Contributors

Saturnino Luz
School of Computer Science and Statistics
Trinity College Dublin
Dublin 2, Ireland
luzs@cs.tcd.ie

1

Restructuring multimodal interaction data for browsing and search

Saturnino Luz

Trinity College Dublin

CONTENTS

This paper presents recent work on techniques for indexing and structuring recordings of interactive activities, such as collaborative editing, computer-mediated and computer-assisted meetings, and presentations. A unifying architecture is presented which encompasses modality translation (e.g. speech recognition and image analysis) as well as the underlying recording under a single linked time-based model. This model is illustrated through case studies and implemented prototypes, which support remote multiparty collaboration and co-located recorded meetings. Requirements and the issues they pose for the design of this kind of systems are discussed.

1.1 Introduction

Consider a remote collaboration between two people who were given the task of organising a final-year weekend trip for staff and students. The trip planners communicated through a dedicated audio conferencing system which included

a real-time audio channel, a text chat channel, and a shared editor on which the participants could write as well as highlight and point to objects such as sentences and paragraphs while talking [41]. During the meeting, the participants used the shared surface as a common note book where they could jot down their thoughts and options considered (or to be considered), record decisions and "action points", share references to external data (such as URLs containing useful resources), and so on. The excerpt in Table 1.1 is an example of a typical text that remained on the shared editor once the meeting was concluded.

TABLE 1.1
Traces of textual interaction on a shared (synchronous) text editor.

t_1.	Organisation of Final-Year/Staff Week-end
t_2.	How much will it cost per person... 124
t_3.	budget of 3000 from the student union
t_4.	maybe charge people more?
t_5.	travel 1500
t_6.	hotel 6120
t_7.	sat night :1375
t_8.	Saturday afternoon (Greyhound racing) - 825
t_9.	total 9820 - 3000 = 6820 / 55 = 124
t_{10}.	Booking Hotel...
t_{11}.	The [X] clube hotel - 240 for double
t_{12}.	4400 for students
	⋮
t_{13}.	**Information about the week-end...** Weekend in Kilkenny City
t_{14}.	Travel by hired bus
t_{15}.	Staying in [Y] Hotel.
t_{16}.	Saturday afternoon's activities: Visit Castle and then go Greyhound racing!
t_{17}.	Saturday night: [Z]'s Steak & Ale House, traditional cuisine and music!
t_{18}.	Total cost- 124 per person (subsidised by the student union)

This excerpt, by itself, can hardly be regarded as a complete record of the discussion, or even the part of it that the text spans. However, while

producing a structured meeting minutes was probably not the intention of the participants, a certain amount of structure remains and can be identified in the text. It seems clear, for instance, that the text up to line t_{12} was used during the decision making phase of the meeting, where options were considered, discussed, and in some cases rejected. From line t_{13} on[1], the text seems to record the final decisions made, costs etc.

In meetings that are recorded for review at a later time, such as the meeting from which the above text has been extracted, there is a possibility of enhancing the textual structure by listening to the recorded audio and watching the video or alternatively, if video is not available, some form of visual playback of relevant actions such as pointing and adding text to the shared surface. Since these complementary modalities are essentially rendered through time-based media, the text can be easily linked to them if the actions that produce and modify it are time stamped. Line t_8 of the text fragment in Table 1.1, for instance, was modified and pointed at while the participants discussed entertainment options for Saturday afternoon. The audio recorded during (or around) these typing and pointing actions contained the dialogue shown in Table 1.2.

Having listened to the audio, a person reviewing the meeting record would arguably gain an understanding of why greyhound racing made it into the programme, how the programme was decided, how much it will cost, what other options were considered (the abbey, which did not appear in the final textual record but was later referred to by B, who stated that a visit to it would have been her preferred option), and so on.

This fairly obvious combination of media for recording and reviewing meetings has been the basis for most approaches to automatic meeting record creation and access found in the literature [46, 58]. In fact, the idea of "anchoring" the content of transient conversations to permanent textual objects has also been suggested by qualitative observational studies in human-computer interaction [43, 60] and exploited in early applications that attempted to structure conversations temporally around documents, from text chat [14] to real-time audio [57, 40]. The timeline has proven a powerful metaphor for access to multimedia records [49], and enhanced methods of meeting browsing usually take the timeline as a basis on which more sophisticated indexing methods based on modality translation can build [18, 47].

However, for all its intuitive appeal, the timeline may lead to inefficient access to content. To illustrate this point, let us take a closer look at the decision to go to greyhound racing on Saturday, and the statement that the total cost for the night would be 1375 (line s_{11}, in the excerpts above). It is not immediately clear from the text and speech transcripts why it is that,

[1] Note, however, that the actual text log for this meeting contains a considerable amount of text between what is identified here as lines t_{12} and t_{13}. This intervening text has been omitted here for clarity, as indicated by the use of the vertical ellipsis symbol. This convention is used throughout this paper for both text and transcribed speech. Textual content is denoted in this paper by $t_1, t_2, \ldots$ and quoted speech is denoted $s_1, s_2, \ldots$

TABLE 1.2
Dialogue produced while line t_8 of Table 1.1 was being referred to through deictic gestures or modified.

s_1.	A: we have hm two options for the Saturday afternoon... so we've got, like, the greyhound racing... or there is kind of this abbey...
s_2.	B: hmm
s_3.	A [reading]: "It is regarded as one of the most interesting Cistercian ruins... offers a unique insight into the lives of the monks because many of its domestic arrangements are still recognisable"...
s_4.	A: it is 10 per person to wander around there
s_5.	B: and how much is the greyhound racing?
s_6.	A: 15 per person
s_7.	B: I wonder, could we give people a choice? or would that complicate matters.
s_8.	A and B: [discuss the options]
	⋮
s_9.	B [returning to the text after a few minutes]: OK
s_{10}.	A: that's gonna cost 825, is that ok?
s_{11}.	B: ok. So Saturday is going to be... 1375

in line s_{11}, B concludes that the cost will be 1375, rather than 825. A closer look at the timestamps (not shown in the transcripts) reveals that changes in the text referring to the greyhound racing overlapped with a segment of the dialogue which in turn overlapped with text modification and pointing events in line t_{17}. These text actions partially overlapped with a discussion about dinner arrangements that took place several minutes earlier. One can hence conclude that the cost of dinner accounts for the difference.

Although there is a clear relationship between the text and the audio streams, sequentiality only reveals this relationship up to a point. Thus, a browsing interface consisting of a timeline representing speech turns annotated with text, whether extracted from segments of the text record that overlap temporally with the speech or text generated through automatic speech recognition would be of limited use in identifying the sort of relationship described in the previous example.

This paper argues that a model that can minimally account for such recursive inter-media relations based on different levels of text and speech (and optionally video) segmentation is needed. It then goes on to propose a model based on temporal links that induce a graph structure on multimedia records

of multiparty communication and collaboration. This model implies a restructuring of such records and often results in the clustering together of segments located far apart in the original linear structure of the data. Instantiations of this model are illustrated through two case studies on information access in multimedia meeting records.

1.2 Recording and analysis of multimodal communication in collaborative tasks

Once restricted to denoting face-to-face interaction, the word "meeting" has, with the spread of information and communication technologies, taken a much broader meaning. It now generally refers to any form of multiparty interaction mediated through video, audio, synchronous text (including collaborative editing, text chat etc) or, as is most often the case, a combination of these communication media. Technological developments have also encouraged recording of meetings for analysis and information retrieval [58, 9]. In both face-to-face (co-located) and remote meetings, it is now possible to capture a wide variety of information exchanged through different communication modalities. This abundance of data, however, needs to be properly indexed if it is to be useful in review and retrieval contexts.

