A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings.

Denis Lalanne, Stéphane Sire, Rolf Ingold, Ardhendu Behera,
Dalila Mekhaldi and Didier von Rotz
Fribourg University,
Chemin du Musée 3,
1700 Fribourg
Switzerland,
Denis.Lalanne@unifr.ch, Stephane.Sire@epfl.ch

Abstract

The recording, multimodal analysis and archiving of meetings introduce new challenges for research in multimedia information management. Meetings involve multiple media that can be aligned together. This requires a global annotation framework. In particular, meetings often deal with documents, either projected or discussed, which can be aligned with the audio and video streams. This article presents a research agenda for bridging the gap between documents, several types of annotations of documents, and multimodal annotations of audio and video streams. It also presents the task of authoring meeting minutes as a means to evaluate multimodal annotations and the alignment of documents with other media.

1. Introduction

Several research projects aim at archiving meeting recordings in suitable forms for later retrieval. The main purposes of these projects are to advance the research on automatic multimodal content analysis and on multimedia information retrieval. They have resulted into powerful browse-and-query user-interfaces. However, most of these projects do not take into account the classical printable documents that may also be part of the information available during a meeting. Further, these projects are usually missing a real end-user task that could greatly help to evaluate the utility of the annotations produced through the analysis of the audio and video streams.

Traditional meetings ends-up with the production of official "minutes", a textual document that summarizes the meeting. In our application our goal is to design a minutes authoring application that stores the minutes, which keep track of interesting meeting moments, as a special kind of annotation of the audio and video streams recorded from a meeting. We also want to link the minutes with the documents manipulated during a meeting. For that purpose we have done two hypotheses.

The first hypothesis is that finding links between documents and multimodal annotations of the audio and video streams will permit the design of user-interfaces that improve the production of minutes as well as any kind of retrieval tasks. The second hypothesis is that the previous user-interfaces will also be improved if they can visualize the annotations and the alignments between the different kinds of media available into the database.

In this article, we first discuss the existing meeting room projects and present our document-centric meeting room. Then we describe the documents that we take into account and the methods that we have designed for aligning these documents with the other meeting modalities. We also present our first alignment results. Further, we explain the meeting scenario we have chosen into which documents have got a central role. Finally, we define the concept of meeting minutes authoring tool that integrates the annotations of multiple media and some visualizations of their relationships. We claim that this application is a powerful framework for evaluating multimodal annotations. This is the framework we are planning to use for validating our own research on automatic document analysis and alignment.

	Type	Camera number	Simple Streams	Control Streams	Derived Streams	Higher level annotations	Reference
ISL	Meetings	not given	A/V	none	spoken word transcript	utterances (speech acts) speech topics emotions participant's identification	(Bett, 00)
MS	Meetings	5	A/V	pen-strokes	none	participant's identification	(Cutler, 02)
FXPal	Meetings and lectures	3	A/V	pen-strokes	slide changes	none	(Chui, 00)
eClass	Lectures	1	A/V	pen-strokes	phoneme transcript slide changes	none	(Brotherton, 98)
DSTC	Lectures	1	A/V	none	slide changes	slide keywords	(Hunter, 01)
Cornell	Lectures	2	A/V	none	slide changes	none	(Mukhopadhyay, 99)

Table 1: Summary of related meeting capture projects.

2. Related meeting room applications

Media enabled meeting rooms are equipped with sensors for recording multiple streams of different media during a meeting. The most common sensors are:
- microphones (omnidirectionnal mic., directionnal mic., mic. arrays, ceilling mic.);
- cameras (analogical video cameras, firewire webcams, camera "rings");
- projection screen recording (camera, direct video projector output);
- whiteboard recording (digital ink capture devices such as SmartBoard™ or Mimio board™);
- pen-computer recording (handwritten notes on notepad computers).

The table below summarizes 6 meeting capture projects. They differ by the number of camera used and by the targeted meeting types. Some projects only use 1 or 2 cameras as they are targeted at recording lectures where one presenter faces an audience. In that situation, a presenter head-and-shoulders view is recorded together with a whiteboard overview for detecting slide changes. For meetings with multiple participants seating around a table, one or more cameras take overviews of the meeting table. The MS project is the only one to record a head-and-shoulders view of every participant.

