Using Bi-modal Alignment and Clustering Techniques for Documents and Speech Thematic Segmentations

Dalila Mekhaldi Denis Lalanne Rolf Ingold

Département d'Informatique
Chemin du Musée, 3
CH-1700 Fribourg, Switzerland
+41 26 429 66 78 (65 96)

{Dalila.Mekhaldi, Denis.Lalanne, Rolf.Ingold} @unifr.ch

ABSTRACT

In this paper, we describe a new method for a simultaneous thematic segmentation of the meeting dialogs and the documents discussed or visible throughout the meeting. This bi-modal method is suitable for multimodal applications that are centered on documents, such as meetings and lectures, where documents can be aligned with meeting dialogs. Bringing into play this alignment, our bi-modal segmentation method first transforms its results into a set of nodes in a 2D graph space, where the two axes represent respectively the document units and the meeting dialogs units. Secondly, via a clustering method, the most connected regions in the constituted bi-graph are detected. Finally, the denser clusters are projected on the two axes. The two sequences of segments, obtained on both axes, represent the thematic structure of the document and of the meeting dialogs respectively.

We present in this article this bi-modal segmentation technique and its performance compared with two mono-modal segmentation methods.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**] indexing methods; H.3.3 [**Information Search and Retrieval**] Clustering- Search process; I.7.5 [**Document Capture**] *Document analysis;* I.5.3 [**Clustering**] *Similarity measures*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Thematic alignment, thematic segmentation, K-Means clustering.

1. INTRODUCTION

The alignment, or more precisely, the thematic alignment of printable documents with other media data such as audio and video, appears as an important and necessary step for the full document integration into multimedia archive. Since multimedia data are time dependent, and not printable documents, it is necessary to bridge a temporal link between them. Recent researches are focusing on the alignment of meeting documents with the transcription of the speech [9,12]. The final goal is to integrate these documents into meetings archive, create temporal links between them and other media, so that they can be used as thematic and structured interfaces to retrieve multimedia data.

However, beside the documents integration into multimedia archive, the thematic alignment has another benefit, that is not less crucial than the first one, which is the thematic segmentation of the resources being aligned, i.e. the document and the speech transcript. Since the thematic alignment consists in the detection of thematic links, between a set *D* of document units and a set *S* of speech transcript units, then *D* and *S* may share the same theme.

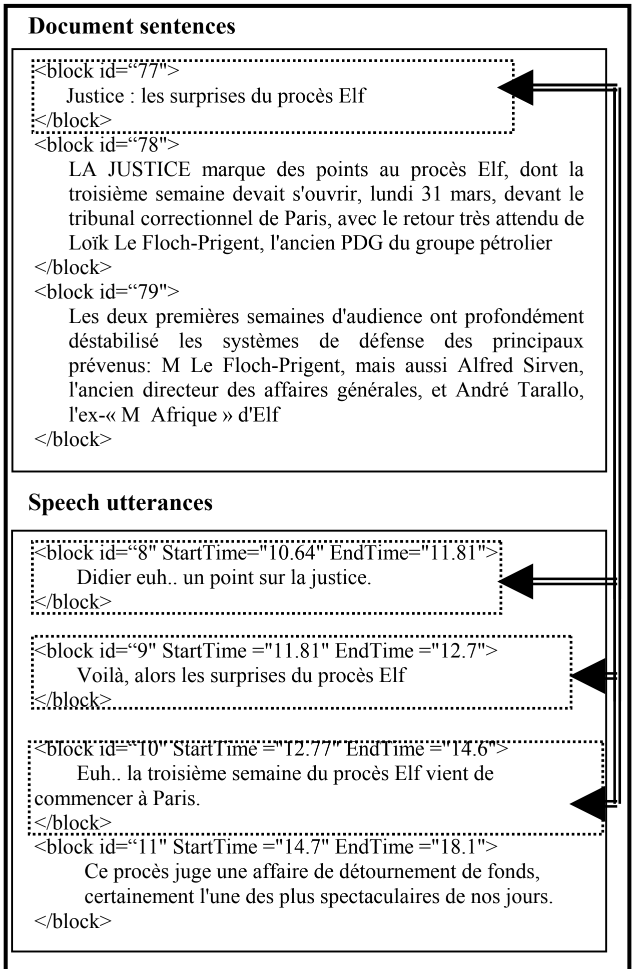
This paper is organized as follow; section 2 is a description of our meetings data. In section 3, a short state-of-the-art of the existing thematic segmentation techniques is presented. In section 4, a brief explanation of our thematic alignment process is given, which is the basis of our bi-modal thematic segmentation method. Then we present our segmentation method along with the evaluation of the obtained results.

2. MEETING RECORDINGS

Given that our project is focusing on multimodal meetings analysis, a recording environment has been installed in our research group meeting room. In collaboration with the University of Applied Sciences of Fribourg, we have equipped the room with 8 camera/microphone pairs, so that one video and audio stream is recorded for each meeting participant. Also, several cameras have been fixed on the ceiling and the walls, in order to capture the documents that are visible throughout the meeting, either projected or standing on the meeting table. In particular, a video projector and a camera have been installed for the projection screen capture. All the audio/video recordings are synchronized and controlled by a master PC [9].

Concerning the meeting recordings that are initially in french, there are document-centric: agenda driven meetings, student projects, job interviews, etc. 22 press reviews meetings have been picked up, with an approximate duration of 15 minutes, and between 3 and 6 speakers. All these meeting recordings are available on a server. The documents that are discussed in these meetings are a French speaking daily newspaper cover page.

CIKM'04, November 8-13, 2004, Washington, DC, USA.