Initial research on automatic indexing of time-based media focused on *modality translation* [52], that is, on translating content originally encoded in transient and sequential form (i.e. audio and video) into a parallel and persistent presentation [30, 51]. Text, key frames, and other static content generated through modality translation may then be used in conjunction with existing retrieval techniques to create a form of content-based indexing. The role time naturally plays in structuring data recorded during meetings, lectures and other such presentations is crucial to these approaches. Thus, content extracted from the audio track through speech recognition, for instance, may provide valuable links into video content. The temporal structure of these data also suggests visualisation and retrieval interfaces that emphasise sequential access, enriching the basic playback metaphor with media and domain specific improvements, such as skimming [3], parsing with compressed data [63], generation of visual summaries [51], text tagging [8], dialogue act annotation [54], and other techniques [46].

1.2.1 Requirements and fieldwork

Observational and ethnographic work has been an influential line of research into the functional requirements for systems to support access to recorded meeting content. Moran et al. [43] looked at how people used audio recordings to review and report on meetings they attended. The researchers defined a

notion of *salvaging* content as "an active process of sense-making" which may encompass browsing and retrieval but goes beyond these activities. Salvaging content involves sorting through the artefacts manipulated and produced during the meeting, and the activities performed by the participants, in order to reassemble them so as to make the meeting record more accessible to "potential consumers" of its content [43].

It is assumed that this salvaging activity is targeted at particular kinds of consumers and therefore may vary considerably in terms of its goals, from producing short summaries (minutes) to explaining the reasons behind the decision making as illustrated in our earlier example. It is also assumed that salvagers would have been aided by tools designed specifically for the purpose of recording and relating activities and artefacts. While the study reported in [43] mainly aimed at studying this salvaging activity as done manually by humans, the prospect of automating the process of identification of key activities, events and artefacts in the multimedia record of a meeting, and allowing the content consumers themselves to specify the goals of the content (re)structuring process has motivated much work on meeting analysis and browsing [9].

Following this line of work, Whittaker et al. [60] conducted ethnographic studies to investigate how meeting participants make records. They categorised such records as *public* or *private*, and identified several shortcomings of existing meeting browsers in supporting the production of both types of records. Shortcomings in support for public records relate to the focus on individual meetings (i.e. the lack of ways to link a meeting record to another) and the mismatch between the output produced by automatic speech recognition and the formality requirements commonly associated to public records such as minutes, which have archival and sometimes legal implications. From the perspective of private records, the authors note that existing browsers fall short in terms of support for extraction of personal actions, focusing instead on low level annotation such as speech turns and key events such as slide changes.

Other qualitative studies confirmed the importance of multimedia records for sense making. Jaimes et al. [25] conducted a survey with people who regularly participate in meetings and grouped their goals when using multimedia (video) records into the following broad categories: verifying what was said by a particular participant, understanding parts that were missed or not understood during the meeting, reexamining contents under a different context, record keeping, and recalling ideas not explicitly discussed. They then conducted a study with 15 participants to determine which facts relating to meetings that had been attended by the participants were more easily remembered and which were more easily forgotten. Perhaps unsurprisingly, items more easily forgotten were dates and times, participants dress, posture and emotional expressions. On the other hand, items such as seat positions, table layout, participant roles and major topics discussed were easily remembered even three weeks after the meeting took place. Based on these goal categories, the au-

thors proposed an interesting framework for retrieval anchored on visual cues anchored on those items they found to be more easily recalled.

1.2.2 Corpora and meeting browsers

Popescu-Belis et al. [46] summarise the findings of a number of studies (observational, questionnaire- or interview-based, and laboratory-based) aimed at eliciting requirements. They highlight the importance of topic lists [4], summaries, and observations of interest which formed part of a proposed *browser evaluation test* [59].

These studies aimed at understanding the goals, tasks and needs of potential users of meeting browsers and happened in connection with research projects focused on meeting technologies. These efforts include the ICSI Meeting Recorder project [27], the European funded AMI/AMIDA projects [48], the ISL Meeting Room project [11], the M4 project [42], VACE-II [12] and the ECOMMET project [29, 37]. These projects have produced a wealth of recorded meeting data and, together with the NIST Meeting Corpus [21], have helped lay the foundations for automatic analysis of meeting contents. In addition to producing corpora, these projects contributed to advancing the state of the art in a number of technologies deemed necessary to satisfy some of the needs identified through fieldwork and user studies.

In order to be able to index recorded audio content according to speaker, for instance, a necessary first step is *speech diarisation* (who said what). Speech diarisation is usually performed through change detection with the Bayesian information criterion [13] followed by clustering of audio feature vectors. The best performing techniques employ Gaussian mixture models as emission probabilities for continuous density Hidden Markov Models [1]. While diarisation can still be rather error prone, specially in noisy environments, great advances have been made in this area [18, 55],

Following the segmentation of the audio stream according to speakers and speech activity performed through diarisation, it is necessary to segment the dialogues into topics. This task has been approached in different ways, and no dominant approach seems to have emerged in the literature. Most approaches to topic segmentation employ a combination of features. Commonly used features are lexical features (or "bags of words") obtained from the output of a speech recogniser, conversational features (lexical cohesion statistics as well as dialogue structure, vocalisation and silence statistics) [20], prosodic features [50], video features [16, 24], and other contextual features such as dialogue type and speaker role [24]. The generation of such features usually pose their own challenges in terms of machine learning and signal processing techniques, and can be performed reliability only to a certain extent. High word error rates for automatic speech recognition, for instance, would hinder the use of lexical features and lexical cohesion statistics, in spite of the fact that topic modelling is resilient to moderate word error rates [24]. An approach to topic

segmentation that avoids speech recognition input altogether is presented in [37] and tested in the ECOMMET and in the AMI corpora [36].

Large vocabulary continuous speech recognition (LVCSR) is nevertheless generally regarded as an essential component of any meeting record indexing system, as a source of input features to topic segmentation as described above, as a source of keywords for annotation of audio segments and video sequences [30, 8, 46], and as general transcripts for browsing and information retrieval [58, 48].

In some cases, summaries of the transcription may be more effective as an aid to meeting browsing [62] than simple annotation by keywords. An addition, other high level functions can arguably only be performed if LVCSR attains a minimum level of accuracy. Stolcke et al. [54], for instance, propose a system that attempts to recognise *dialogue acts* (i.e. to group utterances into classes such as *statements*, *yes-no-question*, *wh-question*, *quotation*, etc) in spontaneous speech. However, the best accuracy achieved by their system was only 65% for automatically recognised words, compared to a chance baseline accuracy of 35%.

Other techniques for recognising high-level events identifiable in meetings, and regarded as useful in the creation of personal records [60] have also been proposed. Hsueh and Moore propose a method for detecting decision points in a discussion [23]. In a similar vein, Dielmann and Renals employ dynamic Bayesian networks to segment and categorise dialogue acts [17], according to the annotation scheme used by the AMI project. Despite relatively good performance in segmentation, classification error for dialogue acts is still quite high for this challenging task.

