Talking about documents: revealing a missing link to multimedia meeting archives.

Denis Lalanne, Dalila Mekhaldi and Rolf Ingold
Université de Fribourg
Chemin du Musée, 3
1700 Fribourg

ABSTRACT

In the context of multimedia meeting recordings and analysis, we introduce a new kind of multimedia alignment, which aims at reunifying documents with all kind of temporal media. The alignment proposed in this article uses the similarities that exist between the documents' content and the speech transcript's content in order to provide temporal indexes to printable documents. Several document content alignment strategies are discussed in this article and evaluated at various levels of granularity.

Keywords: Meeting recordings, multimodal analysis, document content alignment, document multi-layered structure.

1. INTRODUCTION

Classical documents (agenda, reports, transparencies, etc) constitute important information handled before, during and after meetings. They can be either discussed, or projected, or written, or pointed or simply visible on the meeting table. Therefore, they must be captured, indexed and integrated in a multimedia database. However, those documents have not yet fully been considered for inclusion into meeting archives. This is because they do not provide immediate means for being time stamped in reference to a global meeting time clock. We will see in this article that document content and image analysis provide such means. This task not only deals with document structure recognition and understanding, as well as content analysis, but more important with finding links with other medias (video, audio, etc.).
While document alignment, particularly multilingual [6], is a well-known problem, few studies exist of multimodal alignment. Further, multimedia synchronization, a task of coordinating various time-dependent media, has been explored in on-line education [1]. However, Documents are not time-dependent. We propose in this article a method for bridging the gap between documents and other media, through the alignment of documents' content and speech transcript's content. This bi-modal alignment (printed documents vs. speech transcripts) will combine research techniques from several domains: multilingual alignment [6], dialog analysis [19,20], document analysis [7,8,12,17], semantics and information retrieval [9,15]. Their alignment will allow building document-based multimedia browsing interfaces (see figure 1) and will improve query and retrieval of multimedia data. In general it will help answering two types of questions: (a) When were the various parts of a document discussed? (b) What was said about the documents or a part of it? We expect this work to boost automatic document structuring and temporal indexing thanks to information from real-time spoken interaction. Conversely, this work shall support the production of enriched transcriptions of meeting recordings, with explicit information about the contents and the timing of what was said about documents.
We first briefly present in this paper the current meeting room projects. Second we introduce the two different document alignments we have discovered. We then focus our presentation on document content alignment and more specifically on thematic alignments. Further, we detail our methods of segmentations and similarity measures. Finally we conclude with the evaluation of the various thematic alignments we have implemented.

2. RELATED MEETING ROOM PROJECTS

Two groups of meeting room systems emerge from a quick overview (see [14] for more details). They differ in the type of user interfaces they support for retrieval:

a) The first group is focused on document related annotations such as handwriting and slide analysis: MS [5], FXPal [4], eClass [3], DSTC [10] and Cornell [16]. 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.

b) The second group of systems is based on speech related annotations such as the spoken word transcript: ISL [2] and eClass [3]. 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. These channels being integrated in real life, we propose to bridge links between them and integrate them in meeting archives and into related user-interfaces (see figure 1). Further, we suggest considering both the visual and verbal links with documents in order to fully align them with temporal data.

3. FULL DOCUMENT INTEGRATION WITH TEMPORAL ALIGNMENT

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 of 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 a 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.

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 have decided to focus our research on temporal alignment of classical printed documents, and temporal alignment of projected documents at a finer grain than the slide level, to take into account scrolling, zooming and animating effects. We call "document temporal alignment" the operation of extracting the relationships between a document excerpt, at variable granularity levels, and the meeting time. There are two distinct ways to align documents with other temporal media:

- Document content analysis for extracting temporal indexes through the alignment of documents with speech transcripts,
- Document image analysis for extracting time stamps associated with visible state changes (projected documents).

Document content alignment gives information on which document content was being discussed within a specific time interval. Whereas, document image alignment provides information on which document content was visible within a specific range of time. Treating both document content and document image alignment will consolidate and enrich document temporal alignment. Further, the two alignments will support each other. The document image alignment can for example help segmenting the speech transcript in thematic episodes, whereas the document content alignment will reinforce the document image identification phase.

