Multimedia Systems (2007) 12:439–457
DOI 10.1007/s00530-006-0066-5

REGULAR PAPER

Meeting browsing

State-of-the-art review

Matt-M. Bouamrane · Saturnino Luz

Published online: 24 October 2006

Abstract Meeting, to discuss and share information, take decisions and allocate tasks, is a central aspect of human activity. Computer mediated communication offers enhanced possibilities for synchronous collaboration by allowing seamless capture of meetings, thus relieving participants from time-consuming documentation tasks. However, in order for meeting systems to be truly effective, they must allow users to efficiently navigate and retrieve information of interest from recorded meetings. In this article, we review the state of the art in multimedia segmentation, indexing and browsing techniques and show how existing meeting browser systems build on these techniques and integrate various modalities to meet their users' information needs.

Keywords Multimedia segmentation · Indexing and retrieval · Multimodal meeting browsers

1 Introduction

The complexity of many projects performed in the workplace means that most tasks need to be carried out on a daily basis by teams involving people with various responsibilities and fields of expertise, sometimes residing in different places. Phases of individual work are punctuated by meetings of some sort to discuss progress, share ideas, take decisions, allocate tasks, etc. Meeting is thus a central aspect of professional activity. As computers have become ubiquitous tools for communication, possibilities for synchronous collaboration have been greatly enhanced and the complete capture of meetings could in principle free participants from distracting and time-consuming tasks such as note-taking and minute production. Indeed, there are many reasons why one would want to capture and archive meetings. An Internet survey carried out by [32] and involving more than 500 respondents sheds some light on some of the many reasons for storing and reviewing past meetings: to keep accurate records, check the veracity and consistency of statements and descriptions, revisit portions of the meeting which were misunderstood or not heard, re-examine past positions in the light of new information, obtain proofs and recall certain ideas are cited as the main reasons. However, recording meetings only solves part of the problem and as the number of recorded meetings grows, so does the complexity of extracting meaningful information from such recordings. To provide access to multimodal recordings, one is faced with the challenge of structuring and integrating orthogonal (space and time based) media in an intuitive way for the users. Continuous media, such as audio and video, are difficult to access for lack of natural reference points. Navigation in these media is time-consuming and can be confusing. Summarisation is a non-trivial process. A study of users browsing and searching strategies when accessing voicemail messages, sometimes of very short duration (30 s), showed that people had serious problems with local navigation of messages and difficulties remembering message content [75]. Many users performed time-consuming sequential listening of messages in order to find relevant information and often reported taking notes to remember content. In contrast, users

M.-M. Bouamrane (✉) · S. Luz
Department of Computer Science, Trinity College Dublin, Dublin, Ireland
e-mail: Matt.Bouamrane@cs.tcd.ie

S. Luz
e-mail: Saturnino.Luz@cs.tcd.ie

displayed improved browsing performance, playing less audio when speech recognition transcripts were available as audio indexes in the user interface [31]. Thus, to be truly effective, conferencing capture systems need to offer users efficient means of navigating recordings and accessing specific information.

There has been growing research interest in producing applications for visual mining of multimodal meeting data in order to support users' meeting browsing requirements. Interaction modalities used in meetings and thus the nature of the recorded media will typically be dictated by the meetings' physical setup (e.g. purpose-built meeting room [13,39] vs. Internet-based environments [15,19]) and meeting capture capabilities. In what follows, we review the state of the art in multimodal meeting browsers. We will use the taxonomy introduced by [65] where browsers are classified according to the focus of the browsing task, or primary mode used for the meeting data presentation. The three main browser categories are: *audio* browsers, *video* browsers and *artefact* browsers. The first two categories focus on communication modalities used in meetings and the contents they convey, while the third focuses on objects produced or manipulated during the meeting, such as notes, slides, drawings, plans etc.

We first review the state of the art in the various existing segmentation, indexing and searching techniques for speech and video data. We then present some of the main existing multimedia and meeting browser applications, and discuss evaluation methods for such applications.

2 Speech browsing

Unlike space-based media, such as text and images, where one can quickly visually scan a page of text or a set of picture thumbnails to get a general impression of a document's content, audio is a medium that does not lend itself well to visualisation. Audio recordings can be of very long duration, and multimedia databases may contain large numbers of such recordings. Accessing specific parts of audio documents is therefore particularly challenging because we often do not know *what* constitutes relevant information or *where* it is until we have actually heard it. Listening to entire audio recordings is however extremely time-consuming and in some cases simply not feasible. In what follows, we describe the state of the art in techniques for structured speech browsing. We define structured audio as the acoustic signal supplemented by an abstract representation which provides an overview of the recording, indications on the nature or importance of specific parts of the audio, and access to any location within the recording. Other comprehensive surveys of audio and speech access techniques can be found in [20,27,37].

2.1 Speaker segmentation

Visually representing audio in a meaningful manner is a particularly difficult task as there is no obvious or intuitive way of doing so. Graphically displaying an audio recording as a waveform would generally be inappropriate because, for most users, the audio signal spectrum offers no information about content. Some level of structuring can, however, be attained by common signal processing techniques. A frequently used strategy is to visually segment a meeting's audio track according to participants contributions over time. This technique is known as *speaker segmentation* [30,44]. When audio of various participants is recorded on a single track, speaker identification needs to be carried out prior to speaker segmentation. Speaker identification is the process of automatically distinguishing between participants' voices in order to determine when the various talkers are active [77]. Audio browsers based on speaker segmentation will typically display a visual representation of talk spurts as horizontal bar over a timeline, identifying participants through thumbnail pictures, colours, etc. Clicking on a bar will play the corresponding audio segment. A user can choose to listen to neighbouring audio segments or specific contributions.

There are a number of limitations to speaker segmentation as a browsing modality. First of all, typical meetings will contain hundreds of speech exchanges, the majority of which have very short duration.

In order to visually distinguish between contributing speech sources, speaker segments will be represented on windows of short duration (e.g. a few minutes) whose timescales will be stretched in comparison to the overall meeting. Browsing through the audio file therefore implies scrolling across a large amount of these audio segment windows. This can be confusing and might make it difficult for a user to develop a clear picture of the structure of the audio recording. The other limitation of speaker segmentation lies in the fact that individual contributions may be rendered meaningless without the context (other participants' contributions) in which they were said. In accordance with the natural structure of discourse, it is reasonable to assume that audio segments in close time proximity are relevant to one another. This phenomenon can be interpreted in terms of the question–answer pair paradigm [70] where adjacent speech exchanges are considered more informative jointly than in isolation. Therefore, segmenting conversations by topic seems a more appropriate choice for a general audio browsing task. However, speaker segmentation can be useful when additional information is available, such as a textual description of playback points, which can precisely identify the context of specific audio contributions. Roy and Malamud [56] have developed a system which maps text transcripts of proceedings of the United States House of Representatives to speaker transition in the audio recordings. The transcripts are manually drafted in real time during the House's sittings by a human transcriber and later edited. Participants' precise contributions regarding a particular issue can be pin-pointed by a text-to-audio alignment system which provides pointers to an audio database containing hundreds of hours of recording. When selecting a portion of text from the transcripts, a user is presented with a list of audio contributors to which they can listen.

2.2 Speech skimming

Various time- and frequency-based signal processing techniques can be applied to an acoustic signal in order to alter the play-back rate of an audio recording [3]. Playing audio at a faster rate (time compression) will thus permit a user to listen to more in less time. There is of course a limit to play-back rate increase before audio becomes unintelligible. SpeechSkimmer [4] is a system which combines various speech processing techniques in order to navigate through audio recordings. The user can adjust the speed at which he listens to the recording: slower or faster than the normal rate. Playing the audio at an increased speed is achieved through time compression techniques involving sampling of the original signal, while the entire audio content itself is preserved. Skimming, on the other hand, involves playing only selected sections of the recording. Selection is based on acoustic cues of discourse such as pause, voice pitch and speaker identification. The system offers the user several levels of skimming. The first level plays back the recording at the normal pace. In the second level, audio is segmented by speech detection. Speech pauses (silence) under a certain duration threshold are removed whereas longer ones are reduced to a set value (the duration threshold). The next level attempts to take advantage of the natural structure and properties of discourse. Level 3 identifies salient points in the recording by considering longer pauses as *juncture* pauses, which would tend to indicate either a new topic, or a new speech turn. Hence, the system will only play back (for a certain quantum of time) segments of speech which occurred after juncture pauses before jumping to the next one. Level 4, uses emphasis detection to identify salient segments of the recording. The emphasis detection algorithm is based on the speaker's pitch, or voice fundamental frequency (F0). An adaptive threshold for each speaker is generated which identifies points of highest pitch frames within the recording. The system will then only play sentences containing these highest pitch frames. These segmentation techniques are error-prone and will occasionally miss desired boundaries while mistakenly identifying others. The system compensates for these shortcomings by providing the user with additional navigation tools. These include a skimming backward mechanism which plays back the audio normally but jumps to the previous segment. This functionality enables a user who has heard something of interest to pinpoint its precise location. The user can also jump forward to the next segment if he decides that the current segment is not relevant. Skimming audio through the highest levels of the system may be disorientating, as unrelated speech segments are played in fast successive order. A usability study of the SpeechSkimmer showed that users used the system at the highest skimming levels to navigate through the audio in order to identify general topic locations and then used lower skimming levels (normal play-back or pause compression-suppression) to listen to specific parts of the recording.

2.3 Automatic speech recognition

The field of automatic speech recognition (ASR) has made significant progress in the last decade, evolving from single speaker, discrete dictation systems with limited vocabulary for restricted domains to sophisticated systems that tackle speaker independent, large vocabulary continuous speech recognition (LVCSR) tasks. Unconstrained LVCSR is a difficult task for a number of reasons including speech disfluencies in spontaneous dialogues, lack of word or sentence boundaries, poor recording conditions, crosstalk, inappropriate language models, out-of-vocabulary items and variations in accent and pronunciation. These conditions combined can cause substantial decreases in recognition rates [21].

Speech recognition is the task of automatically identifying a sequence of spoken words according to the speech signal [52,79,36]. In other words, given a sequence of word utterances w, recognition consists of finding the most likely word sequence $\widehat{w}$ given the observed acoustic signal S.

