Time-based Memory Support in Collaborative Meetings: Speech Indexing Without Speech Recognition

Saturnino Luz
Department of Computer Science
Trinity College, University of Dublin
O'Reilly Institute
Dublin 2, Ireland
Tel.: +353 1 6083686
luzs@cs.tcd.ie

October 17, 2002

Abstract

Meeting memories are important tools in computer supported collaborative work. Content indexing may provide effective means for supporting meeting memories, particularly in physically remote, multimodal meetings. Most research on novel technologies for meeting indexing and retrieval has focused on speech recognition and language engineering to enable users to reduce time-based modalities (audio, video) to space-based ones (text, graphics). This paper presents an alternative approach which relies on a much simpler analysis of time-based media logs. We claim that this approach can provide effective and intuitive meeting memory functionality without relying on speech recognition in any essential way.

Keywords Computer Mediated Communication, Meeting Memories, Multimedia Information Retrieval, Speech Indexing, Speech Recognition.

1 Introduction

Providing facilities for participants and reviewers of computer supported collaborative meetings to browse, access and structure meeting contents is an important aspect and and active area of research in computer supported collaborative work (CSCW). These facilities are generally known as *meeting memories*. The construction of effective memory tools is a particularly challenging task in non-collocated multimodal meetings in which audio, (sometimes) video, text and graphics are exchanged.

Minutes, agendas and task allocation tables are examples of traditional meeting memories. Although low-tech memories can be very effective in face-to-face meetings, non-collocated scenarios place extra strains on the participants' cognitive apparatus, which tend to impair one's ability to keep effective records and take part in the meeting at the same time[1]. Multicast technology affords non-collocated meetings and the transport protocols employed allow for those meetings to be fully recorded in great levels of detail. The problem then is essentially one of devising ways of presenting the user with views that highlight relevant information and obfuscate irrelevant details.

Collaborative systems have been proposed, including co-presence systems [6], which enhance traditional memory functionality by supporting the production of *notes*. Notes can be personal or shared *text artifacts* which are offered for discussion in real-time as part of the collaborative process. At the end of a meeting, each participant typically walks away with a textual record produced by synchronising, merging, pruning and improving such notes. This process is often mediated by speech and face-to-face communication. Textual records usually take the form of minutes and action tables. Such records can be called *static records*, due to the fact that they place greater emphasis on meeting "outcomes" (by representing them through modalities that are persistent and parallel) rather than the processes by which such outcomes were attained. Process records are usually lost along with the transient, sequential speech modality that mediates the process.

Any attempt at salvaging this kind of records will need to overcome the main obstacles posed by the nature of the modalities involved. While recording the whole audio track of a series of meetings for sequential presentation at a later time proves ineffectual [16, 12], neglecting speech exchanges in favour of static records seems to be a classical case of throwing out the baby with the bath water.

An ambitious alternative approach has been pursued which places greater emphasis on speech recognition and natural language engineering techniques [25, 23]. This approach involves developing speech recognisers capable of coping with noisy environments as well as large-vocabulary spontaneous speech, detecting communicative acts, and tracking prosody, gestures and facial expressions. State-of-the art speech recognisers have word error rates of around $20-60\%$ in Large Vocabulary Conversational Speech Recognition (LVCSR) tasks, depending on the conditions [24, 1]. Other tasks involved in this kind of approach can be just as demanding as LVCSR, obtaining similar accuracy levels. In [23], for instance, a system is described which 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. The maximum accuracy achieved by their system was automatically recognised words (compared to a chance baseline accuracy of 35%). Even if recognition and dialogue labelling were perfect, there is no guarantee that users of a meeting browser based on labelled dialogue acts would be able to use those labels effectively for information retrieval. This is illustrated by the fact that when dialogue act classification is done by humans inter-annotator agreement is far from perfect ([23] report an 84% rate for tagging by linguistcs students) which shows that dialogue acts may not necessarily

[1]This is sometimes referred to as *the divided attention problem* [26]

be intuitive starting points for meeting browsing.

In this paper we present a rather more modest approach which we nevertheless believe could be quite effective. It consists of keeping timestamps, extracting relevant *features* from textual input and determining *neighbourhoods* of text and speech segments. A prototype system called COMAP (short for "Content Mapper") [14] is presented which illustrate these ideas.

1.1 Paper outline

In what follows we delineate the meeting scenarios in which we envisage COMAP should be most useful, define intuitively as well a formally what we mean by "content mapping", and describe an implementation of the time-based aspect of content mapping. We conclude by introducing an evaluation metric devised to summarise the main characteristics of collaborative activities of the type COMAP aims at supporting and a few examples of its application to the visualisation of collaborative writing data.

2 Time-based indexing

First of all, we would like to point out that the memory support techniques described below do not apply to all types of collaborative meetings. It is unlikely any technique emerging from CSCW or any other research area does. The main types of meetings to which the techniques described in this paper apply are those where small groups collaborate, typically from remote locations, through (at least) two communication modalities: speech and text. We will henceforth refer to these modalities as the *speech channel* and the *text channel*, and to the generic class of meeting in which they are employed as *non-collocated speech-and-text meetings*. Even within these types of meetings, the effectiveness of our approach will vary according to a number of constraints which will become clear in the next sections.

In general, the more active the speech and text channels are, the more effective we anticipate memory support to be. Toward the end of the paper, we present a formal approach to predicting and evaluating these factors. First, however, we will detail our strategy and the types of meetings it aims at supporting.