The continuous streams recorded with the different kind of devices cited above are processed in real-time or off-line. It is convenient to categorize them into three types (Brotherton, 98): simple streams, controls streams and derived streams.

Simple streams are captured for the sole purpose of unmodified playback. The simple streams consist of audio and video recordings of meeting participants, with a video recording of the projection screen in some cases. Control streams are generated on the fly through special devices that capture time stamped information. They can be used to index other streams. The only type of control stream we have found, also called digital ink, is a recording of the pen-strokes of meeting participants, either on a shared whiteboard or on individual notebook computers. Some standardization efforts are under way for the representation of digital ink. Derived streams are produced by analyzing other streams after the live event. The derived streams are either phoneme or spoken word transcriptions of audio streams and slide changes time line derived from the projection screen video output. The table 1 summarizes the streams and annotations of the reviewed projects.

Meeting data is processed to produce annotations. The annotations are used to retrieve segments of the media, which are played back in multimedia clips. The following terms refer to the previous processes (Chui, 00):
- **segmentation**: partitioning of continuous media into homogenous regions
- **annotation**: association of text data with particular time locations of a media
- **retrieval**: selection and extraction of segments of a media based on the annotations associated with the segments

The segmentation process in the cited projects applies to the temporal dimension of the simple streams. Most annotations are in fact stored in the control and derived streams. The segments boundaries are defined with time

stamps. In the DSTC project for instance, the slide changes derived stream is stored as VideoSegmentType element in an XML file, StartTime and EndTime elements define the segment; a Slide element identifies the slide and a Keyword element gives some keywords extracted from the title of the slides. In the same way spoken word transcripts are text file that contain the recognized words with time stamps indicating the start time and end time for groups of words, which are attributed to a speaker. In some cases higher-level annotations are derived from all the other types of streams. In the ISL project spoken work transcripts are segmented into utterances, which are labeled with a speech act type such as « question » or « opinion ». Participant identification is also deduced from a skin-color based face tracker. The retrieval process is highly dependent on the user-interfaces.

Two groups of meeting room systems emerge from this quick overview. They differ in the type of user interfaces they support for retrieval:

The first group is focused on document related annotations such as handwriting and slide analysis: MS (Cutler, 02), FXPal (Chui, 00), eClass (Brotherton, 98), DSTC (Hunter, 01) and Cornell (Mukhopadhyay, 99). It proposes meeting-browser interfaces based on visualizations of the slide changes time line, and of the notes taken by participants. In these interfaces, slides and notes are used as quick visual indexes for locating relevant meeting parts and for triggering their playback.

The second group of systems is based on speech related annotations such as the spoken word transcript: ISL (Bett, 00) and eClass (Brotherton, 98). It proposes meeting-browser interfaces based on keyword search in these transcripts. In that context, higher-level annotations such as speech acts or thematic episodes can also be used to display quick indexes of selected meeting parts.

The document-centric and the speech-centric applications correspond respectively to the visual and to the verbal communication channels of a meeting. However, none of the above projects is considering both channels at the same time. These means of communication being integrated in real life, we propose to integrate them in meeting archives and into related user-interfaces.

