# Documents statiques et multimodalité

## L'alignement temporel pour structurer des archives multimédias de réunions

**Denis Lalanne et Rolf Ingold**

*DIVA/DIUF*
*Université de Fribourg*
*Chemin du Musée 3*
*1700 Fribourg*
*Denis.Lalanne@unifr.ch, Rolf.Ingold@unifr.ch*

RÉSUMÉ. *Cet article illustre le rôle central que peuvent jouer les documents statiques dans des applications multimodales et en particulier dans l'analyse et l'indexation d'enregistrements de réunions. L'article montre ainsi l'apport des structures de documents pour segmenter des réunions, tâche complexe en utilisant uniquement l'audio et la vidéo, et propose quatre étapes afin de lier temporellement les documents et des données multimédias de réunions. D'abord, une installation permet d'enregistrer toutes les modalités d'une réunion (audio,* vidéo*, document, etc.). Une méthode hybride d'analyse crée ensuite une représentation multicouche des documents électroniques. Les documents, ainsi représentés, sont ensuite alignés avec la parole et la vidéo. Finalement, une interface utilisateur, basée sur les documents et bénéficiant de tous les liens temporels construits, permet de naviguer sur des archives multimédias de réunions.*

ABSTRACT. *Printable documents play a central role in multimodal applications such as meeting recording and browsing. They provide a variety of structures, in particular thematic, for segmenting meetings, structures that are often hard to extract from audio and video. In this article, we present four steps for bridging a temporal link between documents and multimedia meeting archives. First, a document-centric meeting environment is presented, then, a document analysis tool, which builds a multi-layered representation of documents and creates index that are further used by document/speech and document/video alignment methods. Finally, a document-enabled browsing system, putting all the alignments together, is described.*

MOTS-CLÉS : *Alignement temporel, document numérique, multimodalité, segmentation thématique, structure physique et logique, recherche d'information.*

KEYWORDS: *Document temporal alignment, digital document, multimodality, thematic segmentation, layout and logical structure, information retrieval.*

## 1. Introduction

Les documents multimédias que manipulent les systèmes d'informations modernes peuvent contenir de nombreuses modalités : du texte, des graphiques, des images, mais aussi des animations, de la vidéo ou encore du son. Dans notre étude, nous nous intéressons à des documents qui peuvent être imprimés, c'est-à-dire restitués sur papier. Nous les appelons documents statiques, dans le sens de atemporel, par opposition aux documents dynamiques qui contiennent des modalités temporelles (son, vidéo ou animations). De plus, il est très important selon nous de les distinguer des données multimédias, telles que l'audio ou la vidéo, qui possèdent une temporalité intrinsèque et imposent au lecteur un ordre et une vitesse de lecture.

Les documents statiques, que nous considérons dans cet article, sont des objets graphiques qui peuvent être imprimés et qui possèdent du texte, des images, des figures ou des liens vers d'autres médias. Ces documents, par exemple des rapports scientifiques, des revues, des journaux, des quotidiens, etc., sont à prédominance textuelle et possèdent des structures physiques et logiques complexes. Cette définition de document statique n'exclut pas la prise en compte de la composante évolutive d'un document (son historique, les versions successives, etc.).

L'objectif du travail présenté dans cet article est de créer des liens temporels entre les documents statiques et des données multimédias. La création de ces liens devrait permettre de temporaliser les documents statiques et donc d'utiliser ce médium, hautement thématique, comme interface d'accès à des données multimédias. Citons à titre d'exemple, les archives multimédias de conférences dans lesquelles il devient courant de trouver pour chaque article sa forme numérique au format PDF, le diaporama correspondant et enfin un enregistrement audio/vidéo de la présentation de l'article par l'auteur. La création de liens document/image et document/parole, présentés dans cet article, permettra de lier temporellement tous les documents, aussi bien statiques (article PDF) que multimédias (présentation PPT et enregistrement audio/vidéo), et d'utiliser les documents statiques comme interfaces d'accès à des données multimédias.

## 2. Analyse de réunions et documents statiques

Les recherches en informatique connaissent actuellement un engouement important pour l'enregistrement et l'analyse de réunions, principalement parce que les réunions sont hautement multimodales, ce qui est une caractéristique essentielle afin de concevoir et d'implémenter les systèmes de communication à venir. De nombreux projets de recherche se concentrent actuellement sur l'analyse et l'annotation de réunions afin d'améliorer l'indexation et donc la navigation sur des corpus multimédias de réunions. Cependant, la plupart de ces projets ne prennent pas en compte les documents statiques traditionnels, qui sont pourtant une partie intégrante de la grande majorité des réunions.

Le document est depuis des siècles le vecteur principal permettant à des humains de communiquer et de stocker de l'information. Avec les progrès récents dans le multimédia et les applications multimodales, d'autres modalités, telles que l'audio ou la vidéo, apparaissent afin d'échanger des informations. Ces progrès consolident le rôle des documents traditionnels, qui co-existent aussi bien dans le monde physique que dans le monde digital. Les documents sont hautement thématiques et structurés, facilement indexables et récupérables, et peuvent donc constituer des vecteurs, ou interfaces, naturelles et thématiques pour accéder et naviguer sur des archives multimédias. Pour cette raison, il est essentiel de mettre en valeur les liens qui unissent les documents avec des médias temporels tels que la vidéo ou l'audio. De plus, nous pensons que ces liens faciliteront notablement l'accès aux enregistrements de réunions, ainsi que la conception d'interfaces pour les utilisateurs qui amélioreront la navigation et la recherche à travers des corpus multimédias.

De nombreux projets de recherche visent à archiver les enregistrements de réunions dans une forme adéquate pour la recherche et la navigation. Les objectifs généraux de ces projets sont l'avancement de la recherche sur (a) l'analyse de données multimodales et (b) la recherche d'informations multimédias. Deux directions émergent de l'état de l'art actuel des projets de recherche sur l'analyse des réunions (Lalanne *et al.*, 2003).

Le premier groupe se concentre sur des annotations de documents de type prise de notes, ou analyse de diaporama : MS (Cutler *et al.*, 2002), FXPal (Chiu *et al.*, 2000), eClass (Brotherton *et al.*, 1998), DSTC (Hunter *et al.*, 2001) et Cornell (Mukhopadhyay *et al.*, 1999). Ces recherches proposent des interfaces pour les utilisateurs permettant de naviguer sur des réunions, en utilisant une visualisation des changements de diapositives dans les diaporamas comme outil de navigation, ainsi que les notes prises par des participants. Dans ces interfaces, les diapositives et les prises de notes sont des index visuels permettant de localiser rapidement des segments de réunions intéressants afin de rejouer les séquences audio/vidéo correspondantes.

Le second type de recherche se concentre sur les annotations de la parole, comme par exemple la transcription des dialogues : ISL (Bett *et al.*, 2000) et eClass (Brotherton *et al.*, 1998). Ce type de recherches propose des interfaces utilisateurs permettant de faire des recherches par mots-clefs dans les transcriptions de la parole. Dans ce contexte, des annotations de plus haut niveau, telles que les actes de dialogue (Stolcke *et al.*, 1998) ou les épisodes thématiques peuvent aussi être utilisés afin d'obtenir des index temporels pour accéder rapidement à des morceaux choisis de réunion.

Les applications basées sur les documents et les systèmes basés sur la parole correspondent respectivement aux modalités de communication visuelles et verbales d'une réunion. Ces modalités étant intégrées dans le monde réel, nous proposons de créer des liens entre elles, et de les intégrer dans une archive de réunions et dans les

interfaces de navigation correspondantes. De plus, nous proposons de considérer conjointement a) les liens linguistiques entre le contenu des documents statiques et la transcription de la parole et b) les similarités graphiques entre les documents statiques et les enregistrements vidéo. La construction de ces liens devrait ainsi aboutir à un alignement complet entre les documents statiques et des données temporelles.

Dans la suite de cet article, nous présentons une application multimodale, dans laquelle des réunions sont enregistrées, archivées, indexées puis interrogées. L'objectif de cette application est d'annoter les enregistrements de réunions, et de créer des liens temporels entre les documents statiques et les autres médias, afin de :

- Produire semi-automatiquement des procès verbaux multimédias, sorte de comptes-rendus interactifs ;
- Construire des interfaces qui utilisent les documents statiques comme vecteurs thématiques et structurés pour naviguer sur des archives de réunions.

Rappelons qu'un compte-rendu de réunion est traditionnellement un document statique qui synthétise ce qui s'est passé durant une réunion, pour action, diffusion, etc. Dans notre application, la création d'un compte-rendu dynamique de réunion, qui permette d'organiser l'accès aux enregistrements de celle-ci, est un objectif à long terme. Viser une solution entièrement automatique n'est pas réaliste pour l'instant. Par contre, la création d'un navigateur permettant à un scribe de générer semi automatiquement un procès verbal sera envisagée à moyen terme (voir section 6).