As regards the visual modality, techniques developed for browsing of video data such as highlight extraction, static key frame selection, shot boundary detection, and other methods that combine various source modalities (e.g. high motion and pitch for detection of discussion activities) [53]. Such combinations have been employed with some success in identifying significant meeting actions. In [42], for instance, a technique that uses low-level audio and visual features (speech activity, energy, pitch and speech rate, face and hand blob) is employed to characterise a meeting as a continuous sequence of high-level meeting group actions (monologue, presentation, discussion, etc) using Hidden Markov Models.

1.2.3 Issues and open questions

Despite the fact that recording a vast range of information from several communication modalities is well within current technological capabilities, that the analytic methods reviewed in this section are improving and that in some cases they have reached a level of maturity which permits their use in practical multimedia browsing systems, these technologies still have a long way to go before all communication modalities can be properly accounted for.

The issues that may arise are exemplified by consideration of the process

of *grounding*, a core phenomenon in human dialogue. Grounding is the process of updating the shared information needed for dialogue participants to coordinate and maintain their communication activity [15]. Sometimes grounding is explicit, as in utterance s_{11} where "ok" signals agreement and confirms that both participants know the cost, which then become part of their *common ground*. Most often, however, the process is managed either (a) through backchannels, as in when B utters "hmm", in s_2 to signal to A that she is listening and has understood his utterance s_1, or (b) implicitly, as in when B simply initiates a relevant next turn in s_5 by asking a question. While these are examples of positive feedback (evidence that one has been heard and understood), negative feedback is also common. The example shown in Table 1.3, which was extracted from the meeting quoted above, illustrates how explicit negative evidence can be presented as part of the grounding process.

TABLE 1.3
Dialogue illustrative of the used of negative evidence in grounding.

s_{12}.	A: should we work out how much that costs, then?
s_{13}.	B: —so far?
s_{14}.	A: yes. [...]
s_{15}.	A: ... so... 24 rooms, that would be... like... 48 people? [pause]
s_{16}.	B: 55?
s_{17}.	A: yeah, 55.

B utters s_{13} in order to indicate that she has not quite understood A's suggestion, and later utters s_{16} to indicate disagreement with A's figure (s_{15}), which is uttered as a question (a "try marker", followed by a short pause) which invites A to confirm or correct the information just given.

A proper treatment of grounding is important for indexing of multimedia records of dialogues and meetings because grounding activities are good indicators of segments of a dialogue where indexable terms (keywords), such as references and verbatim descriptions [15, pp 227–230] are uttered and resolved. The shortcomings of automatic speech transcription for this kind of task are evidenced by the fact that transcription alone will fail to identify patterns such as the try marker-correction exchange in s_{15}-s_{16} or positive confirmation evidence provided by verbal backchannels such as "hmm", "m" etc. While prosody extracted from acoustic features might help restore question marks and other relevant information [34] the best performing approaches are still quite inaccurate. In addition, grounding does not necessarily take the form of vocalisations. Eye contact, gestures (e.g. pointing as a way of identifying an object) and monitoring of facial expressions are all techniques unconsciously employed by interlocutors for maintaining common ground.

1.3 Modelling interaction

The requirements and issues discussed suggest that, in addition to technologies capable of capturing and analysing interaction data such as speech, gestures, facial expressions, text editing etc, one needs to be able to integrate these capabilities under a unified information structure.

The timeline provides a natural structure for interaction data. However, as we have seen, additional structure is needed in order to link activities that are discontinuous on the sequential record but are semantically connected. These connections often have a hierarchical structure, as illustrated by the following fragment of a remote meeting of three participants who are planning a HCI course (see Tables 1.5 and 1.4). The participants collaboratively write a course syllabus on a shared whiteboard. The shared surface of this whiteboard is also the focus of deictic gestures that are visible to all participants and thus serve to identify parts of the text as referents which form part the participants' common ground. The text produced (also part of the participants common ground) includes the excerpt shown in Table 1.4.

TABLE 1.4
Text written on a shared editor as part of a course design meeting.

t_{19}.	course assessment. Project to mirror flow of lectures: – 1) model users/tasks/ human context... paper based interaction design... 2) working prototype... 3) evaluation ... user manual...3 chunks 20 30 15 marks??? group or individual??
t_{20}.	web site resources
	⋮
t_{21}.	3- Designing for people
t_{22}.	4- Modelling users and their tasks
t_{23}.	5- Designing user interfaces
	⋮
t_{24}.	8- Experimental evaluation
t_{25}.	9- User support materials
t_{26}.	10- Advanced topic: e.g. Computer supported Cooperative Work [...]

Line t_{26} was edited, highlighted or pointed at 14 distinct and discontinuous time time intervals. Presumably, the speech exchanged at each of these intervals is potentially related to t_{26}. The first interval, for instance, contains

the exchange s_{18}-s_{19}, between participants C and D (Table 1.5). Utterance s_{18} overlaps in time with the pointing gestures that target t_{21}, t_{22} and t_{23}. The first of these text objects, for instance, is active at another 12 time intervals scattered over the duration of the meeting. They will therefore overlap with other speech segments that will contain, for instance, information about high-level course organisation which is not shown in the text, such as the speech turn s_{20}, in Table 1.5.

TABLE 1.5
Dialogue from the course design meeting.

s_{18}.	C: well, but I mean... specially for someone like me who has done CSCW work it would be nice to have CSCW there, but we also have to consider how much we will be teaching in these parts [points to t_{21}, t_{22} and t_{23}].
s_{19}.	D: I think the for this type of introduction to HCI course it is good to have a pointer to advanced topics [writes on t_{20}]
	⋮
s_{20}.	D: this group here [highlights t_{22} and t_{23}] deals with a kind of contextual and human side of things. These three lectures are hmm... kind of a chunk as are these two [points at two other lines of text].

These kinds of linkage structures are common in multimodal interaction and can be exploited for information browsing and retrieval. In [39] a basic model was introduced which related chunks of text to speech segments at two elementary levels: temporal proximity and co-occurrence of key words. This model, however, does not extend beyond timestamping and basic LVCSR output, and therefore cannot accommodate richer data sources and automatically extracted features. In the following section, an alternative is presented which arguably captures the above discussed requirements in a more comprehensive manner.

1.3.1 A linked time-based model

The elementary entities in the framework proposed here are (a) interaction *events*, (b) content *segments* and (c) time *intervals*.

Interaction events are discernible occurrences that alter the recorded content. The most common events of interest for multimedia indexing are actions performed by the people participating in the recorded interaction. Therefore one might refer to writing events, speech events, gesture events etc. These are the only types of events to be distinguished in this framework. Of course not all events are (or result from) actions performed by the participants. A

fire alarm may go off during the meeting, for example. Such types of events will only be addressed indirectly via actions performed by the participants in response to the event (e.g. a participant shouts "Fire!").

Segments are sets of results (traces) of interaction events which contain indexable information. At the most basic level, the minimally indexable content of an interaction event is the record of the very fact that it occurred (e.g. participant A was speaking at a certain point in time). Segments may consist of permanent and predominantly static content or transient and predominantly dynamic content. The first group includes textual and graphic segments such as words, sentences, paragraphs, icons, images, drawings etc. The second group includes vocalisations, sounds, gestures, video sequences etc. Super-segments can often be identified by means of clustering and classification methods. Thus in text and speech one can attempt to segment according to topics [50, 24, 5, 22], tasks [37] activities [16] etc. Speech can also be segmented according vocalisations to turns, speech act sequence patterns [61] an so on.