3. Capture environment

Our meeting room is equipped with 10 cameras (8 close-ups, one per participant, 2 overviews), 8 microphones, a video projector, a projection screen and a camera for slide capture. The equipment is lightweight (PCs with firewire webcams) and not intrusive. Camera and microphone pairs' synchronization is guarantied on a per-computer basis. Due to the high bandwidth of each camera (8.8MB/s for 640x480, 15fps), we could not put all the cameras on a single PC. Therefore, we use 4 PCs remotely controlled and synchronized by a master PC, so
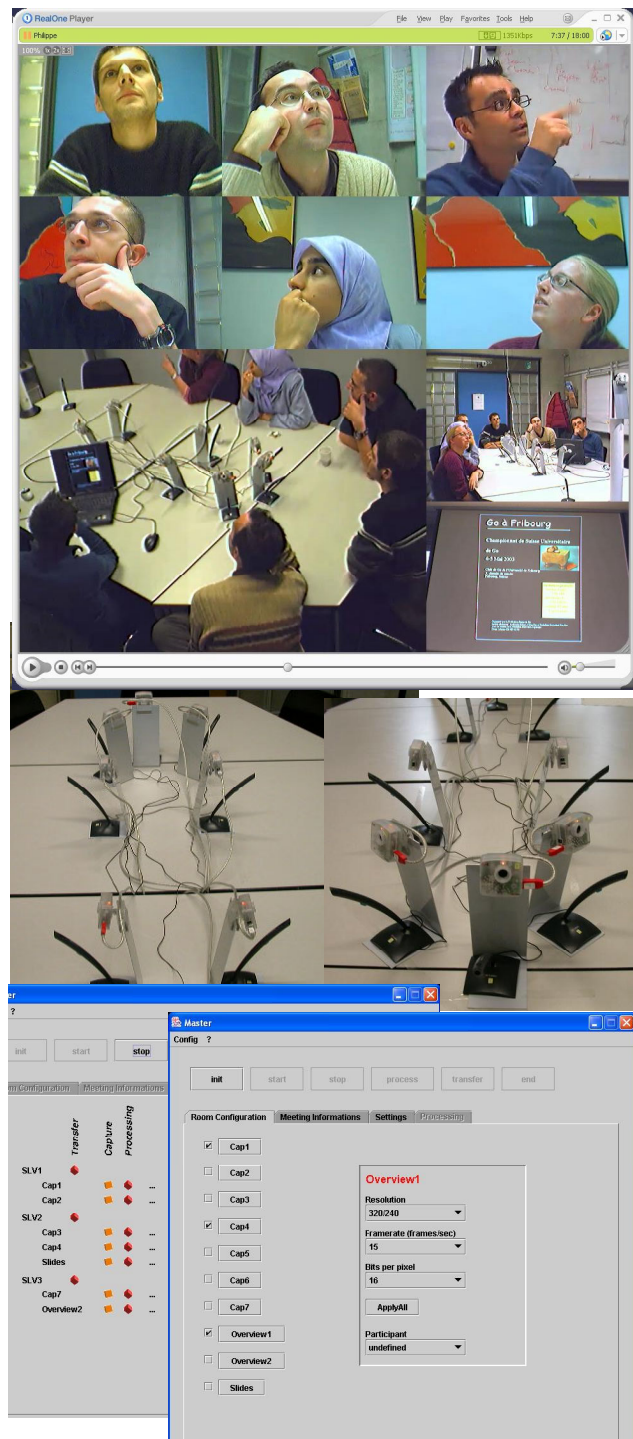

Figure 1 : Our meeting room, some video records and the capture application.

that recordings start simultaneously. The architecture is fully scalable.

A meeting capture application (figure 1), running on the master PC, pilots the slave PCs and their devices. It has a user-friendly interface to start, pause and stop recording, to control post-processing operations such as compression (for streaming and archiving) and to control

file transfers to a server. This application is part of a more general Organizer tool for specifying the cameras and microphones to be used, the participants' position, camera's frame-rate, etc.

The Organizer tool also assists users in the preparation, management and archiving of a meeting. This includes services for: a) registering meeting participants and related information, b) gathering documents before and after the meeting, c) distributing documents to participants, d) archiving documents and any relevant information. The Organizer tool could be extended in the future for real-time interaction with the room during a meeting (roomware and augmented meeting room).

4. Full document integration and alignment

We present in this section different types of documents handled by meeting participants. Some of them **have not yet been fully considered for inclusion into meeting archives** in the previously cited projects. This is because they do not provide immediate means for being time stamped in reference to a global meeting time clock. We will see that document content and image analysis provide such means.

Meeting participants have at least three ways to interact with documents during a meeting. They can 1) take handwritten notes on paper, notebook computers, electronic or classical whiteboards; 2) distribute and share printed documents; 3) project documents onto a shared display. In all the cases above, participants may also verbally discuss documents. All the previous interactions with documents can be described at different document granularity levels. Each of the descriptions holds a relationship with the meeting time that depends on the type of document:

- Handwritten notes can easily be time stamped at the pen strokes level if participants are willing to use special devices such as Anoto pens or tablet PCs or special whiteboard devices.
- Classical printed documents are usually discussed during a meeting, and thus explicitly appear in the speech focus. As the speech transcript contains temporal indexes, if some matches can be found between document extracts' content and the speech transcript, then it's a mean to bring temporality to those document extracts.
- Projected documents are not only discussed but they also appear at specific time in the visual focus, which can be recorded with a camera. Matches between the visual context and some images of document extracts, will convey temporality to those document extracts.