Dans le contexte de cette application, nous présentons quatre étapes pour combler le fossé entre les documents statiques, non temporels, et des données multimédias de réunions. Ces documents permettront de structurer les enregistrements de réunions et serviront d'artefact de navigation. La troisième section présente l'objet des enregistrements ainsi que le dispositif. La section 4 présente une nouvelle approche de reconnaissance et d'indexation de documents électroniques, combinant des méthodes d'analyse d'image avec une extraction du contenu numérique. La section 5 introduit l'alignement temporel de documents, qui permet de synchroniser les documents statiques avec les autres médias enregistrés pendant une réunion. Enfin, la section 6 présente la dernière étape, qui utilise tous les travaux qui précèdent : une interface basée sur les documents, permettant de naviguer sur des archives multimédias d'enregistrements de réunions.

## 3. Enregistrement de réunions

Un environnement d'enregistrement de réunions a été mis en place dans notre laboratoire, en collaboration avec l'Ecole d'Ingénieurs et d'Architectes de Fribourg (figure 1). Cet environnement capture aussi bien des données audio et vidéo, pour

chacun des participants à la réunion, que les documents statiques qui sont projetés, discutés ou simplement présents sur la table durant la réunion. L'équipement a été installé dans une salle de réunion existante, et permet d'enregistrer jusqu'à 8 participants en gros plan. La salle enregistre plusieurs modalités liées aux documents grâce à une douzaine de caméras et huit microphones. Ces périphériques, ainsi qu'un projecteur vidéo, sont connectés à différents ordinateurs personnels, contrôlés et synchronisés par un ordinateur maître. Sur ce dernier, une application ergonomique, dédiée à la capture de réunions, permet de spécifier les sièges, correspondant à un couple caméra/microphone, qui doivent être enregistrés, les périphériques qui doivent être actifs et de nombreuses autres options de contrôle. Une fois l'enregistrement terminé, l'application gère de plus tous les post-traitements (e.g. compressions, analyses, etc.) ainsi que l'archivage des réunions sur un serveur de fichiers.

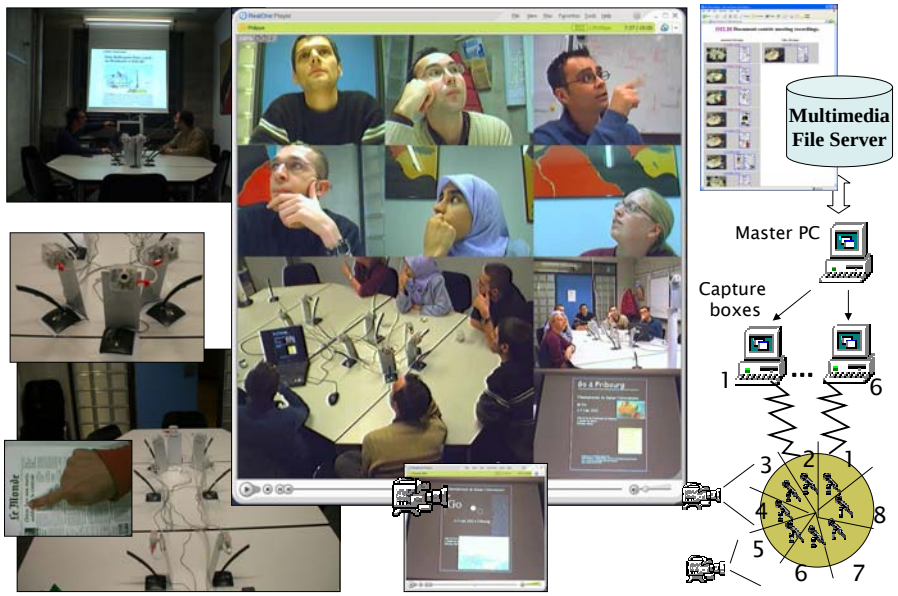


**Figure 1.** *L'environnement d'enregistrement de réunions installé à l'université de Fribourg. A gauche une photo de la salle d'enregistrement. Au centre le résultat de la capture d'une réunion présenté à l'aide de SMIL sous la forme d'une mosaïque de clips* vidéo. *Finalement, à droite l'architecture de la salle, un ordinateur maître contrôle et synchronise les enregistrements sur les stations de capture, sur lesquelles jusqu'à 3 paires de caméra/microphone sont branchées. Une fois l'enregistrement stoppé, tous les fichiers audio/*vidéo *sont transférés automatiquement sur un serveur de fichiers.*

A ce jour, une trentaine de réunions ont été enregistrées : des défenses de projets d'étudiants, des simulations d'entretiens d'embauche, des ateliers de lecture d'articles, etc. En particulier, 22 revues de presse ont été enregistrées et transcrites manuellement. Lors de ces revues de presse, entre 3 et 6 personnes discutent et débattent autour des unes du jour de différents journaux francophones. Les réunions durent en moyenne 15 minutes. Les unes des journaux, originellement en PDF, ont été transformées au format XML, et enrichies de leurs structures physiques et logiques. De plus, chaque enregistrement de réunions est accompagné d'une structure arborescente contenant des informations générales, telles que la date, le titre, la liste des participants et leur siège respectif. Bien évidemment, l'enregistrement contient aussi une vidéo et un fichier audio par participant, ainsi que tous les documents annexés à la réunion, au format PDF et image.

## 4. Analyse de documents

Les documents jouent un rôle important dans les communications journalières. Avec l'accroissement constant de l'Internet, un nombre considérable de documents sont publiés et consultés en ligne. Malheureusement, les différentes structures des documents sont très rarement exploitées, ce qui réduit considérablement les facultés de navigation et de recherche des utilisateurs. De nombreux niveaux d'abstraction sont présents dans les documents, dissimulés dans ses différentes structures : physique, logique, thématique, relationnelle, et même temporelle.

La structure physique désigne souvent la segmentation d'un document en zones homogènes, partageant les mêmes propriétés typographiques pour les zones textuelles, les mêmes propriétés graphiques ou encore contenant des images (Ishitani, 1999). La structure logique, quant à elle, est dérivée de la structure physique, et utilise des modèles de document, afin d'extraire une description symbolique de la structure et du contenu. Les blocs logiques induits peuvent être le résultat du regroupement de plusieurs blocs physiques, et représentent des entités sémantiques de document, comme par exemple un titre, un article, un nom d'auteur, un résumé, etc. (Niyogi *et al.*, 1995). La structure thématique d'un document ne travaille que sur le contenu textuel du document et correspond à une segmentation de ce contenu en différents blocs sémantiquement homogènes, correspondant à des thèmes différents (Salton *et al.*, 1996). Enfin, la structure temporelle d'un document est multiple. Cette dernière vise à temporaliser les différentes structures de document, en considérant toutes les interactions ou opérations qui ont été effectuées sur un document, comme par exemple des modifications, la publication, la projection de ses parties lors d'un diaporama, les discussions sur son contenu, les interactions gestuelles, etc. Nous considérons que pour être intégrées pleinement dans des archives multimédias, il est nécessaire que toutes ces structures co-existent et qu'elles soient considérées conjointement dans une représentation multicouche du document.

Dans la plupart des moteurs de recherche et des systèmes de recherche d'information, cette structure multicouches n'est pas prise en compte, et les documents sont indexés, dans le meilleur des cas, par leur structure thématique ou simplement représentés par un groupe de mots. La mise en page des documents, c'est-à-dire leurs structures physiques et logiques, est sous-estimée et pourrait fournir des indices significatifs sur l'organisation logique des documents. Ainsi, nous pensons que l'extraction des structures des documents peut améliorer considérablement (a) l'indexation et la recherche de documents et (b) leurs associations avec d'autres médias.