2.1 Speech and text meetings

Computer-mediated, synchronous collaborative writing supported by a speech channel is the prototypical activity targeted by COMAP. For the purposes of memory building and information retrieval, the features which distinguish speech-enabled, synchronous collaborative writing from other types of computer-mediated cooperation are: the immediacy of textual interactions, and the ubiquity of transient contributions (i.e. comments, discussions, back channels, jokes, etc, conveyed by the speech mode). The general case of *speech-and-text meetings* will therefore include (to a greater or lesser degree) those types of activities which feature in process models

of collaborative writing such as the one described in [17]. In this category we include the following types of collaborative activities:

- Informal, non-collocated encounters where a shared textual tool or whiteboard acts as a focal point and serves as a medium for exchange of low level details (such as a web or email addresses, formulas, etc).
- Formal meetings, where the shared textual component may act as a focal point (e.g. an agenda tool), or as collaboratively built lasting record of the transient (speech) interaction,
- Document evaluation and revision meetings, particularly those supported by shared real-time editors [19] in which text is loaded onto the collaborative tool at the beginning of the meeting, receiving only minor modifications as the meeting progresses. These modifications may provide focus or act as mnemonic devices for the main arguments and conclusions conveyed by the speech channel.
- The variants of collaborative writing described in [20] as following a *joint writing strategy* in which several group members compose the text together, and even small components of the text are decided by group effort.
- Shared projects, in particular the initial phases of such projects during which group members discuss and agree on an interpretation of the problem, define their goals and plan their work, as well as the *integration phase* during which group members integrate their individual inputs [3, 17].

The prototypical case also bears a relationship with more generic types of collaborative meetings. Even when documents are not the focus of a group activity, in most cases, some form of a text artifact, such as minutes and action tables, is created which later on serves as a means of sharing the contents of the meeting with those who may or may not have attended it. Such documents generally assist people in working on collaborative activities, for instance by reminding them of the responsibilities they may have undertaken during the meeting, or perhaps by providing them with information without which carrying out their group task may not be possible. It is therefore clear that in many group work scenarios, combinations of speech and text play a central role in the interaction process.

We have deliberately omitted video-conferencing from the list above. Although features such as gaze tracking and automatic analysis of facial expressions are exploited in content retrieval systems [25], CSCW studies have shown that the presence of a video channel adds little to the effectiveness of group collaboration [18, 2]. Audio, on the other hand, appears to be the a key component as regards building trust and successful collaboration among geographically dispersed users [9]. Video might be essential to space-based approaches to meeting memory, since information retrieval in those approaches is linked to the system's ability to re-create the formal structure of human dialogues, of which visual cues are arguably an important part. The (time-based) approach described below, on the other hand, assumes that users are able to infer communication structure from contextual information and seeks to maximise user awareness of that information.

There is also another dimension of speech-and-text meetings, orthogonal to channel activity, which must be mentioned in the context of COMAP. It concerns time and intensity of *use* of memory tools. Two typical scenarios scenarios emerge with respect to memory usage:

- Memory as an awareness tool: during the meeting, while the participants interact with each other using speech and text, the system acts as a *listener*, recording the audio and textual contents along with other events such as pointing and selections, and provides subtle feedback on the interaction history by displaying speech turns and text events on a timeline.
- Memory as a content browser: after the meeting users can view the contents of the meeting using a COMAP meeting browser. The system acts as an *assistant* for browsing the meeting document along with the audio communication.

It has been shown [5] that the use of computer-mediated communication tools for collaborative writing tends to generate dissatisfaction among group members and reduce the perceived quality of the result. The usage patterns described above might help extenuate the difficulties encountered by writers who are not able to meet face-to-face [7].

2.2 Content mapping

At the core of our approach to meeting memory is the notion of content mapping between *temporal neighbourhoods* (TN) and *contextual neighbourhoods* (CN). The main assumptions behind TN and CN are that recorded text and speech can be clustered into natural segments, and that text acts as a focus for meeting activities, thus providing an intuitive starting point for memory access. Different levels of analysis will determine different types of text and audio segments. These vary from segments derived exclusively from formatting and markup meta-information to techniques such as the ones presented in [25] and [23] which aim at producing segments modelled on high-level theories of human communication and cognition. Examples of text segments include paragraphs, document sections, items in a list, etc. Examples of speech segments include communicative turns, audio intervals delimited by silences, speech acts, etc. As we are mainly interested in investigating relationships between audio and text neighbourhoods, we will leave the segmentation method unspecified for the moment, and concentrate on defining TN and CN.

Intuitively, these temporal and contextual neighbourhoods may be described as follows:

Temporal Neighbourhood: a segment of audio recording is in a temporal neighbourhood of a text segment if that audio segment (i) was recorded while the section was being created, changed, or discussed by the participants, or (ii) is in a temporal neighbourhood of a related text segment. There could be multiple audio segments in the temporal neighbourhood of a document section, each corresponding to different time intervals during which that section was active.

Contextual Neighbourhood: a segment of the audio recording is in contextual neighbourhood of a document segment when it shares a certain number of keywords (or key-phrases) with that segment. Once again, a document section can have multiple audio segments in its neighbourhood.

These notions can be made a bit more precise if described in terms of functions and sets. Given a set $T = \{t_1, ..., |T|\}$ of text segments, and a set $A = \{a_1, ..., |A|\}$ of audio segments, temporal neighbourhoods can be determined through the recursive method stated in Definition 1.