1.1. Document image alignment

In document image alignment, which is not the main purpose of this article, 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 [16] or any documents. 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),

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

1.2. Document content alignment

In this article we focus our presentation on the document content alignment, where the document tcontent is matched with the speech transcript in order to detect:

a) Citation alignments are pure lexicographic matches between terms in documents and terms in the speech transcription (such as: "The author said << …>>");

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", "the second paragraph in italic", "the caption on the right side", etc).

c) Thematic alignments are content-based similarity links between document units and the dialog structure of speech.

This article presents our results on the third type of alignment, i.e. thematic alignment, and the methods for:

a) Segmenting the documents and the speech transcript (section 4),

b) Measuring the similarity value between the pairs of units resulting of these segmentations (section 5).

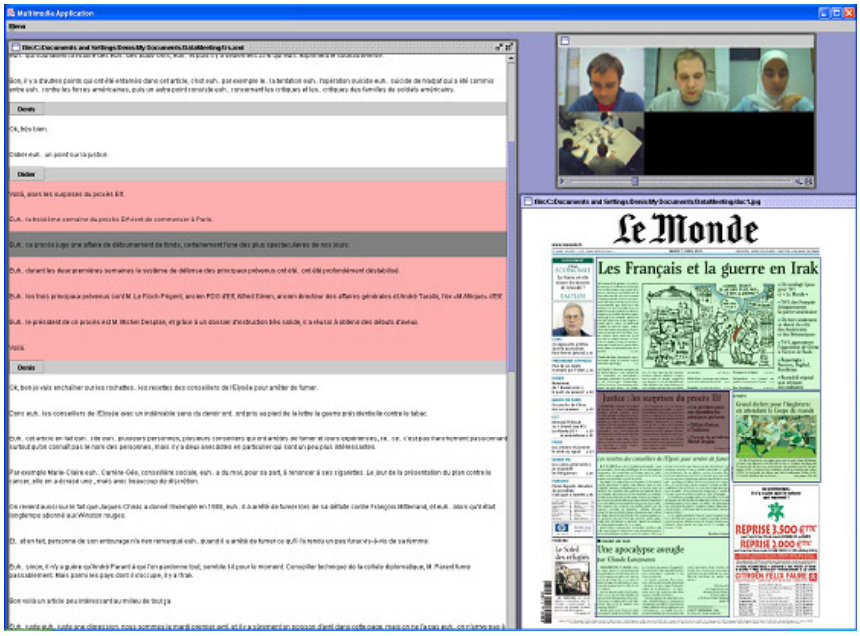

Figure 1: A first prototype of document-based multimedia browsing interface developed in Java (using Batik library for manipulating documents and JMF for audio/video). It highlights thematic links between document blocs and speech units. All the components are synchronized through the meeting time; clicking on a document article automatically plays the corresponding transcription and audio/video. Clicking on a speech utterance or turn highlights the corresponding document article and plays the audio/video. It helps answering two types of question: (a) When were the various parts of a document discussed? (b) What was said about the documents or a part of it?

4. METHODOLOGY

Determining the relations that exist between documents and speech transcript consists in detecting the links between their respective units. For this reason, documents and speech transcripts must be first segmented in various structures. We present in this section various document's structures, that could be integrated in a single multi-layered representation, and briefly introduce the standard structure of the speech transcript. Further in this section, we present various alignment strategies and methods for measuring similarities between units.

4.1. Documents and Speech Transcript segmentation

Both documents and speech are hierarchically structured. In the two following sub-sections we present the various layers of both structures.

4.1.1. A multi-layer document structure

A document can be represented in an important number of structural forms, which we refer to as its structures. We identified various levels in such structure: physical, logical, syntactical, and thematic structure:

a) Standard printed documents generally conform to a certain geometric structure that dictates that the document be composed of a set of interconnecting rectangular printed regions, or blocks, and in some cases (such as text columns, tables, etc.) composite or associated regions. The physical level is often designated as the zoning, segmentation, region forming, geometric analysis, or page analysis level [12].

b) An important aspect of document understanding is document logical structure derivation, which involves knowledge-based analysis of document images to derive a symbolic description of their structure and contents. These logical units can group physical blocks that represent meaningful logical entities, e.g. title, author name, abstract, sequence of sections, etc. [17] Despite the progresses made in this domain [8], the method we have implemented requires that the documents are available in an electronic form, for instance in PDF format, which we use as a pivot representation for document analysis. 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-XML tool, called Xed [7] that will convert a complex PDF document in a linear textual form (in respect to the reading order) and most important that highlights the layout and logical structures of the document (bounding box for each physical block). This tool is based on a novel approach that combines a) low level extraction methods applied on PDF files with b) layout analysis performed on a synthetically generated TIFF image.