A speech recognition process encompasses a number of successive steps based on a property of languages: the use of a limited number of phonemes (the smallest perceptual "building blocks" of words), typically identified from 40 to 60 distinct phones (basic sounds). Phones can be modelled using a Hidden Markov Model (HMM) containing a number of states and connected by transition arcs. These models can be combined together to form word models which in turn can be combined into sentence models. The first step of the recognition task will process the audio signal and extract a number of acoustic features over a certain timeframe duration (typically 10 ms). Features are chosen for extraction according to their ability to discriminate between different phones. The observed acoustic features are subsequently translated into phone probabilities according to an acoustic model. The acoustic model consists of a pronunciation lexicon, where phones are usually divided into three states: beginning, middle and end. The triphone model further adds context to these states whereby individual phones are influenced by the surrounding ones. The *decoding* stage outputs the most likely sequence of words according to the word pronunciation dictionary and a language model. The language model assigns words prior to probabilities according to some grammar inferred from a large corpus. A grammar defines allowable sequence of words and their probabilities. An example of such grammar is the n-gram model, where the presence of a word is deemed to depend only on the $n-1$ previous words. Probabilities of n-grams are thus computed by counting the number of occurrences of n successive words instances in a training corpus (word frequency for unigram model, word pair frequency for bigram model, etc.) As the underlying language model explicitly models inter-word relationships, a misrecognition will often lead to another.

Although LVCSR has produced very encouraging results for certain task-specific applications, serious challenges remain in recognising speaker independent spontaneous speech in unconstrained domains. Current research issues focus on building robust recognition systems by using automatic adaptation techniques, such as adaptation of acoustic models to speakers' voices and speech rate fluctuations, language model adaptation and improved spontaneous speech modelling [22]. Despite the aforementioned shortcomings, ASR is a central component to many audio browsing systems. Typically, the ASR module is used to produce conversation transcripts for convenient user scanning, reading and other text-based information retrieval operations.

2.4 Word spotting

A keyword-based retrieval query offers an alternative paradigm to full LVCSR transcription. Word spotting consists of detecting the presence of a specific word or phrase in a speech corpus. This task is thus computationally far less expensive then generating full transcripts and may also be more appropriate for certain types of applications, such as querying a large audio database. Two types of errors can occur with a word spotting system, a *miss* and a *false-alarm*. A miss consists of not retrieving a particular keyword and a false-alarm of wrongly recognising one. Tuning a system requires finding an acceptable trade-off between correct keyword detection (true-hit) and false-alarm rates. The receiver operating characteristic (ROC) is defined as the percentage of keyword detection as a function of false-alarm rates (in fa/kw/h: false-alarm per keyword per hour). A figure-of-merit (FOM) is calculated as the average value of the ROC curve between 0 and 10 fa/kw/h.

The application of an HMM-based recognition system to keyword spotting will typically require building acoustic and language models for a pre-defined set of keywords and non-keywords, or *fillers* [54]. Spotting a keyword then consists of two phases: hypothesising when a keyword may occur in speech (*putative hit*) and subsequently assigning a score to the hypothesis. The hypothesis is accepted if the keyword score is above a rejection threshold. Thus, the output of a wordspotter would be a set of keywords and their time offsets, with everything in-between considered as background words. Filler modelling is used to match arbitrary non-keywords present in speech and is crucial in the performance of the word-spotter. Appropriate models will reduce the rate of false-alarms, as shown by the comparative studies of filler models in [55]. Another decisive component in the performance of the word-spotter is an appropriate scoring algorithm. DECIPHER [71] assigns a likelihood score to a hypothesised keyword by combining acoustic likelihood probability and language model probability, where the language model is trained from combining task-specific data (with high occurrences of the specific keywords) and task independent data. The main disadvantages of LVCSR-based systems for word spotting is that they are computationally expensive and can only effectively recognise keywords if these are present in their lexicons.

To circumvent some of these shortcomings, an alternative approach to word spotting is off-line speech preprocessing to generate a phone lattice representation. The lattice representation consists of an output of multiple phone hypotheses at every speech frame, along with a likelihood score for the hypothesis [33]. The depth of the lattice can hence be set by preserving only the n best hypotheses. Thus, wordspotting reverts to a keyword pronunciation match against each lattice. The main advantages of the phone lattice representation are that search is fast and that there is no restrictions on keywords. In [16], speech is initially converted offline into a table of phone trigrams with acoustic scores. This is followed by a two-step search, using the keyword phonetic transcription. If the query term does not appear in a pronunciation dictionary, a spelling-to-sound database generates the likely phonetic representation of the word. The first step is a fast coarse match which identifies keyword locations according to the phone trigrams index. In order to reduce the number of false alarms, this is followed by a detailed acoustic match.

2.5 Topic segmentation

Automatic topic segmentation is the process of segmenting a (text or audio) document into regions of semantic relatedness. This is a difficult task for a number of reasons. First of all, as an abstract concept, a topic is difficult to define. Furthermore, it is a subjective notion and topical annotation of documents by humans will often differ from annotator to annotator, particularly in the case of topic shifts. This is evidenced in [29] where seven readers who were asked to find topical boundaries of a text document exposed a variety of judgements. Research on topic detection and tracking (TDT) was originally targeted at newswire and news broadcast and typically involves three distinct phases. The first is to segment data streams into self-contained coherent units. A second phase consists in detecting new (previously unseen) topics. This step can either be performed on-line (as the news are broadcast live) or retrospectively, on a corpus of samples. The final step consists of identifying whether incoming samples are related to a particular (target) topic. In the particular context of audio streams, all these operations are ideally performed using ASR transcripts for full automation. Therefore, the first audio segmentation step can be seen as a text segmentation task. The Dragon system [78] requires a topic model for the segmentation task. A topic is modelled with unigram statistics. A training set is clustered into different topics using a distance metric. If a sample's distance to a given cluster is less than a certain threshold, the sample is included into the cluster and the cluster model is updated. If the distance is above the threshold, a new cluster is created. Once the topic model has been created, segmenting a stream is done by scoring stream frames against the topic model and detecting topic transition. Another approach to segmentation, described in [2], measures shifts in the vocabulary. Each sentence of a text stream is run as a query against a local context analysis (LCA) thesaurus which identifies and returns a number of semantically related words or concepts. Although some sentences of the original text have few or no common words, they may in fact share a number of concepts. The text is then indexed at the sentence level according to these features. A function of the feature offsets is then used as a heuristic measure of change in content. The chief advantage of this approach is that it is unsupervised. Its drawbacks are that it is computationally costly (a database query per sentence) and that LCA results for sentences with poor semantic value (a common feature of speech) are essentially random. Another approach is to model lexical features or "marker words" usually found at start and end of topical segments in order to predict topic changes.

The approaches mentioned earlier make exclusive use of textual features while ignoring some of the specific characteristics of speech such as *prosody*. Prosody (in linguistics) refers to phonological features in speech such as syllable length, intonation, stress and juncture, which convey structural and semantic information. In addition to lexical information obtained from speech recognition, Tur et al. [66] use prosodic features automatically extracted from speech for automatic topic segmentation. A distinctive advantage of using a prosodic model is that it is largely independent of the recognition task and therefore should be robust to recognition errors. The topic segmentation algorithm is implemented in two phases: the speech input is first segmented into sentences (speech units). Then sentence boundaries are analysed to determine whether they coincide with a topical change. In effect, this approach reduces topic segmentation to a boundary classification problem, i.e. estimating the probability of a topic boundary given a word sequence and its set of prosodic features. To this end, a prosodic model needs to be created, built on the feasible extraction of prosodic features for a fully automated solution. A corpus with human labelled topic boundaries was used in order to infer useful prosodic features. Features which were found to be important in identifying topic boundaries include: pause duration at boundaries, pitch or fundamental frequency across boundary, last phone duration before boundary [59]. In addition, non-prosodic features which were available from the speech recogniser, such as speech turns and speaker gender, were also included in the model. The prosodic model was subsequently combined with a language model similar to the one used in [78]. The performance obtained was comparable to that of the best word-based systems.

2.6 Spoken language summarisation

Unlike automatic text summarisation, which has long been a subject of study, spoken language summarisation is a new research domain, with serious issues remaining to be solved. These include how to deal with the presence of speech disfluencies in spoken dialogue, sentence boundaries, information spanning across several speakers and speech recognition. Speech disfluencies include non-lexicalised filled pauses (*um, uh*), lexicalised filled pauses (*like*), repetitions, substitutions and false starts.

DiaSumm [81] is a spoken language summarisation system comprising a number of stages. Audio recordings can theoretically be used as an input. However, the results described below were obtained using human generated transcripts with annotated topic boundaries [80]. The system first runs a part of speech (POS) tagger on the transcripts to identify disfluencies. Repetitions and discourse fillers are subsequently removed through a clean-up filter algorithm. The result of the POS tagger is then fed into the sentence boundary detection component. False starts are then detected and removed. Cross-speaker information detection consists of identifying question–answer pairs, which is first done by detecting questions and then the corresponding answer. Once all these steps are completed, the summarisation mechanisms rank sentences using a term frequency, inverse document frequency-based (TFIDF) MMR ranking within topical segments. This algorithm is intended to extract salient parts of the document while avoiding redundancy. The TFIDF of a term t_k with respect to a document D_j in a set of documents T_r is given by

$$\text{tfidf}(t_k, D_j) = \text{nt}(t_k, D_j) \times \log \frac{|T_r|}{\text{nd}(t_k, T_r)}$$

where $\text{nt}(t_k, D_j)$ is the number of times t_k appears in D_j and $\text{nd}(t_k, T_r)$ is the number of documents from set T_r with at least one occurrence of t_k. TFIDF reflects the intuition that the more a term occurs in a document, the more it is representative of that document, and that the more a term occurs across various documents, the less discriminative it is. Maximum marginal relevance (MMR) [9] rewards "novelty" by allocating increased weight to a document if it is both relevant to the query and has little similarity with previous selected documents.

Valenza et al. [68] present a speech summarisation system which combines inverse term frequency with audio confidence measure from the speech recogniser's output. For a word to be included in a summary it needs to have high probabilities of relevance and correct recognition. The authors stress that in order to produce useful summaries, a certain level of inaccuracy should be acceptable. Giving too much weight to audio confidence risks omitting relevant information from the final summary. The acoustic confidence measure for a particular word is determined by the sum of phone probabilities for that word normalised by word duration. Summaries are generated on a per-minute basis to favour spread content more than punctual information and may be more adapted to the targeted audio (broadcast news). A summary can be a set of keywords (frequently occurring single words), n-grams (n words extracted from audio transcript, with n determined by the user) or utterances (audio segments delimited by speaker or content change). The user interface provides a keyword list, a user-specified summary type as well as full text output. Selecting a keyword causes relevant segments of the summary and full text to be highlighted. The user can also listen to the corresponding audio segment or to larger audio segments. The system thus provides audio indexing and summarisation.

An approach which unlike the previous does not rely on lexical recognition was introduced in [10]. It uses pitch as an energy content to detect emphasis and create summaries by selecting emphasised segments in temporal proximity.