Definition 1 *A* **temporal text-audio mapping** *is a function* $tn : T \rightarrow 2^A$ *defined as follows*

$$tn(t_i) = \{a_j : (st(t_i) \leq st(a_j) \wedge et(t_i) \geq et(a_j)) \vee \ (keyword(t_i, t_k) \wedge a_j \in t(t_k))\}$$

where $st(t)$ *and* $et(t)$ *denote the start and end time of text segment* t*, and* $keyword(t_i, t_k)$ *indicate that text segments* t_i *and* t_k *share at least one keyword (or phrase).*

Once tn has been constructed, one can also recall specific texts segments using audio as a starting point by simply inverting the mapping, or defining an *audio-text mapping* $tn_a : A \rightarrow 2^T$, such that:

$$tn_a(a_i) = \{t_j : a_i \in tn(t_j)\} \tag{1}$$

The relation $\mathcal{T} \subseteq A \times T$ induced by tn is what we call a temporal neighbourhood.

Similarly, one can describe a contextual text-audio mapping cn as in definition 2. The definition of its audio-text conterpart cn_a is analogous to that of equation (1).

Definition 2 *A* **contextual text-audio mapping** *is a function* $cn : T \rightarrow 2^A$ *defined as follows*

$$cn(t_i) = \{a_j : keyword(t_i, a_j)\}$$

where $keyword(t_i, a_j)$ *denote pairs of text and audio segments which share at least one keyword (or phrase).*

A contextual neighbourhood is a relation $\mathcal{C} \subseteq A \times T$ induced by cn.

3 Implementing content maps

From an implementation perspective, the most attractive feature of temporal neighbourhoods is the fact that their extraction from a meeting record does not require any speech recognition at all. Time stamps and keyword spotting are all that is needed. Figure 1 illustrates a text and an audio widget linked via time mappings. The timeline widget indicates presence of speech by means of horizontal bars strechning along the time axis. The words and phrases highlighted on the text pane are the starting point of the browsing activity. When related clusters of text are selected, the speech event viewer highlights the relevant segments.

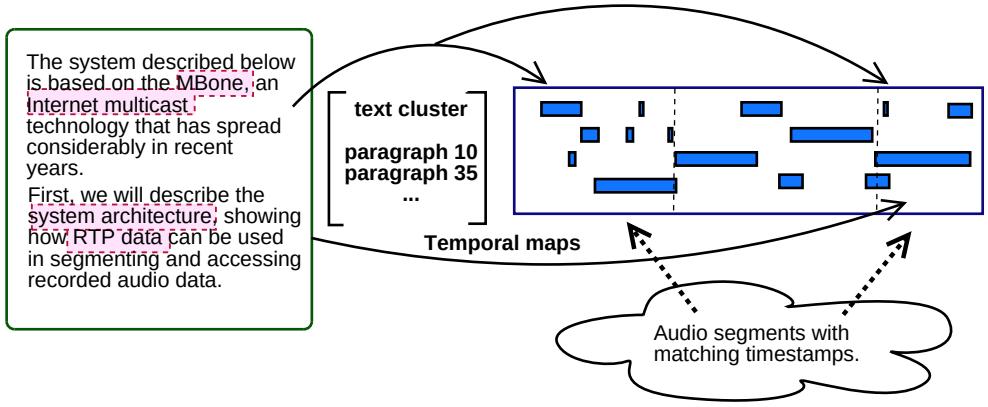

Figure 1: A temporal neighbourhood

Contextual neighbourhood inference, on the other hand, demands either full meeting transcription or, at the very least, keyword extraction from text followed by word spotting on audio via speech recognition. Manual speech transcription would be impractical, while automatic transcription of conversational speech tends to suffer from a number of problems, including noise, disfluency, false starts, overlaps [22] which result in high word error rates for even the best recognisers currently available. The situation is further complicated if one considers the fact that proper names and other "named entities", which are likely to constitute an important class of the information one would need to extract from the audio track, tend to be out of the vocabulary of most systems. A combination of keyword extraction and automatic word spotting appears therefore to be the best way to exploit the intrinsic characteristics of speech-and-text meetings. In this paper, however, we will concentrate on methods for inferring TN mappings. CN inference will be the subject of another paper.

3.1 Meetings over IP MUlticast

In order to illustrate the content mapping technique we have set up a software environment where our prototypical target activity (i.e. synchronous collaborative writing) can be fully supported. The configuration we have implemented should also be flexible enough to cover the more specialised scenarios described above. In our current setup, communication takes place through native IP Multicast in local area networks (LAN), and through the MBONE [13] in the Internet. A typical configuration is shown in Figure 2.

The architecture basically consists of a *meeting memory server* and workstations which support real-time shared editors and multi-party audio conferencing. The protocol used for both text and audio is the Real Time Protocol, RTP [21]. RTP provides the timestamps needed for mapping temporal neighbourhoods as part of its built-in packet delivery mechanism.

During a meeting, both text and audio are multicast among the meeting participants, and the COMAP server acts as a passive RTP listener, decoding RTP packets

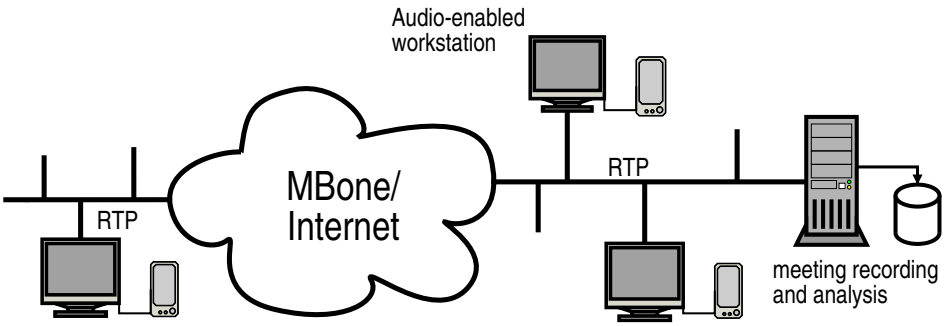

Figure 2: Sample COMAP environment

and recording audio and text on disk. Further processing of RTP audio and text is done mostly off-line, although partial information extraction and word spotting is also possible for ongoing meetings.