We call "document temporal alignment" the operation of extracting the relationships between a document excerpt, at variable granularity levels, and the meeting time. In the presentation of the existing meeting room

projects we have seen that the document-centric applications have focused on temporal alignment of handwritten notes and of slides. Thus we focused our research on un-handled types of document and alignments:

- a) Temporal alignment of classical printed documents,
- b) Temporal alignment of projected documents at a finer grain than the slide level, to take into account scrolling, zooming and animating effects.
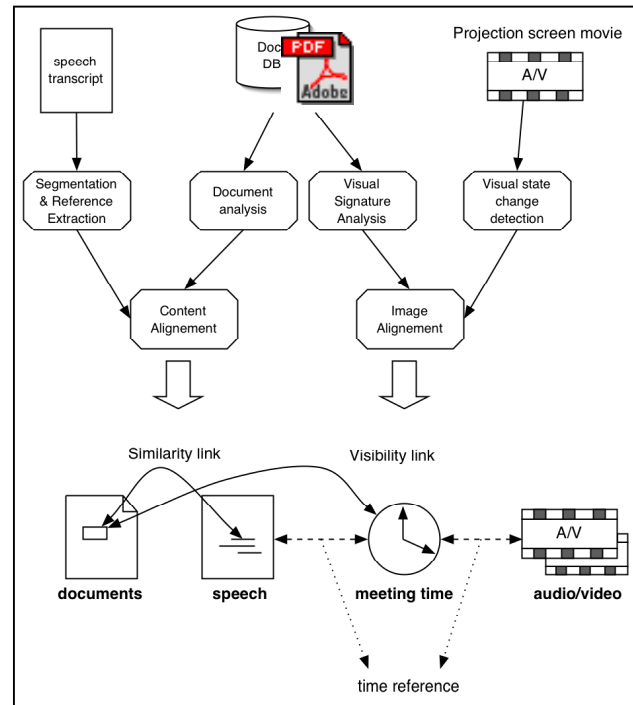

Figure 2: Document temporal alignment links document extracts with the time at which they enter into the speech focus (content alignment) and/or into the visual focus (image alignment).

These research objectives respectively deal with (see figure 2):

- a) Document **content** analysis for extracting temporal indexes through the alignment of documents with speech transcripts, which holds timestamps for each speech utterances and speaker turns.
- b) Document **image** analysis for extracting time stamps associated with visible state changes.

The methods describe bellow requires that 1) the documents are available in PDF, which we use as a pivot representation for document analysis, and that 2) the speech has been transcribed (with speaker turns and timestamps for each speech utterances).

Although PDF (Portable Document Format) is a standard for exchanging electronic documents, its format is relatively unstructured (similar to postscript). For further treatment, we are developing a PDF-to-Text tool that will convert a complex PDF document in a linear textual form

(in respect to the reading order) an we are also developing a PDF2XML tool that will provide the layout structure of the document (bounding box for each physical block).

4.1 Document content alignemnt

In document content analysis, textual content is matched with speech transcript in order to detect:

a) **Citation** alignments are pure lexicographic matches between terms (tuple) in documents and terms in the speech transcription (such as: "The author said << Recording and analyzing meetings introduce new challenges for researchers >>");

b) **Reference** alignments establish links between printed documents and structured dialogs through the references that are made to documents in speech transcript (such as: "the article written by XY", "the second paragraph", "the caption on the right side", "the excerpt in italic", "the editorial on SRAS" or a combination). There are two challenges related to this goal that can benefit from each other: (a) construct a multi-layered representation (physical, logical, relational, thematic, etc.) of printed documents, and (b) recognize the spoken words that refer to elements of the documents and link them accordingly (Popescu-Belis, 2003).

c) **Thematic** alignments are alignments between sections of documents (sentences, paragraphs, logical blocks, etc.) and the dialog structure of speech (utterances, turns, and thematic episodes).