3 Video browsing

A video document essentially consists of a succession of images over time, but will often contain additional modalities such as sound and text. Therefore, indexing a video document can be approached from the various modalities it may contain. Because we have covered techniques for audio document browsing in detail in the previous section, here we will describe visual and multimodal approaches to video indexing from a meeting recording point of view. Current approaches to video browsing for the most part employ techniques which rely on domain knowledge of video types and thus make a number of assumptions about features of the recording which limit their applicability to meeting browsing. The availability of closed-captions in news broadcast, for example, may be used to generate summaries using text-based techniques. Sports video indexing techniques and highlight extraction generally use heuristics valid only in the context of the rules, grammar and semantics of a specific sport (though more generic approaches to sports videos have also been investigated [28]). High motion, high pitch, increased audio volume may be used to identify action scenes in feature films but would be of limited use in recordings of typical meetings. In [62], techniques are reviewed which regard video from an author's perspective, assuming a process of production and editing which defines documents with clear structure and semantics, where scenes and transitions can be identified. In [61], selection of static frames preceded by scenes of camera motion, or zooming are among a number of heuristics used to choose frames of importance. Such assumptions, while valid in a production environment, are mostly inadequate in the case of automatic meeting recordings, which typically contain raw data captured from a number of unmanned fixed cameras. In [40], a survey of browsing behaviour for various types of video content concludes that as information in conferences is essentially audio centric, visual features only offer users minimal cues on content. In what follows and unless otherwise stated, we present a number of techniques suitable for indexing, segmenting and browsing meeting recordings captured by unmanned static cameras in a conference room.

3.1 Visual indexing

A scene or *shot* can be defined as a succession of images which have been continuously filmed and constitutes an intuitive fundamental unit in video. One common technique for automatic video segmentation is automatic *shot boundary detection* which is achieved through the non-trivial task of measuring similarity (or rather dissimilarity) between successive frames over a certain number of features of the image (colour, texture, shapes, spatial features, motion, etc.) A number of reliable methods have been proposed to this end [1]. When a video document is created through a production process, changes between shots may not necessarily be clear cut but a fading, dissolving or wiping transition. In order to detect gradual transitions, algorithms for boundary detection generally include a dissimilarity accumulation function with a boundary found if the functions goes over a certain threshold. Once a video has been segmented into scenes, these can be characterised by a single image, chosen according to certain heuristics (e.g. choose first frame, frame containing a face or significant object). Time-varying spatial information can thus be translated into the spatial domain for convenient scanning through the use of *keyframes*, whereby each scene of a video can be represented with a single image, offering a visual summary of the recording. Although these methods can be valuable for feature films or video database indexing, their application to meeting recordings is certainly limited as significant visual changes in a common meeting scenario are likely to be minor (e.g. drawing on the board); thus their discriminating power is weak and their semantics rather limited without additional (audio) information.

Another promising research area in automatic meeting segmentation and indexing is concerned with identifying specific and significant meeting *actions*. McCowan et al. [49] propose using low-level audio and visual features (speech activity, energy, pitch and speech rate, face and hand blob) to model meetings as a continuous sequence of high-level meeting group actions (monologue, presentation, discussion, etc.) using an HMM model based on the interactions of individual participants.

3.2 Video summarisation and skimming

Similar to the familiar fast-forward feature of standard video players, time-compression can be used to increase the speed of image play-back in digital video recordings. However, speed increase is inversely proportional to a viewer's ability to understand the recording and this technique will quickly result in a serious degradation of comprehension. It is also cumbersome for browsing lengthy recordings with no specific reference points. Another technique consists of displaying frames separated by a fixed time interval (e.g. 30 s) but this process is essentially random, might skip over crucial information, is generally confusing and remains time consuming. In the study of video browsing behaviour carried out in [40], users navigated conference presentation recordings using essentially time compression and speech-based silence removal techniques. An interesting alternative to these fast-forwarding techniques is used in CueVideo [63]. Sequences with low motion are sampled and played with fewer frames thus faster than sequences with higher motion levels in order to quickly skip over scenes with little content information and jump to significant ones. This technique seems particularly well adapted to meeting recordings as they often contain long shots with little or no significant motion. The drawback of this technique is that faster video sequences cannot be synchronised with audio in an intelligible way. It remains, however, an efficient tool for navigation. The Informedia^{TM} [61] project at Carnegie Mellon University offers search and retrieval of video documents from an online digital library. The system integrates image analysis and speech and language processing techniques to produce *skims* of video documents. Keywords are identified through audio transcripts and closed captions (where available) using TFIDF. A compressed audio track is then generated according to the location and duration of these keywords. The number of keywords retrieved therefore defines the duration of the skim. Once the audio track has been created, a corresponding video skim is generated. To avoid redundancy within close proximity, a keyword cannot be selected twice within a certain number of frames. A minimum (2 s) of matching video frames is played along with the keywords from the audio track for clarity, but the video segments are not necessarily time-aligned with the audio. Alternative frames of a corresponding scene are picked according to certain heuristics. These include prioritising introduction scenes, frames with human faces, static frames preceded by camera motion, zoom, etc. Compaction ratio is typically 10:1 but a ratio as high as 20:1 has been found to preserve essential information.

We have alluded to the fact that many domain knowledge assumptions used in broadcast news, feature films and sports video browsing may not fare well when used with unstructured meeting recordings. However, regular meetings in a given organisation may also follow a well-structured grammar, which can then be exploited for meeting summarisation. VidSum [57] uses regular patterns occurring in weekly staff forums (e.g. introduction by first speaker, presentation by second speaker, applause, questions and answers) to generate concise summaries of presentation recordings. Content analysis first extracts a number of visual features, which are then matched against a presentation library in order to find the most likely presentation structure. The next step is to populate a video template called *summary design pattern* (SDP). Slots in the SDP are filled according to priority criteria (e.g. introduction, conclusion) until a pre-determinate structure and the time constraints are met. The result is a concise, well-structured and pleasant to watch summary. However, evaluating the summarisation process remains difficult because, as remarked in [57], different templates may produce significantly different summaries.

4 Artefact browsing

This study is essentially concerned with reviewing automated solutions for accessing meeting recordings; therefore we are primarily interested in tools and techniques which require no additional effort from the participants during the actual meetings. However, there is another important category of meeting access tools, which we will refer to as *meeting minutes systems*. Minutes systems provide support for note-taking, meeting annotation and thus for later access to meeting recordings through active effort by the participants *during* the meeting process. Although a comprehensive review of all meeting support tools is beyond the scope of this article, we will describe a number of meeting minutes systems and analyse some implications of note-taking for browsing.

4.1 Meeting minutes systems

NoteLook [12] is a client-server system deployed in a media-enriched meeting room to support multimedia note-taking. Participants take notes during the meetings through the NoteLook client, which runs on wireless pen-based notebook computers. Presentation material displayed by a room projector, images of the whiteboard, video of the speaker standing at a presentation podium and room activity are some of the data which participants can incorporate into their personal minutes, either as still images or video streams (recorded by the server). The users can select which live video channel is displayed on the client, and still images can be incorporated into the notes either as thumbnails or as the page's background image. For slide presentations, NoteLook provides an automatic note-taking option which captures any new slide transition and generates thumbnails of room activity at regular intervals during a slide duration. Images and pen strokes are timestamped and can therefore be later used to access the video recordings of the meetings. LiteMinutes [11] is an applet-based note-taking application running on a wireless personal computer. Meetings take place in a media-enriched conference room, and video, audio and slide images are a number of potential multimedia items captured and stored at a server. Notes taken during the meeting are timestamped. They can be viewed in real time by other meeting participants (if a designated person acts as a scribe) and can also be revised later on. Notes taken on different laptops are handled separately. After the meeting, notes can be e-mailed to designated recipients and are also accessible through the capture server, which hyperlinks the notes to related media (slide, video) if these were active at the time of writing (smart-link). MinuteAid [39] is a meeting support system which enables participants to request and embed meeting multimedia items within a Word document during a meeting. Multimedia items which can be requested by the MinuteAid client running on a participant's personal computer include projected slides, audio recordings, omni-directional video and whiteboard images. Slides can be obtained in real time, audio tracks require a 15 s delay whereas video can only be obtained once the meeting has ended and the video recording has been processed by the server. Once all data requests have been processed, participants can manipulate the minutes as a standard multimedia document.

4.2 Implications for browsing

In most meeting scenarios, participants will interact with artefacts of some sort, to present and share information (slides), express and clarify ideas (whiteboard) or as personal minutes (note-taking). Thus, actions associated with artefacts will generally be associated with significant meeting events and will convey strong semantic content. A number of researchers have investigated participants' interactions with meeting artefacts as a means of segmenting, indexing and structuring meetings. Filochat [76] is a digital notebook which enables audio indexing of collocated meetings through note-taking. Time indexed handwritten notes allow users to listen to concurrent segments of audio. An important and unforeseen result of a usability study of the device was that some users made explicit indexing notes during meeting when hearing subjects of potential interest in order to revisit these specific points later on. Audio indexing according to note-taking activity is also implemented in the Audio Notebook [64], a paper notebook with cordless pen coupled with a digital audio recorder. Audio indexing is complemented by speech skimming functionalities through speed control of audio play back, phrase detection (which prevents audio from being played from the middle of a sentence) on topic shift detection, based on acoustic features of the audio recording (pitch, pauses and energy). Users of the Audio Notebook were able to use the functionalities provided by the system to successfully review recorded information, clarify ambiguous or misunderstood notes, and retrieve portions of audio which had been intentionally bookmarked. Classroom 2000 [8] is an educational system which aims to give students post-access to the content of university lectures. The system provides access to audio and video recordings as well as additional information, such as web documents, visited during the lecture and note written on an electronic whiteboard. There are several levels of access to a lecture: slide transitions, which provide access to the audio for a duration of each slide, pen-stroke level, which provides access to audio for the writing duration and word level. To facilitate navigation of recorded lectures, the system displays a timeline indexed with all significant events captured during the lecture.