A typical RTP payload definition splits the transported streams into 10 millisecond packets. Timestamps play the role of helping multicast clients reconstruct and synchronise signals. This is specially important for media such as audio and video, but less relevant in text, and obviously too fine-grained for the purposes to temporal mapping. The Network Text Editor (NTE) [8], which we have used in the initial phases of data collection of our project, defines larger abstract data units (ADU) than those used for the audio payload. However, the timing information extracted from NTE's packets still needed prepocessing in order to be usable by COMAP. For simplicity we assumed text segments to be paragraphs (including headings) and generated annotation relating units of text, time and user action using an annotation scheme coded in the Extensible Markup Language (XML). The overall structure of the annotation is described in the class diagram of Figure 3.

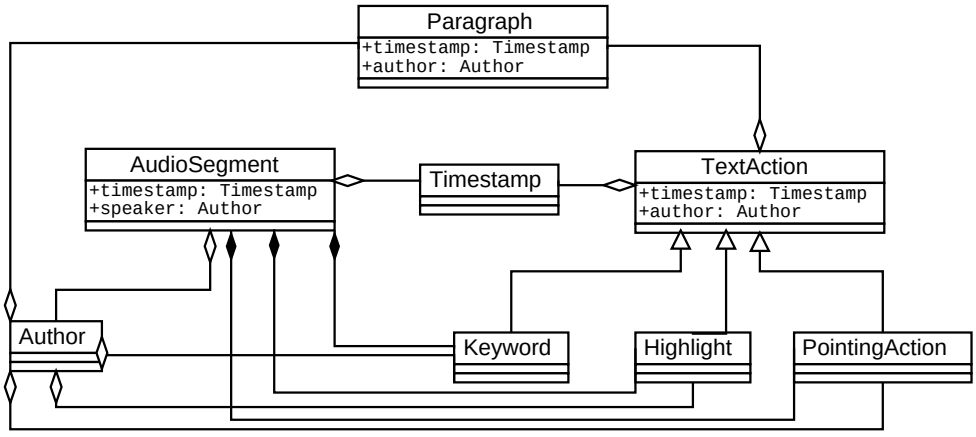

Figure 3: COMAP annotation objects

We are currently working on a new real-time collaborative editor which supports the class hierarchy of Figure 3 and provides automatic XML annotation and times-

tamps as part of its native communication protocol. The COMAP editor has been written in Java and uses a ligh-weight version of RTP as its transport medium.

3.2 Meeting indexing

Once the text produced during the meeting has been properly timestamped, temporal mappings can be calculated. As seen above, temporal neighbourhoods are determined by recursively linking keyword-related text segments to partially co-occuring audio segments. Partial co-occurrence can be immediately extracted from XML markup. Keyword-related segments are harder to determine.

First of all, not all words (and phrases) should be regarded as keywords. If they were so regarded, nearly all text segments would be interrelated, rendering the temporal mapping technique useless. All relevant segments would be *recalled* but the user would be overloaded with information of very low *precision*.

In order to select the most relevant words and phrases we use a module comprising part-of-speech tagging (POS), stop-word removal, collocation analysis and feature extraction, as shown in Figure 4.

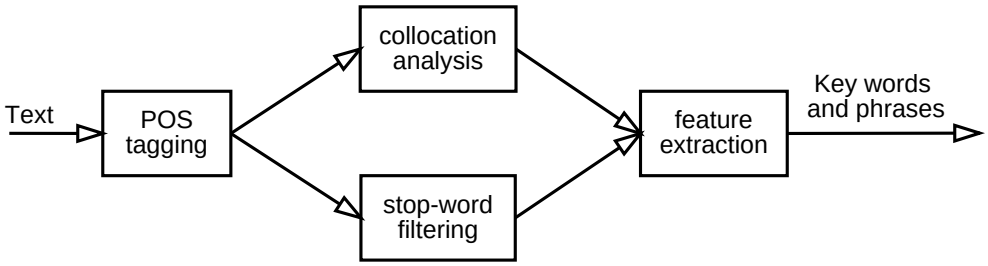

Figure 4: Text processing architecture

POS tagging assigns each word in the text a grammatical category. This phase is necessary as a pre-processing stage to collocation analysis and the removal of very common words and closed class words, such as determiners, auxiliaries, conjunctions etc. We use the transformation-based algorithm of [4] which yields an overral tagging accuracy of around 96.5%.

Stop-word removal consists simply of table lookup. Collocation analysis aims at finding phrases which may be selected as representative features of a text segment. The approach used in this module is a POS filtering algorithm adapted from [10]. It consists of selecting POS sequences that are likely to form phrases. Good candidate patterns include nouns followed by nouns (e.g. "speech recognition", "meeting browser"), adjectives followed by nouns or proper nouns (e.g. "partial information", "passive RTP"), adjectives followed by adjectives and nouns (e.g. "Gaussian random variable"), and many others.