c) The syntactical structure of a document is a description of the document as a sequence of textual components, e.g. words, sentences, paragraphs, etc., without the concern about geometrical and typographical properties of these components.

d) In [18], Salton and al. found that document decomposition should involve more meaningful text units than logical elements. They propose to detect the overall organization of the text into themes or topics, and the identification of text segments that are semantically homogeneous text pieces. That is the so called thematic structure. The document's thematic structure we extracted, using TextTiling [9,13,18], is not yet satisfactory for the type of document we are handling. This preliminary evaluation thus concentrates on other document structures. In the future, treating in parallel the text-tiling and alignment should improve both processes.

4.1.2. Speech transcript structure

The classical structure of a speech transcription is a sequence of thematic episodes. Each thematic episode involves one or more speakers engaged in a dialog about a specific topic. Each episode is composed of turns where each turn is defined as a contiguous part of speech from one speaker. Finally, each turn can be decomposed into utterances with an utterance being a small coherent part of one speaker's speech (more or less corresponding to a sentence in a written document), to which can be further associated a dialog act (question, answer, acknowledgment, agreement, etc.) [19,20]. In the evaluation presented later in this article, the speech recordings have been manually transcribed. Future works includes testing our alignment methods with a) results of automatic speech recognizer and b) manual transcription with added noise.

4.2. Document and Speech alignment

In this section, we explain how the two structures presented above can be aligned. We first discuss two different strategies for bridging links their respective units. Finally, we present methods for estimating their similarities.

4.2.1. Alignment strategies

One-way alignment

The first alignment technique we have implemented was oriented. A source file is aligned with a target file and for each unit of the source file, a most similar unit in the target file is found. Thus, this alignment is asymmetrical; if a unit t1

from a document D1 is aligned with a unit t2 from a speech transcript T2, it does not mean that t2 will be aligned with d1. For this reason, the alignment orientation is to take into consideration: a) from documents to speech transcript and b) from speech transcript to documents.
Many alignment levels can be explored considering the numerous segmentations available for documents and speech transcripts and the two different directions of alignment. However, most of them do not provide any significant benefit to the alignment process. For example, aligning documents' logical blocks with speech utterances will not be informative because only one best utterance will be found for each logical block, which generally contains several paragraphs and numerous sentences. This limitation could be easily solved by not only considering the best alignment for each source unit but all the alignments that overcome a certain threshold (K-bests similar units discussed in next paragraph). We could also consider other metrics for comparing units, such as membership and ownership. However, simple alignments must be first evaluated and the previous limitation imposes that the source units should be smaller or equal to the target units in order to find useful alignments. The following units will be thus considered in the source file: a) utterances and turns in the speech transcript and b) sentences and thematic units for documents.

K-bests: a symmetrical alignment
The one-way alignment considers only one best match for each alignment. However, all the units overcoming a certain similarity 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 (i.e. a logical block of a newspaper's cover page). Considering all the possible units' matches make the alignment being symmetrical. Indeed, the relationships detected, overcoming a certain similarity threshold, are the same in both alignment directions. Considering that units within a same source (i.e. document or speech) are not connected, the similarities found between document's units and speech's units will create a bi-polar graph and detecting the most connected regions in this graph should help discovering thematic regions (figure 2.a). Further, documents and speech transcript are both hierarchically structured. Therefore, a framework should be defined in order to compare and merge the individual alignments at various levels of both sources trees in a single alignment (figure 2.b).