5 Meeting browsers

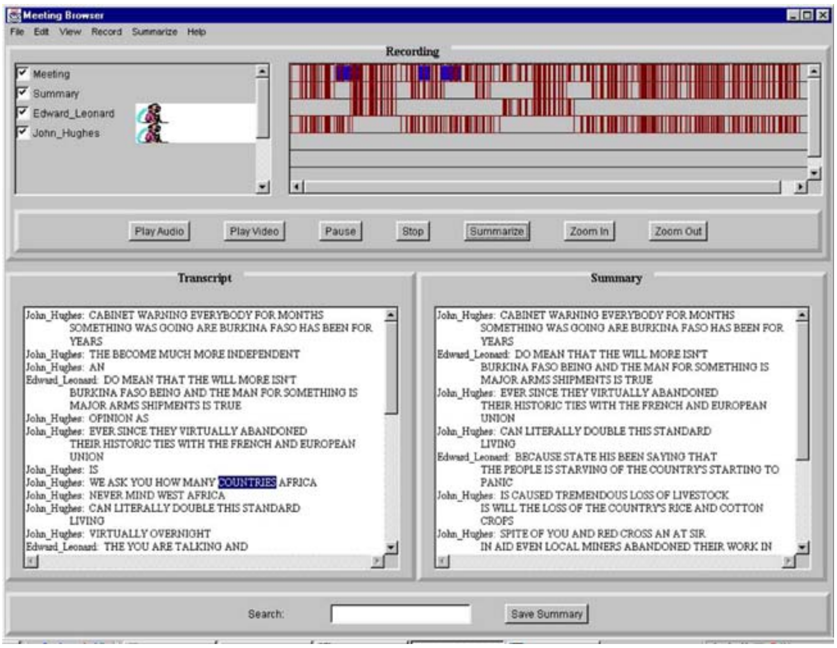
Meeting browsers are systems that integrate some or all of the previously described technologies in order to provide information seekers with a unified interface to multimedia meeting archives. Such integration was usually based on *ad hoc* frameworks built around particular technologies, especially LVCSR. Although general models of multimedia storage have been proposed which explicitly tackle the integration issue, use of those models in meeting browsers has been somewhat limited. Multimedia modelling has, however, been an active research field and today's meeting browsers owe a great (although sometimes unacknowledged) deal to early work on *media streams* [26]. An elaboration of the concept of streams that combines object oriented and relational database techniques into a unified model has been presented in [17]. Research by Jain and collaborators on image retrieval has been extended to define a unified semantics [58] which incorporates elements of interaction and context and thus lends itself well to modelling of more general multimedia data, including meeting data [60]. A unifying effort along similar lines is presented in [50], and an approach that originated of meeting storage concerns is described in [43]. Models continue to be investigated and standards, such as MPEG-7 [45], are starting to emerge which target descriptive annotation and structuring of multimedia data. Nevertheless, of the systems reviewed below, only COMAP, HANMER and Meeting Miner can be said to be fully based on a general model. They employ the content mapping model proposed in [43]. As the area of meeting browser research matures we expect models to play a more prominent role.

5.1 The Meeting Browser

The Meeting Browser [69] displays meeting transcripts time aligned with corresponding sound or video files. The browser comprises a number of components, including a speech transcription engine and an automatic summariser. The summariser attempts to identify salient parts of the audio and present the result to the user as a condensed script, or *gist* of the meeting. The summariser takes a textual transcript as an input. This transcript is either generated manually or from a speech recognition run. The summarisation algorithm works as follows: identify the most common stems present in the transcript and then weight all speech turns accordingly. The turns with the highest weights are then included in the summary. These most commons stems are then removed and the process is repeated over turns not previously included until the summary has reached a pre-defined size. Several experiments were designed in order to evaluate the summarisation system. The first task involved asking users to categorise 30 dialogues into a certain number of pre-defined categories according to a ten turn summary of the dialogues. The authors report a precision of 92.8%. Another task was to ask users to answer a number of questions based on a summary of the dialogue. The dialogue transcript used in this case for summarisation was generated by speech recognition. The user could decide (and increase) the number of turns included in the summary. With the number of correct answers increasing with the number of speech turns included, the authors claimed that this demonstrated the potential use of speech recognition output for summarisation while conveying important points of a dialogue (Fig. 1).

5.2 The SCAN system

The spoken content-based audio navigation (SCAN) is "a system for retrieving and browsing speech documents



Transcript

```
John_Hughes:  CABINET WARNING EVERYBODY FOR MONTHS
            SOMETHING WAS GOING ARE BURKINA FASO HAS BEEN FOR
            YEARS
John_Hughes:  THE BECOME MUCH MORE INDEPENDENT
John_Hughes:  AN
Edward_Leonard:  DO MEAN THAT THE WILL MORE ISN'T
            BURKINA FASO BEING AND THE MAN FOR SOMETHING IS
            MAJOR ARMS SHIPMENTS IS TRUE
John_Hughes:  OPINION AS
John_Hughes:  EVER SINCE THEY VIRTUALLY ABANDONED
            THEIR HISTORIC TIES WITH THE FRENCH AND EUROPEAN
            UNION
John_Hughes:  IS
John_Hughes:  WE ASK YOU HOW MANY COUNTRIES AFRICA
John_Hughes:  NEVER MIND WEST AFRICA
John_Hughes:  CAN LITERALLY DOUBLE THIS STANDARD
            LIVING
John_Hughes:  VIRTUALLY OVERNIGHT
Edward_Leonard:  THE YOU ARE TALKING AND
```

Summary

```
John_Hughes:  CABINET WARNING EVERYBODY FOR MONTHS
            SOMETHING WAS GOING ARE BURKINA FASO HAS BEEN FOR
            YEARS
Edward_Leonard:  DO MEAN THAT THE WILL MORE ISN'T
            BURKINA FASO BEING AND THE MAN FOR SOMETHING IS
            MAJOR ARMS SHIPMENTS IS TRUE
John_Hughes:  EVER SINCE THEY VIRTUALLY ABANDONED
            THEIR HISTORIC TIES WITH THE FRENCH AND EUROPEAN
            UNION
John_Hughes:  CAN LITERALLY DOUBLE THIS STANDARD
            LIVING
Edward_Leonard:  BECAUSE STATE HIS BEEN SAYING THAT
            THE PEOPLE IS STARVING OF THE COUNTRY'S STARTING TO
            PANIC
John_Hughes:  IS CAUSED TREMENDOUS LOSS OF LIVESTOCK
            IS WILL THE LOSS OF THE COUNTRY'S RICE AND COTTON
            CROPS
John_Hughes:  SPITE OF YOU AND RED CROSS AN AT SIR
            IN AID EVEN LOCAL MINERS ABANDONED THEIR WORK IN
```

Fig. 1 The Meeting Browser user interface

from large audio corpora" [14,74]. SCAN uses machine learning techniques over acoustic and prosodic features of 20 ms long audio segments to automatically detect intonational phrase boundaries. Intonational phrases are subsequently merged into intonational paragraphs, or *paratones*. The result of the intonational phrases segmentation is then fed into a speech recogniser (around 30% word-error rate) which uses the automatic transcripts generated for a document retrieval system based on the vector space model of weighted terms. SCAN introduces several interesting mechanisms as an information retrieval system. The first one is called *query expansion* which adds related words (located within high-ranking documents) to users' short queries. The second one is called *document expansion* and attempts to compensate for some of the errors due to speech recognition. It uses the best recognition output for a given audio document as a query on the audio database. The top 25% of words present in the original document word lattice (and not included in the final transcript) and in at least half of the highest ranking documents retrieved by the query are subsequently added to the original document transcript. Both techniques improved the information retrieval tasks. SCAN's user interface has three components: search, overview and transcript. The search component retrieves audio documents based on users' queries match against the ASR transcripts of the documents contained in the database. The ten highest ranking documents are displayed along with the number of hits (number of terms of the query contained in the document transcript). The overview displays the audio document segmented along the paratones mentioned earlier, with their width proportional to their duration. Terms from the user query contained in the speech segments are represented by a colour coded rectangle, whose height is proportional to term frequency. The transcript view displays the paratones' ASR transcripts. Clicking on them will play the corresponding audio segment (Fig. 2).

5.3 Video Manga

Video Manga [5,67] is a video system which automatically creates pictorial summaries of video recordings. The system was primarily tested and evaluated on recordings of collocated meetings in a conference room but it was found to also work well with other video genres (films, commercials). Although recordings were not edited after the meetings, an operator was in control of meeting capture and could pan and zoom as well as switch between a number of cameras and other displays. This would naturally tend to encourage the capture of highlights, which would not occur with unmanned fixed cameras. Video Manga generates summaries of a meeting as a chronologically ordered compact set of still images, similar to a comic strip, hence the name. Clicking on a specific keyframe will play the corresponding video segment. The keyframe extraction technique does not simply rely on shot boundary detection but on a colour histogram-based hierarchical clustering technique which identifies groups of similar frames, regardless of timing. Once video segments have been identified, an importance metric is used to reward segments if they are both long (a heuristic suited to the specific manned capture environment) and unusual. Segments which score less than one-eighth (empirical threshold) of the

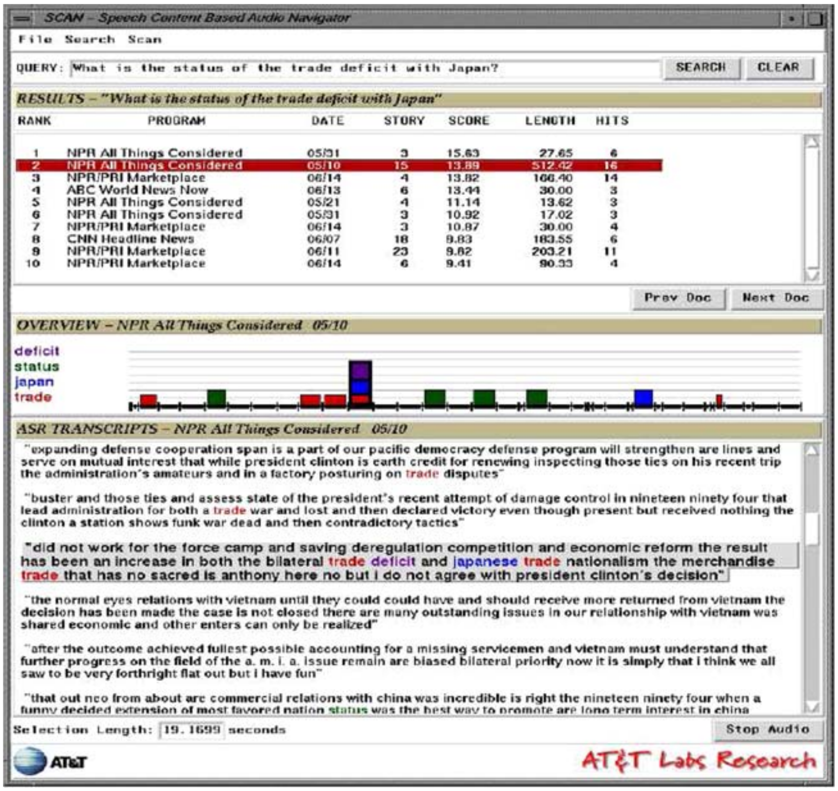


Fig. 2 The SCAN user interface

QUERY: What is the status of the trade deficit with Japan?