Finally, the problem of selecting those terms that best characterise a text segment can be recast as a problem widely studied in the machine learning literature: *feature selection* [15]. The task can be described as a reverse classification task where each text segment represents a category, and one wishes to find the main features of that

category. We have employed an information theoretic measure known as *expected mutual information*, or *information gain* which allows aggressive reductions of feature sets while preserving classification accuracy [27]. Information gain can be calculated through formula (2), where w_i is a word, s_j a segment, $p(w_i)$ is the probability of w_i (i.e. the probability that a randomly chosen word in the text is w_i), $p(s_j)$ is the probability associated with segment s_j, and $p(w_i, s_j)$ is the joint probability of those two variables.

$$G(W,S) = \sum_{w_i \in W} \sum_{s_j \in S} p(w_i, s_j) log \frac{p(w_i, s_j)}{p(w_i)p(s_i)} \tag{2}$$

The output of the information extraction module is a *word table* containing keywords and key phrases, along with the text segments in which they occur. TN-based meeting indexing is largely a matter of cross referencing word tables and the timestamps extracted from audio and text media in the manner prescribed by Definition 1.

4 Interleave factor

The approach to meeting indexing described above rests on the assumption that co-occurrence of events provides valuable clues for information retrieval. In order to assess the impact that temporal mapping might have as a meeting memory device one needs metrics to somehow quantify the the extent to which an action conveyed by the speech medium is accompanied by another in text. In this section we present a metric based on the degree of "interleaving" of text and speech moves during meetings. We call it *Interleave Factor* (*IF*) [11].

The meeting browser software keeps track of speech and text events and stores its records on the meeting server (Figure 2). Based on those records one may draw charts of speech and text on a time axis, grouping them into clusters of, say, 10 seconds in lenght. Figure 5 shows a snapshot of a speech-and-text interaction between two users: *P1* and *P2*. Solid filled boxes represent voice events (over 30-second intervals) while dotted boxes represent text events. In this section we use point charts of small intervals in order to explain the *IF* formula. Real meeting charts exhibit a much more complicated picture, as will be seen in the next section. *IF*scores, however, are insensitive to meeting duration.

We are now able to estimate *IF*. We propose an estimate founded on a continuous probability distribution, where the probability of a speech event a in a meeting m, for instance, is given by equation (3)[2].

$$P(a) = \frac{et(a) - bt(a)}{et(m) - bt(m)} \tag{3}$$

Similar probabilities should be estimated for text events. The basic idea behind *IF*is that the interleave factor will be determined by the probability that a speech and

[2]As before $bt(.)$ and $et(.)$ denote start time and end time.

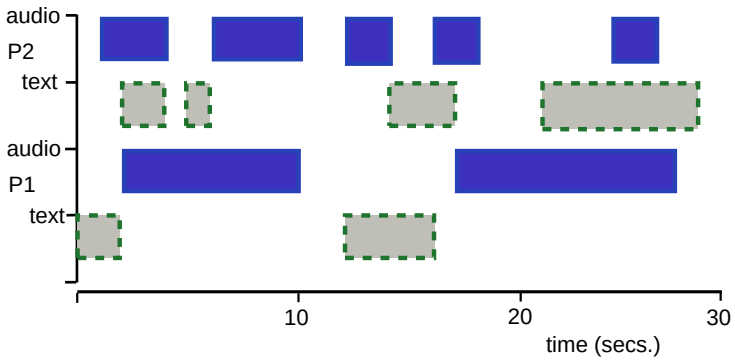

Figure 5: Extended meeting interaction profile

a text event overlap in a certain time interval. However, one cannot simply consider the meeting as a whole, calculate the probabilities of each speech and text event and then say that *IF* is the intersection of those. The reason why this would not work is that one needs to be able to distinguish segments such as the 30-second snapshot shown in Figure 5, in which participant P1 has a high *IF*, from the meeting depicted in Figure 6, in which there is a low level of interveaving, even though the audio and text would have about the same joint probability.

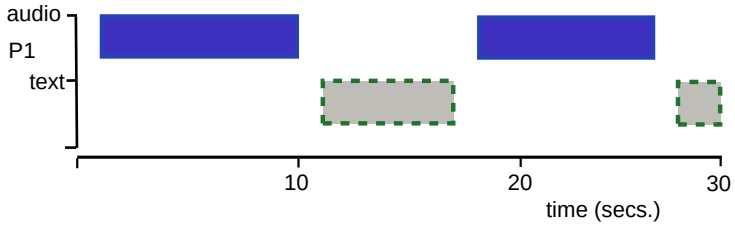

Figure 6: Low *IF* segment

The meeting of Figure 6 would be representative of cases where people talk, then work for a while on the text (mostly in silence), and then talk again. A way to get around this problem is to split the meeting into sensibly small intervals, estimate partial *IF*'s for those intervals separately, and then add up these partial *IF*'s to arrive at the final figure. Let's assume a meeting of total duration M seconds split into sections of L-second long intervals $I_1, ..., I_i$, where $1 \leq i \leq M/L$, as shown in Figure 7.

For each of these intervals we calculate the probability that speech and text events (respectively S_i and T_i) in the interval will overlap:

$$P(S_i \cap T_i | I_i) = \frac{\sum_{j=1}^{k} a_{ij} \sum_{j=1}^{m} t_{ij}}{L^2} \quad (4)$$

In equation (4), m is the number of text segments and k the number of speech segments. Speech and text segments are represented by a_{ij} and t_{ij}, which correspond