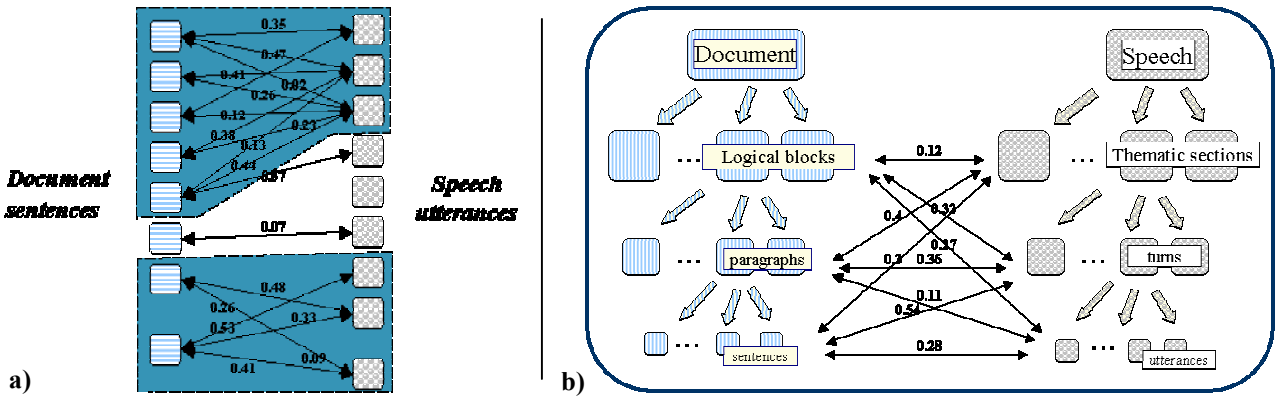

Figure 2: a) The symmetrical alignment is a bi-polar graph and it can help discovering hidden thematic structures. b) Because of the hierarchical aspect of both sources, alignments can be merged in a single framework.

4.2.2. Similarity measures

Assuming that each document unit and speech unit is represented as a bag of weighted terms, and after proper filtering of stop-words and stemming, it is possible to compute pairwise similarity between them. There are various state-of-the-art similarity metrics based on the co-occurrences of terms in the respective units. For two vector representations x and y, and n distinct terms, where $w_{t,v}$ is the weight assigned to a term t in vector v, we have implemented the cosine, Dice and Jaccard's similarity distances as follow:

$$cos(x, y) = \sum^{n}_{t=1} w_{t,x}\, w_{t,y} / \sqrt{\sum^{n1}_{t1=1} w^2_{t1,x} * \sum^{n2}_{t2=1} w^2_{t2,y}}$$
$$dice(x, y) = \sum^{n}_{t=1} 2 * w_{t,x}\, w_{t,y} / (\sum^{n1}_{t1=1} w^2_{t1,x} + \sum^{n2}_{t2=1} w^2_{t2,y})$$
$$jaccard(x, y) = \sum^{n}_{t=1} w_{t,x}\, w_{t,y} / ((\sum^{n1}_{t1=1} w^2_{t1,x} + \sum^{n2}_{t2=1} w^2_{t1,y}) - \sum^{n}_{t=1} w_{t,x}\, w_{t,y})$$

Jaccard's coefficient penalizes a small number of shared terms more than the Dice coefficient does [15]. Further, the cosine coefficient gives better results than Dice when the sizes of the two units are very different. We have decided to use whole three coefficient in our evaluation.
Later in the document, the need for other metrics appears. The evaluation, presented at the end, uses only these similarity measures. However, we observed the need for two other measures: 1) membership (is part of) and 2) ownership (contain). A more complete study should as well consider a thesaurus in order to extend the comparison to synonyms, meronyms and hyper/hyponyms.