RESULTS – "What is the status of the trade deficit with Japan"

RANK	PROGRAM	DATE	STORY	SCORE	LENGTH	HITS
1	NPR All Things Considered	05/31	3	15.63	27.65	6
2	NPR All Things Considered	05/10	15	13.89	512.42	16
3	NPR/PRI Marketplace	06/14	4	13.82	166.40	14
4	ABC World News Now	06/13	6	13.44	30.00	3
5	NPR All Things Considered	05/21	4	11.14	13.62	3
6	NPR All Things Considered	05/31	3	10.92	17.02	3
7	NPR/PRI Marketplace	06/14	3	10.87	30.00	4
8	CNN Headline News	06/07	18	9.83	183.55	6
9	NPR/PRI Marketplace	06/11	23	9.82	203.21	11
10	NPR/PRI Marketplace	06/14	6	9.41	90.33	4

OVERVIEW – NPR All Things Considered 05/10

ASR TRANSCRIPTS – NPR All Things Considered 05/10

"expanding defense cooperation span is a part of our pacific democracy defense program will strengthen are lines and serve on mutual interest that while president clinton is earth credit for renewing inspecting those ties on his recent trip the administration's amateurs and in a factory posturing on trade disputes"

"buster and those ties and assess state of the president's recent attempt of damage control in nineteen ninety four that lead administration for both a trade war and lost and then declared victory even though present but received nothing the clinton a station shows funk war dead and then contradictory tactics"

"did not work for the force camp and saving deregulation competition and economic reform the result has been an increase in both the bilateral trade deficit and japanese trade nationalism the merchandise trade that has no sacred is anthony here no but i do not agree with president clinton's decision"

"the normal eyes relations with vietnam until they could could have and should receive more returned from vietnam the decision has been made the case is not closed there are many outstanding issues in our relationship with vietnam was shared economic and other enters can only be realized"

"after the outcome achieved fullest possible accounting for a missing servicemen and vietnam must understand that further progress on the field of the a. m. i. a. issue remain are biased bilateral priority now it is simply that i think we all saw to be very forthright flat out but i have fun"

"that out neo from about are commercial relations with china was incredible is right the nineteen ninety four when a funny decided extension of most favored nation status was the best way to promote are long term interest in china

Selection Length: 19.1699 seconds

maximum scoring segment are discarded (another option is to precisely select the number of segments included in the summary). For meeting recordings, this threshold led to discarding around 75% of the frames. In order to give higher visual importance to better scoring segments, a keyframe size in the final summary varies on a scale from one to three according to importance score. The selected frames are further reduced by removing consecutive frames from the same clusters and similar frames which are separated by only one single frame from another cluster (e.g. in dialogues). The frames importance score can also be weighted according to human, groups, slides shots detection. Documents, such as slides, web pages, transparencies, displayed in the meeting room are captured every 5 s. The text from these documents is timestamped and can be used to label corresponding shots in the pictorial summary (Fig. 3).

5.4 The Portable Meeting Recorder: MuVie

The Portable Meeting Recorder [18,38] is a system that captures a panoramic view of meetings and detects in speaker location in real time. Post-meeting processing of the recorded data generates a video skim which focuses on participants according to speech activity. As there would normally be minimal motion during a typical meeting, the authors argue that segments of higher motion potentially indicate significant events such as a participant joining the meeting or doing a presentation. Similarly, segment of higher speech volumes may point to phases of intense discussions, particularly when coupled with information about speakers' locations (high number of exchanges). The MuVIE (Meeting VIEwer) user interface thus provides among other information about the meeting (keyframes, transcripts) a visual representation over a timeline of audio and visual activities and speakers' turns. A meeting summary can also be generated by playing back in time order the video segments containing the highest visual or audio activity and highest ranking keywords extracted from the meeting transcripts (Fig. 4).

5.5 The MeetingViewer

The MeetingViewer [24] is a client application for browsing meetings recorded with the TeamSpace [25,53] online conferencing system. The TeamSpace client provides low-bandwidth video for awareness feedback and supports the use of a number of artefacts such as sharing and annotating slide presentations, creating and editing agenda and meeting action items and inserting bookmarks. In addition to session events (joining, leaving meeting) all interaction events performed on the client are automatically recorded and timestamped by the server. These events are subsequently used to index the meeting and are displayed on a timeline on the

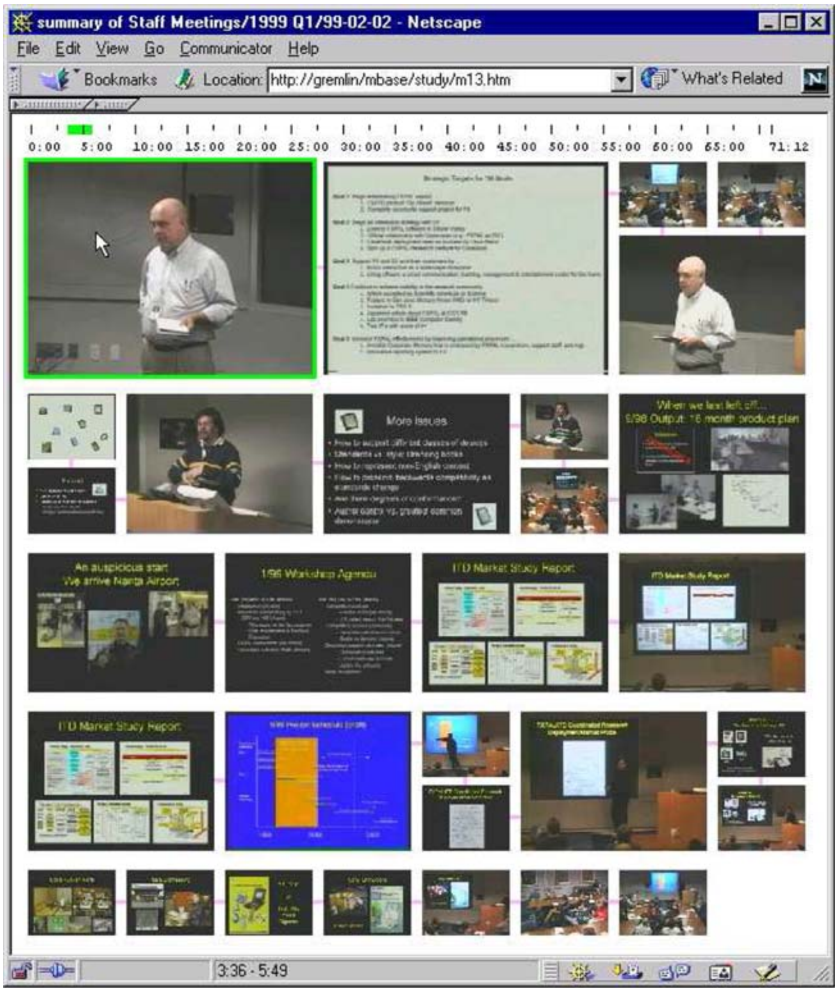


Fig. 3 A meeting summary produced by Video Manga

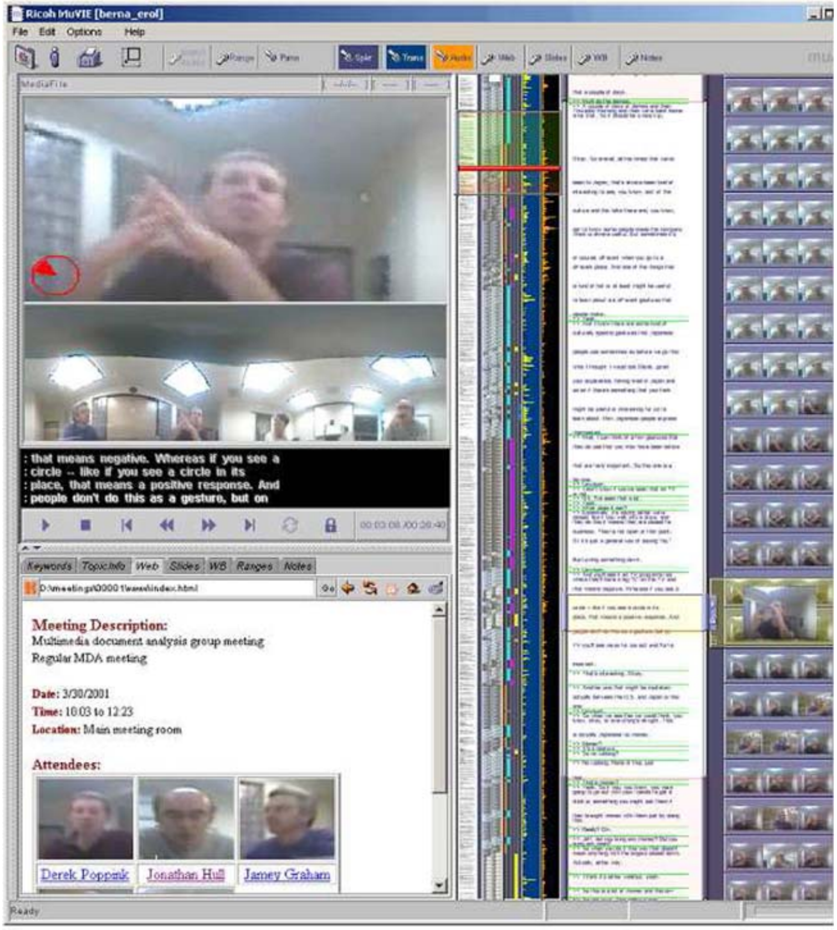
MeetingViewer interface to facilitate navigation. The user can thus choose relevant sections of the meeting. Playback will play corresponding segments of the audio and video recording along with all concurrent meeting events. Specific artefacts may be picked for viewing through the use of a tabbed pane (Fig. 5).

5.6 COMAP and HANMER

The COMAP (COntent MAPper) [46,43] is a system for browsing captured online speech and text meetings using the concepts of *temporal neighbourhoods* and *contextual neighbourhoods*. These concepts are based on viewing meeting data as a collection of discrete events, or segments. A temporal neighbourhood is defined as concurrent media events as well as segments related to these events. Segments are in a contextual neighbourhood if they share some content features (keywords). The system takes as input an audio recording along with an XML file containing detailed metadata about participants' edits and gesturing (telepointing) actions. These action metadata are automatically generated by RECOLED [47], a shared-text editor designed for this purpose. The user interface displays the textual outcome of the co-authoring task along with mosaic timeline views of the participants' speech and editing activities. To browse a meeting, a user can click on a portion of text which will highlight audio segments in the temporal neighbourhood of that text segment. The user can listen to an audio segment by clicking on it, which in turn may highlight potential concurrent editing operations. An *interleave factor* (IF) metric [41] measures levels of concurrent media activity, with intervals of greater activity deemed to be of greatest significance. A summary view of a recording can be generated through IF ranking. HANMER (HANd held Meeting browsER) [48,42] provides the same functionalities as COMAP but was designed for portable devices (Fig. 6).