Time intervals can be regarded as anchor points to which information extracted from different media sources are linked through the interaction events that produced that information. Of particular interest are the following temporal logic relations [2] between the time intervals associated with two segments: *equal* (e), *meets* (m), *overlaps* (o), *during* (d), *starts* (s), and *finishes* (f). These relations are depicted in Figure 1.1. They are the subset of the possible 13 basic interval relations (and 2^{13} possible indefinite interval relations [33]) that encompass continuity and concurrency. We take these to be the basis for relating the contents of two segments. Note that the segments in question may come from any of the media source.

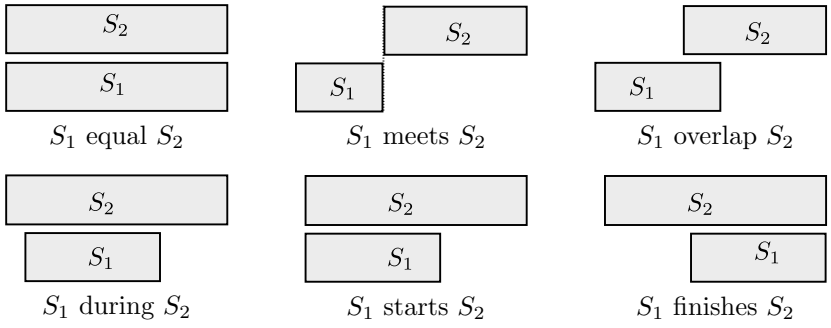

FIGURE 1.1
Time relations of co-occurrence between segments S_1 and S_2 spanning the time intervals represented by the respective rectangles.

Content segments can be linked to time intervals by means of *timestamps*. A timestamp is an annotation of a time interval $[a, b]$ on a content segment. We will represent the set of timestamps of a segment by means of a function $\tau : \mathcal{S} \rightarrow \wp(\mathbb{R}^2)$, so that $\tau(S_i)$ is the set of all intervals in which events related

to $S_i \in \mathcal{S}$ took place (e.g. the intervals in which any participant points at a paragraph of text). A segment made up exclusively of transient events such as speech or gestures has a single timestamp corresponding to the duration of the event(s) that produced the segment. Temporal links between segments can then be defined as follows:

Definition 1 Temporal link*: a segment S_1 is temporally linked to a segment S_2, written $S_1 L S_2$, if there is at least one pair of intervals $i_1 \in \tau(S_i)$ and $i_2 \in \tau(S_2)$ such that $i_1 R i_2$ where R is one of the temporal relations listed above ($e, m, o, d, s,$ or f) or m_1, the inverse relation of m. R is symmetric and irreflexive.*

Given a strategy for individuating segments and the above defined link relation one can proceed to define a meta-structure for a multimedia record of an interactive activity as a graph.

Definition 2 *A* temporal interaction graph *is a tuple $(\mathcal{S}, \mathcal{L})$ where $\mathcal{S}$ is a set of segments extracted from a multimedia record and $\mathcal{L}$ is a set of pairs (S_i, S_j) such that $S_i, S_j \in \mathcal{S}$ and $S_i L S_j$.*

The definitions above do not specify what constitutes a segment or how specifically those segments are time stamped. Our aim here is simply to set out a general framework into which different segmentation and timing techniques can be incorporated and tested. A simple example should help illustrate this point. The setting is the one in which the course planning example was recorded. Participants were remotely located and communicated through an networked audio tool and a shared (virtual) editor/whiteboard. Speech segments were individuated through simple adaptive average-energy analysis of separate streams of the speech signal (each participant wore a lapel microphone, which generated individual speech streams and obviated the need for speaker diarisation [7]). Straightforward synchronisation ensured that each segment was properly time stamped with respect to text events. Text segments were defined as chunks of text enclosed by paragraph-marking boundaries (similar to "boxes" in TEX [32]) and time stamped through monitoring of the participants' activities on the shared editor. Figure 1.2 shows the a temporal interaction graph constructed from the recorded interaction data for text chunk t_{20} of the course syllabus discussion fragment presented above (Table 1.4. Note the connection with the text labelled t_{19} in Table 1.4 as well as the connection to speech turn s_{19}, from Table 1.5.

One can promptly observe that t_{19} is linked to a number of speech segments both directly (adjacently in the graph) and indirectly (connected by a path of more than one edge). Since it extends beyond the node neighbourhood of the segment under consideration (in this case text segment t_{19}), the temporal interaction graph can potentially capture semantic relationships between segments that are both spatially (as in the placement of text and graphics on a page, slide or whiteboard) and temporally discontiguous, and

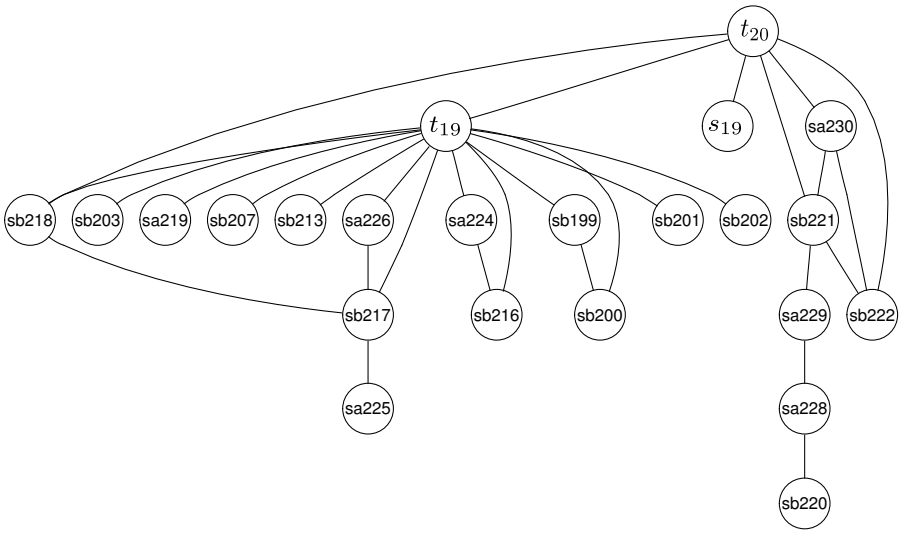

FIGURE 1.2
Temporal interaction graph for segment t_{20}, from the example given in Table 1.4, showing its connections to t_{19} and s_{19} as well as other (arbitrarily labelled) segments.

thus would appear to an observer who scanned the recording sequentially (e.g. by playing back a video) to be semantically unrelated. This underlying structure may therefore help to produce more useful summaries than the current transcription-summarisation approaches adopted for instance by "meeting browsers structured around main topics [...] or action items" [46]. Summarisation techniques that gather facts scattered in the time-based media in order to reassemble the relevant discussions, rationales, counter-arguments behind a decision or action item would come closer to automating the activity of content salvaging which, as we have seen, meets important requirements of the tasks of reviewing and reporting based on multimedia meeting records [43]. In addition to browsing, the graph structure may also be useful for retrieval. Algorithms for extraction of keywords related to a target segment [8, 56] could, for instance, extend the set of candidate words to include words in each of the segments reachable from the target in its temporal interaction subgraph and use link analysis to rank these words [31]. The temporal link structure may also be naturally complemented by spatial context analysis for records that include still images or video sequences [44, 45].