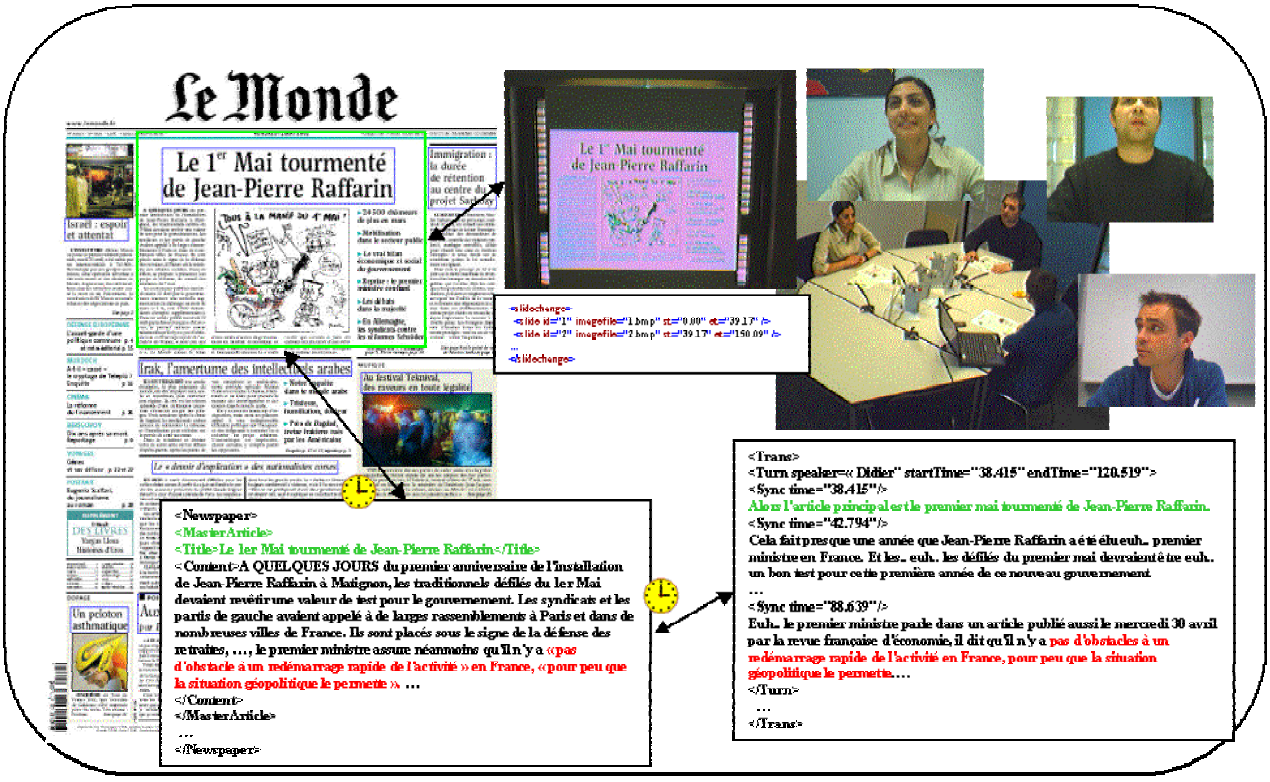

Figure 3: The press review scenario and the multiple document alignments. The newspaper document can be enriched with temporal indexes coming from a) projected documents' events, such as slide changes, and b) timestamps associated with speech utterances and turns.

5. EVALUATION

5.1. Test Data

The first step for validating the integration of documents into multimedia archives, and to measure the document alignments, is to build corpuses of meeting recordings based on scenarios where participants have a high interaction with documents. 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, as represented on figure 3). Newspapers contain several small articles with heterogeneous topics. Thus, press reviews follow a structured agenda that should fit well document temporal alignment through document content thematic alignment with speech transcripts.
Another scenario could be considered for highlighting references alignments, i.e. links between printed documents and structured dialogs through the references that are made to documents in speech transcript (such as: "the article written by", "the caption on the right side", etc). In this scenario, inspired from the MAP Task [11], a person will describe verbally the document layout so that another person can reproduce it. We expect this scenario to bring to light all the verbal expressions associated with the document layout and logical structure.

We have already recorded about 20 press reviews of roughly 15 minutes each, involving generally 3 to 5 participants. For each meeting, a directory has been created on our media file server. It contains audio and video files for each participant, the manual speech transcript of the meeting, the PDF file of each discussed and/or projected document, and for each of those files: a) its linear textual version, b) an XML file holding the manual logical structure and c) the text segmented in sentences (syntactic structure).
In the following sections, we present the results of various alignments, at some fixed levels of the document and speech transcript structures. We studied in particular eight meetings, with a total of 572 utterances and 228 turns. The eight documents studied, newspaper's cover pages, are composed of 90 logical units (newspaper articles or topics), with a total of 1409 sentences.

5.2. Metrics for evaluating alignments

In order to evaluate our algorithms, we have prepared a manual ground-truth containing all the possible alignments that we want to automate. The usual notion of recall and precision helps evaluating the quality of a given alignment in respect to the manual ground-truth:

- Recall = Number of correct alignments found /Number of correct alignments that should be found
- Precision = Number of correct alignments found/ Number of alignments found

The F-measure combines recall and precision in a single efficiency measure:

F = 2* (precision * recall) / (precision + recall)

Alignments that are null in the ground-truth are not considered in this evaluation, as well as units that contain only stop words.

5.3. Alignment results

We have observed that in most of the incorrect alignments discovered using the cosine similarity metrics, which should have been null according to the ground-truth, the similarity value was inferior to 0.1. For this reason, we have fixed this value as a threshold to filter the generated alignments of all three metrics' results (cosine, Dice and Jaccard). However, since the similarity value is highly related to the type of similarity measure used, the threshold should be dynamically calculated. Further, the similarity value is based on terms weight, in respect to their frequency in their units, and thus it is heavily influenced by the unit's length. In the future, this fixed threshold will be statically calculated, or using simple heuristics according to various variables (units' length, membership, etc).