Document sentences

<block id="77">
Justice : les surprises du procès Elf
</block>
<block id="78">
LA JUSTICE marque des points au procès Elf, dont la troisième semaine devait s'ouvrir, lundi 31 mars, devant le tribunal correctionnel de Paris, avec le retour très attendu de Loïk Le Floch-Prigent, l'ancien PDG du groupe pétrolier
</block>
<block id="79">
Les deux premières semaines d'audience ont profondément déstabilisé les systèmes de défense des principaux prévenus: M Le Floch-Prigent, mais aussi Alfred Sirven, l'ancien directeur des affaires générales, et André Tarallo, l'ex-« M Afrique » d'Elf
</block>

Speech utterances

<block id="8" StartTime="10.64" EndTime="11.81">
Didier euh.. un point sur la justice.
</block>
<block id="9" StartTime ="11.81" EndTime ="12.7">
Voilà, alors les surprises du procès Elf
</block>
<block id="10" StartTime ="12.77" EndTime ="14.6">
Euh.. la troisième semaine du procès Elf vient de commencer à Paris.
</block>
<block id="11" StartTime ="14.7" EndTime ="18.1">
Ce procès juge une affaire de détournement de fonds, certainement l'une des plus spectaculaires de nos jours.
</block>

Figure 1. *Multiple-alignments* strategy between document sentences and speech utterances

The benefit of this kind of documents is mainly the heterogeneity and the small size of its articles. However, we prospect to consider other document types such as agenda and slides, and other languages, such as the English language.

3. RELATED WORKS

A thematic segmentation, i.e. the decomposition of a given text into topics or thematically homogeneous segments, has been the subject of many research works. Salton and al. [15] used a text relationship map that establishes similarity links between the text excerpts (sentences or paragraphs), which are represented as nodes in the map. In order to define textual thematic segments, all the triangles are located in the full relationship map. Many triangles can be merged when the similarity between their corresponding vectors centroïds exceeds a defined threshold. This map provides information about homogeneity of the text, so that if there are many links between adjacent paragraphs, this proves the homogeneity treatment of topics.

Hearst's *Texttiling* method divides the text into tokens –individual lexical units- [7]. The adjacent pairs of blocks, i.e. sequences of tokens, are compared using a similarity method.

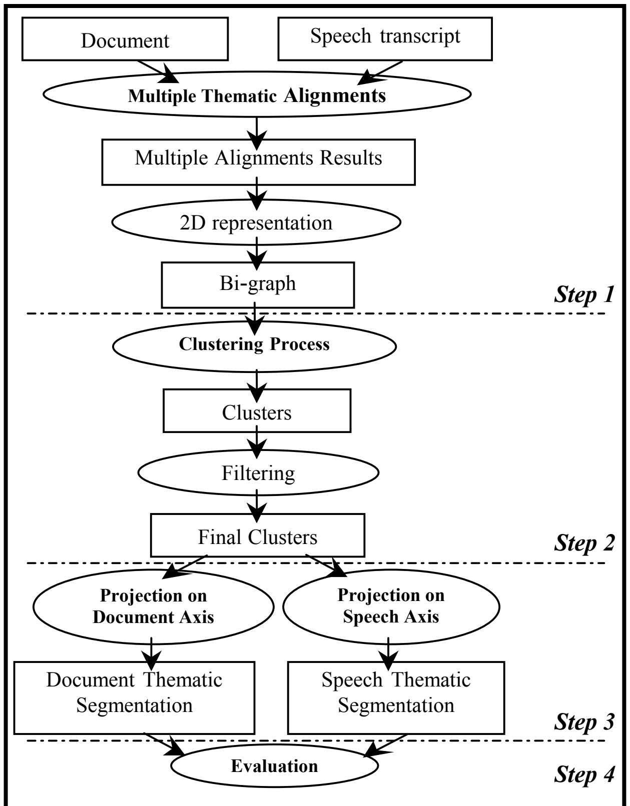

Document
Speech transcript
Multiple Thematic Alignments
Multiple Alignments Results
2D representation
Bi-graph
Step 1
Clustering Process
Clusters
Filtering
Final Clusters
Step 2
Projection on Document Axis
Projection on Speech Axis
Document Thematic Segmentation
Speech Thematic Segmentation
Step 3
Evaluation
Step 4

Figure 2. Illustrative diagram of our segmentation process

The topics boundaries are then defined according to the change in the sequence of similarity scores.

Ferret has used a similar method that based on the boundaries detection and on similarity measure between the adjacent units [4]. Ferret's method is enriched with a lexical co-occurrence network built from a large corpus. This work reinforces the descriptors (vectors representing the units) by linking words that have semantic relationships.

4. A BI-MODAL APPROACH

4.1 Thematic alignment

The thematic alignment of documents with the speech transcript of the meeting dialogs is highly related to the quality of the various segmentations of both document and speech transcript into a set of segments or units. Our alignment process that was presented in [9,12] is described as follow: first, each one of the two resources is segmented. The logical structure [6] (article, section, title) and the syntactical structure (sentences, paragraphs) for the document are first extracted. On the other hand, the speech recording is currently transcribed manually, and then it is segmented into turns and utterances. In the near future, the speech recordings will be transcribed automatically using a state-of-the-art speech recognizer, which may increase drastically the word error rate and decrease our alignment performances. Secondly and after proper stop-words removal and stemming, both documents and speech transcript units, represented as vectors of weighted terms, are compared via similarity metrics. These metrics (*Cosine*, *Dice* and *Jaccard*) count the co-occurrences of terms, in respect to their respective weights, and compute a similarity distance.