Nous avons choisi d'analyser des documents au format PDF principalement car le PDF est devenu le format pivot pour échanger des documents statiques et parce qu'il préserve la mise en page. L'utilisation du PDF est souvent limitée à l'affichage sur un écran et à l'impression, malgré le bénéfice que pourrait apporter les structures sur la recherche et la récupération de documents. En effet, nous pensons que l'extraction des structures physiques et logiques des documents pourrait grandement enrichir l'indexation des fichiers PDF et leur liaison avec d'autres médias. En particulier, dans le cadre d'applications multimodales, telles que l'enregistrement et l'analyse de réunions, l'extraction des structures de documents permet de lier les documents PDF avec la transcription de la parole et avec l'image des documents dans les enregistrements vidéo de la réunion. Pour cette raison, nous avons proposé récemment une approche hybride qui consiste à fusionner (a) des méthodes d'extraction de bas niveau, basées sur la forme électronique du document, à (b) des méthodes d'analyse d'image du document, converti au format TIFF à partir du PDF, permettant d'extraire la structure physique du document (Hadjar *et al.*, 2004). A première vue, il serait plus naturel d'extraire la structure physique directement à partir du fichier PDF, en se servant de sa structuration interne. Notre expérience nous a toutefois montré que cela pose de grandes difficultés parce que les informations structurelles ne sont pas toujours fiables. Dans des documents multi-colonnes, l'ordre d'apparition des blocs de texte ne reflète en général pas l'ordre de lecture. Pire, il arrive que des portions de phrase ou des mots isolés n'apparaissent pas dans leur contexte mais de manière isolée à la fin d'un fichier. Nous avons de bonnes raisons de penser que ce type d'artefact dépend de l'historique du document et des logiciels qui ont servi à le produire. L'analyse à partir de l'image TIFF présente pour nous l'avantage de considérer une représentation quasi universelle.

Dans le domaine de l'analyse de documents, la segmentation de l'image vise à morceler une image de document en différentes zones homogènes possédant des propriétés graphiques similaires, e.g. texte, image, graphiques, etc. Notre algorithme de segmentation commence par extraire les filets, les trames et les lignes de texte, puis sépare les zones de texte des zones d'image, puis finalement fusionne les lignes de texte dans des blocs homogènes. L'algorithme prend en entrée une image au format TIFF, générée à partir d'un fichier PDF, et renvoie en sortie un fichier au

format XML, qui décrit la segmentation physique du document en zones, comme mentionné précédemment.

En parallèle, les différents objets contenus dans le document PDF, aussi bien textuels que graphiques, sont extraits en manipulant directement le contenu du fichier électronique. D'abord, le fichier PDF est clarifié, c'est-à-dire que les ambiguïtés sont éliminées. Puis, les différentes représentations sont homogénéisées, et le contenu du document PDF est projeté dans un arbre, qui peut ensuite être transformé soit en SVG soit dans une forme canonique et structurée, qui décrit complètement la structure physique du document et de son contenu.

Finalement, les objets extraits du document PDF sont mis en correspondance avec les résultats de l'analyse de la structure physique du document, afin de construire une représentation arborescente du document. Par exemple, les positions du texte extrait de la forme électronique sont comparées aux positions des boîtes de chaque bloc physique, afin d'associer à chaque bloc physique son contenu textuel (Hadjar *et al.*, 2004).

Cette approche n'a pas encore été évaluée quantitativement. La base de données considérée jusqu'à présent était composée de documents PDF à structures complexes, des unes de quotidiens principalement. Les résultats de l'extraction que nous avons obtenus sur une centaine de unes de journaux francophones, anglophones, italophones et arabes sont satisfaisants à l'oeil nu en ce qui concerne la préservation de la mise en page. Afin de s'assurer que l'extraction du contenu du document est correcte et qu'elle préserve effectivement la mise en page, il serait possible de prévoir une évaluation automatique en superposant l'image du document PDF avec celle obtenue par l'extraction et de calculer le taux de ressemblance à travers un calcul de distance. Finalement, afin d'évaluer les performances d'extraction des structures physiques et logiques, la constitution de vérités-terrains[1] (ground-truth) devra être considérée, à l'aide d'outils semi-automatiques supervisés par un opérateur humain.

En ce qui concerne l'analyse des journaux, notons qu'il est souvent impossible de récupérer les fichiers produits avant impression par les « quotidiens » ; cela nécessiterait des accords difficiles à négocier. La méthode que nous proposons se veut générale et fonctionne directement sur un document PDF, qui est le format pivot actuel.

Afin de produire des vérités-terrains fiables pour nos données de revues de presse, la segmentation de documents a été complétée manuellement. Les documents PDF correspondant aux unes des différents journaux francophones, discutées pendant les enregistrements de réunions, ont tout d'abord été convertis automatiquement dans une forme canonique et structurée, contenant la structure

1. Aussi connu sous le terme « étalon » dans la communauté linguistique mais nous avons préféré le terme « vérités-terrains » pour faire ressortir le fait qu'il s'agit de données réelles et non de données synthétisées sur mesure pour étalonner le système.

physique complète de chacun des documents, en utilisant la méthode hybride d'analyse de fichiers PDF que nous venons de présenter. La structure logique des documents a ensuite été annotée manuellement et liée à la structure physique extraite précédemment. Afin d'automatiser la phase de structuration logique, une connaissance de tous les modèles de documents, contenus dans l'archive multimédia de réunions, serait nécessaire. A long terme, l'objectif de notre analyse de fichiers PDF, est d'automatiser tout le processus. La DTD, résumée dans la figure 2, présente la structure logique d'une une de quotidien que nous avons utilisée.

```
Newspaper      ->     Date, Name, MasterArticle, Highlight*, Article+,
                      Other*, Filename
MasterArticle  ->     Title, Subheading?, Summary*, Author*, Source?,
                      Content?, Reference?, Other*, JournalArticle*
Article        ->     Title, Subtitle?, Source?, Content, Author*,
                      Summary*, Reference*, Other?
JournalArticle ->     Title, Source?, Summary*, Content?, Reference+
Highlight      ->     Title, Subtitle, Reference+
```

**Figure 2.** *DTD de la structure logique d'une une d'un journal. Les éléments terminaux contiennent du texte (#PCDATA). Les éléments MasterArticle, Article, JournalArticle et Highlight ont tous un attribut ID.*

## 5. Alignement entre des documents statiques et des données temporelles

Afin de naviguer sur des archives multimédias au travers des documents, il est tout d'abord nécessaire de construire des liens entre les documents statiques et d'autres médias, qui sont eux temporels, tels que l'audio ou la vidéo. Nous appelons « alignement temporel des documents » l'opération qui consiste à extraire les relations entre des portions de documents, à différents niveaux de granularité, et le temps de présentation dans la réunion. L'alignement temporel de documents crée des liens entre des extraits de documents et les intervalles de temps dans lesquels ils étaient soit (a) dans le discours, soit (b) dans le champ visuel soit (c) dans le champ de l'interaction gestuelle d'une réunion. Il est donc possible de mettre en correspondance des extraits de documents avec des extraits audio et vidéo, et par extension avec des annotations de la parole, de la vidéo et/ou des gestes. Nous avons identifié trois modalités qui peuvent être associées et alignées temporellement avec les documents :

- La parole : le contenu textuel des documents est comparé avec la transcription de la parole, qui renferme des index temporels pour chaque tour de parole des interlocuteurs et pour chaque énoncé de parole. Les tours de parole sont des monologues, i.e. des segments du dialogue où un seul interlocuteur s'exprime, divisés en énoncés de parole. Un énoncé de parole est une partie cohérente d'un monologue à laquelle peut être associé un acte de dialogue tel qu'une question, une réponse, un remerciement, un désaccord, etc. (Stolcke *et al.*, 1998). Un énoncé de parole correspond plus ou moins à une phrase dans un document statique.
- La vidéo et l'image : les documents électroniques sont comparés avec les images extraites des enregistrements vidéo de documents (e.g. la vidéo des diaporamas projetées sur un écran) afin d'identifier les différents documents visibles dans les vidéos et d'associer aux documents concernés des index temporels liés à leurs périodes d'apparition dans le champ visuel des participants.
- Les gestes : les interactions gestuelles avec des documents sont capturées et analysées (e.g. pointer du doigt un document projeté sur un écran) afin d'en déduire à quel moment et quelle partie de document était dans le champ d'interaction gestuelle des participants.

Nous n'avons pas encore démarré les travaux sur l'alignement document/geste et donc ne présenterons en détails dans les sections qui suivent que l'alignement avec la parole puis celui avec les enregistrements vidéo.

L'alignement avec les gestes ne se différencie pas nettement de l'alignement document/vidéo puisqu'il ne s'agit pas d'analyser directement les gestes mais les enregistrements vidéo dans lesquels des interactions gestuelles apparaissent. La méthode que nous prévoyons d'utiliser pour résoudre cet alignement combine des techniques de deux domaines bien établis : a) l'interaction gestuelle et b) l'analyse de documents. L'analyse des postures et des gestes mène à des annotations de haut niveau sur les gestes (e.g. pointer, entourer, souligner, etc.) avec les marqueurs temporels associés de début et de fin d'interaction. En outre, les techniques d'analyse de documents, comme présentées dans la section 4, fournissent des méthodes pour extraire les structures physiques et logiques de documents électroniques, tels que des fichiers au format PDF, ce qui permet de déterminer quel bloc du document est pointé, encerclé ou souligné par un utilisateur. L'interaction gestuelle avec des documents n'a, à notre connaissance, été que très peu traitée, et devrait mener à des annotations d'un type nouveau, et aboutir à des applications temps-réel qui utilisent les documents papier comme moyen d'accéder à des données numériques et multimédias (Klemmer et al., 2003) (Wellner, 1993).