In the following section we briefly examine two examples of meeting recording and browsing activities and the technologies that support these activities in relation to the above described model.

1.4 Case Studies

Two different cases of meeting activities are investigated: one based on meetings recorded in our laboratory (including both normal work meetings among members of the research group and scenario-based meetings) using tools specifically implemented for this purpose [8, 7], and another based on "real-world" meetings held regularly at a busy teaching hospital [29].

1.4.1 Recording and browsing artefact-based interaction

The goal of this study was to investigate the semantic relationships between the contents of interaction events on orthogonal modalities (visual and auditory) as delivered by different media. A dedicated software environment was implemented to support remote meetings and enable multiparty real-time audio communication, text interaction through a shared editor, and gestures (pointing, highlighting, circling) performed on the text. The effect of a participant's gestures on his remote collaborator's screen is represented in the system though a colour-coded overlay that remained visible long enough to attract the collaborator's attention [7]. Thus, the multimedia record comprises three types of data: permanent visual data (text), transient auditory data (speech, audio), and transient visual data (gestures).

A total of 31 meetings were recorded. The data contain a basic form of segmentation provided by the recording environment: audio was automatically segmented into talk spurts and text was segmented into paragraph chunks, as mentioned in the preceding section.

Meeting browsers were then built whose design drew on two different approaches to structuring the browsing activity. The first was based on the idea of identifying segments of the record which exhibit the greatest levels of *interleaving* between text and speech streams and within speech streams [35], and highlighting these segments on the timeline [38]. This approach provided higher-level segmentation by clustering together talk spurts and text chunks in the regions of high interleaving but such groupings were restricted to contiguous regions of the recording. User trials of the browsing prototype suggested that users explored the record non-sequentially, alternating between reading and listening to the recorded speech. This observation motivated the development of the second meeting browser [10], which was based on the "neighbourhoods" model presented in [39] for structuring and presenting information.

This segment neighbourhood browser followed the typical design of speech-centric browsers, according to the categorisation of [46], displaying a text component coupled with a timeline on its main window, as shown in Figure 1.3. This coupling of the two main components meant that while the audio was played back (middle panel in Figure 1.3), the actions performed on the text segment as the participants spoke were displayed on the text component

(bottom panel in Figure 1.3). The browser structured these data as a tree which encoded text-speech segment links implicitly defined according to action timestamps. Automatic speech recognition was also performed but the output was deemed too inaccurate to be read as running text. Instead, these imperfect transcripts were time stamped and used to augment the text segments temporally linked to them for the purpose of keyword extraction. The keywords extracted in this manner were used to annotate the timeline (top panel in Figure 1.3) thus providing an index into the audio record. Keyword-based search was also available on the interface, as a separate component.

As can be gleaned from this brief description, the information structuring approach employed by this browser can be readily accommodated by the model described in Section 1.3.1. Definition 1 is general enough to cover both text and speech segments. In the specific case of the neighbourhood browser, the definition could be further specialised to reflect the text-speech linking patterns by making the relation $L = L_t \cup L_s$ a union of a relation $L_t \subseteq \mathcal{T} \times \mathcal{S}$ linking text segments (in $\mathcal{T}$) to speech segments to a relation $L_s \subseteq \mathcal{S} \times \mathcal{T}$ where L_s and L_t are not required to be reflexive. The temporal interaction graph could then be constrained to having a tree structure by defining a partial order on L.

Furthermore, the model could be used to extend that approach in a number of useful ways. First of all, the temporal graph could be used to produce dynamic summaries adapted to different contexts. Selection of a text segment, for instance, would cause the entire subgraph for this segment to be produced and displayed on the user interface. This could take the form of highlights on the text component and on the timeline panel which the user would be able to select for closer inspection, thus updating the subgraph. Keyword extraction and the search function could also be improved by link analysis techniques as mentioned above. Finally, static textual topic summaries could be generated by combining information from the different segments in the temporal interaction graph. Unlike the topic summaries generated by most meeting browsers, summaries based on the model presented here would not be constrained to encompassing only data from adjacent segments.

Similarly, other meeting browsers, both speech- and document-centric [46], can also be modelled in terms of this linked time-based model. There is reason to suppose that the techniques described above can also benefit those browsers in terms of improved adaptation to context, mitigation of the negative effects of speech recognition errors on the browsing task, and production of more effective summaries. Improved adaptation to context would be obtained by providing the user with better support for content salvaging, since there is evidence that switching between text and audio modalities appears to better reflect what users actually do [43, 38] than sequential reading of transcripts or audio listening. Similarly, poor speech recognition output can be compensated for by giving the user prompt access to related (temporally linked) text. Finally, new forms of content summarisation can be implemented by employing the temporal relations as a segment clustering criterion.

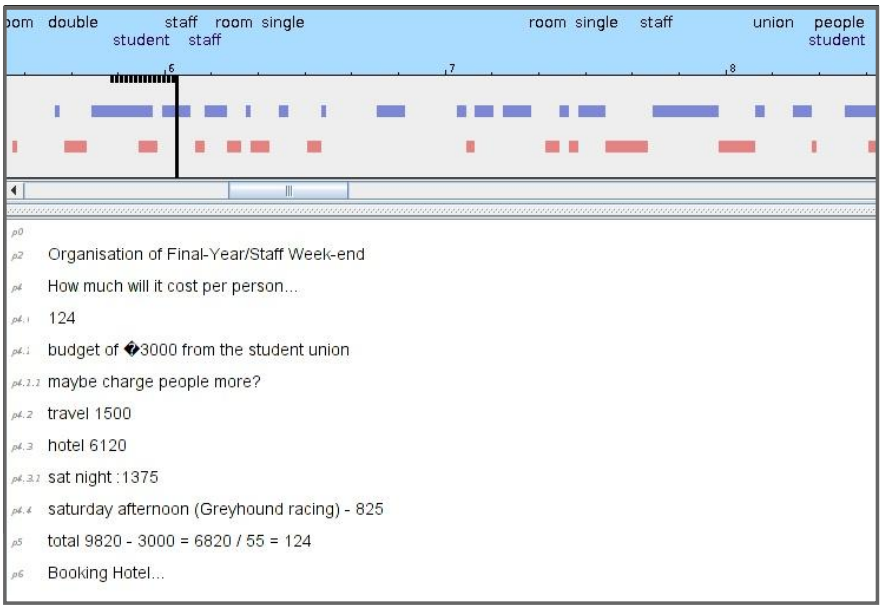

FIGURE 1.3
Main user interface components of a speech-centric meeting browser.

1.4.2 Structuring medical team meetings

Nearly all meeting corpora available to researchers, including the corpus described in Section 1.4.1, have been gathered in laboratory, either under controlled conditions with meeting participants recruited as experiment participants [48, 11] or under naturalistic conditions but in meetings among researchers [27]. While data gathered under such conditions are necessary and useful to researchers since they allow for standardised comparisons among different research efforts on multimedia indexing, machine learning algorithms, natural language processing techniques and other quantitative methods, these data offer little information to guide the use of these methods in realistic application situations.