5.3.1. Aligning documents with the speech transcript

Our first goal was to answer the question: When was discussed the various parts of the document? Thus, we started aligning document's content with the speech transcript content. In that alignment direction, we have considered the document's sentences as the unit to be matched with the speech utterances, and then with the speech turns, as showed in Table 1. The tables below presents 8 meetings where daily's newspaper where discussed. Most of them were relatively stereotyped meetings. Indeed, the newspapers' articles were presented rather than debated. In meetings 03.04 and 04.04 various newspapers' cover pages were discussed respectively 3 and 4 newspapers. In meeting 02.05, newspapers' articles were not closely followed, and meeting participants were much more arguing about the news (55 speaker turns for 94 utterances: ratio > 1/2) in comparison to other meetings (average 20 speaker turns for 60 utterances: ratio 1/3). It thus gives a good indication of how well perform our method in a realistic meeting.

		Dice			Jaccard			Cosine		
Date	Pairs	R	P	F	R	P	F	R	P	F
1.04	117	0.89	0.57	0.7	0.84	0.71	0.77	0.9	0.57	0.7
2.04	129	0.8	0.54	0.64	0.75	0.66	0.7	0.86	0.57	0.68
2.05	110	0.9	0.52	0.66	0.88	0.66	0.75	0.88	0.51	0.64
3.04	359	0.82	0.34	0.48	0.76	0.49	0.6	0.83	0.32	0.47
4.04	359	0.79	0.3	0.44	0.73	0.42	0.53	0.81	0.29	0.43
21.03	101	0.84	0.63	0.72	0.78	0.67	0.72	0.84	0.62	0.71
23.04	114	0.84	0.59	0.69	0.78	0.73	0.75	0.86	0.57	0.69
26.03	120	0.95	0.66	0.78	0.91	0.79	0.85	0.97	0.63	0.76
		0.85	0.52	0.64	0.8	0.64	0.71	0.87	0.51	0.63

		Dice			Jaccard			Cosine		
Date	Pairs	R	P	F	R	P	F	R	P	F
1.04	117	0.69	0.88	0.78	0.36	0.83	0.5	0.86	0.72	0.78
2.04	129	0.61	0.83	0.7	0.36	0.96	0.52	0.89	0.78	0.83
2.05	110	0.74	0.53	0.62	0.57	0.59	0.58	0.74	0.46	0.57
3.04	359	0.52	0.36	0.42	0.38	0.49	0.43	0.74	0.37	0.5
4.04	359	0.35	0.27	0.31	0.12	0.19	0.14	0.67	0.37	0.48
21.03	101	0.46	0.68	0.55	0.33	0.83	0.47	0.77	0.7	0.73
23.04	114	0.66	0.73	0.69	0.38	0.69	0.49	0.75	0.7	0.73
26.03	120	0.46	0.55	0.5	0.18	0.42	0.25	0.75	0.67	0.71
		0.56	0.60	0.57	0.33	0.62	0.42	0.77	0.60	0.67

Table 1: Aligning document's sentences with a) speech utterances and b) speech turns

The alignment between sentences and utterances is very promising because it highlights at low cost the underlying thematic structure of documents. Indeed, this alignment brings to light a linear arrangement of the alignments: a group of spatially successive sentences corresponds to a group of temporally successive utterances (sentence *i* with utterance *j*, sentence *i+1* with utterance *j+2*, sentence *i+2* with utterance *j+1*, etc.). Grouping sentences that correspond to neighboring utterances could lead to the document thematic structure, which is not yet fully automated. This result also suggests that segmentation and alignment are to be considered together, in a single processing loop. When matching sentences with turns, the similarity threshold is a bit higher than when comparing sentences with utterances, mainly because the words in common are fewer in comparison to the all collection size. To avoid the elimination of correct alignments, membership should be as well considered.

5.3.2. Aligning the speech transcript with documents

Our second goal was to answer the question: What was said about the various parts of the document? For this reason, we tried to align documents and speech in the reverse order. We have considered in this case the smallest speech transcript unit, i.e. utterance, and matched it with two document units: a) sentences, and b) logical units. Table 2 shows the results of those two alignments.