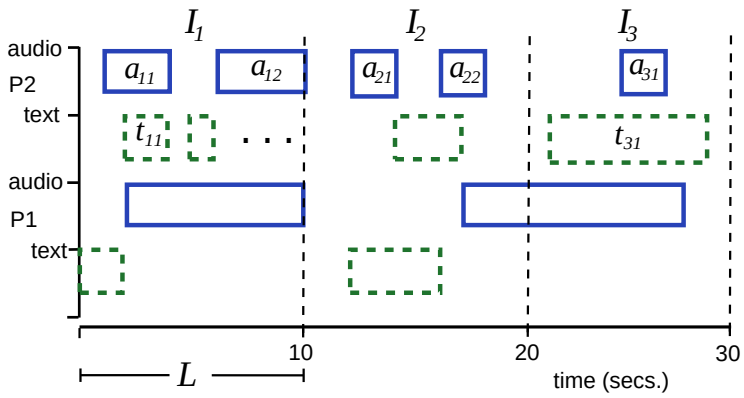

Figure 7: Meeting intervals

to the codes inside the activity bars in Figure 7 (a_{11}, a_{12}, t_{11}, etc). The probability thus expressed gives a measure of overlapping over for a single participant with respect to text and speech channels. The formula can be further generalised to cover an arbitrary number of activity channels, and an arbitrary number of participants. The former is a theoretical possibility which we have not investigated empirically so far. The latter is necessary for calculating the overall IF for speech-and-text meetings.

Generalising (4) to an arbitrary number of channels would be useful if, for instance, a video or a collaborative web browsing channel were added to the set of meeting support tools. A straightforward generalisation is obtained by assuming audio, video, text and other channels to be members of a set $C = \{c_1, ..., c_n\}$ and multiplying over interval activity as shown in (5).

$$P(\bigcap C_i|I_i) \quad = \quad \frac{\prod_{l=1}^{n} \sum_{j=1}^{k} c_{l_{ij}}}{L^n} \tag{5}$$

There are two ways of calculating $P(S_i \cap T_i|I_i)$ for two or more participants, each of which reflects a particular aspect of interaction profile. The first involves arranging the intervals for each stream and each participant serially, and the resulting (extended) intervals as in (4) except that the denominator will now be pL^2, where p is the number of participants. This gives rise to what we call *serial IF*, or IF_s for short. IF_s reflects the average degree of concurrency per participant. The second way of calculating event overlap probabilities involve merging active intervals for all participant streams before summing them up. We call the metric based on this way to estimate event overlap *parallel IF*, or simply IF. This gives us a measure of global activity interleave. IF is obtained by combining the conditional probabilities and normalising the result though the number of intervals, as shown in (6).

$$IF \quad = \quad \frac{\sum_{i=1}^{M/L} P(S_i \cap T_i|I_i)}{M/L} \tag{6}$$

Two examples are shown in Figure 8. Sections 1 (left) and 2 (right) have identical total audio and text time but yield very different IF values, as required.

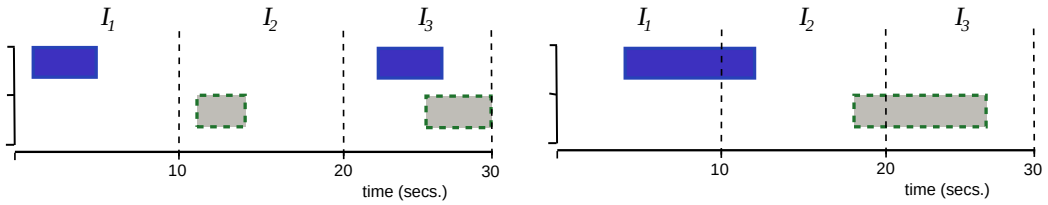

Figure 8: High (left) and low (right) IF sections

For section 1, which visual inspection suggests has the greatest degree of interleaving between audio and text we have $IF = .32$, whereas for section 2, $IF = .02$.

If the text event probability space is enriched by pointing and gesturing information, one can regard *IF* as a rough indicator of the degree to which COMAP is likely to be useful as a tool for information retrieval in speech-and-text collaborative meetings. Further research will concentrate on experimental analysis of the hypothesised correlation between *IF* scores for a number of meetings and effectiveness in information retrieval tasks supported by temporal mappings in the same meetings.

In the next section we illustrate *IF* by estimating the scores for four 2-participant collaborative editing tasks.

4.1 Four meetings and their interleave factors

We conducted and recorded four meetings in order to investigate how their *IF* scores relate to the perceived specificity of temporal neighbourhoods in those meetings. Pairs of users took part in physically remote meetings whose goals were to plan, discuss and produce short documents. The duration of those meetings varied from approximately 20 minutes to 1.5 hours. All events were timestamped. A detailed summary of the interaction can be seen in Table 1.

Participant	A	B	C	D	E	F	G	H	Avg.	Stdev.
No. utterances (/ 2700s)	225	192	121	125	214	163	126	155	165	39
Duration speech	1681	1784	1932	2032	1202	751	499	587	1308	589
Duration speech %	62	66	72	75	45	28	18	22	48	22
No. text events	34	74	99	72	37	10	36	63	53	27
Duration text activity (secs)	256	605	598	349	426	52	427	533	406	175
Duration Text Activity %	9	22	22	13	16	2	16	110	15	6
Total duration speech and text activity	1937	2389	2530	2381	1628	803	926	1120	1714	655
Total no. speech and text events	259	266	220	197	251	173	162	218	218	37
No. simultaneous speech and text events	42	96	85	74	48	6	36	27	52	29
Simultaneous speech and text activity (duration in secs)	189	341	562	329	226	17	99	69	229	167
Simultaneous speech and text activity (%)	7	13	21	12	8	1	4	3	8	6
Periods of inactivity (duration in secs)	952	651	732	648	1298	1914	1873	1649	1215	507
Periods of inactivity (%)	35	24	27	24	48	71	69	61	45	19

Table 1: Summary of speech and text activities for participants A to H

The profile of those speech, typing and pointing events was then plotted into charts such as the one shown in Figure 9.