In order to investigate the production and use of meeting data in a real-world situation we carried out extensive naturalistic observation of a multidisciplinary medical team in their meetings over a period of two years and collected video recordings of their interactions for analysis [29]. Due to the stringent constraints of medical work (time pressures, confidentiality, assurances given to the ethics committee that the research would not interfere with the work of the staff in any way, and so on) the recordings took place under less than ideal conditions. The data were collected from two media sources. The first was an S-VHS recording facility available through the teleconferencing system used by the team, which recorded audio and the screen display being

broadcast to the meeting. The majority of participants were co-located at the main hospital but the teleconferencing equipment was always used for image presentation and display of patient data even when there were no remote participants connected to the system. When remote participants joined the meeting, outgoing video data were captured through the picture-in-a-picture view that was displayed on a TV monitor during the conferences. A second high-end camcorder mounted on a tripod was placed at the back of the room and captured participants' gestures, direction of gaze and activities (e.g. note taking) in addition to audio, through a directional microphone. Two wall-mounted cardioid condenser boundary microphones were also used to capture the speech stream. These media streams were synchronised and annotated with a dedicated media annotation tool.

As in the previous examples of meeting data, interaction and interleave between and within modalities was found to be ubiquitous in medical team meetings. Participants of different specialities (radiologists, pathologists, oncologists, surgeons, nurses) gather in these meetings to present evidence and discuss patient cases. Therefore it is common for some specialities (notably radiologists and pathologists) to augment the description of their findings by images [28, 19]. Others interact with text by taking personal notes or updating the patient sheet which is sometimes visible to the entire group on the teleconferencing system's screen. There is empirical evidence (from studies using laboratory collected corpora) that personal notes can be an effective aid to the creation of meeting summaries [6]. These facts suggest that it makes good sense to gather synchronised data from different modalities also in medical meetings.

Due to the high levels of noise in the speech record, the basic segmentation unit adopted in the analysis of these meetings was not talk spurts as in the previous case, but rather *dialogue states* defined in terms of *vocalisation* events. These states are defined in [37] as follows:

- *Vocalisation*: the event that a speaker "has the floor". A speaker takes the floor when he begins speaking to the exclusion of everyone else and speak uninterruptedly without pause for at least 1 second. The vocalisation ends when a silence, another individual vocalisation or a group vocalisation begins. Talk spurts shorter than 1 second are incorporated into the main speaker's vocalisation.

- *Group vocalisation*: the event that occurs when an individual has fallen silent and two or more individuals are speaking together. The group vocalisation ends when any individual is again speaking alone, or a period of silence begins. Individual speaker identities are lost when a group vocalisation state is entered.

- *Silence*: are periods when no speech is produced for over 0.9 seconds between vocalisations (including group vocalisations). A silence ends when an individual or group vocalisation begins. A silence can be further classified as:

- a *pause*: a silence between two vocalisations by the same participant,
- a *switching pause*: a silence between two vocalisations by different participants,
- a *group pause*: a silence between two group vocalisations, or
- a *group switching pause*: a silence between a group vocalisation and an individual vocalisation.

The temporal interaction graph in this case also included the speaker's (medical) *role* as symbolic information. The structure of these vocalisations augmented with role information has proved quite effective at segmenting medical team meetings into high-level topics. A Bayesian approach based on these structures alone, without transcribed speech, has been shown [37] to match the levels of accuracy obtained by other state-of-the-art topic segmentation algorithms that use transcribed speech (among other features) in their data representations [24].

In brief, this approach consists in defining vocalisation events in terms of neighbouring vocalisations and their speaker labels, as in equation equation (1.1), where V_i is a nominal variable denoting the speaker role (or a pause type or group speech, in the cases of silences and vocalisations by more than one speaker, respectively) and L_i is a continuous variable for the duration of the speech (or silence) interval.

$$s = (V_0, L_0, V_{-1}, L_{-1} \ldots, V_{-n}, L_{-n}, V_1, L_1 \ldots, V_n, L_n) \tag{1.1}$$

Segmentation can then be implemented as a classification learning task to identify instances s to be marked as boundary dialogue states that start new topics. The approach successfully adopted in [36, 37] combined the conditional probabilities for the nominal variables into multinomial models and modelled the continuous variables Gaussian kernels. In the full model, the probabilities to be estimated are simplified through Bayes' rule and the conditional independence assumption to:

$$\begin{aligned} P(b|S=s) &\propto P(V_0 = v_0, \ldots, L_n = l_n|b) \\ &= \prod_{i=1}^{n} P(V_i = v_i|b)P(L_i = l_i|b) \end{aligned} \tag{1.2}$$

where S denotes a random variable ranging over the vector representation of vocalisation events, as defined in equation (1.1). In its simplest form, a classification decision employing a maximum a posteriori criterion will mark as a topic (patient case discussion) boundary all vocalisations s such that $P(b|S=s) \geq P(\neg b|S=s)$.

Another application of these linked time-based data structures is in supporting automatic categorisation of patient case discussions. Patient case discussions are higher level segments of medical team meetings and can be broadly categorised into two groups: *medical* and *surgical* case discussions.

Once such case discussion segments are identified, they can be represented as *vocalisation graphs* through the temporal links of vocalisation event segments. These graphs encode patterns of speech duration and transitions between vocalisation events which can be quantified and normalised for comparison. A typical vocalisation graph is shown in Figure 1.4, where the node labels correspond to the participant's role (e.g. 'sur' for surgeon, 'rad' for radiologist, and so on) or to a general vocalisation event (e.g. 'Group' for group vocalisation, 'Floor' for silence, and so on).

In this representation, each patient case discussion is represented as a directed graph $G = (V, E)$ where V is a set of vertices or nodes and E a binary relation on V. Elements of V are labelled by pairs $(s, p(s))$ representing the probability p_s that the dialogue is in state s (e.g. a vocalisation or a silence) at any given instant. Edges are labelled by conditional probabilities. A probability $p(t|s)$ labelling an edge corresponds to the likelihood that a dialogue state t (the terminal vertex) immediately follows dialogue state s (the initial vertex). Thus, in Figure 1.4, the numbers labelling nodes correspond to the steady state probabilities for those nodes.

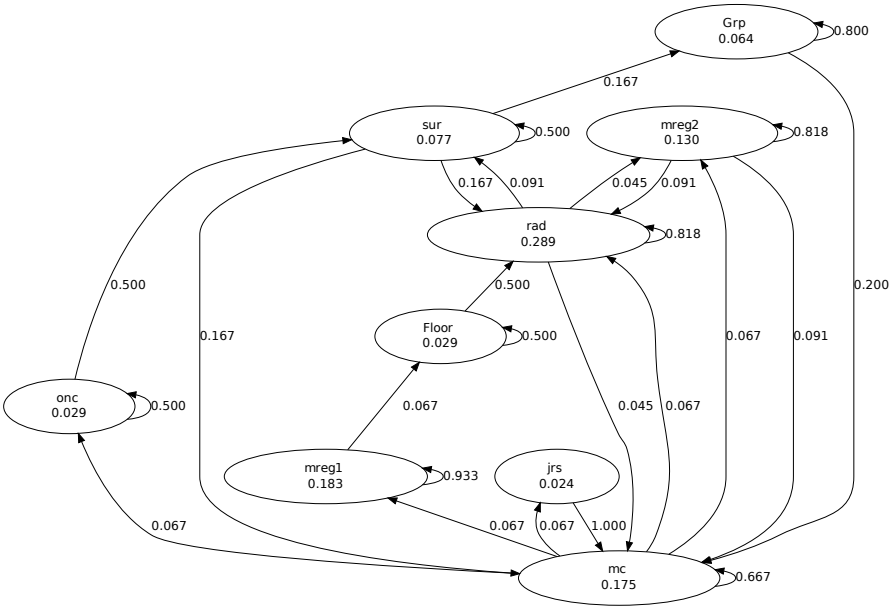

FIGURE 1.4
Vocalisation graph for a patient case discussion extracted from a medical team meeting.