5.7 WorkspaceNavigator

WorkspaceNavigator [35] is designed to provide access to information on loosely structured collaborative design projects which lasted over a long period of time



: that means negative. Whereas if you see a
: circle -- like if you see a circle in its
: place, that means a positive response. And
: people don't do this as a gesture, but on

00:03:08 /00:26:40

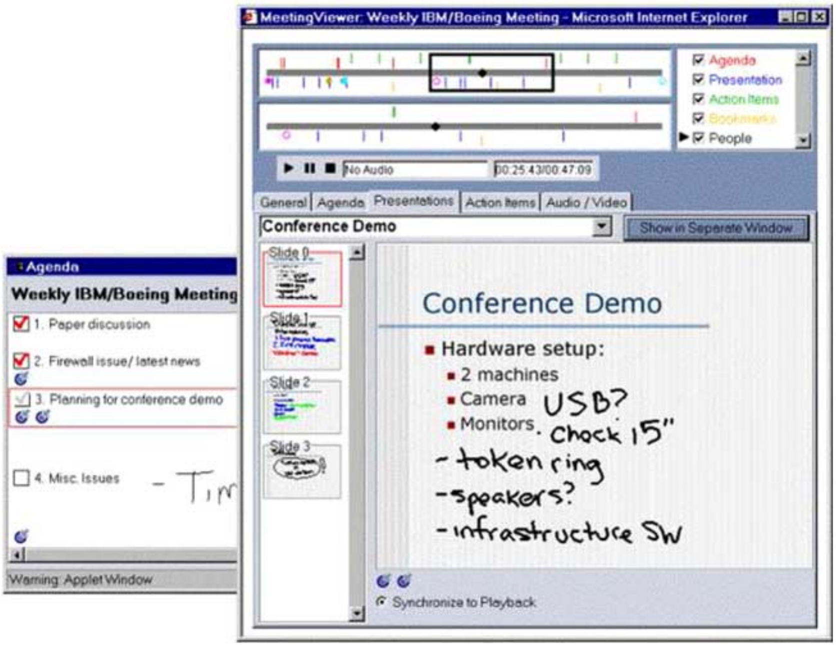
D:\meetings\00001\www\index.html

Meeting Description:
Multimedia document analysis group meeting
Regular MDA meeting

Date: 3/30/2001
Time: 10:03 to 12:23
Location: Main meeting room

Attendees:
Derek Poppink | Jonathan Hull | Jamey Graham

Fig. 4 The Muvie client user interface



☑ Agenda
☑ Presentation
☑ Action Items
☑ Bookmarks
☑ People

No Audio 00:25:43/00:47:09

Conference Demo

Agenda

Weekly IBM/Boeing Meeting

☑ 1. Paper discussion

☑ 2. Firewall issue/ latest news

☐ 3. Planning for conference demo

☐ 4. Misc. Issues – Tim

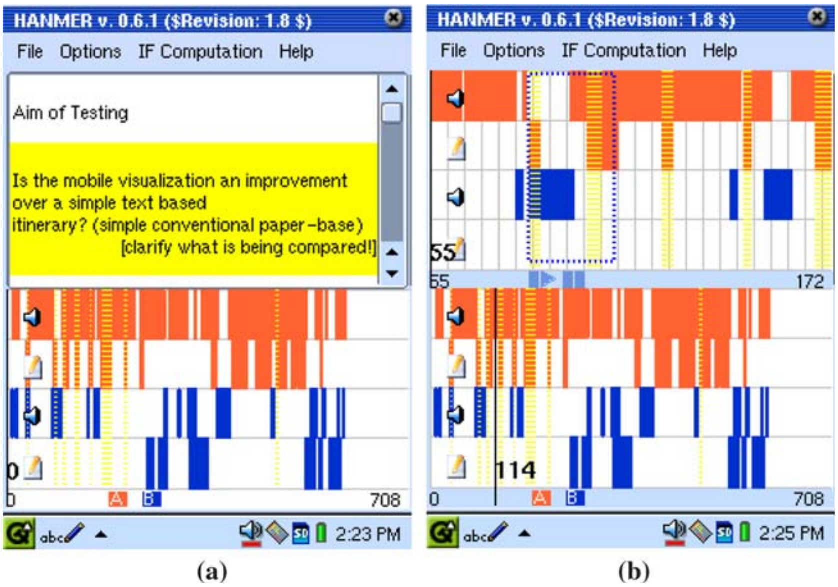
Conference Demo

- Hardware setup:
  - 2 machines
  - Camera USB?
  - Monitors. Check 15"

- token ring
- speakers?
- infrastructure SW

Fig. 5 The MeetingViewer user interface

Fig. 6 Hanmer user interface



Aim of Testing

Is the mobile visualization an improvement over a simple text based itinerary? (simple conventional paper-base) [clarify what is being compared!]

in a designated workplace. Unlike most of the other systems described here, the data recorded for meeting documentation do not include audio and video media streams but rather discrete events. This design choice is motivated by the fact that (1) given the long duration of the design process, recording live streams of all activities would produce a prohibitive quantum of data and (2) the assumption that still images are often sufficient to jog participants memories. Information on the design process is captured implicitly but participants can also explicitly capture specific events should they wish to do so, for later reference. Implicit data capture is performed every 30 s and include an overview image of the activity in the workplace, motion events, computer screenshots as well as opened files and web resources and shots of operations performed on the whiteboard. In addition, participants can choose to capture the state of the whiteboard and integrate images and annotations to the project's documentation at any time. A number of usability studies performed on WorkspaceNavigator demonstrated the usefulness of implicit discrete information capture for design process documentation, data recovery and specific information item retrieval (Fig. 7).

5.8 The Ferret Media Browser

The Ferret Media Browser [72] is a client-server application for browsing recorded collocated multimodal meetings. Recorded data include video, audio, slides (on computer projection screen) as well as whiteboard strokes and individual note-taking (digital pen-strokes), which are timestamped. A tabletop microphone array permits speaker identification. Upon starting the Ferret browser, the user can pick a combination of any available media for display and synchronised play-back. ASR transcripts, a keyword search and speech segmented according to speakers' identity are also available. The user can zoom on particular parts of the meeting. Media streams can be dynamically added to or removed from the display during the browsing task. Other data sources can also be accessed through the Internet (Fig. 8).

5.9 The Meeting Miner

Meeting Miner [6] is a tool designed for navigating recordings of online text-and-speech collaborative meetings. Meetings are recorded through a lightweight collaborative writing environment which can easily be installed on a personal computer and which was specially designed to capture editing activities [7]. Temporal information from the logs of actions captured on self-contained information items (paragraphs of text) is used to uncover potential information links between these semantic data units. Access to the audio recordings can be performed through exploration of a hierarchical tree structure generated for each paragraph through audio and activity linkage, a navigation scheme based on participants' discrete space-based actions, keyword indexing displayed above participants speech exchanges on the timeline, and keyword and topic search (Fig. 9).

6 Meeting browsers evaluation

Meeting browser systems are notoriously hard to evaluate. Unlike speech recognition and spoken document retrieval, for which the TRECs 6–8 (Text REtrieval

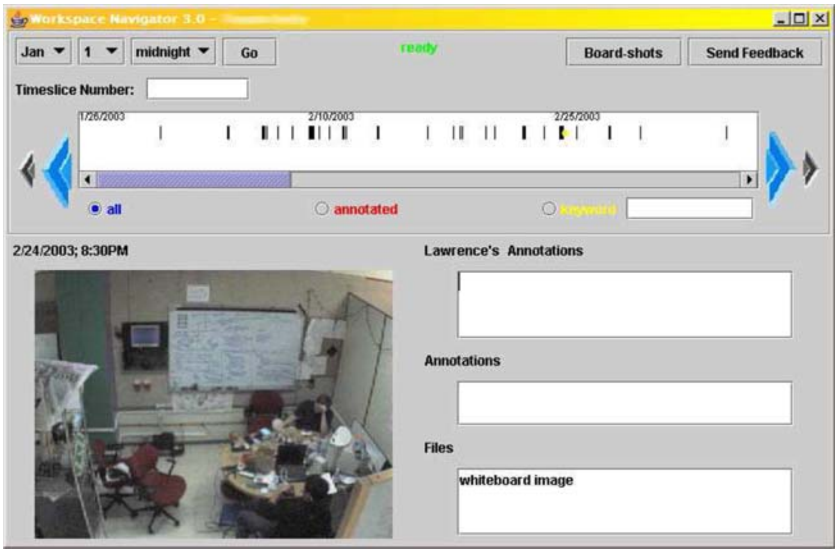


Fig. 7 WorkspaceNavigator user interface

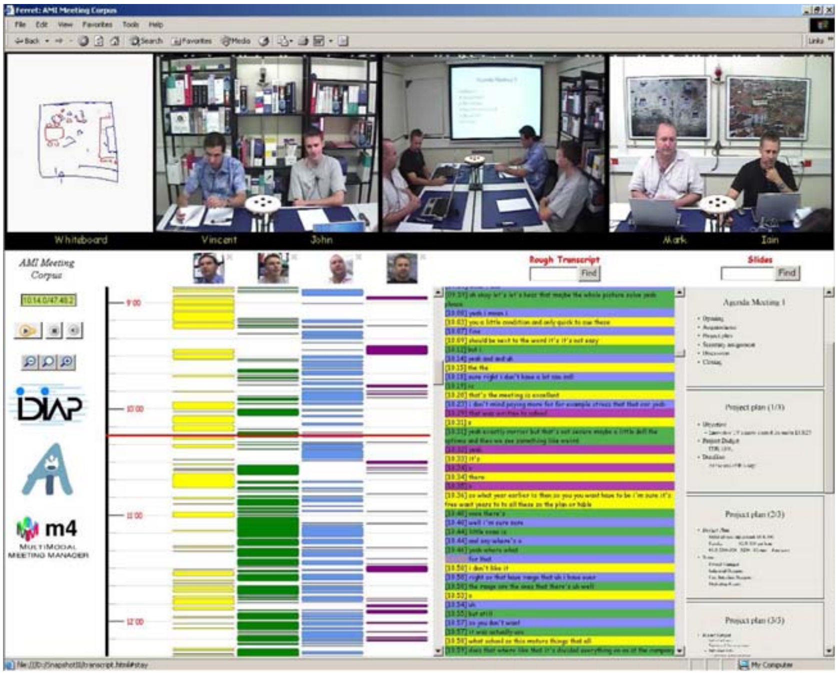
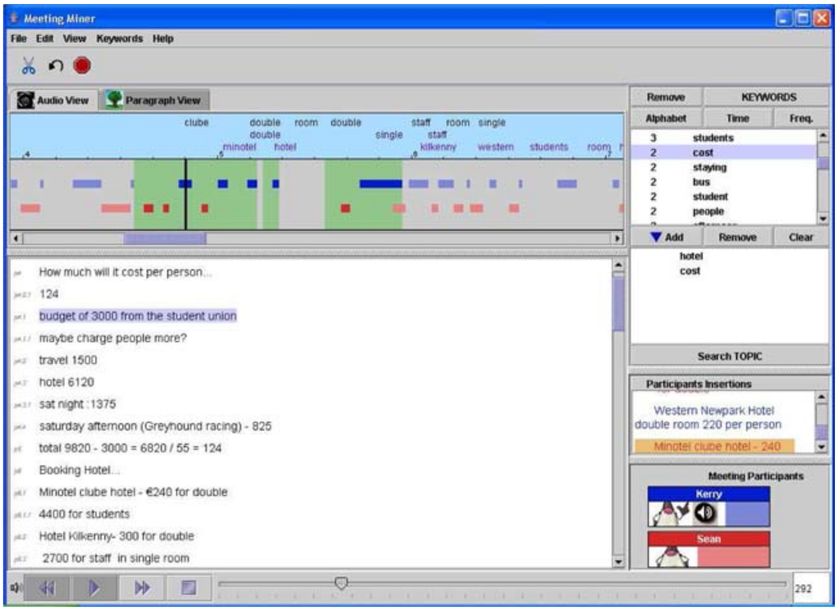


