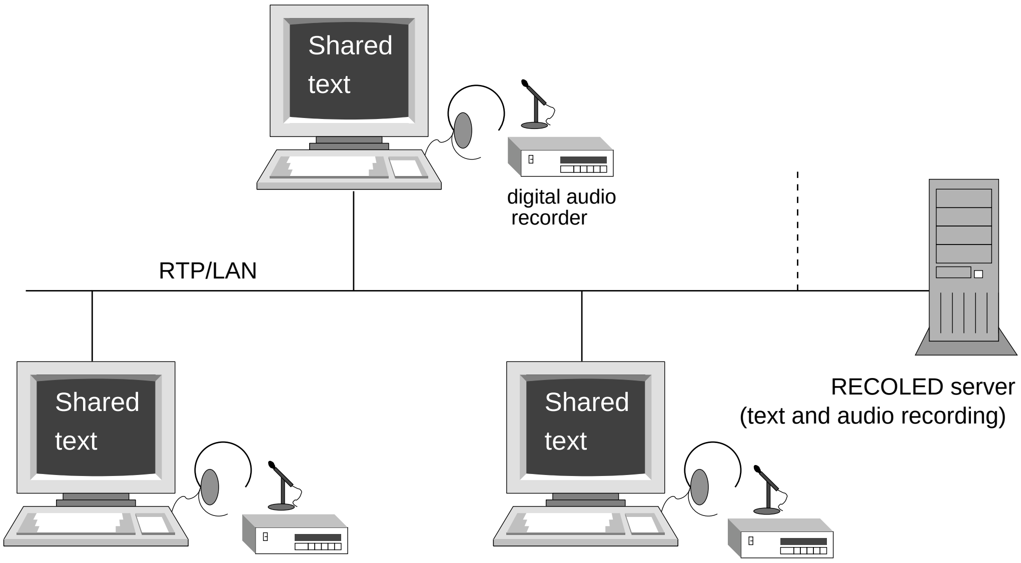

Figure 4: Meeting Recording Architecture

as various conferencing events. RTP data packets include packet header information such as synchronisation source (participant) identifiers and contributors (in the case of mixed-source packets), sampling time stamps, packet sequence number and payload type (Schulzrine et al., 1999). Speech profiles are derived directly from such recorded packets through monitoring of channel activity (packet arrival rate). A speech profile is extracted for each participant and merged with the summary of text time stamps into an MAM. An application programming interface has been specified and a reference library (REXPLORE) has been implemented which unify the handling of speech profiles, MAMs and XML-encoded text.

For the COWRAT corpus we set the audio format used for transmission over RTP to GSM encoding at a sampling rate of 8 KHz. Although this setting ensures reasonable audio quality over the network, the resulting recordings are inadequate for applications that demand higher quality audio, such as automatic speech recognition. Therefore, we have recently begun to complement RTP recording with local recording of high quality audio. Each participant is provided with a portable digital recorder and clip microphone. Local recordings use a sampling rate of 44Khz and a bit rate of 320 Kbit/s and are later synchronised with the system's audio and text tracks.

The overall meeting and data collection architecture is shown in Figure 4. The data collection tools and library, including RECOLED, RECPLAJ and REXPLORE have been released as free software. Further information can be found in the project's web site (COWRAT, 2006).

5. Data collection sessions

The data collection (recording) sessions have been conducted in two separate usability laboratories, at the University of Waikato, New-Zealand and Trinity College Dublin, Ireland. The meeting recording setup and procedures adopted in each laboratory are similar. Participants cannot see each other and can only communicate via the shared editor and audio conferencing tool. Participants wear noise-cancelling headsets and the networked audio tool is set to perform automatic silence suppression so as to minimise transmission of noise over the RTP channel. As mentioned above, for the last couple of data collection sessions, users were fitted with individual digital audio recorders and additional clip microphones in order to complement RTP audio with a high quality audio track. A passive observer is present at the sessions and takes general notes, intervening only in case a technical problem arises.

In the meetings collected so far, participants were familiar with the use of text editors, but had not previously used a shared editor. Therefore, each session began with a demonstration of the software. The participants were then encouraged to explore its functionality and to ask questions they might have as to how the system operates. Once the participants indicated that they were confident in using the system, they were required to perform a specific task and the meeting was recorded. No specific time limit was imposed but the meetings were generally between half-an-hour and an hour long, depending on the tasks.