		Dice			Jaccard			Cosine		
	Pairs	R	P	F	R	P	F	R	P	F
1.04	42	0.87	0.84	0.86	0.87	0.84	0.86	0.87	0.82	0.84
2.04	63	0.89	0.83	0.86	0.89	0.89	0.89	0.91	0.82	0.86
2.05	94	0.9	0.72	0.8	0.88	0.78	0.83	0.9	0.7	0.79
3.04	100	0.65	0.55	0.6	0.65	0.59	0.62	0.68	0.56	0.62
4.04	83	0.7	0.58	0.63	0.7	0.61	0.65	0.74	0.61	0.67
21.03	60	0.83	0.68	0.75	0.81	0.69	0.74	0.75	0.6	0.67
23.04	64	0.98	0.83	0.9	0.98	0.9	0.93	0.95	0.81	0.87
26.03	66	0.84	0.77	0.8	0.84	0.78	0.81	0.84	0.77	0.8
		0.83	0.72	0.77	0.83	0.76	0.79	0.83	0.71	0.76

		Dice			Jaccard			Cosine		
	Pairs	R	P	F	R	P	F	R	P	F
1.04	42	0.61	0.77	0.68	0.46	0.81	0.59	0.86	0.86	0.86
2.04	63	0.8	0.83	0.76	0.35	1	0.52	0.91	0.91	0.91
2.05	94	0.67	0.97	0.79	0.36	1	0.52	0.87	0.91	0.89
3.04	100	0.48	0.42	0.45	0.33	0.6	0.43	0.85	0.55	0.67
4.04	83	0.68	0.93	0.79	0.32	1	0.48	0.89	0.59	0.71
21.03	60	0.47	0.57	0.51	0.19	0.44	0.27	0.67	0.69	0.68
23.04	64	0.62	0.78	0.69	0.36	0.89	0.51	0.89	0.89	0.89
26.03	66	0.55	0.73	0.63	0.37	0.79	0.51	0.75	0.79	0.77
		0.61	0.75	0.66	0.34	0.82	0.48	0.84	0.77	0.78

Table 2: Aligning speech utterances with a) document's sentences and b) document's logical structure

When trying to align utterances with document units, most of the incorrect alignments were detected because of terms' co-occurrence, even thought the topic discussed was different. Other utterances were imperfectly aligned because of the likeness of topics between some document's units. This problem especially appears when discussing about different documents having a similar content. In this alignment case, Jaccard's coefficient performed slightly better mostly because of its higher precision. To avoid these difficulties, we are willing to take into account more than one response for each utterance, in respect to a well-defined threshold. In this direction as well, we found that, aligning utterances with sentences, could help discovering the thematic structure of the speech transcript.

We have tried to align speech turns with document sentences. However, more than one sentence could be matched with a specific turn and it was impossible for a human solver to decide which sentence was most related to a specific turn. For this reason, it was not possible to make a reasonable ground-truth for this alignment. Further, a turn can contain more than one topic and could be as well aligned with several logical units. This reinforces our decision about considering more than one pertinent unit in each alignment, especially when the source unit is larger than the target unit.

		Dice			Jaccard			Cosine		
	Pairs	R	P	F	R	P	F	R	P	F
1.04	13	0.86	1	0.92	0.71	1	0.83	1	1	1
2.04	17	1	0.89	0.94	1	1	1	1	0.89	0.94
2.05	55	0.72	1	0.84	0.44	1	0.61	0.88	0.92	0.9
21.03	21	0.67	0.75	0.71	0.44	0.67	0.53	0.78	0.78	0.78
23.04	28	0.62	0.67	0.64	0.5	0.8	0.61	0.75	0.71	0.73
26.03	22	0.67	0.67	0.67	0.67	0.86	0.75	0.64	0.78	0.7
		0.76	0.83	0.79	0.63	0.89	0.72	0.84	0.85	0.84

Table 3: Speech turns vs document's logical structure

5.4. Remarks

Among the 8 meetings tested, two of them were discussing about more than one document (meetings 03.04 and 04.04), which is often the case in real meetings. In this case, recall and precision values are inferior in comparison to other meetings results. In the future, considering the k-bests alignments, rather than only one best, should solve the problem, since one similar topic can be discussed in various documents.

In general, Jaccard's coefficient performs best when units to compare are of similar size (e.g. sentences and utterances). In all the other cases, the cosine's coefficient is more accurate. However, a more precise evaluation should consider a different threshold value for each of the coefficient and take into account the units' size.
All those alignments require a common representation format and a common annotation framework, so that they can be combined at the end (figure 2.b). We are currently representing the document as a stream of characters and the various document annotations point on this stream. The same representation is used for the speech transcript, enriched with temporal indexes. We believe that representing both documents and speech transcript as streams of characters will allow merging all the various segmentations and alignments.