Using this kind of graph structure as representations for case discussion segments in a 'medical' vs. 'surgical' categorisation task we have been able to obtain high categorisation accuracy (98.7%) using a k-NN (k nearest neighbour) classifier. This classifier operated in the usual way. In the "training"

phase all training instances were converted into vocalisation graphs and stored in the database. In the classification phase, given a new instance to classify, the algorithm selected from the database the $k = 5$ instances nearest to the unlabelled instance by graph similarity and took a vote among those instances. The majority class label was then assigned to the test instance.

In summary, the linked time-based model successfully characterises speech interactions at medical team meetings for purposes of topic segmentation and classification of patient case discussions. In principle, the vocalisation structures employed in these tasks could be augmented with gesture, image and text data and further acoustic features extracted from this type of meeting records, given appropriate data capture conditions. The proposed model can accommodate such extensions straightforwardly. Gesture and image events can be characterised as temporally linked segments in the same way as speech and text segments, in terms of Definition 1 and incorporated into an expanded temporal interaction graph. We are currently working on implementing and testing some of these extensions.

1.5 Conclusion

The need to combine complementary information coming from different media and sources in order to meet the challenges of multimedia information retrieval has been generally acknowledged in the literature [26]. Few applications illustrate this need better than those that aim to provide support for browsing and retrieving information from meeting records. The difficulties in these tasks stem not only from the divergent nature of the media and modalities involved, some of which result in parallel and permanent modes of access while others constrain the user to sequential and transient access to the content, but also from the lack of flexibility of existing systems to accommodate diverse sets of requirements.

This paper discussed requirements for such systems against the background of complexity behind the apparent simplicity of real-time human communication. Lurking in this background one finds a variety of grounding mechanisms [15] which often manifest themselves in forms that lie beyond the analytic capabilities of current modality translation technologies, such as the use of silence, backchannels, facial expressions etc to negotiate shared meanings. One also finds subtle temporal and semantic relationships among segments within and between modalities and media as well as uncertainties as to what constitutes a segment in the first place. Building on this discussion, this paper presented a simple model of linked events which seems general enough to accommodate the structures employed by current systems and shows promise with regards to the integration of new data sources and modalities. This can be seen as a step towards a characterisation of multiparty interaction that can

account for the semantic aspects of multimedia records of communication and collaboration.

So far the design of meeting browsing systems has been mainly driven by the capabilities of their modality translation and event analysis modules which are presented to the user through familiar time (in speech-centric systems) or space (in document-centric systems) interface metaphors. This paper suggested that such metaphors may limit the system designer's as well as the user's ability to find richer contextual and semantic relations in the recorded media. This is particularly unsatisfactory given the new trends in data gathering technology. As ubiquitous sensors, processing devices, and methods for event detection, emotion recognition, improved speech processing, facial expression recognition and contextual information fusion begin to offer greater possibilities for creating truly rich records of human-human interaction this design perspective may be ripe for a rethink.

1.6 Glossary

Content salvaging: the activity of sorting through the artefacts manipulated and produced during a meeting and the activities performed by its participants in order to reassemble these contents so as to make the meeting record more accessible.

Grounding: the process of establishing common ground, that is, of establishing a set of mutual knowledge, assumptions and beliefs that underpins communication.

Meeting browsing: the activity of visualising multimedia meeting recordings and finding information of interest in such recordings.

Modality translation: the process of rendering an output modality into another. Examples include rendering speech into text by an automatic speech recognition system, text into speech by a speech synthesiser, video into key-frames etc.

Multimedia meeting record: a digital recording of a meeting consisting minimally of a time-based data stream, typically speech, and a space-based data stream, typically text.

Space-based media: (or *static media*) a class of media for which space is the main structuring element. Data conveyed through space-based media are generally of a permanent and serial nature. Examples include: text and static graphics.

Temporal link: A relation between two media segments (e.g. two vocalisations, a vocalisation and a text segment etc) that meet or co-occur in time.

Temporal interaction graph: A graph that encodes the temporal links of a multiparty interaction.

Time-based media: (or *continuous media*) a class of media for which time is the main structuring element. Data conveyed through time-based media are generally of a transient and parallel nature. Examples include: audio and video.

Vocalisation: a period of talk by a speaker (or group of speakers speaking together).

Vocalisation graph: a graph encoding the structure of a dialogue as transition probabilities from speaker to speaker, including group vocalisations and pauses.

Bibliography

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'03*, pages 411–416, St. Thomas, Virgin Islands, Nov. 2003. IEEE Press.

[2] J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:225–255, July 1984.

[3] B. Arons. SpeechSkimmer: Interactively skimming recorded speech. In *Proceedings of UIST'93: ACM Symposium on User Interface Software Technology*, pages 187–196, Atlanta, GA, Nov. 1993. ACM Press.

[4] S. Banerjee, C. Rose, and A. I. Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction (INTERACT'05)*, pages 643–656, Rome, Italy, Sept. 2005. Springer.

[5] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–348, New York, NY, USA, Sept. 2001. ACM.

[6] A. Bothin and P. Clough. Participants personal note-taking in meetings and its value for automatic meeting summarisation. *Information Technology and Management*, 13(1):39–57, Mar. 2012.

[7] M.-M. Bouamrane, D. King, S. Luz, and M. Masoodian. A framework for collaborative writing with recording and post-meeting retrieval capabilities. *IEEE Distributed Systems Online*, page 6pp, Nov. 2004. Special issue on the 6th International Workshop on Collaborative Editing Systems.

[8] M.-M. Bouamrane and S. Luz. Navigating multimodal meeting recordings with the meeting miner. In G. P. Henrik Legind Larsen, D. Ortiz-Arroyo, T. Andreasen, and H. Christiansen, editors, *Flexible Query Answering Systems: 7th International Conference, FQAS 2006*, volume 4027 of *Lecture Notes in Artificial Intelligence*, pages 356–367, Milan. Italy, June 2006. Springer.

[9] M.-M. Bouamrane and S. Luz. Meeting browsing. *Multimedia Systems*, 12(4–5):439–457, 2007.

[10] M.-M. Bouamrane and S. Luz. Uncovering non-verbal semantic aspects of collaborative meetings: iterative design and evaluation of the meeting miner. *Signal, Image and Video Processing*, 2(4):337–353, 2008.

[11] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Seventh International Conference on Spoken Language Processing (ICSLP)*, pages 301–304, Denver, CO, Sept. 2002.

[12] L. Chen, R. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang. VACE multimodal meeting corpus. In S. Renals and S. Bengio, editors, *Proceedings of Machine Learning for Multimodal Interaction (MLMI)*, volume 3869 of *Lecture Notes in Computer Science*, pages 40–51, Bethesda, MD, May 2006. Springer.

[13] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Herndon, VA, Feb. 1998.

[14] E. F. Churchill, J. Trevor, S. Bly, L. Nelson, and D. Cubranic. Anchored Conversations: chatting in the context of a document. In *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems*, pages 454–461, The Hague, The Netherlands, Apr. 2000. ACM Press.