6. Corpus description

A corpus of twenty nine meetings has been collected to date. These meetings have been organised into three distinct sets according to the type of tasks the participants were asked to perform. The main kinds of tasks comprised:

(A) reordering an existing text,

(B) organising a weekend break, and

(C) discussing a research project.

In addition, pilot studies were carried out prior to the implementation of RECOLED in which the developers took part in and recorded various collaborative tasks using the Network Text Editor, NTE (Handley and Crowcroft, 1997), including planning a paper presentation and discussing the syllabus of a course. These pilot meetings have been converted into the standard MAM and XML formats described

Task	no. of meetings	total duration	avg. text length	text actions		gesturing actions	
				total	average	total	average
A	7	295 min.	6635 words	1608	229.7	633	90.4
B	9	224 min.	3918 words	1698	188.7	89	9.9
C	9	412 min.	2690 words	1165	129.4	103	11.4
O	4	128 min.	1422 words	544	136.0	300	75.0
total	29	17 h. 39min.	14665 words	5015	172.9	1125	38.8

Table 1: Corpus composition according to task

in section 3. and incorporated to the corpus. We refer to this set of recordings as set O. The composition of the various sets is summarised in Table 1.

Set A comprises dyadic writing sessions organised as part of a usability study of the collaborative writing environment (Masoodian et al., 2005). A total of fourteen students from the Department of Computer Science at the University of Waikato took part in seven dyadic writing sessions, and received a book voucher each for taking part in the study. The majority of participants were fourth year computer science students, while a few were postgraduate students. All the participants were familiar with the use of text editors, but had not previously used a shared editor. The group task required the subjects to collaboratively work on a simple childrens story in which the paragraphs had been randomly rearranged and some words blanked out. The task essentially consisted of cooperatively arranging these fragmented paragraphs in a logical order and adding in the missing words to recreate the full story. The task was specifically designed to encourage a high level of interaction and communication between the participants, and that is reflected in the extensive use of gesturing observed for this set (Table 1).

Set B contains recordings of meetings in which participants were asked to organise a (fictional) weekend social function for final year university students and staff. This scenario included a number of sub-tasks such as booking a hotel, making travel arrangements, proposing daytime and night-time activities, assessing costs, etc. Sets of handouts containing information relevant to the task were distributed to the participants before the meeting. During the meeting, participants had access to that information, some of which was shared, some of which private. This task was specifically designed to encourage interaction and communication between the participants and to ensure that for certain tasks, each person had to rely on the other participant's critical information. A total of twenty four participants performed this task in dyadic sessions. The majority were postgraduate computer science students, from Trinity College, Ireland, but a few first year undergraduate students also took part in this experiment.

The remaining set (C) consists of recordings of student-supervisor meetings relating to ongoing final (fourth) year students projects in the Department of Computer Science at Trinity College. Meeting collection is ongoing with between one or two meetings being recorded a month. The recording of this set aims at exploring interaction in a real-life collaborative situation and keeping track of the evolution of the collaborative process as users become more accustomed to synchronous remote collaboration in general, and the RECOLED tool in particular.

The corpus is currently undergoing manual annotation which will complement the interaction data gathered by automatic means. Annotation being manually added include full transcription of contents and tagging of dialogue acts. The total duration of the audio recordings in the COWRAT corpus is currently over seventeen hours.

7. The COWRAT corpus and meeting browsing

We have made several observations about the specificity of remote, computer-mediated text and speech meetings, with respect to co-located meetings. These observations suggest the greater importance of acknowledgements and verbal feedback, and the scarcity of speech overlaps in remote meetings. Participants tend not to talk and write at the same time, but generally alternate periods of argumentation and discussion with periods of editing activity, except when gesturing is employed. We have also observed a great degree of semantic relatedness among segments in the same temporal neighbourhood (Luz and Masoodian, 2005) across media streams. Based on the latter, we have devised several meeting browsing interfaces which exploit the idea of temporal neighbourhood to improve meeting browsing and information retrieval (Bouamrane et al., 2006). The COWRAT corpus has been instrumental in supporting the design and subsequent evaluation such interfaces.

Techniques originating from analysis of the corpus include a metric of inter-media activity (Luz, 2002), the concept of non-linear browsing and information visualisation techniques for small screens (Masoodian et al., 2003).

8. Future work

We are currently in the process of improving the existing corpus by complementing its metadata with higher level annotation, and collecting more meeting data.