We have already implemented various thematic alignments, using the cosine metric and considering document and speech units as bags of weighted words. We then compared the content of the following units and measured their similarities: logical block and sentences for document versus turns and utterances for speech. It gave back promising results based on a recall and precision evaluation (table 2). We did not consider yet thematic units for documents because the results of the thematic structure we extracted, using TextTiling (Hearst, 95), was not satisfactory for the type of document we are handling. In the future, treating in parallel the text-tiling and the alignment should improve both processes.

Recall and precision measures are relatively high when matching document logical blocks with speech utterances and turns. When the meeting discussion does not relate much with documents, these values are drastically falling (threshold must be tuned). In any case, the results are encouraging. However, logical blocks are currently manually segmented.

On the other hand, sentences/utterances alignment is less precise but it is based on a fully automatic production. The PDF documents are automatically converted in their textual form, further segmented in sentences, and finally matched with the speech utterances. We believe that those simple automatic alignments can help structuring the documents and improve existing

algorithms for discovering the logical structure of documents (Hadjar, 2001).

The alignment presented in this table considers only one best match for each alignment. However, all the units whose similarities overcome a certain threshold should be considered, especially when the source unit's size is higher than the targeted one. For example, more than one speech utterance can be matched with a document article. Considering all the possible units' matches make the alignment symmetrical. Indeed, the relationships detected, overcoming a certain similarity threshold, are the same in both alignment directions.

Further, documents and speech transcript are both hierarchical. Thus, the similarities found between both trees' nodes will create a bi-polar graph. Finally, detecting the most connected regions in this graph should help discovering thematic regions.

	Pairs	R	P	F
Sentence/Utterance	1409	0.87	0.51	0.63
Sentence/Turn	1409	0.78	0.60	0.67
Utterance/Sentence	572	0.83	0.71	0.77
Utterance/Logical unit	572	0.84	0.77	0.80
Turn /Sentence	228	0.86	0.69	0.77
Turn/Logical unit	35	0.88	0.81	0.85

Table 2: Thematic alignment of document versus speech units.

4.2 Document image alignemnt

In document image analysis, low-resolution document images (such as video capture of projected slides) are matched against electronic image signatures of document available in a database. Different algorithms have already been described that identify slides (Mukhopadhyay, 1999) or any document. We are working on extending them to:

a) Document identification and partial document identification.

b) Detection and identification of fine grain state change (animation, scrolling, zooming, etc) and

c) Identification of occluded documents (speaker in front for instance) and identification of pointed document parts.

We have already implemented a new technique for detecting slide changes based on stability detection rather than change detection, i.e. the period during which the slideshow display the same document image.

First of all, our algorithm slices the slideshow movie in several queues of N frames. Then, the N frames are converted to gray scale images, low-passed filtered for noise reduction, and finally converted to binary image by global thresholding. The first frame in the queue is then compared with the other frames in the queue in order to calculate a dissimilarity value. When this value overcomes a certain threshold, based on statistical calculations (see [Behera, 2003] for more technical details), a slide change is signaled.

In our case, the queue of frames corresponds to 2 seconds of video. Our assumption is that there is no relevant slide change during the 2 seconds following a slide change. Indeed, in real world presentation, slides that are visible less than 2 seconds should not be considered because people do not have time to read them.

The fade-in fade-out transition can generate slide images overlapping, i.e. the combination of two successive slide images. This is due to the relatively high frame rate of our acquired video. Thus, in order to detect the precise slide change, we must take into account the first non-overlapped frame, just before the stabilization phase due to the auto-focusing nature of webcam.

Furthermore, when a slide change is signaled, the animation detection method starts working. The main differences with the slide change detection method is that it uses a finer threshold and that each frame in the queue is compared with the previous and next frame, because animation is a local feature.

In order to evaluate our method, we built a corpus compound of more than 3 thousands slides (in PDF) and of 30 movies of projected slideshows. The slideshows were automatically produced using SMIL (actually realPix from realmedia), with randomly picked up images and randomly generated timestamps for the slide changes, which provided us a perfect ground-truth.

We compared the performance of our new technique with all other existing methods: Cornell (Mukhopadhyay, 99), color and gray histograms. It shows the best results on a recall and precision basis (figure 3), mainly because of the precision increase without having to perform slide identification (like in Cornell's method), which would imply a significant additional processing. This method also overcomes the auto focusing and non-uniform lightning nature of webcams.