6. CONCLUSION

We proposed in this article a method for integrating printable documents, non-temporal data, to multimedia meeting archives. We described various strategies for aligning their content and structures, using the speech transcription of the meeting. In this preliminary study, we have noticed that document alignment is closely related to the preceding segmentation phases. We have also discovered that alignments can help discovering hidden document structures, such as the thematic structure. Our future work will concentrate on reunifying the segmentation and alignment process. We also plan to evaluate our methods with other types of documents and document-oriented meetings.

REFERENCES

1. Abowd G. et.al., Teaching and learning as Multimedia Authoring: The Classroom 2000 Project, Proc. ACM Multimedia '96, Boston MA, November 1996, pp. 187-198.
2. Bett, M., Gross, R., Yu, H. Zhu, X. Pan, Y., Yang, J. & Waibel, A. (2000) "Multimodal Meeting Tracker", Proceedings of RIAO2000, Paris, France.
3. Brotherton J. A., Bhalodia J. R., Abowd G. D., Automated Capture, Integration, and Visualization of Multiple Media Streams. In the Proceedings of IEEE Multimedia '98, July 1998.
4. 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.
5. Cutler, R. et al. (2002) "Distributed Meetings: a Meeting Capture and Broadcasting System", proceedings of the ACM Multimedia 2002 Conference.
6. Ghorbel H, Ballim A, Coray G, Rosetta: Rhetorical and semantic environment for text alignment. In Proceedings of Corpus Linguistics 2001, Editors: P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja, pp: 224-233, March-April 2001, Lancaster.
7. Hadjar, K., Rigamonti, M., Lalanne, D. and Ingold, R. Xed: a new tool for eXtracting structures from Electronic Documents. Submitted to DIAL 2004, International Workshop on Document Image Analysis for Libraries, January 23-24, 2004, Palo Alto Research Center (PARC), Palo Alto, CA, USA
8. Hadjar K., Hitz O., Robadey L., Ingold R.: 2(CREM) Une méthode de reconnaissance interactive pour les documents à structure complexe CIFED'2002: Colloque International Francophone sur l'Ecrit et le Document, Hammamet, Tunisie, October 2002
9. Hearst M., Multi-Paragraph Segmentation of Expository Text. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA, June 1994
10. Hunter, J. & Little, S. (2001) "Building and indexing a distributed multimedia presentation archive using SMIL", ECDL'01, Darmstadt.
11. Isard, A. and Carletta, J. Transaction and Action Coding in the Map Task Corpus, HCRC/RP-65.
12. Ishitani Y., Document image analysis with cooperative interaction between layout analysis and logical structure analysis, Document Layout Interpretation and its Applications (DLIA99), Bangalore,India
13. Kaufmann S. (1999) Cohesion and collocation: Using context vectors in text segmentation. In Proceedings of the 37th Annual Meeting of the Association of for Computational Linguistics (Student Session), p 591-595, College Park, USA, June. ACL.
14. Lalanne, D., Sire, S., Ingold R., Behera, A., Mekhaldi, D and von Rotz D. (2003) "A research agenda for assessing the utility of document annotations in multimedia databases of meeting recordings", 3rd International Workshop on Multimedia Data and Document Engineering (MDDE-2003), Berlin, Germany. To be published in a book.

15. Manning, C. and Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts, 1999.
16. Mukhopadhyay, S., and Smith, B. (1999) Passive capture and structuring of lectures, proceedings of the seventh ACM international conference on Multimedia (Part 1), Orlando, Florida.
17. Niyogi D. and Srihari S.N. Knowledge-based derivation of document logical structure. In Proceedings of ICDAR, Montreal, Canada, August 1995
18. Salton G., Singhal A., Buckley C. and Mitra M. Automatic Text Decomposition Using Text Segments and Text Themes. In Proceedings of the Hypertext '96 Conference, Washington D.C., USA
19. Stolcke A., and Shriberg E., Automatic Linguistic Segmentation of Conversational Speech, proceedings of ICSLP '96, Philadelphia, PA, 1996.
20. Stolcke, A.; Shriberg, E.; Bates, R.; Coccaro, N.; Jurafsky, D.; Martin, R.; Meteer, M.; Ries, K.; Taylor, P.; and Van Ess-Dykema, C. 1998. Dialog act modeling for conversational speech. In Proceedings of the AAAI-98 Spring Symposium on Applying Machine Learning to Discourse Processing, 1998.