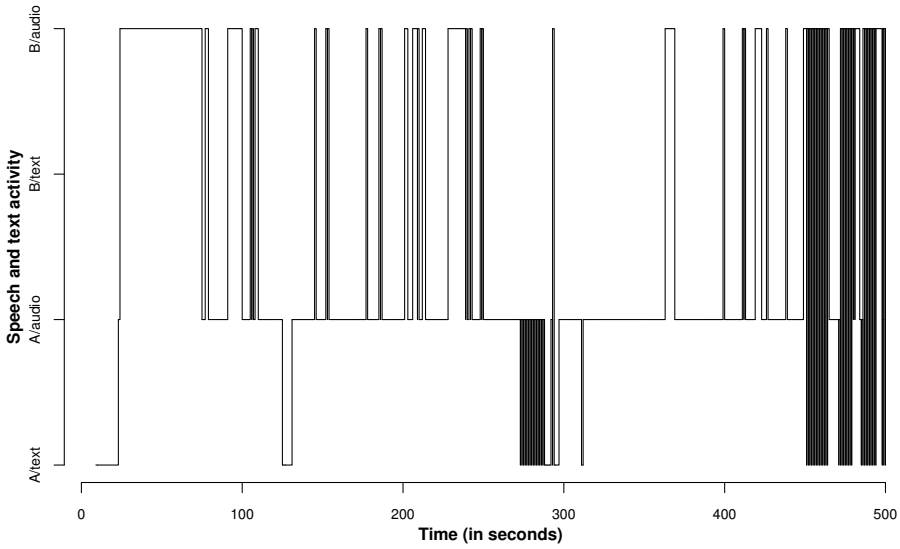

Figure 9: Speech-audio interleave over a period of 500 seconds

These are step-function plots in which the top of the vertical line defines the point at which an event occurred. The ticks labelled "A/text", "A/audio", "B/text", and "B/audio" indicate text and audio produced by meeting participants A and B respectively. Continuous horizontal lines represent stretches of non-overlapping activities, while thick vertical lines represent high levels of concurrency. Figure 9, for instance, shows a clear predominance of speech during the first 450 seconds[3], and a more balanced pattern of speech and text from that point on.

Figure 10, on the other hand, shows a snapshot of a meeting profile which exhibits a much greater degree of concurrency. The *IF*scores for user A in these plots are the following:

- $IF = 0.097$ for Figure 9, and
- $IF = 0.303$ for Figure 10.

The latter was assigned a much greater score than the former, in agreement with what we had expected.

It is interesting to observe from Figure 9 that participant B hardly ever used text, as indicated by the absence of horizontal segments on the "B/text" line. The *IF* score for participant B in that meeting was therefore 0. The overall *IF* score, however, takes into account the activities of both participants. It seems clear that, regardless of whether it was edited or pointed at by a single participant, a segment of text located in a neighbourhood of intense speech activity such as the one extending

[3]In general this is due to the fact that the participants tend to spend the initial phase of this type of meeting establishing contact, exchanging greetings, fine-tuning the audio hardware, etc.

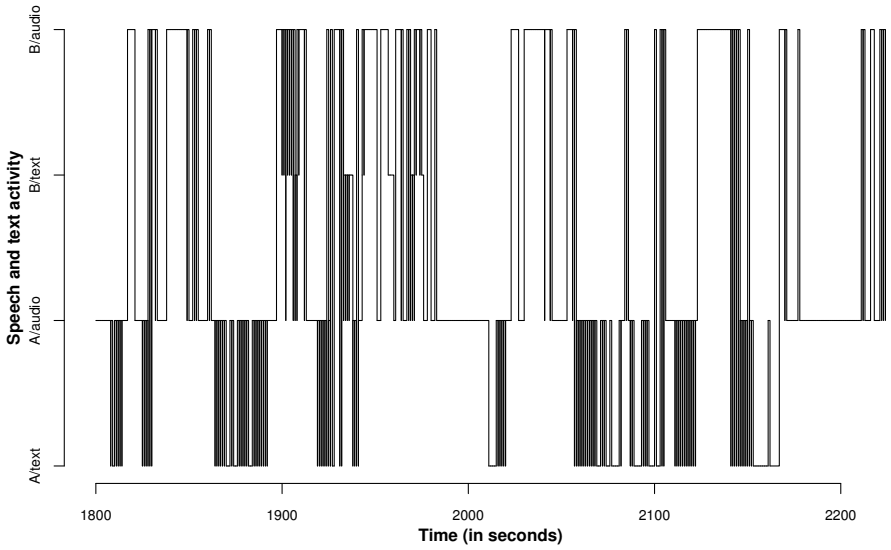

Figure 10: Highly interleaved meeting snapshot

from 450 to 600 in Figure 9 is likely to be higly indicative of the contents conveyed through speech during that interval.

5 Conclusion and further work

We presented an approach to structuring and accessing collaborative meeting memories which, unlike recent approaches based on speech recognition technology, proposes time rather than space as the unifying dimension for speech and text indexing. This approach provides a stepwise strategy for implementation as well as a conceptual framework for experimental research. It builds on the notions of contextual and temporal neighbourhoods.

From an application point of view, we described COMAP, a system for temporal content mapping which relies on shallow language engineering techniques and timestamps to support meeting recording and browsing in IP multicast environments. We also defined metrics for assessing the degree of audio-text interleaving in collaborative activity. Meeting data were presented which illustrate the techniques described. A speech-and-text meeting browser which uses *IF* scores as well as step function plots as a basis for its interface components is currently under development. Further implementation work will experiment with more sophisticated speech and text processing technology as well as further exploration of the data collected using data mining techniques.