### 5.1. *Alignement parole / documents*

Dans l'alignement parole/document, les contenus de la parole et des documents statiques sont comparés, contenus qui peuvent être aussi bien textuels que structurels, afin de détecter des citations ou des paraphrases, des références ou encore des liens thématiques. Les citations sont des concordances lexicographiques exactes entre les mots écrits dans les documents et les mots prononcés dans la parole. Les paraphrases correspondent à une mise en parole, souvent une reformulation, d'une phrase écrite. Les références établissent des liens entre des expressions de la transcription structurée des dialogues d'une réunion et des éléments de la structure logique d'un document (e.g. « l'article à propos de l'Irak », « le titre de la une du journal », etc.). Pour conclure, les alignements thématiques sont des mises en correspondance sémantiques entre des unités de documents (phrases, blocs logiques, etc.) et des unités de la transcription de la parole (énoncés, tour de parole, etc.). La détection des citations et des paraphrases n'ayant pas encore été implantée, nous ne présentons dans la suite de cet article que l'alignement thématique ainsi que la résolution des références aux documents dans les dialogues.

Cet alignement document/parole permettra de répondre à deux questions :

- Quand fut discuté, ou référencé, un document ?
- Qu'est-ce qui a été dit à propos d'un document ?

#### *5.1.1. Alignement thématique*

Un alignement thématique robuste a déjà été implanté, en utilisant plusieurs métriques de similarité telles que la mesure du cosinus, Jaccard ou Dice (Manning, 1999) et en considérant les unités de documents et de la transcription de la parole comme des ensembles de mots. Après avoir supprimé les mots-outils, i.e. les mots les plus courants (« stopwords »), et après avoir analysé les flexions des autres afin de les réduire à leur radical (« stemming »), par suppression des formes conjuguées et des pluriels principalement, le contenu des différents éléments des documents est comparé avec le contenu des différentes unités de la transcription de la parole.

Nous avons considéré 8 réunions afin d'évaluer notre méthode à différents niveaux de granularité de la structure des documents et de la transcription manuelle des dialogues. Les réunions ont été transcrites manuellement à l'aide de *Transcriber*, un logiciel d'aide à l'annotation de signaux de parole qui offre une interface graphique simple permettant à un utilisateur non informaticien de segmenter des enregistrements de longue durée, de les transcrire et de marquer les tours de parole, la segmentation thématique et les conditions acoustiques (Barras *et al.*, 2000). *Transcriber* était utilisé conjointement avec une mosaïque de tous les enregistrements vidéo de la réunion, une vidéo par participant (au centre de la figure 1), ce qui aidait grandement le scribe à transcrire la réunion dans les cas où

plusieurs personnes parlent en même temps (« speech overlapping »). Ces réunions avaient en moyenne une durée de 15 minutes chacune, et impliquaient de 3 à 6 participants. Un nombre total de 572 énoncés et de 228 tours de parole ont été alignés avec les documents. Et un nombre total de 90 blocs logiques (principalement des articles) et de 1409 phrases, extraites des unes de journaux, ont été alignés avec la transcription de la parole.

Afin d'évaluer nos méthodes d'alignement, nous avons mesuré les valeurs de rappel et de précision par rapport aux vérités-terrains produites manuellement. La valeur de rappel correspond au nombre d'alignements corrects détectés par le système sur le nombre d'alignements présents dans les vérités-terrains et la précision correspond au nombre d'alignements corrects détectés par le système sur le nombre total d'alignements trouvés par le système.

Les valeurs de rappel et de précision sont relativement bonnes lorsque les énoncés de parole sont mis en correspondance avec le contenu des blocs logiques de document (e.g. article, titre, etc.). En utilisant la mesure du cosinus, la valeur de rappel est de 0.84 et la valeur de précision de 0.77, ce qui représente des résultats encourageants. Lorsque les tours de parole sont mis en correspondance avec des blocs logiques de document, la valeur de rappel reste à 0.84 et la valeur de précision atteint 0.85 (Mekhaldi *et al.*, 2003).

D'un autre côté, l'alignement entre les énoncés de parole et les phrases des documents est moins précis, mais il est plus intéressant car il ne requiert pas l'extraction de la structure logique des documents. En utilisant la mesure de similarité de Jaccard, la valeur de rappel est de 0.83 en moyenne, et la précision est de 0.76. Les documents PDF sont automatiquement convertis en texte, puis segmentés en phrases, et finalement mis en correspondance avec les énoncés de parole. Nous pensons que cette méthode simple d'alignement automatique peut aider à structurer conjointement les documents et la transcription de la parole.

La plupart des réunions testées étaient relativement stéréotypées ; les articles des journaux étaient présentés plutôt que discutés et débattus. Dans certaines réunions cependant, les participants ne suivaient pas de près le contenu des articles, débattant davantage de l'actualité de la journée dans le monde. Nous avons considéré qu'une réunion est non-stéréotypée lorsque le nombre d'énoncés de parole qu'elle contient est deux fois supérieur au nombre de tours de parole, alors que pour qu'une réunion soit considérée comme stéréotypée, ce rapport doit être inférieur à 2 (une moyenne 60 énoncés de parole pour 20 tours de parole dans notre corpus). Les réunions non-stéréotypées donnent donc une bonne indication des performances de notre méthode dans des cas réalistes. Dans ce dernier cas, les valeurs de rappel et de précision tombent significativement pour les paires énoncés/phrases (rappel : 0.74 et précision : 0.66) et restent relativement stables pour les paires énoncés/blocs logiques de document. Davantage de détails et de résultats peuvent être trouvés dans (Lalanne *et al.*, 2004).

Les épisodes thématiques, i.e. des portions de texte sémantiquement homogènes, n'ont été considérés ni pour les documents ni pour la transcription de la parole, principalement parce que les résultats des segmentations thématiques, de chacune des deux sources, en utilisant des méthodes classiques de l'état de l'art, n'étaient pas satisfaisants (Hearst, 1994). Nous avons implanté récemment une technique bi-modale qui segmente conjointement les documents et la transcription de la parole en épisodes thématiques (Mekaldi *et al.*, 2004). L'idée consiste à détecter les régions les plus connectées dans le graphe biparti constitué par l'alignement des documents avec la parole. Les groupes les plus denses, regroupés grâce à des techniques de « clustering », sont ensuite projetés sur chaque axe, respectivement sur l'axe des documents et sur l'axe de la transcription de la parole. L'évaluation de cette technique bi-modale de segmentation thématique a montré qu'elle était plus performante que des méthodes uni-modales, spécialement dans le cas où les réunions sont moins structurées thématiquement. Dans ce dernier cas, les documents fournissent une structure naturellement thématique sur laquelle la segmentation de la réunion peut s'appuyer (Mekaldi *et al.*, 2004).

### *5.1.2. Résolution des références aux documents dans les dialogues*

Pendant les réunions, les participants se réfèrent souvent à des documents ou à des parties de document. Afin de résoudre ces références, il est nécessaire de trouver les liens entre chacune des expressions référentielles [2] (ER) et les éléments de documents qui correspondent. Par exemple, si un participant dit : « Je ne suis pas d'accord avec le titre de notre dernier rapport », alors « notre dernier rapport » réfère à un document qui peut être retrouvé dans une archive de réunions, et « le titre de notre dernier rapport » se réfère à son élément *Titre*, une zone textuelle qui peut être extraite du document correspondant.

Nous avons ainsi implémenté un algorithme, qui s'inspire du travail sur la résolution des anaphores, qui essaye de résoudre ce type de références. Nous ne présentons pas en détails, dans la suite de cette section, les méthodes développées ni tous les résultats obtenus ; pour une étude complète, nous vous invitons à lire les travaux réalisés en collaboration avec l'université de Genève (Popescu-Belis & Lalanne, 2004). Dans l'étude présentée ci-dessous, les expressions référentielles ont été détectées manuellement. Elles ont été ensuite analysées afin d'en dériver des expressions régulières qui les décrivent, et qui pourraient être utilisées pour les reconnaître automatiquement.

Dans le cadre des réunions de type revue de presse, l'algorithme parcourt la transcription de la parole d'une réunion en suivant l'ordre chronologique, et à tout moment met à jour le document courant, i.e. la une d'un quotidien, et l'article courant. Chaque expression référentielle est d'abord associée au document le plus

2. Parmi les expressions référentielles, il y a non seulement les références aux documents, i.e. des ERs intermodales, mais aussi les références aux propos tenus dans le discours, i.e. des ERs intramodales. Seules les références aux documents ont été considérées dans cette étude.

cité dans la transcription et présent dans la liste des documents associés à la réunion. Les expressions référentielles qui contiennent explicitement le nom d'un quotidien sont référencées au quotidien respectif ; les autres sont supposés référer au document courant, c'est-à-dire qu'elles sont des anaphores.