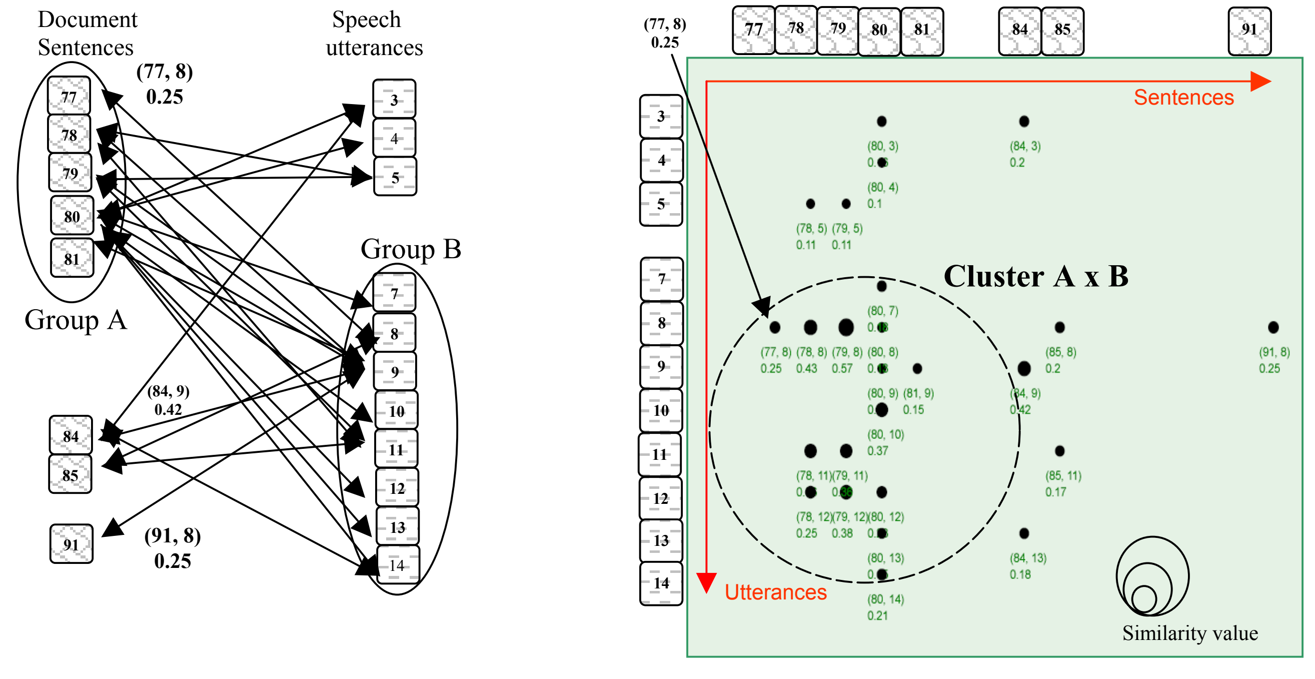


Figure 3. a. Bi-graph representing the *multiple-alignments* strategy b. 2D representation of the *multiple- alignments*

According to the selection criteria of the generated links, two alignment strategies have been defined: the *one-best* strategy and the *multiple-alignments* strategy. In the *one-best* strategy, the best link, from a source unit to a target unit, is chosen, i.e. the link that corresponds to the maximum similarity value. In the *multiple-alignments* strategy, all the best links whose similarity value overcomes a defined threshold are retained. This last strategy generates a symmetrical alignment, i.e. the alignment results obtained from the document to the speech transcript are the same as the ones obtained in the other direction. This property constitutes the basis of our bi-modal segmentation method.

4.2 Alignment vs. thematic segmentation

The thematic segmentation of texts is still a difficult task that is not completely resolved [4,7,15]. In our thematic alignment process [9,12] the thematic segmentation of the resources being aligned is required, even if our preliminary experiments of the alignment techniques are based on simple segmentations. For this reason, some aspects have been deduced from the preliminary results of this thematic alignment, especially from the *multiple-alignments* strategy. The *multiple-alignments* strategy considers all the best alignments, i.e. the best similarity links computed through the *Cosine* and *Jaccard* metrics, between the document's units and the speech transcript's unit (figure 1).

4.2.1 Bi-graph construction

Since an elementary unit (a sentence or an utterance) very rarely belongs to two different thematic regions, our first contribution to solve this problem is based on the thematic alignment results between sentences and utterances, which are respectively the smallest units for documents and speech transcript. In the future, other pairs will be considered, such as the sentences alignment with turns, or turns alignment with logical blocks, etc.

Looking at figure 3.a, which is a visualization of the generated *multiple-alignments* results for a given meeting. Each document unit (respectively speech transcript unit) is represented by a node. The similarities between these units are represented by edges between the corresponding nodes, where the edge weight value represents the alignability value (e.g. sentence 77 has a similarity value of 0.25 with utterance 8). Thus, the generated graph is a bi-graph, since each node belongs to one of the two different sets of nodes.

When analyzing this bi-graph, it appears that some regions are denser than others (e.g. group A and group B on figure 3.a). The nodes on each side of the bi-graph are respecting their appearance order, spatial order or adjacency for the document units (e.g. sentences), and temporal order for the speech units (e.g. utterances). Thus, the denser regions can be explained by the fact that a group of successive units from the document is thematically linked to a group of successive units from the speech transcript (e.g. group A with group B). That means that the groups A and B may share the same theme. More than this, they may represent respectively thematic segments of the document and the speech transcript.

On the basis of the bi-graph presented in figure 3.a, the first step in our thematic segmentation method is to extract the denser regions in this bi-graph. This can be achieved by isolating them from the entire bi-graph. Once these denser regions are detected, the next step is to extract, by projection, the corresponding thematic segments for both document and speech transcript.

Figure 2 illustrates the overall process, from the thematic alignment until the evaluation of the generated thematic structures of both documents and speech transcript.

4.3 Extraction of denser regions

Our first attempt to solve this thematic regions extraction problem was based on the intersection graphs [5], which are generated by the projection of the bi-graph on each side. An intersection graph for the document and another for the speech transcript, are thus generated. These intersection graphs group each two units from one resource, which are related to a same intermediate unit from the second resource. At first, we thought that the edges weight could be fortified by detecting the intersection graphs until a fixed *nth* level, which means: from a source vertex *v1* (i.e. a node from the source file, document or speech transcript), we have to go through *n* intermediate vertexes from the other file before reaching the target vertex *v2*. Unfortunately, we have noticed that this method is not efficient. The main reason is that in the generated intersection graph, all the nodes are related. Another reason for abandoning this representation is that it is not containing all the information available in the original alignment, i.e. the information offered by the thematic links to the other part of the bi-graph, which may be useful in the segmentation process.