Finally, considering the high performance of this new method, the slide show video can be used as a control stream, segmenting the meeting according to the documents in visible focus. In the near future, simple classification methods will be implemented for further document identification and slide thematic alignment.

5. A document-centric meeting scenario

The first step for validating the integration of documents into multimedia archives is to build corpuses of meeting recordings based on scenarios where participants have a high interaction with documents. We have designed some scenarios that address different types of meetings: structured meetings agenda driven, professor lectures with a class book, reading clubs, administrative meetings for writing text laws, etc. Finally, we have decided to concentrate our efforts on press reviews (i.e. meetings where participants discuss the cover page and the content of the newspapers of the day). Newspaper can be matched with speech transcript through citation, reference and thematic alignment. They contain small articles easy to
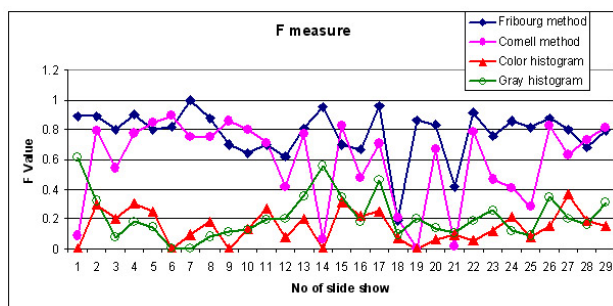

Figure 3. Slide change detection performance through a recall & precision evaluation.

segment. Thus, press reviews follow a structured agenda that should fit well document temporal alignment through document content alignment with speech transcripts. Further, when discussing different newspapers, thematic alignment can also be performed between the newspapers themselves. We have already recorded 20 meetings of about 10 minutes each. For each meeting a directory has been created on our media file server containing:

- The meeting global descriptor (XML file holding general information about the meeting, participants' name, position, A/V setup, location, date, etc);
- SMIL files to play all audio/video together.
- And the following directories:
  - o Audio: audio files for each participants;
  - o Video: one head-and-shoulder movie for each participant, one movie for each of the two overview cameras and the movie of the projected slides;
  - o Document: The PDF file of each discussed and/or projected document, and for each of those files: its linear textual version, and an XML file holding the manual logical structure;
  - o Speech: manual speech transcript and various other speech annotations;
  - o Image: one keyframe for each video.

In a near future we plan to integrate projected slides in our press reviews. This will be a necessary step in order to compare results on document temporal alignment from content analysis and results from image analysis.

6. Meeting minutes authoring

Document temporal alignment links document extracts with the time at which they enter into the speech focus and/or into the visual focus of a meeting. Thus it is possible to align document extracts with audio and video extracts, and by extension with any annotation of audio and/or video. This gives an opportunity to develop an application for assessing the utility of multimodal annotations and the usability of diverse visualizations of those annotations for a real end-user task. We have chosen the task of a scribe writing minutes, which offers a

manual ground-truth procedure. The assessment task consists in measuring the ability for a scribe, who did not attend the meeting, to produce minutes using the authoring tool and to compare them with the production of a scribe who attended the meeting. The quality of the production with or without a certain set of annotation could also be measured (for example speech acts).

Figure 4 represents a mock-up of the minutes authoring application (Photoshop montage with SMIL movie playbacks and a modified version of Transcriber in Tcl/Tk) with the following components: the kaleidoscope visualization, the documents in focus, the audio/video movies, the structured speech transcription and the audio signal. All the representations are synchronized, meaning they all have the same time reference. Clicking on one of them, for example an utterance of the speech transcript, causes all the components to visualize their content at the same time index.

The kaleidoscope visualization at the top left of figure

4 represents the complete meeting's duration. The green arrow turns around the ring like a needle around a clock watch. Each layer of the ring stands for a different temporal annotation: speech turns, document blocks, discourse types, silences, topics and dialog acts. Other annotations could be displayed depending on the meeting type (slide visible state change, pen-strokes for handwritten notes, etc.). Those temporal annotations are currently stored in the form of XML files, which hold timestamps for each state change (i.e. new speaker, new topic, etc.) and spatial information for documents. For example, 1) the document temporal annotation and 2) the speech transcript can be expressed as follow:

```
1) <TempDoc docref="lemonde040403">
      <zone  id="d2"  label="article_content"  startTime="8.2"
      endTime="45.3"      startChar="32"      endChar="245"
      fontsize="12"  fonttype="serif"  height="368"  width="100"
      x="124" y="132"/> …
   </TempDoc>
```