L'algorithme tente ensuite d'associer un élément du document (article, titre, auteur, etc.), déduit de la structure logique du document (voir section 4, figure 2), à chaque expression référentielle. L'algorithme décide d'abord si l'expression référentielle est anaphorique [3] ou non, en la comparant avec une liste d'anaphores typiques, telles que « l'article », « cet article », « il », « l'auteur », etc. Si l'expression référentielle courante est une anaphore, alors son référent est simplement l'article courant du quotidien courant. Si l'expression référentielle courante n'est pas une anaphore, c'est-à-dire qu'elle introduit un nouveau référent, alors une procédure de comparaison est appliquée afin de sélectionner l'article du document courant qui correspond le mieux. La procédure compare (a) le contenu de l'expression référentielle, plus le contenu de son contexte gauche (i.e. les mots situés à gauche de l'ER dans l'énoncé) avec (b) les articles du document courant, pour lesquels titre, auteurs, et contenu sont considérés séparément. Le référent de l'expression référentielle est finalement l'article qui obtient le score de comparaison le plus élevé.

Les premiers résultats obtenus en utilisant cet algorithme sur un ensemble de 14 revues de presse, et 322 expressions référentielles annotées manuellement, sont encourageants. L'identification du document référencé par chaque expression référentielle est correcte dans 98% des cas, et la précision pour associer chaque expression référentielle à des éléments de document est de 64%. Ce qui doit être comparé avec les résultats d'une méthode de base, comme par exemple « toutes les ERs réfèrent à la une » (16% de précision) ou « toutes les ERs réfèrent à l'article principal (MasterArticle) » (18% de précision). De plus, si les anaphores ne sont pas considérées dans le processus de résolution, c'est-à-dire si la comparaison ER/article est effectuée pour toutes les ERs, alors le score tombe à 54% de précision, ce qui prouve l'utilité de la détection des anaphores. D'un autre côté, et pour conclure, si les contextes, qui entourent les ERs, ne sont pas considérés lors de la comparaison, alors le score tombe à 27%.

Finalement, nous travaillons actuellement sur un modèle permettant de fusionner les différents types d'alignements présentés dans cet article, i.e. citations, références et thématiques. Nous espérons que ce modèle permettra de les comparer et de les corriger, afin d'obtenir un alignement document/parole robuste.

3. Une anaphore est une relation entre une ER antécédente qui définit l'entité de discours, i.e. dans notre cas un élément de document (e.g. « l'article sur la guerre en Irak… »), et une ER anaphorique qui se réfère à l'entité définie par la première (e.g. « cet article… »).

### 5.2. *Alignement documents / vidéo*

L'alignement documents/vidéo construit des liens entre des extraits de documents et des segments de séquences vidéo, qui correspondent soit à des diaporamas filmés soit plus généralement à des films dans lesquels des documents sont visibles. Cette approche s'appuie sur l'observation des évènements liés à des documents, qui sont visibles pendant des réunions, cours ou présentations, comme par exemple les changements de diapositive dans des diaporamas, ou des documents qui circulent sur une table et qui sont pointés. Dans un premier temps, notre méthode détecte tous les évènements liés à des documents (e.g. un changement de diapositive, une animation, etc.) et extrait une image qui correspond à ce segment stable de vidéo ; dans un deuxième temps, elle associe une signature visuelle à cette image basse-résolution, qu'elle compare finalement avec toutes les signatures des images haute-résolution, contenues dans la base de données de référence, afin d'identifier l'image ; enfin, elle l'enrichit avec le contenu textuel associé au document électronique correspondant, en tenant compte des informations structurelles.

Cet alignement vise la résolution de trois types de questions :

- À quel moment un document, ou une partie de document, était-il dans le champ visuel des participants?
- De quel document, ou partie de document, s'agissait-il?
- Quel était le contenu de ce document ?

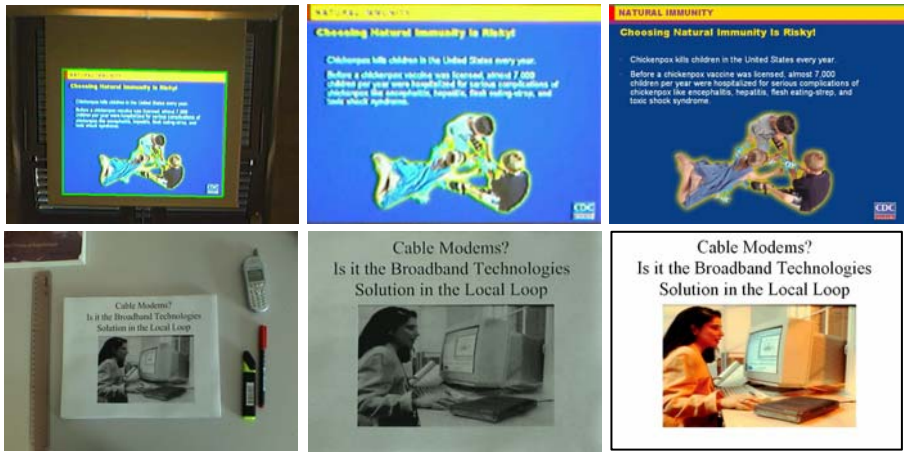


**Figure 3.** *A gauche une image extraite de l'enregistrement vidéo des documents projetés et en dessous des documents sur la table; au centre, une image extraite de la séquence vidéo reconstituée après détection de la zone contenant le document et correction des déformations ; finalement à droite l'image originale au format PDF.*

*5.2.1. Détection et extraction du document dans la* vidéo

Avant de détecter les évènements liés à des documents (e.g. changement de diapositive dans un diaporama), ces documents doivent être d'abord détectés dans les enregistrements vidéo, puis leur emplacement précis doit être calculé (figure 3). Les documents peuvent être de divers types, diapositives ou fichiers PDF pour les documents projetés, ou documents quelconques pour les documents filmés sur la table. A l'heure actuelle, nous utilisons une interface graphique pour définir manuellement la zone graphique où se trouve le document dans la vidéo. Une fois la zone définie, les déformations liées aux problèmes de perspectives sont corrigées et finalement, une vidéo, que nous appellerons *vidéo document* dans la suite de l'article, est reconstituée à partir des images de document redressées.

*5.2.2. Détection des changements de diapositive dans la vidéo document*

Nous proposons dans cette section une méthode permettant de détecter un changement de document dans l'enregistrement vidéo de l'écran de projection (figure 3). Cette méthode permet de répondre à la question présentée dans l'introduction de la section 5.2., i.e. « À quel moment un document était-il dans le champ visuel des participants ? ».

Au lieu d'essayer de détecter automatiquement un changement de diapositive, notre méthode cherche à identifier automatiquement les diapositives stables, c'est-à-dire les périodes suffisamment longues pour être considérées par un spectateur et durant lesquelles une seule et unique diapositive est projetée. Notre algorithme suit pour cela deux étapes distinctes : (a) d'abord il détecte les périodes stables puis (b) il cherche à l'intérieur des périodes instables la position exacte du changement de diapositive.

Les images sont d'abord extraites une à une de la *vidéo document*, ensuite un filtre passe-bas leur est appliqué afin de réduire le bruit et finalement un filtrage adaptatif est utilisé afin d'obtenir des images binaires, ce qui permet ainsi d'éviter les problèmes liés à l'éclairage non-uniforme des diapositives projetées (Mukhopadhyay *et al.*, 1999). La première image $F_S$ extraite de la *vidéo document* est d'abord comparée avec l'image $F_E$ se trouvant exactement 2 secondes plus tard dans la *vidéo document* correspondant au diaporama. Les deux images, $F_S$ et $F_E$, sont considérées identiques si le nombre de pixels noirs en commun dépasse un certain seuil, suffisamment bas (0.6) pour ne pas rater des périodes instables. Si elles sont similaires, la période est considérée comme stable et deux nouvelles images sont prises, pour $F_S$ et $F_E$, une demi-seconde plus en avant dans la vidéo et sont comparées de nouveau. Ce processus continue tant que le couple ($F_S$, $F_E$), pris toutes les demi-secondes, est suffisamment similaire. Lorsqu'une période de dissimilarité démarre, et que donc le couple $F_S$ et $F_E$ n'est plus suffisamment similaire, une file d'images, contenant toutes les images délimitées par $F_S$ et $F_E$, est construite. La première image de la file est ensuite comparée à toutes les autres images de la file afin de détecter la position exacte du changement de diapositive.