Finally, we looked at a more suitable solution in the clustering field. Our two resources are represented on two axes, *X* and *Y*, where the identifiers of their respective units are represented by the *X* values and *Y* values. The alignment links between these units are represented by nodes in this 2D representation (see figure 3.b), so that the grouped regions in the bi-graph (in figure 3.a, groups *A* and *B*) are represented by clusters of nodes (e.g. groups *A* and *B* become one cluster *A* x *B*). Furthermore, the similarity value for each link corresponds to the nodes size; a big node has a heavier weight than a smaller one.

As shown in figure 3.b, this 2D approach is the most representative illustration of the thematic *multiple-alignments* data, since all the alignment information is plotted. The document spatial attributes or adjacency are represented by the *X* values, the speech transcript temporal attributes by the *Y* values. The alignability values are represented by the nodes' size. Thus, many information can be derived from this representation:

- Are there many themes in the meeting?
- What are the main themes?
- When was an article discussed?
- In which temporal order the meeting was played, i.e. in which order the document articles were discussed?
- Is there any overlapped thematic segments, i.e. are there any similar document articles or speech transcript segments? etc.

We will see later in this article that these information help us not only to detect the thematic segments of both document and speech transcript, but also to detect the temporal links between the segments of the document. In the future, our 2D illustration will be improved, so that it might be exploited as an interface for navigation in the meeting recordings space (see figure 5).

4.3.1 Clustering

Thus, the thematic segmentation has been transformed into a clustering problem. As known in this field [1,17], many relative attributes of the objects being clustered should be considered. Such as the document spatial attributes and the speech temporal attributes.

The clustering process is a considerable research area, which is exploited in many fields. Starting with a large amount of data, this process consists in the creation, in an unsupervised way, of many data sets that are called clusters. This categorization is based on the similitude between the various data. Thus, the similarity is computed via a distance metric (*Euclidean*, *Manhattan*, etc.). Clustering methods can be categorized as follow [3]:

- Hierarchical methods, where the clusters are organized as a tree. We distinguish two kinds of hierarchical methods, agglomerative methods and divisive methods. In the agglomerative methods (or bottom-up), each object is initially assigned to one cluster, and then successively, the closest clusters are merged. In the divisive methods (top-down), all the data are initially in the same cluster, and successively, the resulting clusters are split, until the desired number of clusters is obtained.
- Partitioning methods, where the data set is decomposed directly to a set of clusters, so that each datum belongs to only one cluster, e.g. the *K-Means* method [11]. The methods of this category are based on some criterion, such as the maximization of the similarity between data within each cluster, and the maximization of the dissimilarity between the various clusters.

Even if many clustering methods are available, we have chosen the most common one, in order to bootstrap our method, the *K-Means* method. In the future, other methods will be considered. In the next paragraph, the standard version of this method is presented.

4.3.1.1 Standard K-Means algorithm

The objective of this process is to define the denser clusters, and the most separated ones from the others. The original *K-Means* procedure is described as follow:

1. First, the K value, the number of clusters to obtain, must be defined.
2. K points are then generated randomly in the 2D graph, which are considered as the preliminary clusters centroïds.
3. Via a distance formula, e.g. the *Euclidean* distance, each point from our data set is assigned to the nearest centroid.
4. After the clusters construction, the new centroïds are computed using an average formula, to find the mean of the *X* and *Y* values of the clusters' nodes.

Steps 3 and 4 are repeated until a stable state is reached, i.e. there are no large change in the clusters centroïds positions between two successive iterations, according to an Error formula and a prefixed error threshold. However this *K-Means* version has many drawbacks:

1. The K value must be fixed at the beginning, which is not easy since the number of clusters changes from a graph to another, which is considered as the major drawback of this method.
2. The internal criterion are not considered:
   - The compactness of clusters, taking into account the distance of the clusters centroid from the nodes,
   - The density of the clusters, taking into consideration not only the distance to the centroid but the nodes weights (the similarity values) and their number too.

3. The external criterion are not considered:
    - The distance between the clusters centroïds.
    - The average cluster density in case of overlapping.

For all these reasons, the application of this version of the *K-Means* algorithm is not sufficient. Thus, an extended version is presented in [10], which considers the mentioned internal and external criterion. However, this method does not consider neither the nodes weights when assigning the nodes to their clusters, nor the clusters density.

4.3.1.2 Extended K-Means algorithm

In this *K-Means* version, many thresholds that must be initialized by the user are considered:

1. Defining the K centroïds randomly is always followed by a merging method, which checks the non-closeness of the generated centroïds; otherwise it merges the clusters with the closest centroïds. This merging task is based on the distance between the centroïds, using a defined threshold.
2. The number of nodes in a cluster is significant, and thus a value needs to be defined by the user, by observing a sample of clustering. In our work we have fixed this threshold to two nodes per cluster as the minimum size of a relevant cluster.

Taking into account these two extra rules, it is more practical to define a large K value, since this number decreases if the centroïds are close or if the generated clusters are non-significant.

3. A third parameter is considered: the clustering validity measure. The convergence of this clustering method to the best result is measured by the variance formula *XB* [16]. Which checks the variance of the centroïds positions.

With these additional parameters, the clusters internal criteria (compactness) and external criteria (well separated each one from the other), as well the validity indexes are considered.

By applying the extended *K-Means* method, our clustering process takes as input a vector of nodes representing the bi-graph. Three variables are assigned to each node, the x and y values (the identifiers in the document and the speech transcript respectively) and the weight w (similarity value). Based on the *Euclidean* distance between the (x, y) components to compute the similarities between nodes, a set of clusters is generated as output. However, we prospect to consider not only the (x, y) components but also the weight w to compute the distances, which may increase the clustering performances. The clustering process is repeated until the clusters centroïds reach a stable state.