[15] H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

[16] A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic Bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25–36, 2007.

[17] A. Dielmann and S. Renals. Recognition of dialogue acts in multiparty meetings using a switching dbn. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1303–1314, 2008.

[18] J. G. Fiscus, J. Ajot, and J. S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans*, pages 3–34, Baltimore, MD, May 2008. Springer.

[19] O. Frykholm, A. Lantz, K. Groth, and A. Walldius. Medicine meets engineering in cooperative design of collaborative decision-supportive system. In *Porceedings of the 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 116–121, Perth, Australia, May 2010. IEEE.

[20] M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In E. Hinrichs and D. Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan, July 2003.

[21] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi. The NIST meeting room pilot corpus. In *Proc. 4th Intl. Conf. on Language Resources and Evaluation (LREC)*, pages 1411–1414, Lisbon, Portugal, May 2004. ELRA.

[22] M. A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[23] P.-Y. Hsueh and J. D. Moore. Automatic decision detection in meeting speech. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction (MLMI '07)*, volume 4892 of *Lecture Notes in Computer Science*, pages 168–179. Springer, Brno, Czech Republic, June 2007.

[24] P.-Y. Hsueh and J. D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 1016–1023, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[25] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, CARPE'04, pages 74–85, New York, NY, USA, Oct. 2004. ACM.

[26] R. Jain. Multimedia information retrieval: watershed events. In *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 229–236, Vancouver, BC, Canada, Oct. 2008.

[27] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, volume 1, pages 364–367, Hong Kong, China, Apr. 2003.

[28] B. Kane and S. Luz. Information sharing at multidisciplinary medical team meetings. *Group Decision and Negotiation*, 20:437–464, 2011. Springer.

[29] B. Kane, S. Luz, and S. Jing. Capturing multimodal interaction at medical meetings in a hospital setting: Opportunities and challenges. In *Proceedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pages 140–145, Malta, June 2010. LREC.

[30] R. Kazman, R. Al-Halimi, W. Hunt, and M. Mantey. Four paradigms for indexing video conferences. *IEEE Multimedia*, 3(1):63–73, 1996.

[31] D. Kelleher and S. Luz. Automatic hypertext keyphrase detection. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1608–1609, Edinburgh, Scotland, July 2005. IJCAI.

[32] D. E. Knuth. *The TeXbook*. Addison-Wesley, 1986.

[33] M. Li, Y. Sun, and H. Sheng. Temporal relations in multimedia systems. *Computers & Graphics*, 21(3):315–320, 1997. Computer Graphics in China.

[34] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540, Sept. 2006.

[35] S. Luz. 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, Sept. 2002.

[36] S. Luz. Locating case discussion segments in recorded medical team meetings. In *Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09)*, pages 21–30, Beijing, China, Oct. 2009. ACM Press.

[37] S. Luz. The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Transactions on Information Systems*, 30(3):article 17, Aug. 2012.

[38] S. Luz and M. Masoodian. A mobile system for non-linear access to time-based data. In *Proceedings of Advanced Visual Interfaces AVI'04*, pages 454–457. ACM Press, May 2004.

[39] S. Luz and M. Masoodian. A model for meeting content storage and retrieval. In Y.-P. P. Chen, editor, *11th International Conference on Multi-Media Modeling (MMM 2005)*, pages 392–398, Melbourne, Australia, Jan. 2005. IEEE Computer Society.

[40] S. Luz and D. M. Roy. Meeting browser: A system for visualising and accessing audio in multicast meetings. In *Proceedings of the International Workshop on Multimedia Signal Processing*, pages 489–494, Copenhagen, Denmark, Sept. 1999. IEEE Signal Processing Society.

[41] M. Masoodian, S. Luz, M.-M. Bouamrane, and D. King. RECOLED: A group-aware collaborative text editor for capturing document history. In

P. Isaías and M. B. Nunes, editors, *Proceedings of WWW/Internet 2005*, volume 1, pages 323–330, Lisbon, Portugal, Oct. 2005.

[42] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, Mar. 2005.

[43] T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle, and P. Zellweger. "I'll get that off the audio": A case study of salvaging multimedia meeting records. In *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*, volume 1, pages 202–209, San Jose, CA, May 1997. ACM.

[44] P. Mylonas. Understanding how visual context influences multimedia content analysis problems. In I. Maglogiannis, V. Plagianakos, and I. Vlahavas, editors, *Artificial Intelligence: Theories and Applications*, volume 7297 of *Lecture Notes in Computer Science*, pages 361–368. Springer, May 2012.

[45] P. Mylonas, E. Spyrou, Y. Avrithis, and S. Kollias. Using visual context and region semantics for high-level concept detection. *IEEE Transactions on Multimedia*, 11(2):229–243, Feb. 2009.

[46] A. Popescu-Belis, D. Lalanne, and H. Bourlard. Finding information in multimedia meeting records. *MultiMedia, IEEE*, 19(2):48–57, Feb. 2012.

[47] S. Renals. Recognition and understanding of meetings. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 1–9, Los Angeles, CA, June 2010. ACL Press.

[48] S. Renals, T. Hain, and H. Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, pages 238–247, Kyoto, Japan, Dec. 2007. IEEE.

[49] D. M. Roy and S. Luz. Audio meeting history tool: Interactive graphical user-support for virtual audio meetings. In *Proceedings of the ESCA Tutorial and Research Workshop (ETRW) on Accessing information in spoken audio*, pages 107–110, Cambridge, Apr. 1999. Cambridge University.

[50] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154, Sept. 2000.

[51] A. F. Smeaton. Indexing, browsing, and searching of digital video and digital audio information. In M. Agosti, F. Crestani, and G. Pasi, editors,

3rd European Summer School on Information Retrieval, volume 1980 of *Lecture Notes in Computer Science*, pages 93–110, Varenna, Italy, Sept. 2001. Springer-Verlag.

[52] J. R. Smith. Universal multimedia access. *Proc. SPIE, Multimedia Systems and Applications*, 4209(III):21–32, Mar. 2001.

[53] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, Jan. 2005.

[54] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, Sept. 2000.

[55] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, Sept. 2006.

[56] G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tur, J. Dowding, B. Favre, R. Fernndez, M. Frampton, and M. Frandsen. The CALO meeting assistant system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1601–1611, Aug. 2010.

[57] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In D. E. M. Penrose, editor, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pages 281–286, Lansdowne, VA, Feb. 1998. Morgan Kaufmann.

[58] A. Waibel, M. Brett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. 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, Pittsburgh, PA, May 2001. IEEE Press.

[59] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker. A meeting browser evaluation test. In *CHI '05 Extended abstracts on Human factors in computing systems*, pages 2021–2024, New York, NY, May 2005. ACM Press.

[60] S. Whittaker, R. Laban, and S. Tucker. Analysing meeting records: An ethnographic study and technological implications. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 101–113. Springer, Bethesda, MD, May 2006.

[61] Z. Yu, Z. Yu, X. Zhou, C. Becker, and Y. Nakamura. Tree-based mining for discovering patterns of human interaction in meetings. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):759–768, Apr. 2012.

[62] K. Zechner and A. Waibel. DiaSumm: flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of the 18th Conference on Computational linguistics*, pages 968–974, Hong Kong, China, Oct. 2000. ACL.

[63] H. Zhang, C. Low, and S. Smoliar. Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, 1:89–111, Mar. 1995.

Index