L'utilisation d'une webcam pour filmer le diaporama projeté sur un écran introduit de nombreux phénomènes perturbateurs dans la détection du changement de diapositive. Par exemple, un autofocus à chaque changement important de la luminosité implique que la vidéo met un temps non négligeable à se stabiliser après chaque changement de diapositive. Pour cette raison, la dissimilarité entre les images se stabilise graduellement. De plus, à la suite d'un changement de diapositive, l'image de la nouvelle diapositive risque d'être superposée avec la précédente (e.g. fade-in/fade-out), d'autant plus que le temps de transition entre deux diapositives est souvent supérieur au laps de temps entre deux images capturées par la vidéo. Pour cette raison, nous considérons que la position exacte du changement de diapositive est l'image dont la valeur de dissimilarité s'approche le plus de la moyenne entre la dissimilarité minimale et la dissimilarité maximale dans la file.

L'évaluation de la méthode précédente a été effectuée sur des enregistrements générés automatiquement en SMIL (Synchronized Multimedia Integration Language, consortium W3C). Chaque fichier SMIL, correspondant à un diaporama, contient les temps de début et de fin, choisis aléatoirement, pour chaque diapositive prise elle aussi aléatoirement dans un répertoire de diaporamas (Behera & Lalanne, 2004). Ces fichiers SMIL, au format XML, peuvent ainsi servir de vérités-terrains pour notre évaluation. 60 diaporamas ont été ainsi générés automatiquement et filmés, chacun contenant en moyenne une vingtaine de diapositives, ce qui représente approximativement un total de plus de 1000 changements de diapositives.

Afin d'évaluer la qualité de notre méthode, nous avons mesuré les performances de rappel et de précision. Le rappel représente le nombre de changements de diapositive correctement détectés sur le nombre de changements dans les vérités-terrains du diaporama (fichier SMIL). Alors que la précision représente le nombre de changements correctement détectés sur le nombre total de changements détectés.

L'évaluation a montré que notre méthode dépasse les performances des techniques existantes dans l'état de l'art pour la détection des périodes stables (rappel: 1.0, précision: 1.0), ainsi que pour la détection de la position exacte du changement de diapositive (rappel: 0.84, précision: 0.82 et rappel: 0.93, précision: 0.90 pour une tolérance de respectivement 1 ou 2 images de décalage). Si la tolérance est augmentée jusqu'à une demi-seconde de décalage, alors la performance de la détection du changement de diapositive devient maximale (rappel : 1.0, précision : 1.0).

### *5.2.3. Identification des documents visibles*

Pour chaque période stable, déterminée par notre méthode de détection du changement de diapositive, une image claire est extraite. L'image est ensuite comparée avec les images des documents originaux, stockés dans une base de

données, afin d'être identifiée. Notre méthode d'identification de documents est basée sur deux processus :

a) L'extraction d'une signature visuelle hiérarchiquement structurée, contenant des caractéristiques globales de l'image (nombre de pixels noirs, profils de projection, etc.) et une décomposition en zones (texte, image, puce, etc.). Cette signature visuelle est générée aussi bien pour l'image basse-résolution extraite de la vidéo, que pour l'image haute-résolution distillée depuis le document PDF. L'extraction de la signature visuelle est basée sur des méthodes d'analyse de l'image d'un document, telles que le RLSA (Run Length Smearing Algorithm [Wahl *et al.*, 1982] [Wong, 1982]), les composantes connexes, les profils de projection, etc.

b) Une comparaison multi-niveaux de ces signatures visuelles, qui suit leurs hiérarchies. Les caractéristiques de plus haut niveau sont d'abord comparées ; toutes les images dans la base de données qui dépassent un certain seuil de similarité, sont conservées. La comparaison continue sur ce sous-ensemble d'images en testant des caractéristiques de plus bas niveau. Lorsque tous les niveaux de l'arborescence de la signature visuelle ont été parcourus et que toutes les caractéristiques correspondantes ont été testées, une comparaison globale, combinant toutes les caractéristiques avec des poids dépendant du niveau dont elles sont issues, est effectuée. Les meilleures images, c'est-à-dire les plus similaires, sont conservées et la comparaison re-démarre à la racine de la signature visuelle avec des seuils de similarité plus hauts ; les comparaisons deviennent donc plus restrictives.

L'avantage principal de cette méthode est qu'elle ne requiert aucune technique de classification. Elle est rapide, principalement car la hiérarchie de la signature visuelle guide la recherche vers des sous-espaces de recherche fructueux. De plus, le fait d'alterner des comparaisons spécifiques à chaque caractéristique avec des comparaisons globales, garantit qu'aucune bonne solution n'est éliminée.

Afin d'évaluer notre méthode, nous avons d'abord capturé 500 images de diapositives projetées. Nous avons ensuite utilisé ces images basse-résolution comme requêtes sur une archive contenant plus de 1000 diapositives haute-résolution. Nous avons ainsi obtenu une valeur de rappel de 0.81 et une précision de 0.99. Les résultats de cette évaluation montrent que notre méthode fonctionne parfaitement sur des diapositives qui possèdent des arrière-plans homogènes, et ne renfermant pas de textures complexes. Afin d'améliorer les performances de notre méthode, nous planifions dans un futur proche, de considérer conjointement la signature visuelle présentée dans cette section, qui représente la structure physique du document, et les informations colorimétriques contenues dans les diapositives. Finalement, dans un futur proche, la méthode devrait être étendue à des documents disposés sur la table ou échangés entre les participants.

*5.2.4. Extraction du contenu des documents*

Aussi bien la signature visuelle des documents, présentée dans la section précédente, que le résultat de l'analyse du PDF, présenté dans la section 3, sont au format XML. Lorsqu'une zone de texte est présente dans l'image du document, extraite de la vidéo, les deux fichiers XML sont comparés afin d'associer un contenu textuel à chaque séquence vidéo correspondant à une diapositive. Il s'agit, dans cette comparaison, d'aligner a) la structure physique extraite du document PDF original avec b) la structure physique superficielle de l'image du document extraite de la vidéo, c'est-à-dire la signature visuelle. Cette procédure permet d'éviter l'utilisation d'un OCR, dont les performances seraient certainement médiocres, vu la qualité restreinte des images vidéo (à peu près 200 par 150 pixels).

*5.2.5. Segmentation et annotation de la* vidéo *basée sur les documents*

La segmentation et l'annotation de la vidéo du diaporama, ou de toute autre vidéo dans laquelle apparaît un document, sont stockées dans un fichier XML. Une fois que la détection du changement de diapositive a été effectuée, ou la détection de tout autre événement lié aux documents, les marqueurs temporels de début et de fin de séquences, ainsi que l'identificateur de la réunion, sont ajoutés au fichier d'annotation de la vidéo. Une fois que le document visible dans la séquence a été identifié, le document original, dans la base de données des réunions, est lié au fichier d'annotation. Finalement, après que le contenu du document PDF a été extrait et associé, le contenu textuel est ajouté à la séquence vidéo. Cette annotation de la vidéo permet ainsi de rechercher une séquence en faisant simplement une recherche à base de mots-clefs.

## 6. Une interface de navigation basée sur les documents statiques

Des projets de recherche ont récemment utilisé des techniques d'analyse d'images et de vidéos afin de créer automatiquement des index visuels et des résumés de séquences vidéo de réunion. Ces résumés visuels, qui ressemblent à des bandes dessinées animées, aident des utilisateurs à percevoir rapidement le déroulement d'une réunion et permettent de naviguer à travers son enregistrement multimédia (Uchihashi *et al.*, 1999). Cependant, ces méthodes sont souvent basées sur des caractéristiques visuelles de bas niveau et elles manquent considérablement d'informations sémantiques.

D'autres projets de recherche utilisent des techniques d'analyse de la parole et du langage, ou l'extraction des sous-titres des enregistrements vidéo en utilisant des OCRs, afin de créer des index plus performants et ainsi de permettre des mécanismes de recherche d'information (Smith *et al.*, 1998). Nous pensons que pour améliorer de tels systèmes de navigation, il est nécessaire de combiner tous les index disponibles dans une seule interface. De plus, nous souhaitons utiliser les documents comme interfaces d'accès puisque dans une grande proportion des

applications multimédias (e.g. cours, présentations, réunions, etc.), les documents statiques jouent un rôle prépondérant dans l'organisation thématique des discussions. De récents projets de recherche ont montré l'importance de la structuration des médias lorsque l'on désire les synchroniser (Tran_Thuong, 2001) et nous pensons que les documents statiques peuvent fournir cette structure. Ils sont naturellement structurés, aussi bien physiquement, logiquement que thématiquement, et, une fois liés temporellement avec les autres médias, peuvent constituer des vecteurs particulièrement adaptés afin de naviguer sur des archives multimédias de réunions.