Fig. 8 The Ferret media browser

Conference) tracks [23] set precise evaluation tasks, with specific evaluation metrics on well-defined corpus collections, the diversity of multimodal meeting recordings and browsing strategies makes defining evaluation metrics and system comparison impractical. System comparisons have been confined to assessments of information retrieval performance on multimodal meeting browsers against a baseline system, typically one based on a tape-recorder interface metaphor, as in [74]. A more common, less constrained but inherently less comparable approach is evaluation by usability testing. Tasks employed in usability testing have included using the browser for identifying the topic of a conversation [81], classifying media items into pre-defined categories [69], answering specific questions about meetings (quiz) [18], locating specific information items [5], and producing meeting summaries. Evaluation focuses on user feedback such as ranking of features of the user interface according to perceived usability [5] and overall impressions on system performance [68] which give a good indication of how fit a system is for general use. In [67], manual minutes generated by a scribe during the meetings are used as a benchmark. Automatically generated meeting summaries were analysed to quantify the



3	students
2	cost
2	staying
2	bus
2	student
2	people

hotel
cost

How much will it cost per person...
124
budget of 3000 from the student union
maybe charge people more?
travel 1500
hotel 6120
sat night : 1375
saturday afternoon (Greyhound racing) - 825
total 9820 - 3000 = 6820 / 55 = 124
Booking Hotel...
Minotel clube hotel - €240 for double
4400 for students
Hotel Kilkenny- 300 for double
2700 for staff in single room

Western Newpark Hotel
double room 220 per person
Minotel clube hotel - 240

Fig. 9 The Meeting Miner

number of significant events they were able to convey. Information contained in the minutes which could not be inferred from the complete meeting recording (e.g. information external to the meeting or outside camera range) was not taken into account in the performance measure as it could not possibly have been included in the video summary.

Recently, more systematic approaches to comparing meeting browser performance have started to emerge. A strategy is described in [73] which suggests using the number of *observations of interest* uncovered by system users in a certain period of time as an evaluation metric. A test is proposed which can be described as follows. Human observers review information of interest in recorded meetings. Test subjects are then asked to answer as many (true–false) questions as they can in a period corresponding to half the duration of the meeting. Although this metric is general enough to be used by most meeting browsers, it alone does not solve issues relating to the diversity of corpora and access modalities, and therefore does not suffice for performance comparison. Standard meeting corpora, such as the ICSI meeting corpus [34], have become available which might help alleviate this problem. A crucial issue relating to usefulness and structure in browsing tasks is that of how to locate and "salvage" (recover with the purpose of creating a summary of the meeting) observations of interest. An interesting study on this issue is presented in [51]. Although that investigation is set in the context of designing meeting capture support tools, it offers valuable insight into how meeting browsers can be evaluated.

7 Conclusion

We have presented an overview of existing methods for segmentation, indexing and searching of captured multimedia meeting data and introduced a number of browsing systems, underlining their individual approaches to integrating various modalities for navigation of meeting recordings. Multimodal meeting browsing is currently an active research area with many open issues. While it would have been impossible to review all contributions in this area, we hope this survey will prove useful in indicating general trends in multimedia search and retrieval, information visualisation, seamless integration of multiple modalities, meeting interaction modelling, and evaluation methods which are essential to today's meeting browsing systems.

Acknowledgments The authors would like to thank the Meeting Browser, ScanMail, Video Manga, Portable Meeting Recorder: MuVie, Teamspace-MeetingViewer, WorkspaceNavigator and Ferret development teams for kindly giving us permission to reproduce screenshots of their systems. We would also like to thank the reviewers of the article for their valuable comments and suggestions. This work has been supported by Enterprise Ireland through a Basic Research Grant.

References

1. Aigrain, P., Zhang, H., Petkovic, D.: Content-based representation and retrieval of visual media: a state-of-the-art review. Multimed. Tools Appl. **3**, 179–202 (1996)
2. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: final report. In: Proceedings of the DARPA broadcast news transcription and understanding workshop (1998)

3. Arons, B.: Techniques, perception, and applications of time-compressed speech. In: Proceedings of conference of American voice I/O society, pp. 169–177 (1992)
4. Arons, B.: Speechskimmer: a system for interactively skimming recorded speech. ACM Trans. Comput. Hum. Interact. **4**(1), 3–38 (1997)
5. Boreczky, J., Girgensohn, A., Golovchinsky, G., Uchihashi, S.: An interactive comic book presentation for exploring video. In: Proceedings of CHI'00: human factors in computing systems, pp. 185–192. ACM Press (2000)
6. Bouamrane, M.M., Luz, S.: Navigating multimodal meeting recordings with the Meeting Miner. In: Proceedings of flexible query answering systems, FQAS'2006, LNCS, vol. 4027, pp. 356–367. Springer, Berlin Heidelberg New York (2006)
7. Bouamrane, M.M., Luz, S., Masoodian, M., King, D.: Supporting remote collaboration through structured activity logging. In: Hai Zhuge, G.C.F. (ed.) Proceedings of 4th international conference on grid and cooperative computing, GCC 2005, LNCS, vol. 3795, pp. 1096–1107. Springer, Berlin Heidelberg New York (2005)
8. Brotherton, J.A., Bhalodia, J.R., Abowd, G.D.: Automated capture, integration, and visualization of multiple media streams. In: Proceedings of the international conference on multimedia computing and systems, ICMCS '98, p. 54. IEEE Computer Society (1998)
9. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st ACM sigir conference on research and development in information retrieval, SIGIR '98, pp. 335–336. ACM Press (1998)
10. Chen, F., Withgott, M.: The use of emphasis to automatically summarize a spoken discourse. In: Proceedings of IEEE conference on acoustics, speech, and signal processing, ICASSP'92, vol. 1, pp. 229–232 (1992)
11. Chiu, P., Boreczky, J., Girgensohn, A., Kimber, D.: Liteminutes: an Internet-based system for multimedia meeting minutes. In: Proceedings of the 10th international conference on World Wide Web, WWW '01, pp. 140–149. ACM Press (2001)
12. Chiu, P., Kapuskar, A., Reitmeier, S., Wilcox, L.: NoteLook: taking notes in meetings with digital video and ink. In: Proceedings of the 7th ACM international conference on multimedia (Part 1), MULTIMEDIA '99, pp. 149–158. ACM Press (1999)
13. Chiu, P., Kapuskar, A., Wilcox, L., Reitmeier, S.: Meeting capture in a media enriched conference room. In: CoBuild '99: Proceedings of the 2nd international workshop on cooperative buildings, integrating information, organization, and architecture, pp. 79–88. Springer, Berlin Heidelberg New York (1999)
14. Choi, J., Hindle, D., Pereira, F., Singhal, A., Whittaker, S.: Spoken content-based audio navigation (SCAN). In: Proceedings of the ICPhS-99 (1999)
15. Cutler, R., Rui, Y., Gupta, A., Cadiz, J.J., Tashev, I., wei He, L., Colburn, A., Zhang, Z., Liu, Z., Silverberg, S.: Distributed meetings: a meeting capture and broadcasting system. In: ACM multimedia, pp. 503–512. ACM Press (2002)
16. Dharanipragada, S., Roukos, S.: A multistage algorithm for spotting new words in speech. IEEE Trans. Speech Audio Process. **10**(8), 542–550 (2002)
17. Dionisio, J.D.N., Cardenas, A.F.: Unified data model for representing multimedia, timeline, and simulation data. IEEE Trans. Knowl. Data Eng. **10**(5), 746–767 (1998)
18. Erol, B., Lee, D.S., Hull, J.J.: Multimodal summarization of meeting recordings. In: Proceedings of international conference on multimedia and expo, ICME '03, vol. 3, pp. 25–28 (2003)
19. Erol, B., Li, Y.: An overview of technologies for e-meeting and e-lecture. In: IEEE international conference on multimedia and expo, pp. 1000–1005 (2005)
20. Foote, J.: An overview of audio information retrieval. In: ACM multimedia systems, vol. 7, pp. 2–10 (1999)
21. Furui, S.: Automatic speech recognition and its application to information extraction. In: Proceedings of the 37th annual meeting of the association for computational linguistics, pp. 11–20. ACL (1999)
22. Furui, S.: Robust methods in automatic speech recognition and understanding. In: Proceedings EUROSPEECH, vol. III, pp. 1993–1998 (2003)
23. Garofolo, J.S., Voorhees, E.M., Auzanne, C.G., Stanford, V.M.: Spoken document retrieval: 1998 evaluation and investigation of new metrics. In: Proceedings of ESCA ETRW on accessing information in spoken audio, pp. 1–7 (1999)
24. Geyer, W., Richter, H., Abowd, G.D.: Making multimedia meeting records more meaningful. In: Proceedings of international conference on multimedia and expo, ICME '03, vol. 2, pp. 669–672 (2003)
25. Geyer, W., Richter, H., Fuchs, L., Frauenhofer, T., Daijavad, S., Poltrock, S.: A team collaboration space supporting capture and access of virtual meetings. In: Proceedings of the 2001 international conference on supporting group work, GROUP '01, pp. 188–196. ACM Press (2001)
26. Gibbs, S., Breiteneder, C., Tsichritzis, D.: Data modeling of time-based media. ACM SIGMOD Record. **23**(2), 91–102 (1994)
27. Goldman, J., Renals, S., Bird, S., de Jong, F., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D., Stewart, C., Wright, R.: Accessing the spoken word. Int. J. Digit. Libr. **5**(4), 287–298 (2005)
28. Hanjalic, A.: Generic approach to highlights extraction from a sport video. In: Proceedings of international conference on image processing, ICIP 2003, vol. 1, pp. 1–4. IEEE Press (2003)
29. Hearst, M.A.: Multi-paragraph segmentation of expository text. In: Proceedings of the 32nd annual meeting of the association for computational linguistics, pp. 9–16. ACL (1994)
30. Hindus, D., Schmandt, C.: Ubiquitous audio: capturing spontaneous collaboration. In: Proceedings of the 1992 ACM conference on computer-supported cooperative work, CSCW '92, pp. 210–217. ACM Press (1992)
31. Hirschberg, J., Whittaker, S., Hindle, D., Pereira, F., Singhal, A.: Finding information in audio: a new paradigm for audio browsing and retrieval. In: Mani, I., Maybury, M.T. (eds.) Proceedings of the ESCA workshop: accessing information in spoken audio, pp. 117–122 (1999)
32. Jaimes, A., Omura, K., Nagamine, T., Hirata, K.: Memory cues for meeting video retrieval. In: CARPE'04: Proceedings of the the 1st ACM workshop on continuous archival and retrieval of personal experiences, pp. 74–85. ACM Press (2004)
33. James, D.A., Young, S.J.: A fast lattice-based approach to vocabulary independant worspotting. In: Proceedings of international conference on acoustics, speech, and signal processing, ICASSP-94, vol. 1, pp. 377–380 (1994)
34. Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., Wrede, B.: The ICSI meeting project: resources and research. In: NIST ICASSP meeting recognition workshop (2004)
35. Ju, W., Ionescu, A., Neeley, L., Winograd, T.: Where the wild things work: capturing shared physical design workspaces. In: CSCW '04: Proceedings of the 2004 ACM conference on computer supported cooperative work, pp. 533–541. ACM Press (2004)