Future research will involve collecting a larger corpus of speech-and-text meeting data in order to further establish the relationship between *IF* and effectiveness of information access under varied circumstances as well as the characteristics of temporal mapping in the presence of richer interaction logs. Techniques for establishing reliable contextual mappings and improving key word and phrase detection

as well as more structured information extraction from textual records are also being investigated.

References

[1] *NIST Automatic Meeting Transcription, Data Collection and Annotation Workshop* (2001). http://www.nist.gov/speech/test_beds/mr_proj/.

[2] Apperley, M., and Masoodian, M. An experimental evaluation of video support for shared work-space interaction. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems* (1995), vol. 2 of *Short Papers: Multimodal Interfaces*, pp. 306–307.

[3] Bikson, T. K., and Eveland, J. D. The interplay of work group structures and computer support. In *Intelectual teamwork: social and technological foundations of cooperative work*, J. Galegher, R. E. Kraut, and C. Egido, Eds. Lawrence Erlbaum Inc., Hillsdale, NJ, 1990, pp. 245–290.

[4] Brill, E. Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence. Volume 1* (Menlo Park, CA, USA, July 31–Aug. 4 1994), AAAI Press, pp. 722–727.

[5] Galegher, J., and Kraut, R. E. Computer-mediated communication for intellectual teamwork: an experience in group writing. *Information Systems Research 5*, 2 (1994), 110–139.

[6] Greenberg, S., Boyle, M., and LaBerge, J. PDAs and shared public displays: Making personal information public, and public information personal. *Personal Technologies 3*, 1 (March 1999).

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

[8] Handley, M., and Crowcroft, J. 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)* (New York, Sept. 14–18 1997), vol. 27,4 of *Computer Communication Review*, ACM Press, pp. 197–208.

[9] Jensen, C., Farnham, S. D., Drucker, S. M., and Kollock, P. The effect of communication modality on cooperation in online environments. In *Proceedings of CHI 2000 Conference on Human Factors in Computing Systems* (2000), vol. 1, ACM, pp. 470–477.

[10] Justeson, J. S., and Katz., S. M. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering 1*, 1 (1995), 9–27.

[11] Luz, S. Interleave factor and multimedia information visualisation. In *Proceedings of Human Computer Interaction 2002* (London, 2002), H. Sharp, P. Chalk, J. LePeuple, and J. Rosbottom, Eds., vol. 2, pp. 142–146.

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

[13] Macedonia, M. R., and Brutzman, D. P. MBone provides audio and video across the internet. *IEEE Computer 27*, 4 (1994), 30–36.

[14] Masoodian, M., and Luz, S. A content mapper for audio-mediated collaborative writing. In *Usability Evaluation and Interface Design* (New Orleans, LA, USA, Aug. 2001), M. J. Smith, G. Savendy, D. Harris, and R. J. Koubek, Eds., vol. 1 of *Proceedings of HCI International 2001*, Lawrence Erlbaum, pp. 208–212.

[15] Mitchell, T. M. *Machine Learning*. McGraw-Hill, 1997.

[16] Moran, T. P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., and Zellweger, P. "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* (1997), vol. 1 of *PAPERS: Enhancing, Finding, & Integrating Audio*, pp. 202–209.

[17] Neuwirth, C. M. Computer support for collaborative writing: A human-computer interaction perspective. In *The Second International Workshop on Collaborative Editing Systems* (Philadelphia, Pennsylvania, USA, 2000), J. D. Campbell, Ed., ACM CSCW'2000 Workshop.

[18] Ochsman, R. B., and Chapanis, A. The effects of 10 communication modes on the behavior of teams during co-operative problem-solving. *International Journal of Man-Machine Studies 6*, 5 (1974), 579–619.

[19] Olson, J. S., Olson, G. M., Storrosten, M., and Carter, M. How a group-editor changes the character of a design meeting as well as its outcome. In *Proceedings of ACM CSCW'92 Conference on Computer-Supported Cooperative Work* (1992), The Power of Simple Shared Workspaces, pp. 91–98.

[20] Posner, I. R., and Baecker, R. M. How people write together. In *Readings in Computer Supported Collaborative Work*, R. M. Baecker, Ed. Morgan Kaufmann, 1993, pp. 239–250.

[21] Schulzrine, H., Casner, S., Frederick, R., and Jacobson, V. RTP: A transport protocol for real-time applications. IETF Internet Draft draft-ietf-avt-rtp-new-04, February 1999.

[22] Shriberg, E., Stolcke, A., and Baron, D. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of EUROSPEECH* (Aalborg, Denmark, 2001), vol. 2, pp. 1359–1362.

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

[24] University of Maryland. *Proceedings of the 2000 Speech Transcription Workshop* (2000), NIST. http://www.nist.gov/speech/publications/tw00/.

[25] Waibel, A., Bett, M., Finke, M., and Stiefelhagen, R. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the Broadcast News Transcription and Understanding Workshop* (Lansdowne, Virginia, Feb. 1998), D. E. M. Penrose, Ed., Morgan Kaufmann, pp. 281–286.

[26] Wiberg, M. Knowledge management in mobile CSCW: evaluation results of a mobile physical/virtual meeting support system. In *Proceedings of HICSS-34* (2001), IEEE Press.

[27] Yang, Y., and Pedersen, J. O. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), D. H. Fisher, Ed., Morgan Kaufmann Publishers, San Francisco, US, pp. 412–420.