Un prototype d'interface utilisant les documents statiques comme outil d'interaction, et permettant de naviguer sur des archives multimédias de réunions, est présenté sur la figure 4 puis sur la figure 5. Tout d'abord, la figure 4 représente une visualisation de tous les articles de journaux présents dans l'archive de revues de presse qui répondent à une requête donnée (e.g. « Bush, guerre, Sharon »).

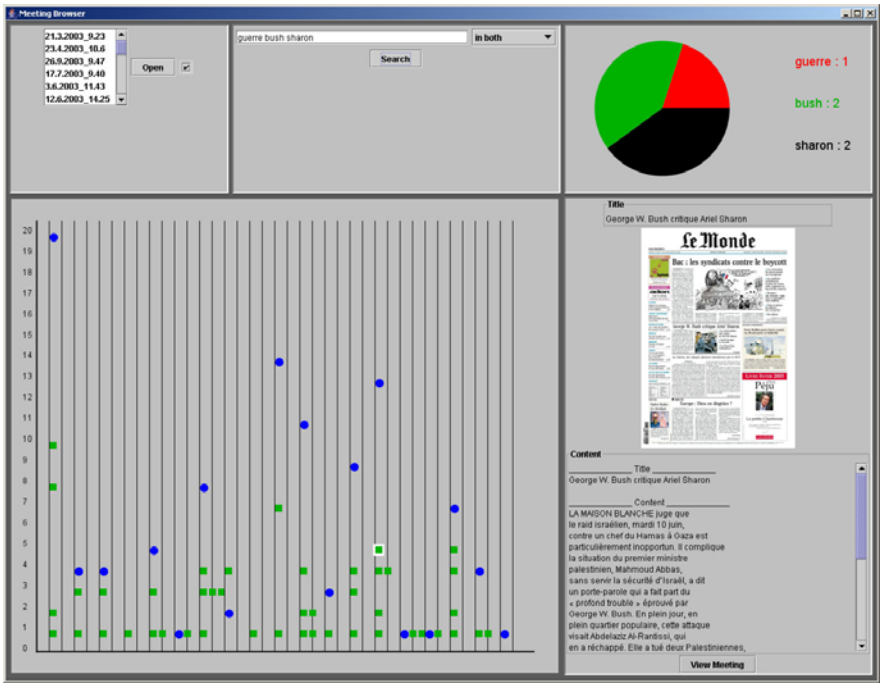


**Figure 4.** *Visualisation des articles de journaux pertinents pour un ensemble de mots-clefs. A droite, le contenu de l'article sélectionné est affiché ainsi que l'image du document dans lequel il se trouve. En dessus, le score de l'article, pour chacun des mots-clefs spécifiés par l'utilisateur, est représenté par un camembert.*

Les articles les plus pertinents sont retournés par le système et organisés spatialement selon les mots-clefs spécifiés. Plus un article, représenté sous la forme d'un carré, est haut dans la visualisation, plus il contient de mots-clefs et donc plus il répond à la requête. L'axe horizontal représente, quant à lui, la date de la réunion dans laquelle l'article a été discuté et indique ainsi l'évolution d'un thème dans le temps. Sur la même visualisation la transcription des dialogues de chacune des réunions est représentée, par un rond, suivant le même procédé. En résumé, cette application cross-réunions permet de visualiser rapidement un grand nombre d'articles, selon un ensemble de mots-clefs défini par l'utilisateur, et surtout favorise une navigation thématique sur l'ensemble des réunions, en utilisant comme points d'accès non seulement le contenu de la transcription des réunions mais aussi le contenu des documents discutés ou visionnés durant les réunions.

Lorsque l'utilisateur clique sur l'un des articles, l'enregistrement de la réunion correspondante est ouvert dans un navigateur dédié à l'instant où l'article est discuté (figure 5), ainsi que toutes les données liées à cette réunion, telles que les séquences audio/vidéo de chaque participant, la transcription des dialogues, les documents et diaporamas de la réunion, ainsi que toutes les annotations liées à ces données. Le navigateur présenté sur la figure 5 est ainsi constitué des composants suivants : la visualisation *sliderBar* tout en bas, les documents discutés à gauche, les diaporamas en bas à droite, les séquences audio/vidéo au centre et la transcription des dialogues à droite. Toutes ces représentations sont synchronisées, ce qui signifie qu'elles ont toutes la même référence temporelle : le temps de la réunion. Lorsque l'utilisateur clique sur l'une de ces représentations visuelles, par exemple sur un article de journal ou sur un énoncé de parole de la transcription textuelle des dialogues, tous les autres composants se synchronisent, i.e. se positionnent au même moment dans la réunion, et affichent leur contenu à cet instant. Par exemple, cliquer sur un article d'un document place les séquences audio/vidéo à l'instant où l'article était discuté, positionne la transcription au même instant et affiche le document qui était projeté. Ces liens visuels sont une illustration directe des alignements documents/parole et documents/vidéo.

La visualisation *sliderBar* en bas de l'écran représente la durée complète de la réunion. Chaque couche symbolise une annotation temporelle différente : les blocs thématiques des documents discutés, les diapositives visibles à chaque instant, les tours de parole, et les énoncés de parole. D'autres annotations temporelles pourraient être ainsi affichées, suivant le type de réunions, les données capturées, ou encore suivant les outils d'analyse disponibles (les actes de dialogue, les prises de note, les gestes, etc.). Ces annotations temporelles sont pour l'instant stockées sous forme de fichiers XML, qui contiennent les marqueurs temporels de début et de fin de chaque changement d'état (i.e. nouvel interlocuteur/tour-de-parole, nouveau thème, changement de diapositive, etc.), ainsi que des informations topologiques pour les documents. Par exemple, la transcription de la parole contient des tours de parole, c'est-à-dire des segments de parole où un seul interlocuteur s'exprime, divisés en énoncés de parole, avec les temps respectifs de début et de fin.

La visualisation *sliderBar* est de plus interactive ; les utilisateurs peuvent ainsi cliquer sur n'importe quelle partie d'une couche afin d'accéder à un moment spécifique de la réunion, une diapositive spécifique ou à tout ce qui a été dit concernant un article spécifique d'un document. Toutes les synchronisations entre les documents statiques, les données vidéo et la transcription des dialogues, sont le fruit des alignements temporels de documents présentés dans cet article. Le *sliderBar*, ainsi que d'autres visualisations similaires, révèlent les relations potentielles entre des ensembles d'annotations, mettent en évidence des synergies possibles ou des conflits, et peuvent ainsi permettre de découvrir de nouvelles méthodes afin d'améliorer la génération automatique d'annotations.

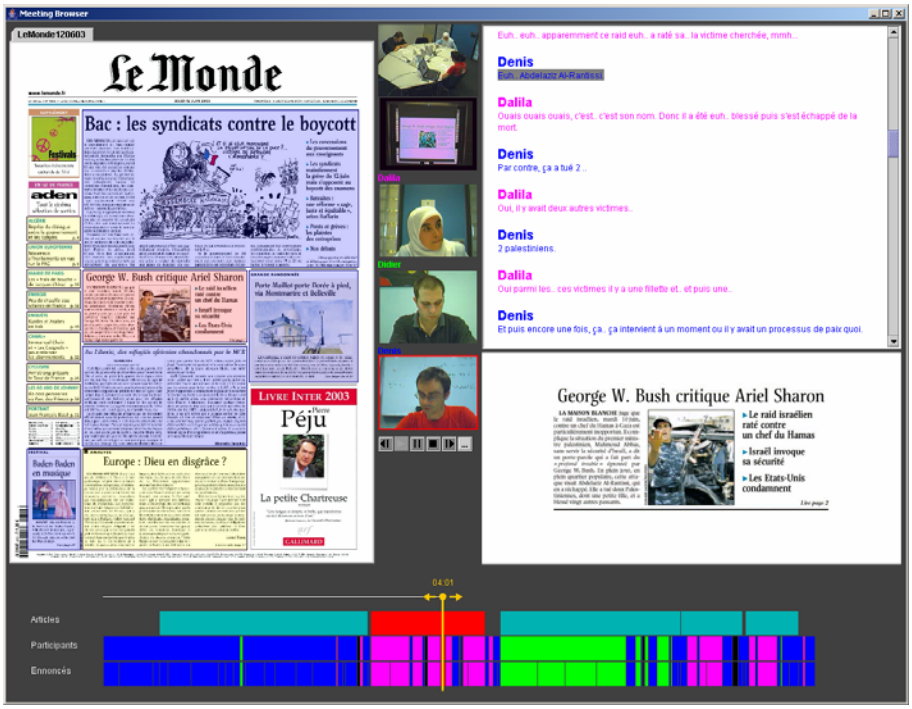


**Figure 5.** *Ce prototype d'interface de navigation, basé sur les documents, a été implanté en Java (JMF et Batik). Tous les composants, audio,* vidéo, *transcription, documents, visualisations, sont synchronisés sur le temps courant de la réunion, grâce aux alignements temporels de documents.*

A l'heure actuelle, 22 réunions, d'une quinzaine de minutes chacune, ont été intégrées dans cette interface de navigation basée sur les documents. Une évaluation a été effectuée par 8 utilisateurs. L'objectif était de mesurer l'utilité des alignements de documents statiques pour naviguer et rechercher des informations sur des

archives multimédias de réunions. Les performances des utilisateurs pour répondre à des questions, aussi bien unimodales que multimodales (e.g. « Quels articles de la une du *Monde* ont été discutés par Didier ? »), ont été mesurées aussi bien d'un point de vue qualitatif que quantitatif (e.g. durée, nombre clics afin d'accomplir la tâche, satisfaction de l'utilisateur, etc.).