36. Jurafsky, D., Martin, J.H.: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice-Hall, Englewood Cliffs (2000)
37. Koumpis, K., Renals, S.: Content-based access to spoken audio. IEEE Signal Proc. Mag. **22**(5), 61–69 (2005)
38. Lee, D.S., Erol, B., Graham, J., Hull, J.J., Murata, N.: Portable meeting recorder. In: Proceedings of the 10th ACM international conference on multimedia, MULTIMEDIA '02, pp. 493–502. ACM Press (2002)
39. Lee, D.S., Hull, J., Erol, B., Graham, J.: Minuteaid: multimedia note-taking in an intelligent meeting room. In: IEEE international conference on multimedia and expo, vol. 3, pp. 1759–1762. IEEE Press (2004)
40. Li, F.C., Gupta, A., Sanocki, E., wei He, L., Rui, Y.: Browsing digital video. In: CHI '00: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 169–176. ACM Press (2000)
41. Luz, S.: Interleave factor and multimedia information visualisation. In: Sharp, H., Chalk, P. (eds.) Proceedings of human computer interaction, vol. 2, pp. 142–146 (2002)
42. Luz, S., Masoodian, M.: A mobile system for non-linear access to time-based data. In: Proceedings of the working conference on advanced visual interfaces, AVI '04, pp. 454–457. ACM Press (2004)
43. Luz, S., Masoodian, M.: A model for meeting content storage and retrieval. In: Proceedings of the 11th international multimedia modelling conference, MMM'05, pp. 392–398 (2005)
44. Luz, S., Roy, D.: Meeting browser: a system for visualising and accessing audio in multicast meetings. In: Society, I.S.P. (ed.) Proceedings of the international workshop on multimedia signal processing (1999)
45. Martinez, J., Koenen, R., Pereira, F.: MPEG-7: the generic multimedia content description standard, part 1. IEEE Multimedia **9**(1070-986X), 78–87 (2002)
46. Masoodian, M., Luz, S.: Comap: A content mapper for audio-mediated collaborative writing. In: Smith, M.J., Savendy, G., Harris, D. Koubek, R.J. (eds.) Usability evaluation and interface design, vol. 1, pp. 208–212. Lawrence Erlbaum, Hillsdale (2001)
47. Masoodian, M., Luz, S., Bouamrane, M.M., King, D.: Recoled: A group-aware collaborative text editor for capturing document history. In: Proceedings of WWW/Internet 2005, vol. 1, pp. 323–330 (2005)
48. Masoodian, M., Luz, S., Weng, C.: Hanmer: A mobile tool for browsing recorded collaborative meeting contents. In: Kemp, E., Philip, C., Wong, W. (eds.) Proceedings of CHI-NZ '03, pp. 87–92. ACM Press (2003)
49. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 305–317 (2005)
50. Meghini, C., Sebastiani, F., Straccia, U.: A model of multimedia information retrieval. J. ACM **48**(5), 909–970 (2001)
51. Moran, T.P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., Zellweger, P.: "I'll get that off the audio": a case study of salvaging multimedia meeting records. In: Proceedings of ACM conference on human factors in computing systems, CHI 97, vol. 1, pp. 202–209 (1997)
52. Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall, Englewood Cliffs (1993)
53. Richter, H.A., Abowd, G.D., Geyer, W., Fuchs, L., Daijavad, S., Poltrock, S.E.: Integrating meeting capture within a collaborative team environment. In: Proceedings of UbiComp '01, pp. 123–138. Springer, Berlin Heidelberg New York (2001)
54. Rohlicek, J., Russell, W., Roukos, S., Gish, H.: Continuous hidden Markov modeling for speaker-independent word spotting. In: Proceedings of international conferenceof acoustics, speech, and signal processing, ICASSP-89, vol. 1, pp. 627–630 (1989)
55. Rose, R.C., Paul, D.B.: A hidden Markov model based keyword recognition system. In: Proceedings of international conference on acoustics, speech, and signal processing, ICASSP-90, vol. 1, pp. 129–132 (1990)
56. Roy, D., Malamud, C.: Speaker identification based text to audio alignment for an audio retrieval system. In: Proceedings of the 1997 IEEE international conference on acoustics, speech, and signal processing, ICASSP '97, vol. 2, pp. 1099–1102. IEEE Computer Society (1997)
57. Russell, D.M.: A design pattern-based video summarization technique: moving from low-level signals to high-level structure. In: HICSS '00: Proceedings of the 33rd Hawaii international conference on system sciences, vol. 3, p. 3048. IEEE Computer Society (2000)
58. Santini, S., Gupta, A., Jain, R.: Emergent semantics through interaction in image databases. IEEE Trans. Knowl. Data Eng. **13**(3), 337–411 (2001)
59. Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur, G.: Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun. **32**(1–2), 127–154 (2000)
60. Singh, R., Li, Z., Kim, P., Pack, D., Jain, R.: Event-based modeling and processing of digital media. In: Proceedings of CVDB'04: computer vision meets databases, pp. 19–26. ACM Press (2004)
61. Smith, M.A., Kanade, T.: Video skimming and characterization through the combination of image and language understanding techniques. In: Proceedings of workshop on content-based access of image and video database, pp. 61–70. IEEE Computer Society (1998)
62. Snoek, C.G.M., Worring, M.: Multimodal video indexing: a review of the state-of-the-art. Multimed. Tools Appl. **25**(1), 5–35 (2005)
63. Srinivasan, S., Ponceleon, D., Amir, A., Petkovic, D.: What is in that video anyway?: in search of better browsing. In: Proceedings of IEEE conference on multimedia computing and systems, vol. 1, pp. 388–393 (1999)
64. Stifelman, L., Arons, B., Schmandt, C.: The audio notebook: paper and pen interaction with structured speech. In: Proceedings of CHI'01: Human factors in computing systems, pp. 182–189. ACM Press (2001)
65. Tucker, S., Whittaker, S.: Accessing multimodal meeting data: systems, problems and possibilities. In: Samy Bengio, H.B. (ed.) Machine learning for multimodal interaction: first international workshop, MLMI 2004, vol. 3361, pp. 1–11. Springer, Berlin Heidelberg New York (2005)
66. Tur, G., Hakkani-Tur, D., Stolcke, A., Shriberg, E.: Integrating prosodic and lexical cues for automatic topic segmentation. Comput. Linguist. **27**(1), 31–57 (2001)
67. Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video manga: generating semantically meaningful video summaries. In: MULTIMEDIA '99: Proceedings of the 7th ACM international conference on multimedia (Part 1), pp. 383–392. ACM Press (1999)
68. Valenza, R., Robinson, T., Hickey, M., Tucker, R.: Summarisation of spoken audio through information extraction. In: Proceedings of the ESCA workshop: accessing information in spoken audio, pp. 111–115 (1999)
69. Waibel, A., Bett, M., Finke, M., Stiefelhagen, R.: Meeting browser: tracking and summarizing meetings. In: Penrose,

D.E.M. (ed.) Proceedings of the broadcast news transcription and understanding workshop, pp. 281–286. Morgan Kaufmann (1998)

70. Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., Zechner, K.: Advances in automatic meeting record creation and access. In: Proceedings of the international conference on acoustics, speech and signal processing, pp. 597–600 (2001)
71. Weintraub, M.: Keyword-spotting using SRI's decipher large-vocabulary speech-recognition system. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, ICASSP-93, vol. 2, pp. 463–466 (1993)
72. Wellner, P., Flynn, M., Guillemot, M.: Browsing recorded meetings with Ferret. In: Bengio, S., Bourlard, H. (eds.) Proceedings of machine learning for multimodal interaction: first international workshop, MLMI 2004, vol. 3361, pp. 12–21. Springer, Berlin Heidelberg New York (2004)
73. Wellner, P., Flynn, M., Tucker, S., Whittaker, S.: A meeting browser evaluation test. In: CHI '05 extended abstracts on human factors in computing systems, pp. 2021–2024. ACM Press (2005)
74. Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., Singhal, A.: Scan: designing and evaluating user interfaces to support retrieval from speech archives. In: Proceedings of the 22nd ACM SIGIR conference on research and development in information retrieval, SIGIR'99, pp. 26–33. ACM Press (1999)
75. Whittaker, S., Hirschberg, J., Nakatani, C.H.: Play it again: a study of the factors underlying speech browsing behavior. In: CHI '98: CHI 98 conference summary on human factors in computing systems, pp. 247–248. ACM Press (1998)
76. Whittaker, S., Hyland, P., Wiley, M.: Filochat: handwritten notes provide access to recorded conversations. In: Proceedings of the ACM conference on human factors in computing systems, pp. 24–28. ACM Press (1994)
77. Wilcox, L., Kimber, D., Chen, F.: Audio indexing using speaker identification. In: Proceedings of conference on automatic systems for the inspection and identification of humans, pp. 149–157 (1994)
78. Yamron, J., Carp, I., Gillick, L., Lowe, S., van Mulbregt, P.: Event tracking and text segmentation via hidden Markov models. In: Proceedings of IEEE workshop on automatic speech recognition and understanding, pp. 519–526 (1997)
79. Young, S.: Large vocabulary continuous speech recognition: a review. In: Proceedings of the IEEE workshop on automatic speech recognition and understanding, pp. 3–28 (1995)
80. Zechner, K.: Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In: Procedings of the conference on research and development in information retrieval, SIGIR'01, pp. 199–207. ACM Press (2001)
81. Zechner, K., Waibel, A.: DiaSumm: flexible summarization of spontaneous dialogues in unrestricted domains. In: Proceedings of the 18th conference on computational linguistics, pp. 968–974. ACL (2000)