4.3.2 Clusters filtering

Since the extended *K-Means* algorithm does not consider the nodes weights, we have enriched our clustering process by a filtering method that filters the weak clusters, in regards to their density. This density is based on the clusters' nodes weights, nodes number, and nodes distances (*Euclidean* distance) from the final clusters' centroïds.

$$density = \Sigma_{i=1}^{Size} (w_i / distance_i(C)) * (distance_{max}(C)) * Size$$
$$density_{relative} = (density - density_{min}) / (density_{max} - density_{min})$$

Where *density* and $density_{relative}$ are respectively the cluster density and the relative density of this cluster according to the overall clusters. *Size* is its size and *C* is its centroid. $distance_i(C)$ is the *Euclidean* distance between a node *i* and its corresponding cluster centroid *C*. w_i is the weight of a node *i*. $density_{min}$ and $density_{max}$ are respectively the minimum and maximum clusters densities.

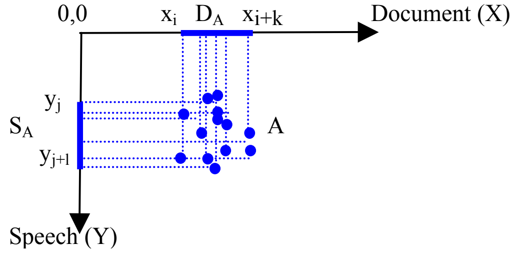


Figure 4. Cluster projection

This way, a cluster with a given number of nodes in a large surface, is less significant than a cluster with the same number of nodes but in a smaller surface. In the same way, a cluster with heavy nodes is more significant than a cluster having the same number of nodes but with lighter weights. Filtering the clusters with weak densities is based on a dynamic defined threshold, according to the average clusters densities.

Within each cluster, another filtering method, which consists in removing the isolated nodes, is useful. This additional process is applied on the light nodes that are very far from the other nodes in the same cluster, according to the distance from the document and speech axes. The gap of these isolated nodes is due in general to the transitional utterances between the spoken articles. Thus, they cause an overlapping of the thematic segments, by making them longer than what they should be (see section 4.7.1) [13]. After the filtering of the isolated nodes, the ratio of the overlapped segments decreases. Hence, the performances of our bi-modal segmentation method increase (see section 4.5.2).

Once the two filtering steps are applied on the clusters, the final remaining clusters may represent the various meeting themes, where each theme links a speech thematic segment to a similar document theme.

4.4 Extraction of thematic structures

Knowing that the final clusters representing the denser regions in the 2D representation (figure 3.b) are composed by many adjacent nodes, decomposing a given cluster *A* by its projection on the two axes (figure 4), generates a group D_A of adjacent x values, and a group S_A of adjacent y values, where:

$D_A = (x_i, x_{i+1}, x_{i+2}, .., x_{i+k})$ is considered as a document thematic segment, delimited by the sentences (x_i) and (x_{i+k}).
$S_A = (y_j, y_{j+1}, y_{j+2}, .., y_{j+l})$ is considered as a speech thematic segment, delimited by the utterances (y_j) and (y_{j+l}).

After running the overall process, the documents and speech transcript segmentations are evaluated. This is presented in the next section.

4.5 Evaluation

4.5.1 Meetings data set

22 press reviews meeting recordings have been tested in this experiment, with a total of 2936 speaker utterances and 3173 document sentences.

The segmentation results are compared to two mono-modal methods: the *Texttiling* [7] method, and a specific baseline method for each one of the document and the speech transcript. The document baseline segmentation method is a *Reflexive* segmentation method, which is based on a reflexive document alignment and clustering, i.e. aligning the document with itself, then, clustering it with itself. The baseline method for the speech transcript corresponds to the speech turns, i.e. each turn whose size overcomes a defined threshold (10 words) is considered as a thematic segment.

4.5.2 Evaluation metrics

In order to evaluate our bi-modal segmentation method, the Beeferman P_k metric [2] is used in respect to a manual ground truth. This metric measures the probability that a randomly chosen pair of units, at a distance of k units apart, is inconsistently classified in respect to the ground truth. Thus, for a perfect segmentation, the metric value is 0. For this experiment, the k parameter has been fixed to 4 units, which corresponds to the minimum size of a relevant thematic segment. This P_k metric is more adequate than the *recall/precision* metric, which has many disadvantages [8,14], especially, because it measures just the correctness of boundaries detection, without considering the distribution within the generated segments.

4.5.2.1 Speech thematic structure

Through the 22 meetings, two scenarios are followed. 9 meetings are stereotyped, and 13 are non-stereotyped. A meeting is considered as being stereotyped, if its ratio utterances/speaker turn is superior to 2. Otherwise, it is considered as non-stereotyped.

Actually, in the stereotyped meetings, the speakers present the various articles rather than debate or discuss them, with few interactions or comments. In this category of meetings, there are 1542 sentences, 617 speaker utterances and 234 turns, with an average ratio of 2.6 utterances per turn, and an average duration of 497 seconds per meeting.

In the non-stereotyped meetings, the participants have a high interactivity between them, and debate the different articles. This increases the number of turns comparing to the stereotyped meetings. Within this category, there are 1631 document sentences, 2319 speaker utterances and 1654 turns, with an average ratio of 1.4 utterances per turn. The average duration in this category is 745 seconds per meeting.

As shown in Table 1, the segmentation results using the three methods depend on the scenario of the meetings tested. With our bi-modal method, the P_k evaluation is in general satisfactory for the two meetings categories (stereotyped and non-stereotyped), in comparison to the *Texttiling* and the *Speaker-Turns* methods. However, our method is relatively more efficient for non-stereotyped meetings, which are closer to realistic meetings. Thus, it should be mentioned that in this category, the second filtering step for the isolated nodes, as seen in section 4.3.2, have not been applied, because it does not improve the results. Indeed, in non-stereotyped meetings, the number of transitional utterances or comments rises, even in the middle of a theme, and at the same time, a topic can be composed of many small turns (the average utterances/turn ratio is 1.4).