Pour cette évaluation, deux versions du navigateur ont été produites : une version complète dans laquelle les alignements de documents étaient disponibles, et une version sans les alignements. Dans la seconde version, au contraire de la première, lorsque l'utilisateur cliquait sur un article, les autres modalités ne se synchronisaient pas. De même si une unité de parole, i.e. énoncé ou tour de parole, était sélectionnée, l'article courant n'était pas mis en valeur. De plus, dans la seconde version, la visualisation *sliderBar* n'indiquait ni les changements d'articles discutés, ni les changements de diapositives. Finalement, dans les deux versions l'utilisateur pouvait cliquer sur un article pour voir son contenu textuel.

Les 8 utilisateurs ont résolu 76% des questions posées lorsqu'ils avaient à disposition les alignements de documents, i.e. en utilisant le premier prototype, et seulement 66% des questions lorsqu'ils n'avaient pas les alignements, i.e. en utilisant la deuxième version de l'application. Ces différences de performance sont devenues particulièrement apparentes pour les questions multimodales, i.e. qui nécessitaient des informations aussi bien contenues dans la transcription de la parole que dans les documents projetés ou discutés. Dans ce cas, 70% des questions ont été résolues lorsque les alignements étaient disponibles et seulement 50% des questions lorsqu'ils n'étaient plus à disposition.

## 7. Conclusion

Cet article propose quatre étapes qui permettent de combler le fossé entre des documents statiques et des données multimédias de réunions. L'analyse de documents permet tout d'abord de construire une représentation multicouche des documents et de créer des index utiles à l'alignement avec d'autres modalités. En particulier, les alignements documents/parole et les alignements documents/vidéo ont été présentés ainsi que leurs évaluations. Ces alignements permettent de « temporaliser » les documents. Finalement, une interface de navigation, basée sur les documents et assemblant tous les alignements dans une seule plate-forme, a été présentée. Une évaluation par des utilisateurs est actuellement en préparation permettant de mesurer l'utilité des alignements de documents pour naviguer et rechercher des informations sur des archives multimédias de réunions.

## 8. Remerciements

Nous tenons à remercier l'Ecole d'Ingénieurs et d'Architectes de Fribourg qui nous a aidé à mettre en place notre environnement d'enregistrement de réunions et tout particulièrement Didier von Rotz. Nous souhaitons aussi remercier Dalila Mekhaldi, Ardhendu Behera et Andrei Popescu-Belis pour leurs contributions. Finalement, nous tenons aussi à remercier les relecteurs anonymes de cette revue qui nous ont grandement aidé à améliorer le contenu de cet article.

## 9. Bibliographie

Barras C., Geoffrois E., Wu Z., Liberman M., « Transcriber: development and use of a tool for assisting speech corpora production », *Speech Communication special issue on Speech Annotation and Corpus Tools*, Vol 33, No 1-2, January 2000.

Behera A., Lalanne D., Ingold R., « Looking at projected documents: Event detection & document identification », *Proceedings of ICME 2004, IEEE International Conference on Multimedia and Expo*, Taiwan, 2004.

Bett M., Gross R., Yu H., Zhu X., Pan Y., Yang J., Waibel A., « Multimodal Meeting Tracker », *Proceedings of Conference on Content-Based Multimedia Information Access, RIAO'2000*, Paris, France.

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, p. 54.

Chiu P., Kapuskar A., Reitmeier S., Wilcox, L., « Room with a rear view. Meeting capture in a multimedia conference room », *IEEE Multimedia*, Volume 7, Issue 4 (October 2000), p. 48-54.

Cutler R., Rui Y., Gupta A., Cadiz JJ., Tashev I., He L., Colbum A., Zhang Z., Liu Z., Silverberg S., « Distributed Meetings: a Meeting Capture and Broadcasting System », *Proceedings of the ACM Multimedia* 2002 Conference, p. 503-512.

Hadjar K., Rigamonti M., Lalanne D., Ingold R., « Xed: a new tool for eXtracting hidden structures from Electronic Documents », *International Workshop on Document Image Analysis for Libraries*, Palo Alto, California, 2004, p. 212-224.

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, p. 9-16.

Hunter J., Little S., « Building and indexing a distributed multimedia presentation archive using SMIL », *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2001*, Darmstadt, Germany, September 4-9, 2001, p. 415-428.

Ishitani Y., « Document image analysis with cooperative interaction between layout analysis and logical structure analysis », *Document Layout Interpretation and its Applications* (DLIA99), Bangalore, India, 1999.

Klemmer S., Graham J., Wolff G., Landay J., « Books with Voices: Paper Transcripts as a Tangible Interface to Oral Histories », *Proceedings of the conference on Human factors in computing systems* (CHI), Ft. Lauderdale, Florida, USA, 2003, p. 89-96.

Lalanne D., Mekhaldi D., 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*, San Jose, 2004, p. 82-91.

Lalanne D., Sire S., Ingold R., Behera A., Mekhaldi D, von Rotz D., « 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*, Berlin, Germany, 2003.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, 1999.

Mekhaldi D., Lalanne D., Ingold R., « Thematic alignment of recorded speech with documents », *ACM Symposium on Document Engineering 2003*, Grenoble, France, 2003, p. 52-54.

Mekhaldi D., Lalanne D., Ingold R., « Thematic Segmentation of Meetings Through Document/Speech Alignment », *in Proceedings of ACM Multimedia 2004, 12th Annual Conference*, October 10-16, 2004, New York City, Columbia University, p. 804-811.

Mukhopadhyay S., Smith, B., « Passive capture and structuring of lectures », *Proceedings of the seventh ACM international conference on Multimedia*, Orlando, Florida, p. 477-487.

Niyogi D., Srihari S.N., « Knowledge-based derivation of document logical structure », *Proceedings of ICDAR*, Montreal, Canada, August 1995, p. 472-475.

Popescu-Belis A., Lalanne D., « Reference Resolution over a Restricted Domain: References to Documents », *ACL 2004 Workshop on Reference Resolution and its Applications*, Barcelona, Spain, p.71-78.

Salton G., Singhal A., Buckley C., Mitra M., « Automatic Text Decomposition Using Text Segments and Text Themes », *In Proceedings of the Hypertext '96 Conference*, Washington D.C., USA, 1996, p. 53-65.

Smith M., Kanade T., « Video Skimming and Characterization through the Combination of Image and Language Understanding », *Proceedings of the Sixth International Conference on Computer Vision (ICCV-98)*, Bombay, India, 1998, p. 61-70.

Stolcke A., Shriberg E., Bates R., Coccaro N., Jurafsky D., Martin R., Meteer M., Ries K., Taylor P., Van Ess-Dykema C., « Dialog act modeling for conversational speech », *In Proceedings of the AAAI-98 Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998.

Tran-Thuong T., Roisin C., « Structured Media for Authoring Multimedia Document », *International Workshop on Web Document Analysis (WDA'2001)*, Seattle, Washington, USA, 8 Sept., 2001.

Uchihashi S., Foote J., Girhensohn A., Boreczky J., « Video Manga: Generating Semantically Meaningful Video Summaries », *in Proceedings of ACM Multimedia Conference*, 1999, p. 383-392.

Wahl F., Wong K., Casey R., « Block Segmentation and text extraction in mixed text / image documents », *Computer Graphics and Image Processing*, Volume 20, 1982, p. 375-390.

Wellner P., « Interacting with paper on the DigitalDesk », *Communications of the ACM*, 36(7), July 1993, p. 86-96.

Wong K., Casey R., Wahl F., « Document Analysis system », *IBM Journal of Research and Development*, Volume 26, Number 6, 1982, p. 647-656.