Figure 4 : A mock-up of the minutes application as we envision it, with the following components: the kaleidoscope visualization, the document in focus, the audio/video movies, the structured speech transcription and the audio signal. All those representations are synchronized.

```
2) <SpeechTrans>
    <Turn speaker="Denis" startTime="18.2" endTime="45.3"/>
      <Utterance startTime="18.2" endTime="20.5">
           bla bla bla
      </Utterance>…
    </Turn> …
  </SpeechTrans>
```

The kaleidoscope is a visual overview of the overall meeting. Further, it is interactive; the scribe can click on any part of a ring layer in order to access a specific moment of the meeting, a specific topic or a specific document article. Indeed, documents content is aligned with the speech transcript. Clicking on an article places the audio/video sequences at the moment when the content of this document block is being discussed and it highlights the most related articles in other documents. This is a direct illustration of text/speech alignment and of document/document alignment.

The kaleidoscope or other similar visualizations reveal some potential relationships between sets of annotations. For example, this particular visualization indicates that for each topic changes there are often silences just before, or a speaker change or a new document block being discussed. This kind of observations brings to light new methods that could improve the automatic generation of annotations.

The minutes authoring application comes along with a minutes' viewer application, which allows a person to read, skim or search the multimedia database through the meeting minutes. It contains two modules. The first one for generating on-the-fly several models of hypermedia meeting minutes (combination of document extracts, speech transcripts, audio and video snapshots, etc) to answer a user request. This is an extension of the video manga metaphor (Boreczki et al, 2000). The second one for creating multimedia presentations adapted to multi-speakers dialogs.

7. Conclusion and future work

We have seen in this article how to integrate multimodal annotations of meeting archives through temporal alignment. We have illustrated this method for the alignment of documents with other media. This integration, together with advances in visualizations of multimodal annotations, finds a natural application into the meeting minutes authoring framework. In the long term, we plan to use this framework to assess the utility and usability of different set of multimodal annotations, which are becoming common in meeting capture projects.

8. Aknowledgment

Refences

Bett, M., Gross, R., Yu, H. Zhu, X. Pan, Y., Yang, J. & Waibel, A. (2000) "Multimodal Meeting Tracker", Proceedings of RIAO2000, Paris, France.

Behera, A., Lalanne, D. and Ingold, R. Documents in visible focus: another path to multimedia meeting archives. Submitted to MIR 2003, 5th International workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia 2003.

Boreczky J., Girgensohn A., Golovchinsky G., & Uchihashi S. (2000) An Interactive Comic Book Presentation for Exploring Video. In CHI 2000 Proceedings, ACM Press.

Brotherton, J.A. Bhalodia, J.R. & Abowd, G.D. (1998) "Automated Capture, Integration, and Visualization of Multiple Media Streams", In the Proceedings of IEEE Multimedia '98.

Chiu, P., Kapuskar, A., Reitmeier, S. & Wilcox, L. (2000) "Room with a rear view. Meeting capture in a multimedia conference room", IEEE Multimedia, Volume 7 Issue 4.

Cutler, R. Rui, Y., Gupta, A., Cadiz, J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z. & Silverberg, S. (2002) "Distributed Meetings: a Meeting Capture and Broadcasting System", proceedings of the ACM Multimedia 2002 Conference.

Hadjar K., Hitz O., Ingold R. (2001), *Newspaper Page Decomposition Using a Split and Merge Approach*. ICDAR 2001: pp. 1186-1189.

Hearst M. (1995), TileBars: Visualization of Term Distribution Information in Full Text Information Access, Proceedings of the Conference on Human Factors in Computing Systems, CHI'95.

Hunter, J. & Little, S. (2001) "Building and indexing a distributed multimedia presentation archive using SMIL", ECDL'01, Darmstadt.

Mukhopadhyay, S. Smith, B. (1999) Passive capture and structuring of lectures, proceedings of the seventh ACM international conference on Multimedia (Part 1), Orlando, Florida.

Popescu-Belis A. (2003) - Evaluation-Driven Design of a Robust Reference Resolution System. *Natural Language Engineering*, vol. 9, n. 2, p.1-26.