Table 1. Evaluation of the speech transcript thematic segmentation

	Speech Transcript	
	Stereotyped (9)	**Non- Stereotyped (13)**
Bi-modal	0.36	0.44
Texttiling	0.53	0.69
Speaker-Turns	0.44	0.61

Table 2. Evaluation of the documents thematic segmentation

	Documents			
	Mono-document (18)		**Multi-documents (4)**	
	Stereotyped (7)	**Non-Stereotyped (11**	**Stereotyped (2)**	**Non-Stereotyped (2)**
Bi-modal	0.32	0.28	0.24	0.3
Texttiling	0.64	0.6	0.49	0.55
Reflexive	0.45	0.48	0.51	0.63

In general, our bi-modal segmentation method is more accurate in detecting the exact number of thematic segments. Which is not the case for neither the *TextTiling* method nor the *Speaker-Turns* method, which generate many extra segments. This can be explained by the fact that using the document modality limits the number of possible themes, which constrains the segmentation, and thus helps in computing the exact number of the speech thematic segments.

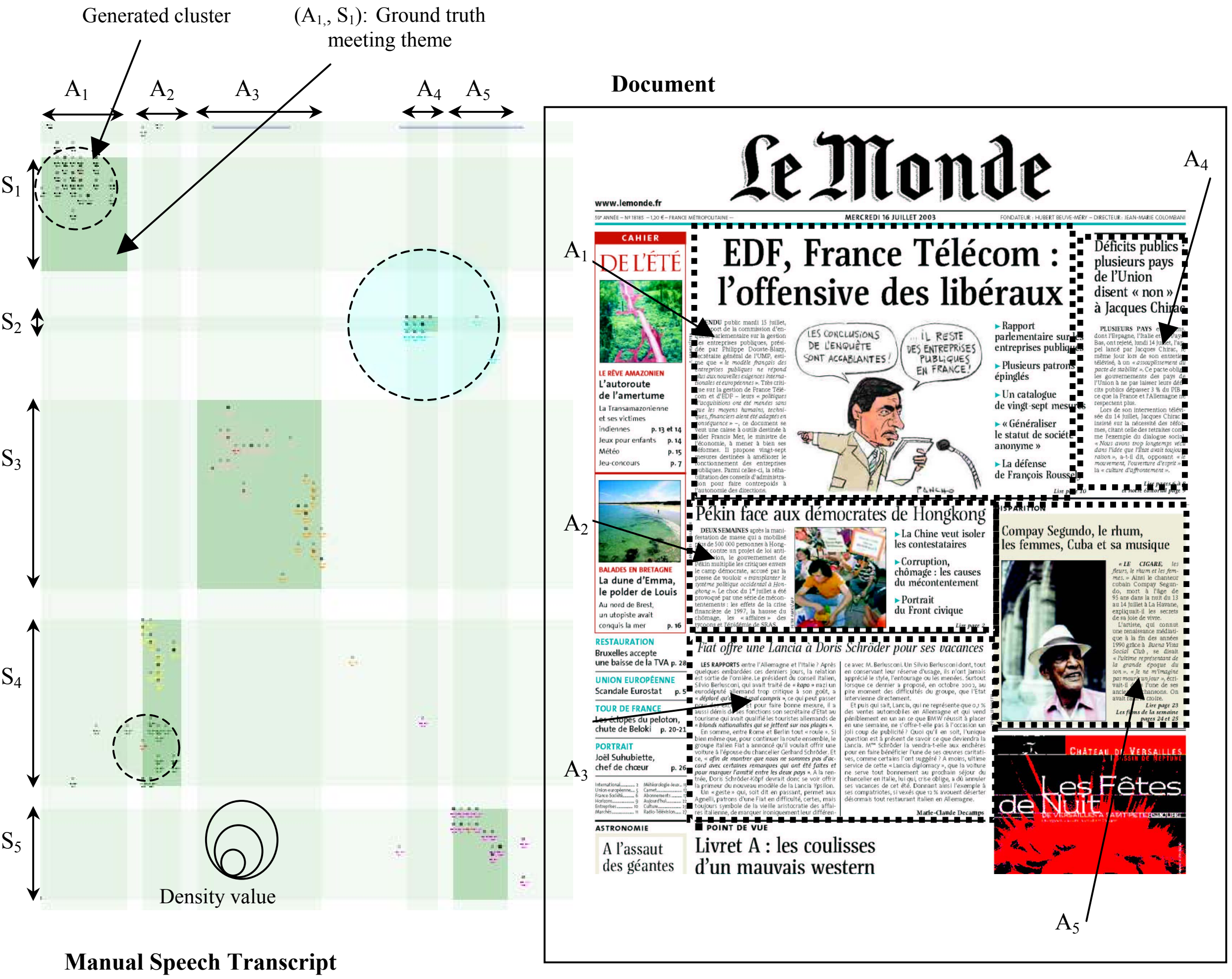
Another benefit of our bi-modal method is the detection of the similar thematic segments within the speech transcript. Which appears when the clusters are aligned vertically. This can be explained by the fact that many speech thematic segments are aligned with the same document thematic segment (e.g. cluster *A* and cluster *B* in figure 6).

4.5.2.2 Documents thematic structure

Among the 22 meetings tested, 18 are mono-document meetings, i.e. only one document is discussed during the whole meeting, with a total of 2354 utterances and 2040 sentences. The 4 other meetings are multi-documents meetings, i.e. more than one newspaper cover page is discussed: 3, 4, 2 and 2 newspapers are discussed in the four respective meetings. In this category, there are 582 speaker utterances and 1133 sentences.

As shown in Table 2, the P_k values using our bi-modal method are better, comparing to the *Texttiling* method and the *Reflexive* segmentation method (reflexive alignment/clustering of the document). The main reason is that our method is more accurate in detecting the thematic segments. Indeed, the *Texttiling* and the *Reflexive* methods generate many extra segments.

Another benefit of our method is to detect the similar articles within a document, or in various documents in the case of the multi-documents meetings. This kind of phenomena happens when some clusters are aligned horizontally in our 2D representation, i.e. when documents articles are aligned with the same speech transcript segment (clusters *A* and *C* in figure 6).