Higher level metadata currently being added consists mainly of transcripts synchronised with text and audio interaction, but will also incorporate tagging of dialogue acts. The ELAN annotation tool (MPI, 2005) has been used for this purpose. Speech transcription, in particular, will play a role in the investigation of the *contextual neighbourhood* component of the model proposed in (Luz and Masoodian, 2005).

Future meeting recording efforts will focus on larger groups (three to five participants) working on a more diverse range of tasks. Since the existing software does not limit the number of participants, extending the corpus beyond dyadic sessions is straightforward. Diversifying the set of tasks, on the other hand, could involve extending the capabilities of the shared editor to handle simple graphics. We have a particular interest in scenarios involving high degree of object manipulation and gesturing, such as collaborative production of architectural plans.

Acknowledgements

This research was partly supported by an Enterprise Ireland Basic Research Grant. The authors would like to thank all volunteers for their participation in the data collection meetings, and David King for his work on an early version of the RECOLED editor.

9. References

Matt-Mouley Bouamrane, David King, Saturnino Luz, and Masood Masoodian. 2004. A framework for collaborative writing with recording and post-meeting retrieval capabilities. *IEEE Distributed Systems Online*. Special issue on the 6th International Workshop on Collaborative Editing Systems.

Matt-Mouley Bouamrane, Saturnino Luz, Masood Masoodian, and David King. 2005. Supporting remote collaboration through structured activity logging. In Geoffrey C. Fox Hai Zhuge, editor, *Proceedings of 4th International Conference in Grid and Cooperative Computing, GCC 2005, LNCS*, volume 3795 / 2005, pages 1096–1107, Beijing, China, Nov. Springer-Verlag GmbH.

Matt-Mouley Bouamrane, Saturnino Luz, and Masood Masoodian. 2006. History based visual mining of semi-structured audio and text. In *Proceedings of Multimedia Modelling, MMM06*, pages 360–363, Beijing, China, January. IEEE Press.

COWRAT. 2006. http://cowrat.berlios.de/.

Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Procs. of the Conference on Computer-Supported Cooperative Work*, pages 107–114, Toronto. ACM Press.

Carl Gutwin and Saul Greenberg. 1999. The effects of workspace awareness support on the usability of real-time distributed groupware. *ACM Transactions on Computer-Human Interaction*, 6(3):243–281.

Mark Handley and Jon Crowcroft. 1997. Network text editor (NTE): A scalable shared text editor for the MBone. In *Proceedings of the ACM SIGCOMM Conference : Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM-97)*, volume 27,4 of *Computer Communication Review*, pages 197–208, New York, September 14–18. ACM Press.

Saturnino Luz and Masood Masoodian. 2005. A model for meeting content storage and retrieval. In Yi-Ping Phoebe Chen, editor, *11th International Conference on Multi-Media Modeling (MMM 2005)*, pages 392–398, Melbourne, Australia. IEEE Computer Society.

Saturnino Luz. 2002. Interleave factor and multimedia information visualisation. In H. Sharp, P. Chalk, J. LePeuple, and J. Rosbottom, editors, *Proceedings of Human Computer Interaction 2002*, volume 2, pages 142–146, London.

Masood Masoodian, Saturnino Luz, and Cheng Weng. 2003. HANMER: A mobile tool for browsing recorded collaborative meeting contents. In E. Kemp, C. Philip, and W. Wong, editors, *Proceedings of CHI-NZ '03*, pages 87–92, Dunedin, New Zealand. ACM Press.

Masood Masoodian, Saturnino Luz, Matt-Mouley Bouamrane, and David King. 2005. RECOLED: A group-aware collaborative text editor for capturing document history. In *Proceedings of WWW/Internet 2005*, volume 1, pages 323–330, Lisbon.

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, March.

Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. 2001. The meeting project at ICSI. In *Procs. of Human Language Technologies Conference*, San Diego.

MPI. 2005. ELAN: Eucido Linguistic Annotator. Max Planck Institute for Psycholinguistics, March. http://www.mpi.nl/tools/elan.html.

Henning Schulzrine, Stephen Casner, Ron Frederick, and Van Jacobson. 1999. RTP: A transport protocol for real-time applications. IETF Internet Draft draft-ietf-avt-rtp-new-04, February.

Simon Tucker and Steve Whittaker. 2005. Accessing multimodal meeting data: Systems, problems and possibilities. In *MLMI '04: Machine Learning for Multimodal Interaction*, pages 1–11. Springer-Verlag GmbH.

Alex Waibel, Michael Brett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. 2001. Advances in automatic meeting record creation and access. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 597–600. IEEE Press.