Generated cluster
(A1, S1): Ground truth meeting theme
A1
A2
A3
A4
A5
S1
S2
S3
S4
S5
Density value
Document
Le Monde
www.lemonde.fr
MERCREDI 16 JUILLET 2003
CAHIER DE L'ÉTÉ
EDF, France Télécom : l'offensive des libéraux
LES CONCLUSIONS DE L'ENQUÊTE SONT ACCABLANTES !
... IL RESTE DES ENTREPRISES PUBLIQUES EN FRANCE !
Déficits publics plusieurs pays de l'Union disent « non » à Jacques Chirac
Pékin face aux démocrates de Hongkong
Compay Segundo, le rhum, les femmes, Cuba et sa musique
Fiat offre une Lancia à Doris Schröder pour ses vacances
Livret A : les coulisses d'un mauvais western
A l'assaut des géantes
Les Fêtes de Nuit
CHÂTEAU DE VERSAILLES

Manual Speech Transcript

```
<Segment id= "S1">                    …
    <block id= "6"  StartTime="47.429" EndTime="61.062" speaker="spk2">
        Alors euh.. mardi 15 juillet, hier, euh.. la commission d'enquête parlementaire en France a rendu un rapport euh.. sur la gestion des
        entreprises.. entreprises publiques.      </block>
          …
    <block id= "8"  StartTime="71.806" EndTime="88.664" speaker="spk1">
        Euh.. ils critiquent très fortement euh.. France Télécom et EDF et leur gestion, et euh.. il est dit aussi que leur politique
        d'acquisition ont été menées sans que les moyens humains, techniques et financiers aient été adaptés en conséquence.  </block>
        …
</Segment>

    <block id= "26" StartTime="148.898" EndTime="151.674" speaker="spk3">
        On passe à l'article suivant? Celui-là sera pas très long non plus.</block>


<Segment id= "S2">
    <block id= "27" StartTime="151.674" EndTime="170.194" speaker="spk3">
        C'est.. les déficits publics: plusieurs pays de l'union disent non à Jacques Chirac. Donc euh.. lundi 14 juillet, Jacques Chirac a fait
        un appel, a lancé en fait euh.. l'appel d'assouplissement du pacte de stabilité, qui a été refusé par plusieurs pays européens.</block>
    <block id= "28"  StartTime="170.194" EndTime="178.896" speaker="spk3">
        Et quel est ce pacte? Ce pacte oblige en fait les gouvernements des pays de l'Union à ne pas laisser leurs déficits publics à dépasser
        3% du PIB.
    </block> …
</Segment>
```

Figure 5. A preliminary interface for meetings navigation

Although, it should be mentioned that our bimodal method is only appropriate for documents almost fully discussed during the meetings; otherwise the document is partially segmented.

4.6 Visualization

Figure 5 represents the results of the clustering process presented above for a given meeting, using an SVG browser. The circle, around each cluster centroïd, represents the cluster density, where the radius increases with its density value. The horizontal bars (A_1, A_2, etc.) correspond to the manual thematic segmentation of the document, which represents the various newspaper articles. The vertical bars (S_1, S_2, etc.) correspond to the manual thematic segmentation of the speech transcript. Furthermore, the pairs (A_i, S_j) generated by the intersection of these bars, i.e. the highlighted rectangles, represent the ground-truth meeting themes. Each rectangle means that the article A_i corresponds to the speech segment S_j.

This visualization technique is an efficient tool to check the validity of the *multiple-thematic* alignments, and to check whether or not the generated thematic segments are lined up with the thematic segments of the ground truth. Moreover, it gives an idea about the reading order of the various articles in the meeting.

4.7 Analysis

Despite the fact that the alignment similarity values are not yet considered in our clustering process, i.e. when assigning the nodes to the clusters, the comparison between our bi-modal method and the two mono-modal methods, tends to prove that, using various modalities improves considerably both the documents and speech transcript thematic segmentation. Moreover, our bi-modal segmentation method represents an important advantage, which is the detection of all the potential thematic links between non-adjacent document segments (respectively speech transcript segments). This happens when these non-adjacent segments are linked to a same speech transcript thematic segment (respectively document thematic segment), especially when discussing many documents containing similar articles. However, the horizontal or the vertical alignment of the clusters may generate overlapped segments during the clusters projection step. This aspect, which can have a negative effect on the segmentation results, is discussed in details in the next paragraph.

4.7.1 Thematic segments overlapping

During the segments extraction step, while projecting the denser clusters, we have faced a problem with the clusters aligned horizontally or vertically, i.e. clusters whose projections on the document axis or the speech axis generate overlapped segments (see figure 6). Given two overlapped segments, two kinds of overlapping exist:

- A segment contains a smaller one (S_c contains S_A)
- Two segments are partially overlapped (D_B with D_A).

In the first kind of overlapping, only the largest segment is considered, because it is extremely probable that they have the same theme. In the second kind of overlapping, depending on the size of the common part, the overlapped segments will be two distinctive segments taking into consideration a pre-defined tolerance margin, in number of units. We did not fully solve this kind of overlapping, and thus we are still working on it.

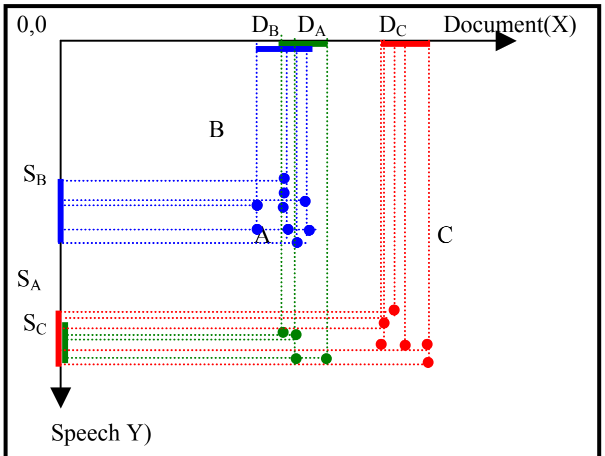


Figure 6. Thematic segments overlapping

Depending on which axis the overlapping is taking place, a different explanation can be given:

- *Overlapping on the document axis:*

This first type (e.g. D_B with D_A) can be explained by the fact that a specific topic *T1* from the document is discussed twice during the meeting, in two nonadjacent moments. For instance, at the time corresponding to the segment S_B, then at the time corresponding to S_A. Another possibility is that two distinct topics from the document (e.g. *T1*, *T2* corresponding respectively to segments D_B, D_C) are similar thematically, so the corresponding speech part, (e.g. S_C) has been aligned with the two topics, which has generated two clusters (*A* and *C*). For example, the two following sentences come from two distinct articles, and have a distinct topic. One deals with the war in Iraq, and the second tackle the effect of Chirac's political position on his popularity in France. However, they share some words (guerre, Iraq) and thus risk to be aligned with similar speech segments:

a- «La prolongation de la guerre en Irak affecte gravement la conjoncture économique, l'indice européen … »

b- «Avec sa position forte sur la guerre en Irak, il s'est produit quelque chose entre Jacques Chirac et les Français,... "

- *Overlapping on the speech axis:*

In this second type, the overlapping within the speech segments (S_A with S_C) can be explained either by the thematic similarity between two topics of the document, such as in the multi-documents meetings case, where the documents may have share similar articles. Another possibility is that one of the two topics may be referenced at the same time when the other was discussed. For instance, the following utterances illustrate this case. In the first utterance, the journalist talks about Iraq after-Saddam, and if it will be liberated or occupied. In the second utterance another speaker refers to another article describing the point of view of Blair about this issue:

a. « Et.. il y a un article aussi dans lequel le journaliste commence à réfléchir à l'après-Saddam. Euh.. voir qu'est-ce qui va se passer, si l'Irak sera occupé ou libéré »

b. « Justement j'ai un petit article sur ce point-là, donc selon Tony Blair, euh.. l'après-Saddam, c'est-à-dire l'Irak de l'après-Saddam va être géré par des irakiens. »

In order to improve our segmentation scores, the overlapping of segments should be fully solved. Several methods should be implemented and evaluated in order to solve jointly document and speech segments overlapping. With the filtering step that consists in the isolated nodes elimination within the clusters, as seen in section 4.3.2, the segmentation performances have been increased for stereotyped meetings. However, more work should be done, in order to solve overlapping in the non-stereotyped meetings.

CONCLUSION AND FUTURE WORK

In this article, a new method has been presented for segmenting thematically not only the meeting dialogs transcripts, but also all the documents discussed during the meeting. The method is based on the result of *multiple-alignments* of meeting documents with the meeting dialogs. The evaluation of our bi-modal thematic segmentation method, compared to other mono-modal methods (*Texttiling* and baseline methods), has shown very promising results. These satisfactory results tend to prove that thematic segmentation and thematic alignment are closely related, and further that combining modalities improves considerably the thematic segmentation scores. However, this bi-modal method seems more effective for speech transcripts segmentation than for documents, partially segmented when partially discussed, which is not the case in our evaluation.

In the near future, the overlapping of segments, currently under study, will be solved, since its effect on the segmentation results is important, especially in the non-stereotyped meetings, where overlapping is frequent. From another side, the nodes weights should be taken into consideration and included in the clustering process. This may improve the thematic segmentation by eliminating the weak and distant nodes.

Finally, other alignment pairs will be considered, such as the alignment between sentences and speech turns, or speech turns with document logical blocks. This should allow building a robust alignment and further a more performant thematic segmentation.

6. REFERENCES

[1] Anil K. J., Richard C. D., Algorithms for Clustering Data, Edition Hardcover, Publisher Prentice Hall College Div; 1988, ISBN 013022278X

[2] Beeferman D., Berger A. and Lafferty J. D., Statistical Models for Text Segmentation, Machine Learning 34(1-3), 1999, p177-210.

[3] Fasulto D., An Analysis of Recent Work on Clustering Algorithms, University of Washington, Technical Report, April 1999.

[4] Ferret O., Grau B. and Masson N.: Thematic Segmentation of Texts: Two Methods for Two Kinds of Text. COLING-ACL, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics 1998, Montreal, Canada.

[5] Golumbic M.C., Algorithmic Graph Theory and Perfect Graphs, Second Edition, 2004, Edition Hardcover, Publisher Academic Press 1997, ISBN: 0-444-51530-5

[6] Hadjar K., Rigamonti M., Lalanne D. and Ingold R., Xed: a new tool for eXtracting hidden structures from Electronic Documents, Proceedings of DIAL'04, USA 2004, p212-224.

[7] Hearst M., Multi-Paragraph Segmentation of Expository Text, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics 1994, Las Cruces, New Mexico.

[8] Kehagias A., Pavlina F. and Petridis V., Linear Text Segmentation using a Dynamic Programming Algorithm, Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003, p171-178.

[9] Lalanne D., Mekhaldi D. and Ingold R., Talking about Documents: Revealing a Missing Link to Multimedia Meeting Archives, Document Recognition and Retrieval XI, IS&T/SPIE's International Symposium on Electronic Imaging, USA, 2004, pp.82-91.

[10] Looney C., Interactive Clustering and Merging with a New Fuzzy Expected Value, Pattern Recognition, August 2002.

[11] McQueen, J. Some methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, p281-297.

[12] Mekhaldi D., Lalanne D. and Ingold R., Thematic Alignment of Recorded Speech with Documents, Proceedings of ACM Symposium on Document Engineering, France 2003, p52-54.

[13] Mekhaldi D., Lalanne D. and Ingold R., Thematic Segmentation of Meetings through Document/Speech Alignment, to be published in ACM Multimedia 2004, 12th Annual Conference, New York, October 2004.

[14] Pevzner L. and Hearst M., A Critique and Improvement of an Evaluation Metric for Text Segmentation, Computational Linguistics 28(1), 2002, p19-36.

[15] 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, USA.

[16] Xie X.L., Beni G., A Validity Measure for Fuzzy Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, N°4, August 1991.

[17] Zhao Y. and Karypis G., Criterion Functions for Document Clustering, University of Minnesota, Technical Report, February 2002.